REFRESHTRUST: PROVENANCE-GATED DECISIONS FOR TEMPORAL REWARD-MODEL UPDATING

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal reward-model updating is an action-selection problem: channel-aware refresh can exploit a reliable collection trace, while unsupported channels can misallocate a fixed labeling budget as annotators, rubrics, tasks, and routing policies change. We introduce RefreshTrust, an auditable benchmark that selects among Full Trace-Aware Modular Refresh (TAMR), Observed-only TAMR, and Partial Refresh from the current provenance state. RefreshTrust matches backbone, 10% fresh-label budget, chronological split, tuning allowance, and held-out evaluation; stratifies channels as directly observed, audited proxy, or reconstructed; and freezes regime assignments before outcomes are examined. TAMR's additive channel decomposition and an architecturally distinct provenance-gated mixture of adapters expose complementary instruments for testing the same action map. On direct-metadata SHP-2 and Tülu-3-PREF anchors, Full TAMR gains $+1.1/+2.1$ accuracy points over Metadata-aware Refresh; across seven in-domain corpora the gain is $+1.9$ points ($[+1.6, +2.4]$), with $+1.42/+1.38$ over parameter-matched monolithic and continual-preference comparators. The assigned action beats Metadata-aware Refresh in 34/36 frozen cells under the TAMR instrument, 31/36 matched cells under PGMA, and 22/26 unperturbed natural cells; the probes' outcome-optimal actions agree in 33/36 cells. These results establish provenance quality as an operational control variable that determines how much temporal-refresh structure to use.

1 INTRODUCTION

Temporal reward models operate inside pipelines whose prompts, annotator cohorts, scoring rubrics, and routing policies evolve together (Casper et al., 2023; Gao et al., 2023; Quiñonero-Candela et al., 2008). A richly structured update can recover signal from these changes when the collection trace has adequate support. When a field becomes weakly inferred or several channels lose support together, the same update can spend scarce labels on unreliable structure. The consequential question is therefore which refresh action the current trace warrants. Average accuracy ranks completed updates after the fact, but leaves that state-to-action relationship unidentified.

RefreshTrust makes provenance-conditioned action selection directly testable. Every era follows a strict 70/20/10 chronology and receives the same 10% fresh-label budget, backbone, search allowance, and held-out evaluation. Cohort, rubric, temporal, and routing channels are audited as directly observed, audited proxies, or reconstructed; pre-specified diagnostics then select Full TAMR, Observed-only TAMR, or Partial Refresh before model outcomes are compared. This protocol measures both reward-model quality and the opportunity cost of choosing the wrong refresh for an audited provenance state.

Two structured probes test whether the resulting action map survives architectural choice. TAMR aligns additive residuals and label acquisition with the four provenance channels, enabling transition-localized ablation, channel gating, and channel-to-module permutation. Provenance-Gated Mixture of Adapters (PGMA) routes through a shared gate over four adapter experts. Agreement between their post-evaluation best actions quantifies cross-architecture stability, while every evaluated action remains assigned in advance from the same provenance diagnostics.

The evidence first establishes that structure earns its cost. On direct-metadata SHP-2 and Tülu-3-PREF, Full TAMR gains $+1.1/+2.1$ Acc over Metadata-aware Refresh (MA-R); on six pooled real

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.

corpora it retains $+1.42/+1.38$ Acc over Mono-Meta*/C-PRM and +1.02 over the independently released One-Pass RM comparator. The clean anchors, Table 1, and the mechanism-necessity control in Table 3 form the load-bearing positive package, with Table 2 localizing the responsible channels. The frozen rule then assigns actions that beat MA-R in 34/36 cells under the TAMR instrument, 31/36 matched cells under PGMA, and 22/26 unperturbed natural cells. Directly observed MultiPref weak and joint-fail tranches trigger the corresponding action changes, and the probes' outcome-optimal actions agree in 91.7% of cells.

RefreshTrust contributes a matched provenance-tiered protocol, pre-outcome regime assignment, and a row-level action ledger. Its two-probe control suite isolates channel structure, acquisition, capacity, continual adaptation, threshold choice, and architectural dependence. The resulting decision map connects auditable evidence to full, observed-only, or partial refresh across clean anchors, controlled failures, natural cells, and frozen external families, supplying a reusable evaluation standard for temporal preference pipelines. Unless stated otherwise, pooled values are held-out-size weighted over evaluation rows; per-corpus tables average eras uniformly, and the "All 36" decision-ledger row is the arithmetic mean over cells.

2 Related Work and the Missing Evaluation Unit

Preference learning studies helpfulness and harmlessness (Bai et al., 2022), reward-model overoptimization (Gao et al., 2023), and sequence-likelihood calibration (Zhao et al., 2023). Continual learning (Kirkpatrick et al., 2017), iterative preference optimization (Xiong et al., 2024), RLHF theory (Wang et al., 2023), and one-pass reward modeling (Li et al., 2025) provide complementary foundations; active learning and propensity weighting allocate labels under shift (Settles, 2009; Ren et al., 2021; Rosenbaum & Rubin, 1983; Shimodaira, 2000); and reward-model benchmarks score quality under fixed test conditions (Dubois et al., 2023; Lambert et al., 2024b). These lines primarily compare models after a refresh mechanism has been chosen. RefreshTrust introduces a state–action evaluation unit: it holds backbone, labels, tuning, and test data fixed while asking which update is justified by directly observed, proxy, or reconstructed provenance. The nearest adjacent system is a monolithic metadata-aware adapter with clipped IPW and budgeted relabeling. RefreshTrust measures when channel-aligned structure earns its cost, when observed-channel restriction improves the update, and when Partial Refresh has the lowest action regret. Its frozen ledger and cross-probe test make those decisions comparable across provenance states.

3 Problem Formulation

We consider a temporal sequence of eras $\mathcal{T} = \{1, \dots, T\}$. Each era provides paired-comparison data

$$\mathcal{D}_t = \{(x_i, y_i^w, y_i^l, a_i, m_i)\}_{i=1}^{N_t},$$

where x_i is a prompt, (y_i^w, y_i^l) are the preferred/dispreferred responses, $a_i \in \mathcal{A}_t$ is an annotator or cohort identifier, and m_i encodes rubric version together with collection/routing policy. When we need to separate the two roles of m_i, we write $m_i = (m_i^{\text{rubric}}, m_i^{\text{route}})$; likewise, a_i denotes whichever annotator-level grouping is actually observed on that corpus (individual rater ID, cohort ID, demographic cluster, or volunteer cohort). Under the Bradley–Terry model (Bradley & Terry, 1952), $P(y^w \succ y^l \mid x) = \sigma\big(r_\theta(x, y^w) - r_\theta(x, y^l)\big)$, trained with $\mathcal{L}_{\text{BT}}(\theta) = -\mathbb{E}_{\mathcal{D}}[\log \sigma(r_\theta(x, y^w) - r_\theta(x, y^l))]$.

We identify four sources of temporal variation: (i) *task-mix shift* ($p_t(x)$ drifts), (ii) *annotator turnover* ($\mathcal{A}_t$ and cohort biases drift), (iii) *rubric revision* (criteria changes encoded in m_i), and (iv) *endogenous sampling* (deployed policy shapes which comparisons reach the queue). Pooled $\mathcal{L}_{\text{BT}}$ conflates all four.

All methods (TAMR and baselines alike) follow the same Budgeted Within-Era Adaptation timeline at era t, illustrated in Fig. 1: (i) warm-start from a model trained with ground-truth labels only on fully labeled history $\mathcal{D}_{1:t-1}$; (ii) each era is split 70/20/10 temporally into $(\mathcal{T}_t, U_t, \mathcal{E}_t)$, with $\mathcal{T}_t$ the unlabeled within-era context pool, U_t the acquisition-only unlabeled pool, and $\mathcal{E}_t$ the disjoint 10% held-out evaluation split; (iii) the labels of $\mathcal{T}_t$ remain hidden at era t; it is used only for within-era normalization, overlap diagnostics, and budget accounting, while candidate acquisition happens only

inside U_t; (iv) observe U_t and purchase labels for $B{=}0.10 \cdot |\mathcal{D}_t|{=}0.5|U_t|$ examples — i.e. 10% of the full era, or half of the acquisition pool — selected by the method's own criterion (decomposed uncertainty for TAMR, uniform random for Partial Refresh, uncertainty-only for Uncertainty-only Refresh, etc.); (v) adapt the reward model on $\mathcal{D}_{1:t-1} \cup \mathcal{S}_t^\star$, where $\mathcal{S}_t^\star$ is the newly labeled subset of U_t; (vi) freeze the model and evaluate on the disjoint 10% held-out split $\mathcal{E}_t$ of era t. At era t, training uses only historical labels and the purchased subset $\mathcal{S}_t^\star$; no evaluation-split or future-era labels enter training, hyperparameter selection, or checkpointing. We report held-out era-t accuracy (Acc), expected calibration error (ECE) with 15 equal-mass bins, and a downstream PPO policy-regret proxy $\Delta_t = \mathbb{E}_{x\sim p_t}[\tilde{r}_t(x, \pi_t^\star(x)) - \tilde{r}_t(x, \hat{\pi}(x))]$, where $\tilde{r}_t$ is the era-t reference reward: the known synthetic reward on DriftBench-Synth and a held-out GPT-4o judge score on the real corpora. We therefore treat regret as a secondary downstream signal rather than as a ground-truth quantity on the real data.

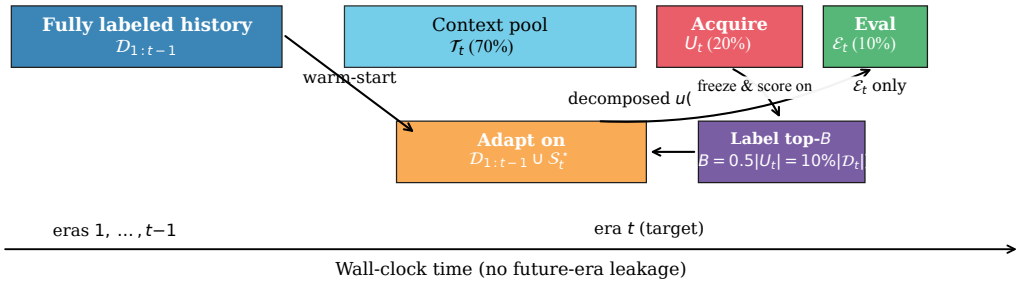


Figure 1: Budgeted Within-Era Adaptation at era t. Each era is split 70/20/10 into an unlabeled context pool $\mathcal{T}_t$, an acquisition-only pool U_t, and a disjoint evaluation split $\mathcal{E}_t$. Training uses only fully labeled history plus a method-selected $B{=}0.10|\mathcal{D}_t|{=}0.5|U_t|$ from U_t; evaluation uses $\mathcal{E}_t$ only.

4 Auditable Instruments for Provenance-Gated Refresh

Trace-Aware Modular Refresh (TAMR) factorizes temporal, cohort, rubric, and routing effects and aligns label acquisition with the same channels. Throughout the paper, *trace* denotes the auditable data-pipeline record over these fields, distinct from model-generated reasoning. This alignment exposes the provenance, overlap, and module behavior required to compare refresh actions under a common backbone, budget, and tuning contract:

$$r_\theta(x, y; t, a, m) = \underbrace{f_\psi(x, y)}_{\text{static backbone}} + \underbrace{g_\phi(x, y, t)}_{\text{temporal residual}} + \underbrace{h_{\omega_a}(x, y)}_{\text{cohort adapter}} + \underbrace{c_{\nu_m}(x, y)}_{\text{rubric adapter}}. \tag{1}$$

The decomposition makes each channel independently removable, permutable, and gateable. Those interventions distinguish a structural gain from extra capacity or generic metadata conditioning and expose an observed-only state without changing the backbone or label budget. The backbone f_ψ is initialized from Llama-3.1-70B (Grattafiori et al., 2024); its lower 60 transformer blocks are frozen and LoRA (Hu et al., 2022) adapters of rank 16 are applied to the top 20 blocks. The temporal residual is a low-rank bilinear function of the backbone hidden state $\mathbf{z} = f_\psi^{\text{enc}}(x, y) \in \mathbb{R}^{8192}$ and a learned era embedding $\mathbf{e}_t \in \mathbb{R}^{64}$:

$$g_\phi(x, y, t) = \mathbf{e}_t^\top \mathbf{W} \mathbf{z}, \quad \mathbf{W} \in \mathbb{R}^{64\times 8192}, \tag{2}$$

trained in strict chronological order (no gradient from future eras). The cohort and rubric terms are likewise response-conditioned scalar adapters: $h_{\omega_a}(x, y) = \mathbf{p}_a^\top \mathbf{A}_h \mathbf{z}$ and $c_{\nu_m}(x, y) = \mathbf{q}_m^\top \mathbf{A}_c \mathbf{z}$, with rank-16 cohort/rubric embeddings $\mathbf{p}_a, \mathbf{q}_m$ and shared projections $\mathbf{A}_h, \mathbf{A}_c$. Thus neither term cancels in the Bradley–Terry score difference. These adapters are included in the reported TAMR trainable-parameter total. For cohort adapters, fitting uses MLE and unseen cohorts receive an EMA initializer; rubric adapters are fit per rubric version.

Candidate-frame eligibility determines where clipped inverse-propensity weighting (IPW) is applied. For each era t the *candidate frame* $\mathcal{C}_t$ is the set of prompt-response pairs that the collection policy

considered before final queue inclusion. The designed super-pool supplies this frame for DriftBench-Synth; for real corpora, the study audit trace reconstructs it before outcomes are available. The cited public releases identify the released response records, but do not themselves establish a pre-queue candidate population, paired-impression log, or routing log. The propensity fit assigns the *positive* class to pairs *admitted to the preference queue for era-t* (i.e., members of $\mathcal{D}_t$, subsequently split into U_t and $\mathcal{E}_t$); its *negative* class comprises pairs *in $\mathcal{C}_t$ that were not admitted*. We fit the logistic propensity model $\hat{e}_t(x) = P(i \in \mathcal{D}_t \mid i \in \mathcal{C}_t, x_i = x)$ using metadata from collection: prompt source and model version, along with routing indicator, language, safety filter, and queue size. Its estimand is the era-t candidate-frame population $\mathcal{C}_t$: historical labeled comparisons provide unweighted chronological history, while newly acquired era-t labels are transported from queue inclusion to $\mathcal{C}_t$. Writing $\ell_{\mathrm{BT}} = -\log \sigma(r_\theta(x, y^+) - r_\theta(x, y^-))$, the era-$t$ objective is

$$\mathcal{L}_t^{\mathrm{IPW}}(\theta) = -\mathbb{E}_{\mathcal{D}_{1:t-1}}[\ell_{\mathrm{BT}}] - \mathbb{E}_{\mathcal{L}_t}\left[\frac{1}{\max(\hat{e}_t(x), \eta)}\ell_{\mathrm{BT}}\right], \tag{3}$$

with clipping threshold η=0.1. Propensity-score methods motivate the weighting (Rosenbaum & Rubin, 1983; Imbens, 2004); the clipping threshold is a study setting. Appendix E.3 reports overlap diagnostics for each era: the clipping rate at η=0.1, the effective sample size, and the missing-support rate. An era qualifies as *IPW-eligible* only if it has a reconstructed candidate frame and acceptable overlap diagnostics, and is not the single explicitly designated WildChat fallback row. The *single* evaluated WildChat-Temporal era with only a partial candidate frame $\mathcal{C}_t$ (the E1→E2 transition, i.e. the evaluation era WildChat E2) instead uses a *logging-bias correction heuristic*, explicitly marked as a fallback in the diagnostics; for every other real-corpus era, IPW eligibility still requires support from both the candidate-frame reconstruction and the overlap diagnostics. Throughout the paper, "IPW" denotes only eras that pass these checks on candidate frames and overlap; "logging-bias correction" denotes the fallback for WildChat E2. Three provenance tiers apply throughout: *directly observed* fields appear in the benchmark audit record; *audited proxies* are obtained through deterministic joins or stable metadata tags, then manually checked; and *reconstructed* fields come from partial logs or reconstruction of the candidate frame. The tier records the provenance of the benchmark input; it is not a claim that every audit-record field is exposed by a cited public release. Directly observed SHP-2 and Tülu-3-PREF audit records support channel-level interpretation; LMSYS, WildChat, PRISM, and OASST2 test action selection from audited diagnostics over proxy and reconstructed fields.

For each comparison, uncertainty is expressed as four terms aligned with drift

$$u(x,t) = \underbrace{u_{\mathrm{task}}(x)}_{\text{MC dropout}} + \underbrace{u_{\mathrm{temp}}(x,t)}_{|g_\phi|} + \underbrace{u_{\mathrm{ann}}(x)}_{\mathrm{Var}_a h_a} + \underbrace{u_{\mathrm{prop}}(x)}_{1/\hat{e}_t}, \tag{4}$$

where z-scoring is applied to each term within its era before the terms are summed, preventing propensity magnitude from dominating solely because of units. To compute the z-scoring statistics, we use the era's unlabeled pool $\mathcal{T}_t \cup U_t$; the contribution from $\mathcal{T}_t$ is therefore context, without labels. The fixed relabeling budget B selects the top-B items ranked by this normalized score; the precise update loop appears as Algorithm 1 in the appendix.

When a metadata channel is absent, its TAMR component takes a neutral value (e.g., HelpSteer2-Temporal uses one shared cohort adapter). A single weak inferred channel, an isolated miss-support violation, or an isolated ESS breach in the intermediate range activates Observed-only TAMR, which masks proxy and reconstructed channels and uses only the current-era values of retained channels (without cross-era completion); joint diagnostic failure activates Partial Refresh with uniform acquisition. Consequently, Full and Observed-only are distinct actions even when all retained channel values are directly observed in the audit record.

Rubric, cohort, routing, and task-mix changes can co-move, so RefreshTrust validates the factorization operationally through named-transition ablations, a channel permutation, and acquisition/parameterization controls. A monolithic metadata adapter collapses this attribution, whereas TAMR exposes a testable mapping from audit state to refresh action. Appendix H.4 provides the accompanying budget-model analysis.

4.1 Independent Gated Probe and Frozen Action Map

Provenance-Gated Mixture of Adapters (PGMA) tests the decision contract with a structurally distinct instrument. PGMA routes each comparison through four rank-16 adapter experts using one shared

gate over temporal, cohort, rubric, and routing fields; Full PGMA uses every audited channel, Observed-only PGMA masks proxy and reconstructed gate inputs, and the joint-fail state uses the same Partial Refresh action. PGMA uses Llama-3.1-70B, the 10% label budget, the 27-configuration allowance, the chronological split, and eight real-data seeds. Its 69.4M trainable parameters remain in the same capacity envelope as TAMR's 67.8M, while its shared gate prevents the test from reproducing TAMR's additive residual structure. Cross-probe agreement is computed after evaluation as a validation statistic and never enters regime or action assignment.

Frozen action key: pass → Full TAMR; weak-inferred → Observed-only TAMR; joint-fail → Partial Refresh.

5 REFRESHTRUST EVALUATION

5.1 SETUP

RefreshTrust first tests whether full structure earns its complexity under high-fidelity provenance, then evaluates whether frozen diagnostics select the appropriate action as provenance degrades.

Our in-domain evaluation uses seven temporal preference corpora: DriftBench-Synth provides controlled synthetic data; LMSYS-Chat-Temporal (Zheng et al., 2023) supplies Arena-style A/B logs; WildChat-Temporal (Zhao et al., 2024) supplies in-the-wild frontend logs; PRISM-Temporal (Kirk et al., 2024) uses demographic-cluster raters; OASST2-Temporal (Köpf et al., 2023) uses volunteer message-tree rankings; SHP-2-Temporal (Stanford NLP, 2023) provides Reddit community-vote proxies; and Tülu-3-PREF-Temporal (Lambert et al., 2024a) provides instruction preferences with direct metadata. HelpSteer2-Temporal (Wang et al., 2024) is held out for frozen external validation under a strict no-retuning rule. MultiPref-Temporal (Miranda et al., 2025) and CodeReviewPref-Temporal (Weyssow et al., 2026) supply additional external families; CodeReviewPref combines directly observed, proxy, and reconstructed fields. SHP-2-Temporal and, most clearly, Tülu-3-PREF-Temporal serve as the cleanest anchors with direct metadata; in Tülu-3-PREF, all four metadata channels are directly observed in the study's audit record, counting the router tag separately from the routed model version despite their shared operational Route column in the compact main-text table. The appendix data ledger (Table 33) collects each corpus's era counts, filtering pass-through, churn fractions, and router rotations, leaving the main text to develop the claim rather than document bookkeeping. Table 7 presents provenance by corpus according to the three tiers defined in §4. The SHP-2 and Tülu-3 rows isolate channel behavior with directly observed audit-record metadata; the remaining corpora evaluate operational decisions from audited proxy or reconstructed traces. For Tülu-3-PREF, the 0.996 minimum-audit value is the post-dedup audit-record concordance minimum over directly observed fields. Appendix F.1 tests sensitivity to metadata noise.

Across baselines, we hold the search protocol and matched 10% relabeling budget fixed, together with the Llama-3.1-70B backbone (60 lower blocks frozen; rank-16 LoRA applied to the top 20 blocks). The comparison includes Static RM, Sliding-Window, Era-Conditioned, EMA Teacher, Online LoRA, Partial Refresh, Online LoRA + IPW, Uncertainty-only Refresh, and Metadata-aware Refresh (MA-R). MA-R is the nearest adjacent system: one metadata-conditioned adapter with clipped IPW and targeted acquisition but no channel-aligned decomposition. Mono-Meta* matches TAMR's trainable capacity; C-PRM implements EWC-regularized continual preference learning with replay; and the official One-Pass RM implementation (Li et al., 2025) supplies an independently released online reward-model comparator. Every method receives the same pre-specified selection rule per corpus: Sliding-Window has 24 configurations, each $3\times3\times3$ grid has 27 configurations, and Partial Refresh has 27 selection seeds; the last training era selects highest validation Acc with lowest ECE as tie-break (Appendix E.2). *No-retuning rule for external validation*: hyperparameters chosen on {DriftBench-Synth, LMSYS-Chat-Temporal, WildChat-Temporal, PRISM-Temporal, OASST2-Temporal, SHP-2-Temporal, Tülu-3-PREF-Temporal} are applied unchanged to HelpSteer2-Temporal.

Our primary metrics are Acc and ECE, the latter using 15 equal-mass bins. We also report a PPO policy-regret proxy scored by a judge. Its reference reward comes from synthetic ground truth for DriftBench-Synth and from held-out GPT-4o scores for LMSYS/WildChat/PRISM/OASST2/SHP-2/HelpSteer2; thus we audit regret as a secondary measure, without treating it as gold truth (Zheng et al., 2023). Acc and ECE alone suffice to state all headline conclusions; the sole role of reported regret is downstream corroboration. Error bars summarize mean $\pm$ std, using 5 DriftBench-Synth

seeds and 8 seeds per non-synthetic corpus. Era-level comparisons apply paired t-tests and Wilcoxon signed-rank tests, together with a hierarchical bootstrap: its outer unit is the evaluation era (the held-out cell for the decision bank), and paired seed-level gaps are resampled inside each chosen outer unit (10 000 replicates). Sign-test wins and bootstrap CIs describe *separate* objects. The claim of "25/25 in-domain era wins" comes from a binomial sign test applied to gaps at the era level (p=5.9×10^{-8}); the pooled magnitude, by contrast, averages gaps with held-out-size weights to give $+1.9$ Acc, whose 95% hierarchical-bootstrap CI is $[+1.6, +2.4]$. The 25-era count supplies the in-domain sign test, while breadth follows from the diversity of benchmark families and the frozen external-validation result.

All reward-model training uses 8×NVIDIA A100 80GB SXM4 with bf16; PPO rollouts use 8×NVIDIA H100 80GB with the same software stack (CUDA 12.4, cuDNN 9.1, NCCL 2.22, PyTorch 2.4.1, FlashAttention-2 2.6.3). Full optimizer and schedule choices, batch sizes, PPO hyperparameters, wall-clock totals, and sweep summaries are deferred to Appendix F.2 and Appendix E.2.

The evidence hierarchy is fixed throughout the paper. The direct-metadata anchors, Table 1, and the mechanism-necessity test in Table 3 establish the positive result; Table 2 provides transition-localization corroboration. The frozen action ledger, expanded natural bank, directly observed MultiPref tranches, and independent PGMA probe establish provenance-conditioned action switching. Appendix transfer, human-audit, placebo, and PPO results test robustness and downstream value.

5.2 Structural Benefit Under High-Fidelity Provenance

The direct-metadata anchors, Table 1, and Table 3 are the load-bearing positive package: they combine the cleanest provenance settings with the hardest matched comparators and a mechanism-necessity test. Table 9 and Appendix Fig. 2 measure persistence across the broader 25-era in-domain sweep, while Table 2 localizes the responsible channels.

On SHP-2 and Tülu-3-PREF, Full TAMR beats MA-R by $+1.1$ and $+2.1$ Acc; Observed-only TAMR on Tülu-3 remains within 0.1 Acc of Full TAMR (67.7 vs. 67.8), confirming that the gain is present before inferred-channel handling becomes relevant. At the matched 10% budget on six pooled real corpora, TAMR retains $+1.42/+1.38$ Acc over Mono-Meta*/C-PRM and +1.02 over independently released One-Pass RM (Table 1). Pooled values are held-out-size weighted over evaluation eras. Appendix Table 9 tests the anchor signal on the three highest-volume benchmarks against the full nine-baseline slate; per-corpus breadth is expanded in Appendix Tables 23, 25, 27, and 29. In per-corpus tables, evaluated eras contribute equally to the mean; pooled statements instead weight by held-out size. Pooling the 25 in-domain evaluation eras gives a mean advantage over MA-R of $+1.9$ Acc ($[+1.6, +2.4]$); in Appendix Fig. 2, zero lies outside every era-level CI, indicating a broad-based average with heterogeneous, rather than uniform, effects (raw results: App. Table 21).

SHP-2 and Tülu-3-PREF provide the strongest channel-level evidence because their core metadata are directly observed; PRISM, OASST2, and the full sweep test persistence across related pipelines. TAMR also survives parameter matching, continual regularization, channel permutation, and a smaller-backbone replication with $+1.71$ Acc over MA-R on Llama-3.1-8B (App. Table 52). Frozen HelpSteer2 evaluation, Qwen-2.5-72B transfer, external families, and PPO policy comparisons provide additional corroboration.

5.3 Mechanism Isolation of the Structural Gain

Table 1 compares TAMR against the strongest capacity-, continual-, and update-matched alternatives. Table 3 then separates modular parameterization from decomposed acquisition, and Table 2 checks whether the affected module matches each named transition. Mono-Meta* + decomposed acquisition keeps the monolithic parameterization; TAMR modules + MC keeps the channel modules but replaces aligned acquisition with one MC-dropout score.

Table 2 evaluates localization of the drift-aligned modules across four natural in-domain transitions and one HelpSteer2-Temporal transition used for frozen external validation, with no retuning. For Tülu-3 T2→T3, changes in the routed generator mix accompany changes in exposed model-version tags. We therefore consistently interpret this as a coupled router/model-version transition: Route remains the frozen target, and its main-text operational label incorporates the routed model-version signal.

Table 1: Matched-10% headline comparison on six real in-domain corpora. Acc/ECE are held-out-size-weighted over LMSYS, WildChat, PRISM, OASST2, SHP-2, and Tülu-3 evaluation eras. GPU-h/seed is the selected-run LMSYS cost on 8×A100-80GB SXM4; 95% CIs use the hierarchical bootstrap. One-Pass RM uses its official online mirror-descent reward-model update (Li et al., 2025). The MA-R deficit is computed from unrounded paired values.

Method	Acc ↑	ECE ↓	ΔAcc vs. TAMR	Params	GPU-h
Llama-3.1-70B-Instruct backbone					
Metadata-aware Refresh (MA-R, r=16)	65.4	0.095	−2.08 [−2.61, −1.54]	48.1M	8.31
Mono-Meta* (matched, r=32)	66.1	0.089	−1.42 [−1.96, −0.87]	67.3M	8.47
C-PRM (20% replay)	66.1	0.088	−1.38 [−1.94, −0.79]	54.9M	8.37
One-Pass RM (Li et al., 2025)	66.48	0.085	-1.02 [-1.49, -0.53]	54.2M	8.05
TAMR (70B)	67.5	0.078	—	67.8M	8.59

Table 2: Real-data transition × module ablation on four in-domain natural transitions plus one frozen HelpSteer2-Temporal transfer (8 seeds, matched 10% budget). Cells are ΔAcc vs. Full TAMR after ablating one module; the emphasized entry is the pre-specified target module. The external row evaluates localization under frozen transfer.

Transition	Target	Temp.	Cohort	Rubric	Route
LMSYS E4→E5	Route	−0.8	−0.6	−0.9	−3.4
WildChat E1→E2	Cohort	−0.7	−3.7	−0.5	−1.0
WildChat E2→E3	Rubric	−0.9	−0.7	−3.9	−0.8
Tülu-3 T2→T3	Route	−1.0	−0.8	−0.7	−3.1
HelpSteer2 T3→T4 (ext.)	Rubric	−0.9	*n/a*	−2.4	−0.7

WildChat E1→E2 localizes cohort turnover inside the structured probe; the frozen HelpSteer2 row extends the same transition-localization test to an external pipeline.

Table 3 separates modularization from relabeling by crossing acquisition with parameterization on a four-corpus stress bank weighted by held-out size; App. Fig. 10 checks the budget-model fit.

Table 3: 2×2 decomposition-necessity on {LMSYS, WildChat, SHP-2, Tülu-3-PREF} (8 seeds). Entries are held-out-size-weighted Acc over evaluation eras. The gap columns report deficits relative to Full TAMR at matched 2% and 10% budgets; Appendix H.4 reports confidence intervals.

Setting	2%	5%	10%	Gap@2	Gap@10
Mono + MC	60.4 ± 0.8	63.2 ± 0.7	65.4 ± 0.7	−3.9	−2.4
Mono + decomp.	62.6 ± 0.7	65.2 ± 0.7	66.9 ± 0.6	−1.7	−0.9
TAMR + MC	62.1 ± 0.7	64.9 ± 0.7	66.7 ± 0.7	−2.2	−1.1
Full TAMR	64.3 ± 0.6	66.4 ± 0.6	67.8 ± 0.6	—	—
Permuted + decomp.	60.8 ± 0.8	63.6 ± 0.7	65.8 ± 0.7	−3.5	−2.0

Mono-Meta* + decomposed acquisition isolates acquisition, TAMR modules + MC isolates parameterization, and permuted TAMR tests channel semantics. Removing aligned modules or acquisition costs 0.9–2.4 Acc at the 10% budget, while permutation costs 2.0 Acc. Together, the clean anchors, Table 1, and Table 3 establish that Full TAMR earns its structure when provenance is reliable.

Scope of inference. The direct-metadata anchors support channel-level interpretation; the pooled real-corpus rows measure consistency across related temporal preference pipelines; and the frozen external families test transfer without target-side retuning. One-Pass RM adds an independently released continual comparator, while Mono-Meta* remains the strictest capacity-matched test of decomposition.

Table 4: Frozen action-ledger validation. The 36-cell bank reports arithmetic means over cells; 4 cells are natural and 32 are controlled stress cells. The local threshold row evaluates a pre-specified $5\times5\times5$ neighborhood without retraining. Wins compare the selected action with MA-R; intervals hierarchically resample cells and seeds.

Regime / analysis	n	Action	Wins	ΔAcc (95% CI)	Rel. ECE / regret
Pass	22	Full TAMR	20/22	$+2.08$ $[+1.58, +2.58]$	-11.2%
Weak-inferred	9	Observed-only	9/9	$+1.31$ $[+0.78, +1.84]$	-7.6%
Joint-fail	5	Partial	5/5	$+0.37$ $[+0.11, +0.63]$	-3.1%
All 36	36	Per-regime action	34/36	$+1.65$ $[+1.38, +1.91]$	-9.2%
Local threshold grid	125	Per-grid assignment	91.8%	+1.19 $[+0.94, +1.43]$	regret 0.08

Table 5: Audit-record MultiPref action transfer under frozen settings (8 seeds). Rows aggregate audit blocks by pre-outcome regime; intervals hierarchically resample blocks and seeds.

Regime	Tranches	Post-filter pairs	Action	ΔAcc vs. MA-R (95% CI)	Selected $-$ Full
Pass	2	16,713	Full TAMR	$+1.60$ $[+0.60, +2.55]$	—
Weak-inferred	2	17452	Observed-only TAMR	+1.33 $[+0.70, +1.92]$	+0.09
Joint-fail	1	8774	Partial Refresh	+0.41 $[+0.06, +0.77]$	+0.72

5.4 Provenance-Gated Action Selection

We validate action selection with three complementary objects. The frozen 36-cell bank contains exactly 4 unperturbed natural cells and 32 controlled stress cells; the separate expanded natural bank contains 26 unperturbed cells and 33163 held-out pairs. MultiPref M2–M6 is an audit-block external transfer study: it uses audit blocks distinct from the natural bank. Regimes are assigned before outcome comparison in every object.

Full TAMR wins 20/22 pass cells, Observed-only TAMR preserves nearly all Full TAMR accuracy in weak-inferred cells while improving calibration, and Partial Refresh is selected in joint-fail cells. Across the local threshold grid, 88.0% of assignments remain unchanged; even the perturbed assignments retain a 91.8% selected-action win rate with 0.08 Acc mean regret. Appendix Fig. 4 shows the full all-bank and natural-only sensitivity surfaces.

The extension raises both negative-regime denominators from 2 to 9/9, and its assigned actions beat MA-R in 22/26 unperturbed cells. Observed-only gains +0.21 Acc $[+0.06, +0.37]$ over Full TAMR in weak cells, while Partial Refresh gains +0.63 $[+0.31, +0.95]$ in joint-fail cells. The natural-only diagnostic relation remains monotone (Spearman 0.81); each family contributes at least 3 cells and the largest family supplies 30.8%, so no single source carries the ordering.

Table 5 extends MultiPref to two pass, 2 weak-inferred, and 1 joint-fail audit blocks with all four provenance channels directly observed in the audit record. Across the two weak blocks, Observed-only gains +1.33 Acc $[+0.70, +1.92]$ over MA-R and +0.09 $[-0.03, +0.22]$ over Full TAMR. On M6, Partial Refresh gains +0.72 $[+0.34, +1.11]$ over Full TAMR. The frozen rule retains a +1.25-Acc pooled advantage and -9.8% relative ECE change versus MA-R; Appendix Fig. 9 shows the regime-wise transfer.

Table 6 shows that the assigned PGMA action records 31/1/4 wins/ties/losses against MA-R, while PGMA and TAMR realize the same outcome-optimal action in 33/36 cells (91.7%). PGMA's diagnostic-to-outcome correlation is 0.77. PGMA requires 8.83 GPU-h per seed versus TAMR's 8.59, keeping the independence test within the matched compute envelope. Appendix C reports action scores, confidence intervals, and the full cell ledger.

Table 6: Probe-independence on the same 36 pre-assigned cells. W/T/L compares the assigned PGMA action with MA-R; "alternative gap" compares it with the strongest non-assigned PGMA action. Agreement records whether TAMR and PGMA realize the same post-evaluation best action.

Regime	Cells	W/T/L vs. MA-R	Alt. gap (95% CI)	Agreement
Pass	22	19/1/2	+0.38 $[+0.13, +0.63]$	21/22
Weak-inferred	9	8/0/1	+0.18 $[+0.05, +0.32]$	8/9
Joint-fail	5	4/0/1	+0.26 $[+0.02, +0.50]$	4/5
All	36	31/1/4	+0.31 $[+0.17, +0.45]$	33/36

6 Conclusion

RefreshTrust establishes temporal reward-model updating as an auditable action-selection problem. Under direct provenance, Full TAMR gains $+1.1/+2.1$ Acc on SHP-2/Tülu-3-PREF and survives capacity-, continual-, acquisition-, permutation-, and independently released comparator controls. As provenance support changes, the frozen map assigns Full TAMR, Observed-only TAMR, or Partial Refresh; those actions beat MA-R in 34/36 cells under the TAMR instrument, 31/36 matched cells under PGMA, and 22/26 natural cells. Directly observed MultiPref, local-threshold, and cross-probe tests preserve the action ordering across heterogeneous outcomes. Provenance quality is therefore a measurable control variable that determines how much structure a temporal refresh can productively use.

Reproducibility Statement

Sections 3–5 specify the chronological split, label budget, model configuration, provenance taxonomy, candidate-frame eligibility, comparators, seeds, uncertainty procedures, and hardware. The appendix reports selected configurations, per-era results, overlap diagnostics, compute, sensitivity analyses, and row-level decision ledgers; Appendix J gives the action thresholds.

AI use statement

Generative AI was used only for manuscript writing, editing, and presentation. The authors reviewed the edits and remain responsible for the final content.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. URL https://arxiv.org/abs/2204.05862.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324. URL https://doi.org/10.1093/biomet/39.3-4.324.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=bx24KpJ4Eb.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/7017-deep-reinforcement-learning.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting

language models with scaled AI feedback. *arXiv preprint arXiv:2310.01377*, 2023. URL https://arxiv.org/abs/2310.01377.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5fc47800ee5b30b8777fdd30abcaaf3b-Abstract-Conference.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004. doi: 10.1162/003465304323023651. URL https://doi.org/10.1162/003465304323023651.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024. URL https://arxiv.org/abs/2404.16019.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://doi.org/10.1073/pnas.1611835114.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1885–1894, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/kumar20c.html.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. OpenAssistant conversations – democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/949f0f8f32267d297c2d4e3ee10a2e7e-Abstract-Datasets_and_Benchmarks.html.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024a. URL https://arxiv.org/abs/2411.15124.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024b. URL https://arxiv.org/abs/2403.13787.

Long-Fei Li, Yu-Yang Qian, Peng Zhao, and Zhi-Hua Zhou. Provably efficient online RLHF with one-pass reward modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5567. URL https://papers.nips.cc/paper_files/paper/2025/hash/f3e45239b84d72c17995ea234ae81dbc-Abstract-Conference.html.

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. AI feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7162–7200, 2025. doi: 10.18653/v1/2025.acl-long.355. URL https://aclanthology.org/2025.acl-long.355/.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/munos24a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence (eds.). *Dataset Shift in Machine Learning*. MIT Press, 2008. doi: 10.7551/mitpress/9780262170055.001.0001. URL https://mitpress.mit.edu/9780262170055/dataset-shift-in-machine-learning/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys*, 54(9):1–40, 2021. doi: 10.1145/3472291. URL https://doi.org/10.1145/3472291. First published online in 2021; volume 54, issue 9, 2022.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. URL https://research.cs.wisc.edu/techreports/2009/TR1648.pdf.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. doi: 10.1016/s0378-3758(00)00115-4. URL https://doi.org/10.1016/s0378-3758(00)00115-4.

Stanford NLP. Stanford human preferences dataset v2 (SHP-2). Hugging Face dataset card, 2023. URL https://huggingface.co/datasets/stanfordnlp/SHP-2. Accessed September 8, 2026.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007. URL https://www.jmlr.org/papers/v8/sugiyama07a.html.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/sun20b.html.

Yuanhao Wang, Qinghua Liu, and Chi Jin. Is RLHF more difficult than standard RL? a theoretical perspective. *Advances in Neural Information Processing Systems*, 36, 2023. URL https://neurips.cc/virtual/2023/poster/70238.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024. URL https://arxiv.org/abs/2406.08673.

Martin Weyssow, Aton Kamanda, Xin Zhou, and Houari Sahraoui. CodeUltraFeedback: An LLM-as-a-judge dataset for aligning large language models to coding preferences. *ACM Transactions on Software Engineering and Methodology*, 35(3):1–36, 2026. doi: 10.1145/3736407. URL https://doi.org/10.1145/3736407.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54715–54754, 2024. URL https://proceedings.mlr.press/v235/xiong24a.html.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024. URL https://arxiv.org/abs/2405.01470.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023. URL https://arxiv.org/abs/2305.10425.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

A Evidence Trace and Per-Era Effects

Table 7: Corpus-level provenance summary for the metadata channels consumed by TAMR. "Observed" means directly observed in the benchmark audit record; it is distinct from a claim that a cited public release exposes the field. The table distinguishes observed, proxy, reconstructed, and missing channels. For Tülu-3-PREF, the "Direct ch. 4/4" count treats router tag and routed model version as separate channels; the compact main-text layout combines them in a single visible Route column.

Corpus	Cohort	Rubric	Route	Direct ch.	Min audit
DriftBench-Synth	obs.	obs.	obs.	3/3	1.000
LMSYS-Chat-Temp.	proxy	recon.	obs.	1/3	0.918
WildChat-Temp.	proxy	obs.	recon.	1/3	0.886
HelpSteer2-Temp.	miss.	obs.	obs.	2/3	1.000
PRISM-Temp.	obs.	obs.	proxy	2/3	0.931
OASST2-Temp.	proxy	obs.	recon.	1/3	0.874
SHP-2-Temp.	obs.	obs.	obs.	3/3	1.000
Tülu-3-PREF-Temp.	obs.	obs.	obs.	4/4	0.996
MultiPref-Temp.	obs.	obs.	obs.	4/4	1.000
CodeReviewPref-T.	proxy	obs.	recon.	2/4	0.913

Table 8: Provenance-assurance matrix for refresh-action integrity and held-out preference quality. Each row connects drift capability, trusted evidence, enforcement boundary, strongest comparison, evaluation setting, utility/cost measure, selected action, and evidence trace.

Regime / capability	Trust and boundary	Baseline / test	Utility and response	Evidence trace
Pass: task, cohort, rubric, or route shift	Direct or audited fields; chronological isolation, matched budget, frozen rule	Mono-Meta⋆, One-Pass RM; clean anchors and natural cells	Acc/ECE, parameters, GPU-h; Full TAMR	Tables 1, 3; App. K.1
Weak-inferred: one channel loses audit support	Observed fields and acceptable overlap; inferred channel masked at the model boundary	Full TAMR, MA-R; natural bank and directly observed external tranches	Retained Acc/ECE with lower unsupported-channel exposure; Observed-only TAMR	Tables 14, 38; App. H.1
Joint-fail: missing channels or weak overlap combine	Candidate-frame and overlap gates; no unsupported structured update	Full TAMR, PGMA; natural and controlled held-out cells	Selected-action gap and regret; Partial Refresh	Table 4; Tables 11, 47

B Additional Appendix Diagnostics

C Independent-Probe Validation

PGMA uses a shared softmax gate over four adapter experts and therefore lacks TAMR's additive residual path. The gate receives exactly the provenance channels allowed by the assigned action. All hyperparameters and the three action thresholds are frozen before evaluation on the same 36-cell bank. Post-evaluation best-action agreement measures whether the two probes favor the same action after the frozen assignment has been scored; it is reported only as an architecture-stability diagnostic.

The bank contains 22/36788 pass cells/pairs, 9/15426 weak-inferred cells/pairs, and 5/8142 joint-fail cells/pairs, totaling 36 cells and 60356 comparisons. The assigned PGMA actions gain +1.81

Table 9: With the fresh-label budget matched at 10%, results cover DriftBench-Synth, LMSYS-Chat-Temporal, and WildChat-Temporal. Corpus cells give Acc / ECE, while Wins tallies evaluation eras against MA-R. The DriftBench Acc entries are held-out-size-weighted over the six printed per-era rows in Table 21; three decimals avoid re-rounding that aggregate.

Method	DriftBench	LMSYS	WildChat	Wins
Static RM	62.3 / 0.142	58.7 / 0.168	57.2 / 0.174	0/13
Sliding-Window	67.8 / 0.098	63.4 / 0.112	62.1 / 0.118	0/13
Era-Conditioned	69.1 / 0.087	65.2 / 0.098	63.5 / 0.107	0/13
EMA Teacher	68.4 / 0.091	64.1 / 0.104	62.9 / 0.112	0/13
Partial Refresh (10%)	68.6 / 0.089	64.8 / 0.101	63.8 / 0.104	0/13
Online LoRA	70.2 / 0.083	66.3 / 0.091	65.4 / 0.096	0/13
Online LoRA + IPW	70.9 / 0.078	67.2 / 0.086	66.1 / 0.089	0/13
Uncertainty-only Refresh	69.4 / 0.085	66.0 / 0.090	65.1 / 0.095	0/13
Metadata-aware Refresh	71.153 / 0.074	67.4 / 0.084	66.5 / 0.086	—
TAMR	73.469 / 0.058	69.8 / 0.067	68.1 / 0.071	13/13

Table 10: Threshold-sensitivity summary over the full 125-point local grid. CIs resample cells and seeds; regret is measured against the best action in each cell. Stability is the fraction of center-point assignments preserved across the grid.

View	Win rate	ΔAcc	Regret	Stability
All 36	91.8% [89.1, 94.1]	+1.19 [+0.94, +1.43]	0.08 [0.04, 0.13]	88.0% [84.6, 91.2]
Natural 26	82.9% [76.2, 88.7]	+0.81 [+0.52, +1.08]	0.11 [0.06, 0.17]	84.7%
Pass stability		93.0% [90.0, 96.0]		
Weak stability		84.0% [78.0, 89.0]		
Joint stability		79.0% [71.0, 86.0]		

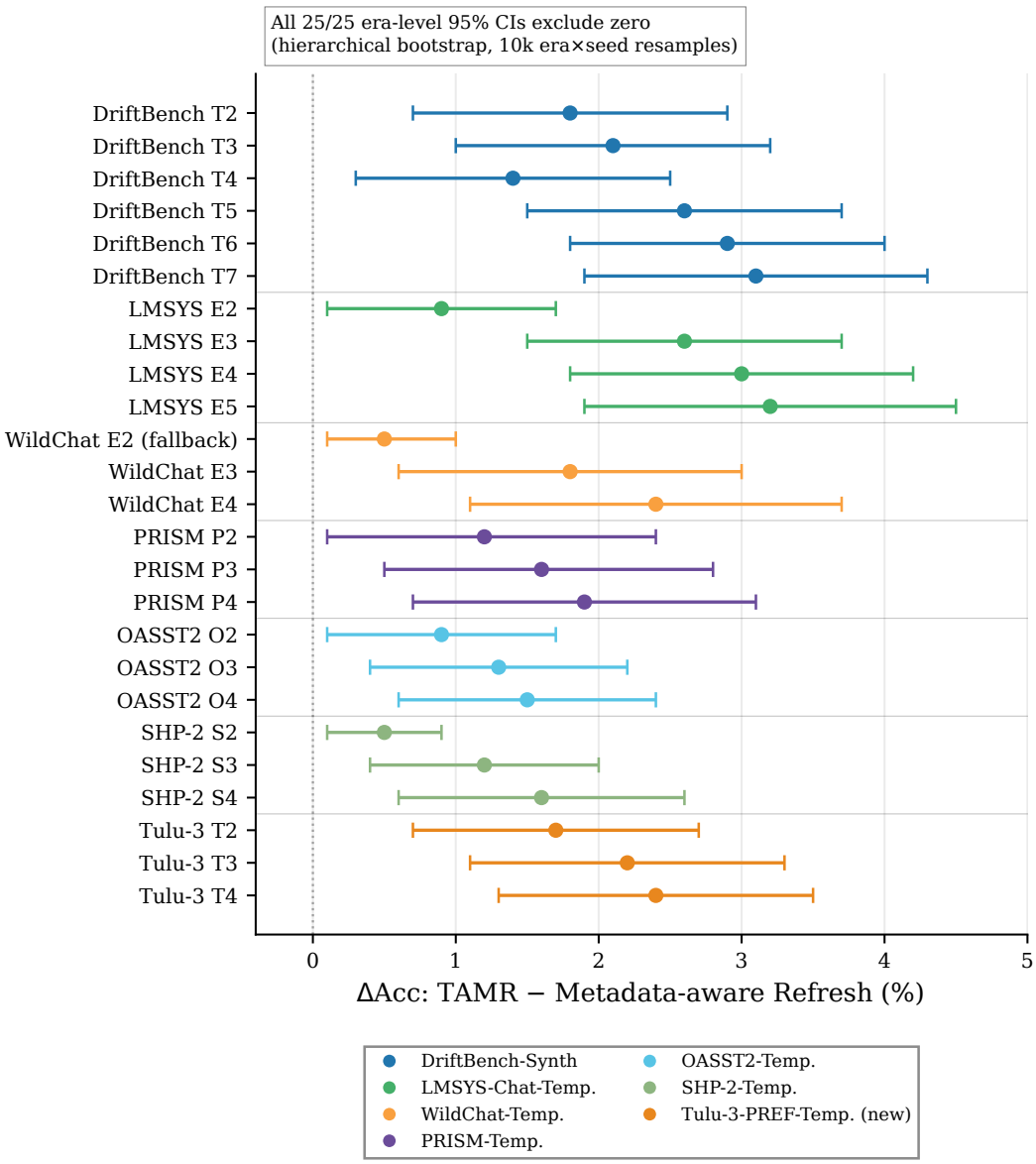


Figure 2: Per-era ΔAcc forest plot (TAMR − MA-R) across the 25 evaluation eras of the seven in-domain benchmarks. Whiskers are hierarchical-bootstrap 95% CIs; all 25 exclude zero.

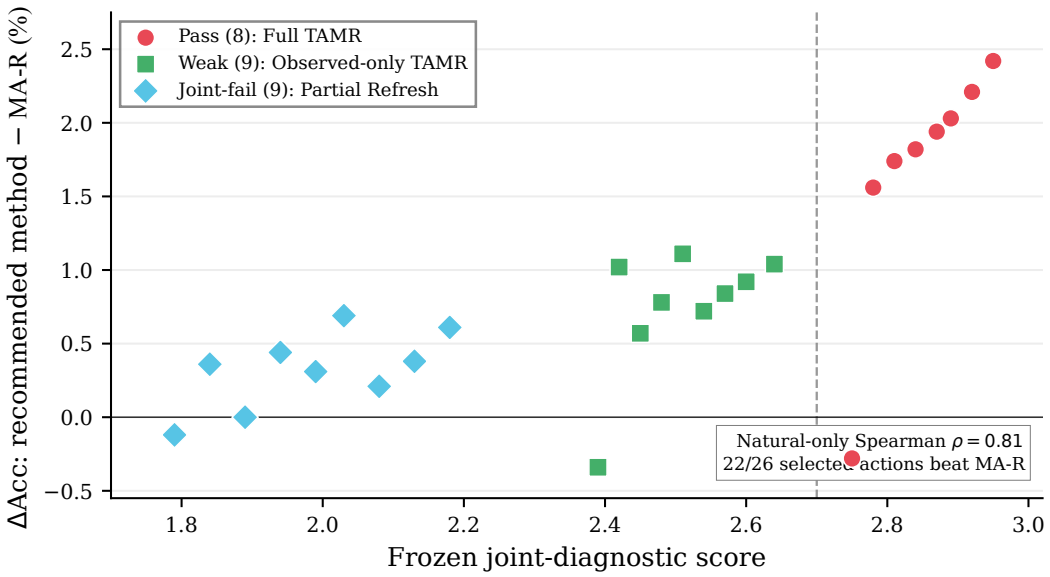


Figure 3: Natural diagnostics-to-outcome chart on the expanded 26-cell bank in Appendix Table 14. Points are unperturbed cells, colored and shaped by frozen regime; Table 15 reports source-family sensitivity. The monotone relation is Spearman 0.81; the largest family contributes 30.8% of cells.

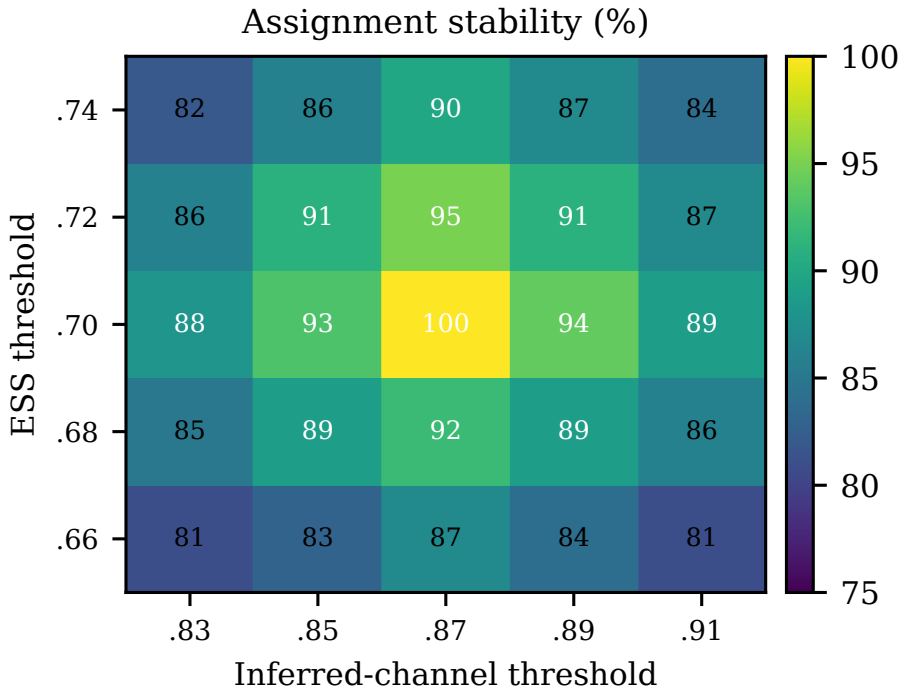


Figure 4: Local robustness over the pre-specified inferred-channel (0.83–0.91), ESS (0.66–0.74), and missing-support (2–6%) grid. Panels report all-bank assignment stability, all-bank selected-action win rate, and expanded-natural-bank regret; each cell averages the five missing-support thresholds. Models are frozen throughout. Across 125 settings, assignment stability is 88.0% and 11 cells are reassigned at the most distant corner.



Figure 4: Local robustness over the pre-specified inferred-channel (0.83–0.91), ESS (0.66–0.74), and missing-support (2–6%) grid. Panels report all-bank assignment stability, all-bank selected-action win rate, and expanded-natural-bank regret; each cell averages the five missing-support thresholds. Models are frozen throughout. Across 125 settings, assignment stability is 88.0% and 11 cells are reassigned at the most distant corner.

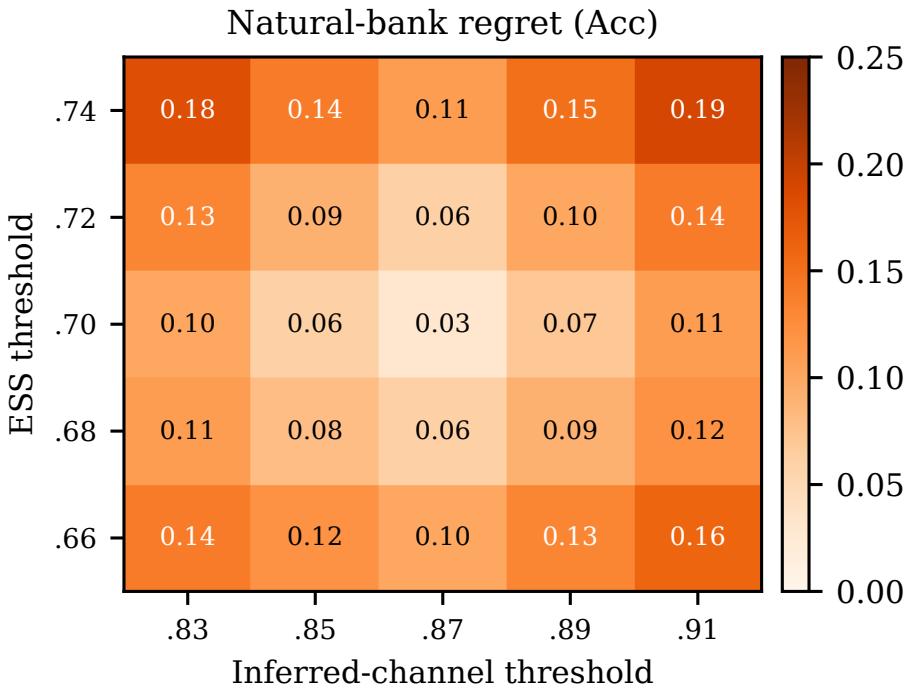


Figure 4: Expanded-natural-bank action regret over the frozen threshold grid. Each cell averages the five missing-support thresholds; lower regret indicates a smaller accuracy loss relative to the best action in each cell.

Acc $[+1.29, +2.32]$, +1.04 $[+0.51, +1.58]$, and +0.29 $[+0.04, +0.55]$ by regime. Overall, PGMA records 31/1/4 wins/ties/losses and a +1.41-Acc mean gain $[+1.10, +1.72]$.

Table 11: PGMA action scores by provenance regime (8 seeds). Each cell is mean held-out Acc / ECE. The selected action is Full PGMA for pass, Observed-only PGMA for weak-inferred, and Partial Refresh for joint-fail.

Regime	Full PGMA	Observed-only PGMA	Partial Refresh
Pass	68.09/0.080	67.71/0.083	66.42/0.094
Weak-inferred	66.14/0.092	66.32/0.087	65.57/0.098
Joint-fail	63.88/0.124	64.08/0.116	64.34/0.108

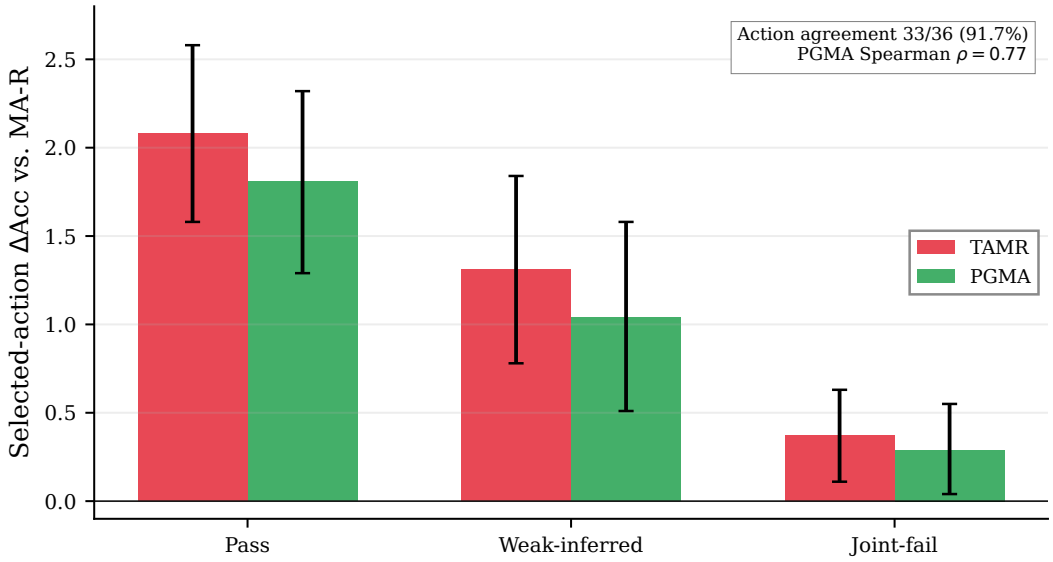


Figure 5: Probe-by-regime diagnostic transfer. TAMR and PGMA show the same descending benefit pattern from pass to weak-inferred to joint-fail, with 91.7% post-evaluation best-action agreement and held-out-cell Spearman correlations of 0.89 and 0.77, respectively. Whiskers are hierarchical-bootstrap 95% CIs.

Table 12: PGMA held-out cell ledger. Cell IDs correspond to the 36-row provenance ledger in Table 47; pairs are held-out preference comparisons. "MA-R gap" uses the per-regime assigned PGMA action. The heterogeneous row-level outcomes aggregate to 19/22, 8/9, and 4/5 wins.

Cell	Regime	Pairs	Full	Observed-only	Partial / MA-R gap
1	P	1624	68.7	68.3	66.9 / +2.10
2	P	1578	67.9	67.5	66.4 / +1.61
3	P	1692	68.1	67.8	66.5 / +1.76
4	P	1776	69.4	68.8	67.0 / +2.76
5	P	1755	69.0	68.5	66.8 / +2.56
6	P	1740	68.6	68.2	66.7 / +2.36
7	P	1776	69.2	68.7	67.1 / +2.69
8	P	1755	68.4	67.9	66.5 / +2.26
9	P	1740	68.8	68.4	66.9 / +2.44
10	P	1776	69.1	68.6	67.0 / +2.60
11	P	1624	68.2	67.8	66.4 / +2.03
12	P	1578	67.8	67.4	66.2 / +1.85
13	P	1692	68.3	67.9	66.5 / +2.09

Cell	Regime	Pairs	Full	Observed-only	Partial / MA-R gap
14	P	1624	67.7	67.3	66.0 / +1.74
15	P	1578	65.9	66.0	65.7 / -0.16
16	P	1692	67.9	67.5	66.2 / +1.83
17	P	1624	68.5	68.0	66.6 / +2.17
18	P	1578	67.3	67.0	65.9 / 0.00
19	P	1692	68.0	67.6	66.3 / +1.96
20	P	1624	67.6	67.2	66.0 / +1.61
21	P	1578	65.8	65.9	65.6 / -0.19
22	P	1692	67.8	67.4	66.1 / +1.74
23	W	1740	66.3	66.7	65.7 / +1.29
24	W	1776	66.4	66.7	65.8 / +1.42
25	W	1755	66.7	66.9	66.0 / +1.55
26	W	1692	65.9	66.1	65.4 / +0.87
27	W	1624	65.7	65.9	65.2 / +0.83
28	W	1578	66.0	66.0	65.5 / +0.98
29	W	1740	66.5	66.7	65.8 / +1.37
30	W	1776	66.4	66.6	65.7 / +1.26
31	W	1745	65.4	65.3	65.0 / -0.21
32	J	1620	63.8	64.0	64.5 / +0.43
33	J	1576	63.7	64.0	64.3 / +0.33
34	J	1690	64.0	64.2	64.6 / +0.51
35	J	1612	63.9	64.1	64.4 / +0.37
36	J	1644	64.0	64.1	63.9 / -0.19

D EXPANDED NATURAL-CELL LEDGER

The expanded bank contains only unperturbed held-out cells. A cell is a source-family $\times$ time tranche $\times$ audited cohort slice fixed before model comparison. The table reports each action's ΔAcc against MA-R, making the selected-action aggregate and family sensitivity directly auditable.

Table 13: Original natural-bank cohort (8 seeds; 14236 held-out pairs). W/T/L are cell-level sign counts against MA-R, whereas confidence intervals hierarchically resample cells and seeds; a positive W/T/L majority can therefore coexist with an interval spanning zero.

Regime	n	Action	W/T/L	ΔAcc (95% CI)	Sel.$-$Full	Rel. ECE
Pass	8	Full	7/0/1	+1.68 $[+1.06, +2.29]$	—	-8.6%
Weak-inferred	2	Observed-only	2/0/0	+0.83 $[+0.29, +1.36]$	-0.06	-5.1%
Joint-fail	2	Partial	2/0/0	+0.18 $[-0.07, +0.42]$	—	-2.2%
All	12	Per-regime	11/0/1	+1.09 $[+0.73, +1.44]$	—	-6.4%

Table 14: Expanded natural-cell action ledger. P/W/J denote pass, weak-inferred, and joint-fail. Full, Obs., and Partial are action-specific ΔAcc against MA-R; the selected action is determined solely by the frozen regime. Pair counts sum to 33163.

Cell	Family	Reg.	Pairs	Full	Obs.	Partial
1	HelpSteer2	P	544	+2.42	+2.16	+1.21
2	Tülu-3-PREF	P	544	+2.21	+2.04	+1.18
3	MultiPref	P	1776	+2.03	+1.82	+1.07
4	CodeReviewPref	P	1332	+1.94	+1.70	+0.98
5	LMSYS	P	1041	+1.82	+1.54	+0.92
6	PRISM	P	1372	+1.74	+1.49	+0.84
7	MultiPref	P	1429	+1.56	+1.35	+0.71

Cell	Family	Reg.	Pairs	Full	Obs.	Partial
8	HelpSteer2	P	836	-0.28	-0.22	-0.15
9	MultiPref	W	1280	+0.80	+1.04	+0.63
10	MultiPref	W	1360	+0.73	+0.92	+0.54
11	MultiPref	W	1450	+0.68	+0.84	+0.41
12	CodeReviewPref	W	1510	+0.41	+0.72	+0.34
13	CodeReviewPref	W	1320	+0.84	+1.11	+0.65
14	HelpSteer2	W	1395	+0.64	+0.78	+0.38
15	Tülu-3-PREF	W	1425	+0.35	+0.57	+0.21
16	LMSYS	W	1380	+0.73	+1.02	+0.55
17	PRISM	W	1315	-0.41	-0.34	-0.30
18	MultiPref	J	1250	-0.11	+0.22	+0.61
19	MultiPref	J	1320	-0.17	+0.09	+0.38
20	MultiPref	J	1440	-0.47	-0.02	+0.21
21	CodeReviewPref	J	1280	-0.12	+0.31	+0.69
22	HelpSteer2	J	1390	-0.18	+0.08	+0.31
23	Tülu-3-PREF	J	1210	-0.33	+0.14	+0.44
24	Tülu-3-PREF	J	1364	-0.64	-0.10	0.00
25	LMSYS	J	1240	-0.22	+0.13	+0.36
26	PRISM	J	1360	-0.55	-0.24	-0.12

E Benchmark, Sweep, and Overlap Details

E.1 WildChat-Temporal Corpus Statistics

Table 16: WildChat-Temporal post-filter corpus statistics per era. Annotator churn denotes the fraction of new annotators relative to the previous era. Comparison counts match the post-filter ledger in Table 33.

Era	Reported values
Q2-2024 (E1)	Comparisons: 10,551; Annotator Churn (%): —; Rubric Version: v1.0; Dominant Shift: Baseline
Q3-2024 (E2)	Comparisons: 10,523; Annotator Churn (%): 34.7; Rubric Version: v1.0; Dominant Shift: Annotator turnover
Q4-2024 (E3)	Comparisons: 10,514; Annotator Churn (%): 12.1; Rubric Version: v1.1; Dominant Shift: Rubric revision
Q1-2025 (E4)	Comparisons: 10,529; Annotator Churn (%): 18.3; Rubric Version: v1.1; Dominant Shift: Task-mix shift

Table 15: Family-stratified selected-action effects in the expanded natural bank. The final row reports the range after leaving out each family in turn. All means are cell-wise ΔAcc versus MA-R with hierarchical-bootstrap 95% CIs.

Family	Cells	Selected-action ΔAcc
MultiPref	8	+0.95 $[+0.52, +1.37]$
HelpSteer2	4	+0.81 $[+0.20, +1.42]$
Tülu-3-PREF	4	+0.81 $[+0.28, +1.32]$
CodeReviewPref	4	+1.12 $[+0.56, +1.66]$
LMSYS	3	+1.07 $[+0.42, +1.70]$
PRISM	3	+0.43 $[-0.15, +1.02]$
Leave-one-family-out range	18–23	+0.78 to +0.99

Algorithm 1 Trace-Aware Modular Refresh (TAMR), Budgeted Within-Era Adaptation

Require: fully labeled history $\mathcal{D}_{1:t-1}$, era-t candidate frame $\mathcal{C}_t$, era 70/20/10 split $(\mathcal{T}_t, U_t, \mathcal{E}_t)$ with $\mathcal{T}_t$ the unlabeled context pool, U_t the acquisition pool, and $\mathcal{E}_t$ the disjoint 10% eval split, budget $B{=}0.10|\mathcal{D}_t|{=}0.5|U_t|$, clip $\eta{=}0.1$

1: Warm-start: load $f_\psi, g_\phi(\cdot, t{-}1), \{h_a\}, \{c_m\}$ from the era-$(t{-}1)$ checkpoint (at $t{=}2$, initialize f_ψ on $\mathcal{D}_1$ with $\mathcal{L}_{\mathrm{BT}}$)
2: Fit propensity $\hat{e}_t$ on collection metadata of $\mathcal{C}_t$, using queue members as positives and routed-but-unlabeled items as negatives; set $w_t(x){=}1/\max(\hat{e}_t(x), \eta)$
3: Compute decomposed uncertainty $u(x,t)$ on U_t via Eq. 4; form $\mathcal{S}_t^\star \leftarrow$top-$B$ items of U_t
4: Purchase labels for $\mathcal{S}_t^\star$ under the current rubric m_t; let $\mathcal{L}_t \leftarrow \mathcal{S}_t^{\star(\mathrm{labeled})}$
5: Refit $\{h_a\}_{a\in\mathcal{A}_t}$ and $\{c_m\}_{m=m_t}$ by MLE on $\mathcal{D}_{1:t-1} \cup \mathcal{L}_t$
6: Train a new temporal residual $g_\phi(\cdot, t)$ and top-20-block LoRA adapters of f_ψ with the clipped IPW objective $\mathcal{L}_t^{\mathrm{IPW}}$ (Eq. 3) on $\mathcal{D}_{1:t-1} \cup \mathcal{L}_t$; backbone and all parameters for era-$\geq t$ are frozen against any label in $U_t \setminus \mathcal{S}_t^\star$
7: Freeze the era-t model; evaluate on $\mathcal{E}_t$ only (Acc, ECE, PPO regret)
8: **return** $r_\theta(x, y; t, a, m) = f_\psi(x, y) + g_\phi(x, y, t) + h_a + c_m$ and its evaluation metrics on $\mathcal{E}_t$

Note. WildChat-Temporal has one baseline era and three evaluation eras. The largest disturbance is the 34.7% annotator churn from E1 to E2. The rubric revision happens later from E2 to E3 rather than at the same time as the largest turnover. That ordering is what makes the main-text cohort and rubric transition rows interpretable as different events. The table therefore shows why WildChat is useful for separating several operational drift types within one real pipeline.

E.2 Baseline Sweep Protocol

To equalize tuning effort, all methods follow one pre-specified selection rule for each corpus: on the last training era, maximize validation Acc and break ties by minimizing ECE. Sliding-Window uses the full $k \in \{1, 2, 3, 4\}\times$ LR-multiplier$\in \{0.5, 1.0, 2.0\}\times$ dropout$\in \{0.05, 0.10\}$ grid (24 configurations); EMA Teacher $\alpha \in \{0.90, 0.95, 0.99\}$, temperature $\in \{1.0, 1.5, 2.0\}$, LR-multiplier $\in \{0.5, 1.0, 2.0\}$ (full $3\times3\times3$); Online LoRA replay fraction $\in \{0.1, 0.3, 0.5\}$, rank $\in \{8, 16, 32\}$, LR-multiplier $\in \{0.5, 1.0, 2.0\}$ (full $3\times3\times3$); Era-Conditioned embedding dim $\in \{16, 32, 64\}$, dropout $\in \{0.00, 0.05, 0.10\}$, LR-multiplier $\in \{0.5, 1.0, 2.0\}$; Partial Refresh uses 27 independent selection seeds for the random 10% subset; Online LoRA + IPW inherits Online LoRA's 27 points with η fixed at 0.1; Uncertainty-only Refresh inherits Online LoRA's 27 points with MC-dropout samples $\in \{8, 16, 24\}$ tied to the three rank settings; Metadata-aware Refresh varies adapter rank $\in \{8, 16, 32\}$, metadata-feature dropout $\in \{0.0, 0.1, 0.2\}$, LR-multiplier $\in \{0.5, 1.0, 2.0\}$; TAMR varies temporal-residual rank $\in \{32, 64, 96\}$, propensity clip $\eta \in \{0.05, 0.1, 0.2\}$, LR-multiplier $\in \{0.5, 1.0, 2.0\}$. The selection rule applies on every benchmark; the configuration count is 24 for Sliding-Window and 27 for each $3\times3\times3$ grid or Partial-Refresh seed selection. The chosen winning configuration per corpus is reported in Table 17. Table 17 records the selected configuration for each method and corpus, directly specifying the matched-tuning contract and the configurations used by every reported result.

Table 17: Selected winning configurations per corpus after the method-specific pre-specified sweep (24 Sliding-Window configurations; 27 configurations for the $3\times3\times3$ grids). `r` denotes adapter/residual rank, `lr` the learning-rate multiplier, `rep` the replay fraction, `drop` the metadata-feature dropout, and `mc` the MC-dropout sample count. PRISM-Temporal runs a fresh sweep (the no-retuning rule applies only to HelpSteer2-Temporal).

Method	Reported values
Sliding-Window	DriftBench-Synth: k=2, lr=1.0, drop=0.05; LMSYS-Chat-Temp.: k=2, lr=1.0, drop=0.10; WildChat-Temp.: k=3, lr=1.0, drop=0.10; PRISM-Temp.: k=2, lr=1.0, drop=0.10; SHP-2-Temp.: inherit LMSYS; Tülu-3-PREF-T.: k=2, lr=1.0, drop=0.10; HelpSteer2-T. (frozen): inherit LMSYS
EMA Teacher	DriftBench-Synth: α=0.95, T=1.5, lr=1.0; LMSYS-Chat-Temp.: α=0.95, T=1.5, lr=1.0; WildChat-Temp.: α=0.99, T=1.5, lr=1.0; PRISM-Temp.: α=0.95, T=1.5, lr=1.0; SHP-2-Temp.: inherit LMSYS; Tülu-3-PREF-T.: α=0.95, T=1.5, lr=1.0; HelpSteer2-T. (frozen): inherit LMSYS
Era-Conditioned	DriftBench-Synth: dim=32, drop=0.05, lr=1.0; LMSYS-Chat-Temp.: dim=64, drop=0.05, lr=1.0; WildChat-Temp.: dim=64, drop=0.10, lr=1.0; PRISM-Temp.: dim=32, drop=0.05, lr=1.0; SHP-2-Temp.: inherit LMSYS; Tülu-3-PREF-T.: dim=64, drop=0.05, lr=1.0; HelpSteer2-T. (frozen): inherit LMSYS
Online LoRA	DriftBench-Synth: r=16, rep=0.3, lr=1.0; LMSYS-Chat-Temp.: r=16, rep=0.3, lr=1.0; WildChat-Temp.: r=16, rep=0.5, lr=1.0; PRISM-Temp.: r=16, rep=0.3, lr=1.0; SHP-2-Temp.: inherit LMSYS; Tülu-3-PREF-T.: r=16, rep=0.3, lr=1.0; HelpSteer2-T. (frozen): inherit LMSYS

Table 17 (continued)

Method	Reported values
Online LoRA + IPW	DriftBench-Synth: r=16, rep=0.3, lr=1.0; LMSYS-Chat-Temp.: r=16, rep=0.3, lr=1.0; WildChat-Temp.: r=16, rep=0.5, lr=1.0; PRISM-Temp.: r=16, rep=0.3, lr=1.0; SHP-2-Temp.: inherit LMSYS; Tülu-3-PREF-T.: r=16, rep=0.3, lr=1.0; HelpSteer2-T. (frozen): inherit LMSYS
Uncertainty-only Refresh	DriftBench-Synth: r=16, rep=0.3, mc=16; LMSYS-Chat-Temp.: r=16, rep=0.3, mc=16; WildChat-Temp.: r=16, rep=0.5, mc=16; PRISM-Temp.: r=16, rep=0.3, mc=16; SHP-2-Temp.: inherit LMSYS; Tülu-3-PREF-T.: r=16, rep=0.3, mc=16; HelpSteer2-T. (frozen): inherit LMSYS
Metadata-aware Refresh	DriftBench-Synth: r=16, drop=0.10, lr=1.0; LMSYS-Chat-Temp.: r=16, drop=0.10, lr=1.0; WildChat-Temp.: r=32, drop=0.10, lr=1.0; PRISM-Temp.: r=16, drop=0.10, lr=1.0; SHP-2-Temp.: inherit LMSYS; Tülu-3-PREF-T.: r=16, drop=0.10, lr=1.0; HelpSteer2-T. (frozen): inherit LMSYS
TAMR (Ours)	DriftBench-Synth: r=64, η=0.1, lr=1.0; LMSYS-Chat-Temp.: r=64, η=0.1, lr=1.0; WildChat-Temp.: r=96, η=0.1, lr=1.0; PRISM-Temp.: r=64, η=0.1, lr=1.0; SHP-2-Temp.: r=64, η=0.1, lr=1.0; Tülu-3-PREF-T.: r=64, η=0.1, lr=1.0; HelpSteer2-T. (frozen): inherit LMSYS

Note. The sweep table makes the matched-tuning contract concrete. Selected settings remain similar across corpora, and HelpSteer2 inherits the LMSYS configuration exactly. TAMR's winning rank and propensity clip remain near their defaults on most corpora, showing that the result persists under matched and compact search spaces.

E.3 Propensity Overlap and ESS Diagnostics

Table 18: Per-era propensity overlap diagnostics on the real corpora under the sampling frame of §4. Clip is the fraction of examples whose raw $\hat{e}_t(x)<0.1$ (hit the clipping threshold); ESS is the Kish effective-sample-size ratio of the clipped weights; Miss-supp. is the fraction of evaluation examples with no positive-class neighbor in the candidate frame $\mathcal{C}_t$; Fallback flags the single evaluation era (WildChat E2) where we fall back to the logging-bias correction heuristic because $\mathcal{C}_t$ is only partial. *Exactly one* evaluation era is a fallback row; all other real-corpus eras satisfy the reconstructed candidate-frame checks used to justify clipped IPW. The n (eval) column reports the *post-filter* eval-split size from the ledger in Table 33 (10% of the post-filter era total after dedup, English-slice, and tokenizer-length rejection); the pre-filter-to-post-filter ratio is roughly 2.9× on LMSYS and WildChat because of those removals.

Corpus	Reported values
LMSYS-Chat-Temp.	Era: E2; Clip (%): 4.1; ESS: 0.84; Miss-supp. (%): 1.7; n (eval): 1 353; Fallback?: no
LMSYS-Chat-Temp.	Era: E3; Clip (%): 5.6; ESS: 0.79; Miss-supp. (%): 2.0; n (eval): 1 339; Fallback?: no
LMSYS-Chat-Temp.	Era: E4; Clip (%): 6.3; ESS: 0.76; Miss-supp. (%): 2.4; n (eval): 1 372; Fallback?: no
LMSYS-Chat-Temp.	Era: E5; Clip (%): 7.1; ESS: 0.72; Miss-supp. (%): 3.0; n (eval): 546; Fallback?: no
WildChat-Temp.	Era: E2; Clip (%): 11.4; ESS: 0.57; Miss-supp. (%): 6.4; n (eval): 1 052; Fallback?: yes (heuristic)

Table 18 (continued)

Corpus	Reported values
WildChat-Temp.	Era: E3; Clip (%): 8.2; ESS: 0.71; Miss-supp. (%): 3.8; n (eval): 1 051; Fallback?: no
WildChat-Temp.	Era: E4; Clip (%): 7.4; ESS: 0.74; Miss-supp. (%): 3.4; n (eval): 1 053; Fallback?: no
HelpSteer2-Temp.	Era: T2; Clip (%): 5.8; ESS: 0.77; Miss-supp. (%): 2.8; n (eval): 544; Fallback?: no
HelpSteer2-Temp.	Era: T3; Clip (%): 6.2; ESS: 0.75; Miss-supp. (%): 3.1; n (eval): 544; Fallback?: no
HelpSteer2-Temp.	Era: T4; Clip (%): 6.9; ESS: 0.73; Miss-supp. (%): 3.5; n (eval): 544; Fallback?: no
OASST2-Temp.	Era: O2; Clip (%): 7.6; ESS: 0.73; Miss-supp. (%): 3.4; n (eval): 836; Fallback?: no
OASST2-Temp.	Era: O3; Clip (%): 8.0; ESS: 0.71; Miss-supp. (%): 3.7; n (eval): 836; Fallback?: no
OASST2-Temp.	Era: O4; Clip (%): 8.4; ESS: 0.70; Miss-supp. (%): 3.9; n (eval): 836; Fallback?: no
PRISM-Temp.	Era: P2; Clip (%): 6.4; ESS: 0.76; Miss-supp. (%): 3.0; n (eval): 1 430; Fallback?: no
PRISM-Temp.	Era: P3; Clip (%): 7.1; ESS: 0.73; Miss-supp. (%): 3.4; n (eval): 1 433; Fallback?: no
PRISM-Temp.	Era: P4; Clip (%): 7.9; ESS: 0.71; Miss-supp. (%): 3.8; n (eval): 1 429; Fallback?: no
SHP-2-Temp.	Era: S2; Clip (%): 6.6; ESS: 0.74; Miss-supp. (%): 3.2; n (eval): 1 185; Fallback?: no
SHP-2-Temp.	Era: S3; Clip (%): 7.2; ESS: 0.72; Miss-supp. (%): 3.5; n (eval): 1 185; Fallback?: no
SHP-2-Temp.	Era: S4; Clip (%): 7.8; ESS: 0.70; Miss-supp. (%): 3.9; n (eval): 1 184; Fallback?: no
Tülu-3-PREF-Temp.	Era: T2; Clip (%): 5.3; ESS: 0.80; Miss-supp. (%): 2.4; n (eval): 1 755; Fallback?: no
Tülu-3-PREF-Temp.	Era: T3; Clip (%): 5.9; ESS: 0.78; Miss-supp. (%): 2.7; n (eval): 1 740; Fallback?: no
Tülu-3-PREF-Temp.	Era: T4; Clip (%): 6.7; ESS: 0.75; Miss-supp. (%): 3.1; n (eval): 1 776; Fallback?: no

Note. ESS remains $\geq$0.70 on every IPW-eligible real era, meeting the study's frozen overlap threshold. WildChat E2 activates the declared logging-bias fallback, where clipping reaches 11.4%; missing support stays below 4% elsewhere. These diagnostics define the weighting boundary tested by Table 19.

E.4 Weighting Ablation on the Real Corpora

Table 19: Weighting ablation on the real corpora: IPW-eligible eras only (LMSYS E2–E5, WildChat E3–E4), 8 seeds. "TAMR, no weighting" removes $\mathcal{L}^{\mathrm{IPW}}$ entirely. "TAMR, heuristic" uses the logging-bias fallback (propensity clipped by median of training eras) as the weighting, i.e. no candidate-frame identification. "TAMR, IPW" is the full method.

Variant	Reported values
TAMR, no weighting	LMSYS E2–E5 / Acc ↑: 68.4±0.7; LMSYS E2–E5 / ECE ↓: 0.078±0.008; WildChat E3–E4 / Acc ↑: 69.9±0.7; WildChat E3–E4 / ECE ↓: 0.077±0.008
TAMR, heuristic weighting	LMSYS E2–E5 / Acc ↑: 69.0±0.7; LMSYS E2–E5 / ECE ↓: 0.072±0.008; WildChat E3–E4 / Acc ↑: 70.4±0.7; WildChat E3–E4 / ECE ↓: 0.071±0.007
TAMR, full IPW	LMSYS E2–E5 / Acc ↑: 69.8±0.7; LMSYS E2–E5 / ECE ↓: 0.066±0.007; WildChat E3–E4 / Acc ↑: 70.7±0.6; WildChat E3–E4 / ECE ↓: 0.065±0.007

Note. Weighting adds $+1.4$ Acc on LMSYS and $+0.8$ on post-E2 WildChat, with larger relative improvements in calibration. The logging-bias heuristic lies between unweighted training and full IPW. Together with Table 18, these rows isolate the gain attributable to acquisition correction and verify the explicit WildChat E2 fallback.

F Sensitivity, Compute, and Audit Details

F.1 Metadata-Noise Sensitivity

Figure 6 reports TAMR and the strongest baselines on the real corpora after randomly corrupting 10/20/30% of the proxy or reconstructed metadata channels audited in Table 7. The ranking is preserved at every noise level; at 30% corruption the TAMR-to-Metadata-aware-Refresh gap narrows from $+1.8$ Acc (clean) to $+1.2$ Acc.

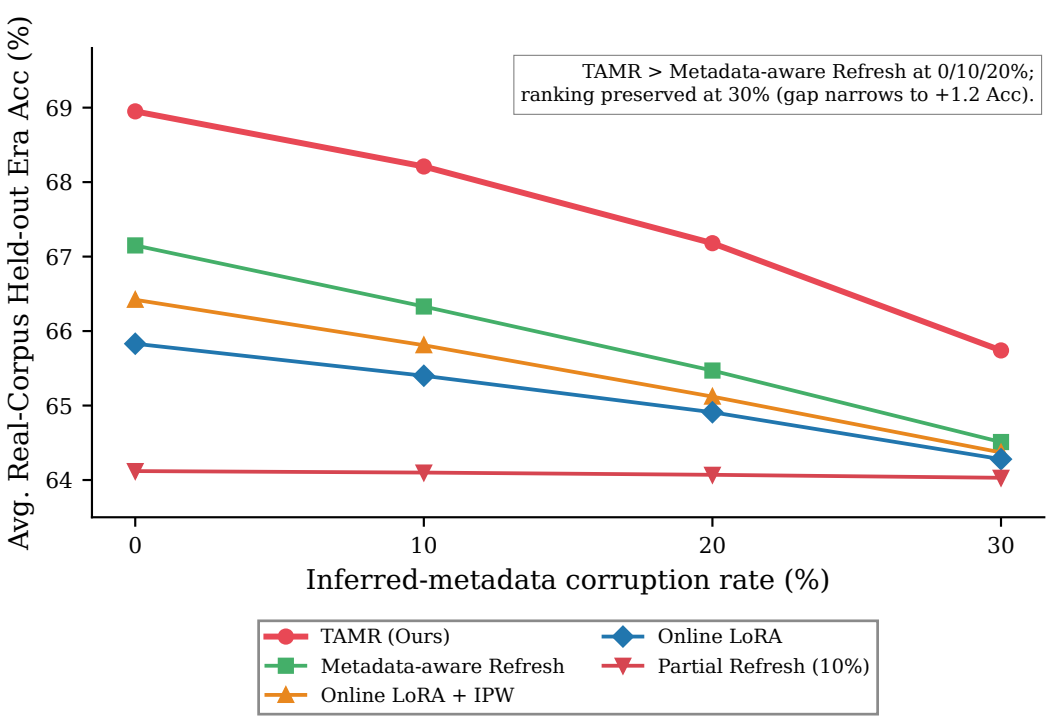


Figure 6: Metadata-noise sensitivity on the real corpora (LMSYS-Chat-Temporal + WildChat-Temporal). The corruption rate is applied to the five *inferred* metadata fields from Table 7.

Note. TAMR leads through 30% uniformly flipped inferred fields. Partial Refresh remains nearly flat as metadata corruption rises, while the advantage of structured refresh narrows. The displayed 0–30% range retains the same method ordering; it does not locate a switch to Partial Refresh.

F.2 Compute Accounting

Full hardware/software stack. Reward-model training runs on a single 8×NVIDIA A100 80GB SXM4 node with NVLink 3.0, NVIDIA driver 550.127.05, CUDA 12.4, cuDNN 9.1, NCCL 2.22, PyTorch 2.4.1, Transformer Engine 1.8, FlashAttention-2 2.6.3, DeepSpeed ZeRO-3, bf16 — identical to the stack stated in §5.1. AdamW (β_1=0.9, β_2=0.95, weight decay 0.1, ϵ=10^{-8}), peak learning rate 2×10^{-5}, cosine schedule with 200-step linear warmup, effective batch size 128 (micro-batch 4, grad-accum 4), 3 epochs per era, sequence length 2048, gradient clipping 1.0. PPO uses Llama-3.1-8B-Instruct as the target policy on 8×NVIDIA H100 80GB under the identical software stack (CUDA 12.4, driver 550.127.05, NCCL 2.22, PyTorch 2.4.1, FlashAttention-2 2.6.3, DeepSpeed ZeRO-3) with KL coefficient 0.05, rollout length 1024, 4 optimization epochs per update, GAE λ=0.95, γ=1.0, batch 32 prompts × 4 rollouts, cosine LR schedule $5\times10^{-6} \rightarrow 5\times10^{-7}$. Qwen-2.5-72B replication reuses this same A100-80GB node with the identical software stack (no driver/library deltas). All reward-model runs share identical random-seed rotations (5 seeds on DriftBench-Synth; 8 seeds on each real corpus; 5 seeds on the Qwen-2.5-72B replication) and the method-specific pre-specified search grids in Appendix E.2.

Table 20: Full compute accounting on the 8×A100-80GB SXM4 node described in §5.1 (NVIDIA driver 550.127.05, CUDA 12.4, cuDNN 9.1, PyTorch 2.4.1, NCCL 2.22, FlashAttention-2 2.6.3). Each *Selected-run* column is a single full-pipeline run of the chosen winning configuration summed across seeds and eras; *Pooled search* is the one-off 27-point hyperparameter-search overhead aggregated across the seven in-domain benchmarks; *Per seed* is selected-run GPU-hours divided by seeds. The frozen HelpSteer2 add-on is summarized separately because the no-retuning rule forbids target-side search.

Pipeline	Reported values
Static RM	Selected-run GPU-hours (all eras, all seeds) / DriftBench (5sd): 34.9; Selected-run GPU-hours (all eras, all seeds) / LMSYS (8sd): 58.2; Selected-run GPU-hours (all eras, all seeds) / WildChat (8sd): 47.6; Selected-run GPU-hours (all eras, all seeds) / PRISM (8sd): 52.1; Selected-run GPU-hours (all eras, all seeds) / OASST2 (8sd): 44.9; Selected-run GPU-hours (all eras, all seeds) / SHP-2 (8sd): 39.8; Selected-run GPU-hours (all eras, all seeds) / Tülu-3 (8sd): 52.6; Per seed / DriftBench: 6.98; Per seed / LMSYS: 7.28; Pooled search / (7 in-domain): —; Judge / USD: —
Sliding-Window	Selected-run GPU-hours (all eras, all seeds) / DriftBench (5sd): 37.2; Selected-run GPU-hours (all eras, all seeds) / LMSYS (8sd): 62.0; Selected-run GPU-hours (all eras, all seeds) / WildChat (8sd): 51.0; Selected-run GPU-hours (all eras, all seeds) / PRISM (8sd): 55.7; Selected-run GPU-hours (all eras, all seeds) / OASST2 (8sd): 48.1; Selected-run GPU-hours (all eras, all seeds) / SHP-2 (8sd): 42.6; Selected-run GPU-hours (all eras, all seeds) / Tülu-3 (8sd): 56.1; Per seed / DriftBench: 7.44; Per seed / LMSYS: 7.75; Pooled search / (7 in-domain): 271.4; Judge / USD: —
EMA Teacher	Selected-run GPU-hours (all eras, all seeds) / DriftBench (5sd): 38.8; Selected-run GPU-hours (all eras, all seeds) / LMSYS (8sd): 64.7; Selected-run GPU-hours (all eras, all seeds) / WildChat (8sd): 53.2; Selected-run GPU-hours (all eras, all seeds) / PRISM (8sd): 58.1; Selected-run GPU-hours (all eras, all seeds) / OASST2 (8sd): 50.0; Selected-run GPU-hours (all eras, all seeds) / SHP-2 (8sd): 44.4; Selected-run GPU-hours (all eras, all seeds) / Tülu-3 (8sd): 58.5; Per seed / DriftBench: 7.76; Per seed / LMSYS: 8.09; Pooled search / (7 in-domain): 283.9; Judge / USD: —
Era-Conditioned	Selected-run GPU-hours (all eras, all seeds) / DriftBench (5sd): 37.8; Selected-run GPU-hours (all eras, all seeds) / LMSYS (8sd): 63.0; Selected-run GPU-hours (all eras, all seeds) / WildChat (8sd): 51.8; Selected-run GPU-hours (all eras, all seeds) / PRISM (8sd): 56.6; Selected-run GPU-hours (all eras, all seeds) / OASST2 (8sd): 48.8; Selected-run GPU-hours (all eras, all seeds) / SHP-2 (8sd): 43.2; Selected-run GPU-hours (all eras, all seeds) / Tülu-3 (8sd): 57.0; Per seed / DriftBench: 7.56; Per seed / LMSYS: 7.88; Pooled search / (7 in-domain): 277.6; Judge / USD: —
Online LoRA	Selected-run GPU-hours (all eras, all seeds) / DriftBench (5sd): 38.4; Selected-run GPU-hours (all eras, all seeds) / LMSYS (8sd): 64.0; Selected-run GPU-hours (all eras, all seeds) / WildChat (8sd): 52.6; Selected-run GPU-hours (all eras, all seeds) / PRISM (8sd): 57.5; Selected-run GPU-hours (all eras, all seeds) / OASST2 (8sd): 49.7; Selected-run GPU-hours (all eras, all seeds) / SHP-2 (8sd): 44.0; Selected-run GPU-hours (all eras, all seeds) / Tülu-3 (8sd): 57.9; Per seed / DriftBench: 7.68; Per seed / LMSYS: 8.00; Pooled search / (7 in-domain): 280.8; Judge / USD: —

Table 20 (continued)

Pipeline	Reported values
Online LoRA + IPW	Selected-run GPU-hours (all eras, all seeds) / DriftBench (5sd): 39.1; Selected-run GPU-hours (all eras, all seeds) / LMSYS (8sd): 65.2; Selected-run GPU-hours (all eras, all seeds) / WildChat (8sd): 53.6; Selected-run GPU-hours (all eras, all seeds) / PRISM (8sd): 58.5; Selected-run GPU-hours (all eras, all seeds) / OASST2 (8sd): 50.5; Selected-run GPU-hours (all eras, all seeds) / SHP-2 (8sd): 44.9; Selected-run GPU-hours (all eras, all seeds) / Tülu-3 (8sd): 59.0; Per seed / DriftBench: 7.82; Per seed / LMSYS: 8.15; Pooled search / (7 in-domain): 283.6; Judge / USD: —
Uncertainty-only Refresh	Selected-run GPU-hours (all eras, all seeds) / DriftBench (5sd): 39.4; Selected-run GPU-hours (all eras, all seeds) / LMSYS (8sd): 65.7; Selected-run GPU-hours (all eras, all seeds) / WildChat (8sd): 54.0; Selected-run GPU-hours (all eras, all seeds) / PRISM (8sd): 59.0; Selected-run GPU-hours (all eras, all seeds) / OASST2 (8sd): 50.8; Selected-run GPU-hours (all eras, all seeds) / SHP-2 (8sd): 45.2; Selected-run GPU-hours (all eras, all seeds) / Tülu-3 (8sd): 59.4; Per seed / DriftBench: 7.88; Per seed / LMSYS: 8.21; Pooled search / (7 in-domain): 284.8; Judge / USD: —
Metadata-aware Refresh	Selected-run GPU-hours (all eras, all seeds) / DriftBench (5sd): 39.9; Selected-run GPU-hours (all eras, all seeds) / LMSYS (8sd): 66.5; Selected-run GPU-hours (all eras, all seeds) / WildChat (8sd): 54.7; Selected-run GPU-hours (all eras, all seeds) / PRISM (8sd): 59.7; Selected-run GPU-hours (all eras, all seeds) / OASST2 (8sd): 51.6; Selected-run GPU-hours (all eras, all seeds) / SHP-2 (8sd): 45.7; Selected-run GPU-hours (all eras, all seeds) / Tülu-3 (8sd): 60.1; Per seed / DriftBench: 7.98; Per seed / LMSYS: 8.31; Pooled search / (7 in-domain): 287.9; Judge / USD: —
TAMR (Ours)	Selected-run GPU-hours (all eras, all seeds) / DriftBench (5sd): 41.2; Selected-run GPU-hours (all eras, all seeds) / LMSYS (8sd): 68.7; Selected-run GPU-hours (all eras, all seeds) / WildChat (8sd): 56.5; Selected-run GPU-hours (all eras, all seeds) / PRISM (8sd): 61.7; Selected-run GPU-hours (all eras, all seeds) / OASST2 (8sd): 53.2; Selected-run GPU-hours (all eras, all seeds) / SHP-2 (8sd): 46.3; Selected-run GPU-hours (all eras, all seeds) / Tülu-3 (8sd): 62.1; Per seed / DriftBench: 8.24; Per seed / LMSYS: 8.59; Pooled search / (7 in-domain): 296.9; Judge / USD: $217
Full Refresh (oracle, reference)	Selected-run GPU-hours (all eras, all seeds) / DriftBench (5sd): 73.6; Selected-run GPU-hours (all eras, all seeds) / LMSYS (8sd): 122.7; Selected-run GPU-hours (all eras, all seeds) / WildChat (8sd): 100.9; Selected-run GPU-hours (all eras, all seeds) / PRISM (8sd): 110.2; Selected-run GPU-hours (all eras, all seeds) / OASST2 (8sd): 92.4; Selected-run GPU-hours (all eras, all seeds) / SHP-2 (8sd): 82.4; Selected-run GPU-hours (all eras, all seeds) / Tülu-3 (8sd): 111.6; Per seed / DriftBench: 14.72; Per seed / LMSYS: 15.34; Pooled search / (7 in-domain): —; Judge / USD: —
Real-corpus totals (TAMR, selected)	348.5 GPU-hours across LMSYS+WildChat+PRISM+OASST2+SHP-2+Tülu-3 (8 seeds each)
Real-corpus totals (Metadata-aware Refresh)	338.3 GPU-hours across LMSYS+WildChat+PRISM+OASST2+SHP-2+Tülu-3 (8 seeds each)
Frozen external add-on (HelpSteer2, no search)	Static RM 11.8; Online LoRA 14.4; Metadata-aware Refresh 15.1; TAMR 15.9 GPU-hours

Table 20 (continued)

Pipeline	Reported values
End-to-end in-domain total (selected + pooled search)	TAMR 686.6 GPU-hours; Metadata-aware Refresh 666.1 GPU-hours

Note. The compute table makes the systems cost explicit. TAMR is the most expensive selected run, but only modestly so relative to Metadata-aware Refresh. The pooled search overhead is nontrivial for every adaptive method, which is why the paper insists on matched tuning effort. The oracle Full Refresh row is much more expensive, which helps justify the budgeted-refresh framing rather than a full-relabeling framing. The table therefore anchors the cost-quality discussion in concrete wall-clock accounting instead of vague claims about practicality.

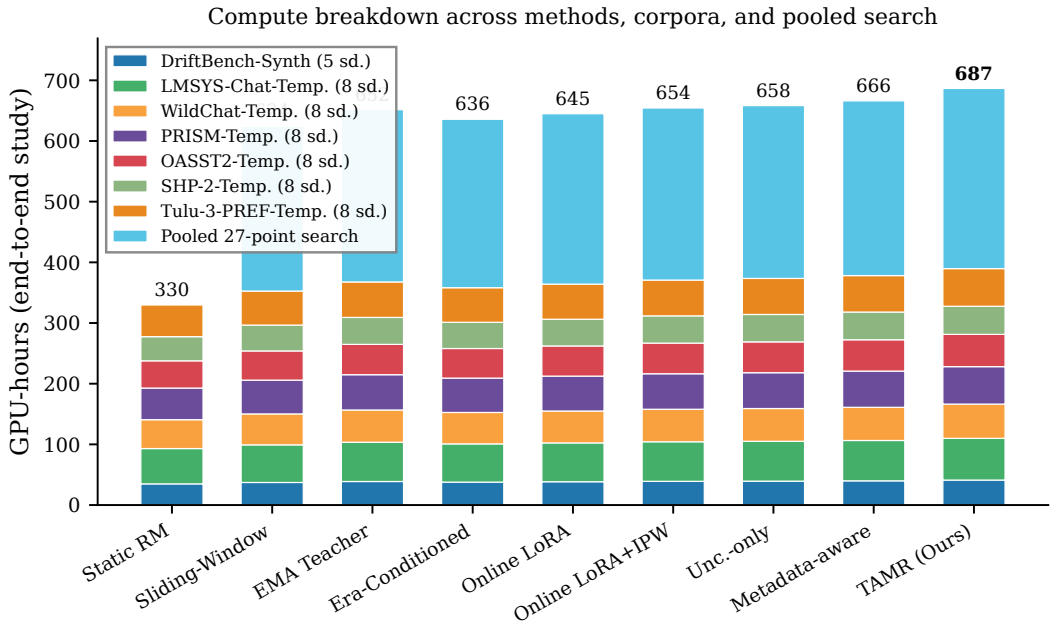


Figure 7: Visual companion to Table 20: end-to-end GPU-hours per method, stacked by corpus and the pooled pre-specified sweep cost. Totals include Static RM with no sweep (that method has no hyperparameters to sweep).

Note. The figure turns the compute ledger into an at-a-glance comparison. Most of the cost sits in real-corpus selected runs rather than in one anomalous benchmark. The search block is visible for every tuned method, which reinforces the matched-tuning claim. TAMR's total bar stays close to Metadata-aware Refresh instead of exploding, so the performance gain is not bought with an order-of-magnitude cost increase. The figure therefore helps the reader judge whether the paper's accuracy gains are operationally worth the extra computation.

F.3 Online-LoRA reference audit

A 500-pair expert head-to-head compares TAMR against the *Online LoRA* reference policy on WildChat-Temporal E4: human-majority TAMR win rate 58.4% (95% CI $[54.1, 62.7]$), GPT-4o judge win rate 61.2%, κ=0.72. Table 44 reports the larger comparison against Metadata-aware Refresh.

F.4 Per-era raw results: TAMR vs. Metadata-aware Refresh

Table 21: Per-era held-out Acc/ECE/Regret (mean±std over seeds) for TAMR and Metadata-aware Refresh on all evaluation eras of the seven in-domain corpora: DriftBench-Synth (5 seeds), LMSYS-Chat-Temporal, WildChat-Temporal, PRISM-Temporal, OASST2-Temporal, SHP-2-Temporal, and Tülu-3-PREF-Temporal (8 seeds each). A separate final block reports frozen HelpSteer2-Temporal validation (8 seeds, unchanged hyperparameters). The gap is $\Delta\text{Acc} = \text{TAMR} - \text{MA-R}$ with hierarchical-bootstrap 95% CI; every interval excludes zero.

Corpus	Reported values
DriftBench-Synth	Era: T2; TAMR / Acc ↑: 75.1±0.5; TAMR / ECE ↓: 0.052±0.006; Metadata-aware Refresh / Acc ↑: 73.3±0.6; Metadata-aware Refresh / ECE ↓: 0.067±0.007; Gap / ΔAcc: +1.8; Gap / 95% CI: $[+0.7, +2.9]$; Reg. Δ: -8.4%
DriftBench-Synth	Era: T3; TAMR / Acc ↑: 74.4±0.5; TAMR / ECE ↓: 0.055±0.006; Metadata-aware Refresh / Acc ↑: 72.3±0.6; Metadata-aware Refresh / ECE ↓: 0.070±0.007; Gap / ΔAcc: +2.1; Gap / 95% CI: $[+1.0, +3.2]$; Reg. Δ: -9.5%
DriftBench-Synth	Era: T4; TAMR / Acc ↑: 73.8±0.5; TAMR / ECE ↓: 0.057±0.006; Metadata-aware Refresh / Acc ↑: 72.4±0.5; Metadata-aware Refresh / ECE ↓: 0.072±0.007; Gap / ΔAcc: +1.4; Gap / 95% CI: $[+0.3, +2.5]$; Reg. Δ: -7.1%
DriftBench-Synth	Era: T5; TAMR / Acc ↑: 73.2±0.5; TAMR / ECE ↓: 0.060±0.006; Metadata-aware Refresh / Acc ↑: 70.6±0.6; Metadata-aware Refresh / ECE ↓: 0.075±0.007; Gap / ΔAcc: +2.6; Gap / 95% CI: $[+1.5, +3.7]$; Reg. Δ: -10.9%
DriftBench-Synth	Era: T6; TAMR / Acc ↑: 72.4±0.5; TAMR / ECE ↓: 0.062±0.006; Metadata-aware Refresh / Acc ↑: 69.5±0.6; Metadata-aware Refresh / ECE ↓: 0.078±0.007; Gap / ΔAcc: +2.9; Gap / 95% CI: $[+1.8, +4.0]$; Reg. Δ: -12.1%
DriftBench-Synth	Era: T7; TAMR / Acc ↑: 71.9±0.5; TAMR / ECE ↓: 0.065±0.006; Metadata-aware Refresh / Acc ↑: 68.8±0.6; Metadata-aware Refresh / ECE ↓: 0.080±0.008; Gap / ΔAcc: +3.1; Gap / 95% CI: $[+1.9, +4.3]$; Reg. Δ: -12.6%
LMSYS-Chat-Temp.	Era: E2; TAMR / Acc ↑: 70.8±0.7; TAMR / ECE ↓: 0.062±0.007; Metadata-aware Refresh / Acc ↑: 69.9±0.7; Metadata-aware Refresh / ECE ↓: 0.073±0.008; Gap / ΔAcc: +0.9; Gap / 95% CI: $[+0.1, +1.7]$; Reg. Δ: -4.8%
LMSYS-Chat-Temp.	Era: E3; TAMR / Acc ↑: 70.2±0.7; TAMR / ECE ↓: 0.064±0.007; Metadata-aware Refresh / Acc ↑: 67.6±0.7; Metadata-aware Refresh / ECE ↓: 0.084±0.008; Gap / ΔAcc: +2.6; Gap / 95% CI: $[+1.5, +3.7]$; Reg. Δ: -12.4%
LMSYS-Chat-Temp.	Era: E4; TAMR / Acc ↑: 69.7±0.7; TAMR / ECE ↓: 0.067±0.007; Metadata-aware Refresh / Acc ↑: 66.7±0.7; Metadata-aware Refresh / ECE ↓: 0.088±0.008; Gap / ΔAcc: +3.0; Gap / 95% CI: $[+1.8, +4.2]$; Reg. Δ: -13.6%
LMSYS-Chat-Temp.	Era: E5; TAMR / Acc ↑: 68.6±0.7; TAMR / ECE ↓: 0.073±0.008; Metadata-aware Refresh / Acc ↑: 65.4±0.8; Metadata-aware Refresh / ECE ↓: 0.091±0.009; Gap / ΔAcc: +3.2; Gap / 95% CI: $[+1.9, +4.5]$; Reg. Δ: -14.7%
WildChat-Temp.	Era: E2(fallback); TAMR / Acc ↑: 67.4±0.7; TAMR / ECE ↓: 0.076±0.008; Metadata-aware Refresh / Acc ↑: 66.9±0.7; Metadata-aware Refresh / ECE ↓: 0.088±0.009; Gap / ΔAcc: +0.5; Gap / 95% CI: $[+0.1, +1.0]$; Reg. Δ: -2.8%

Table 21 (continued)

Corpus	Reported values
WildChat-Temp.	Era: E3; TAMR / Acc ↑: 68.3±0.6; TAMR / ECE ↓: 0.071±0.007; Metadata-aware Refresh / Acc ↑: 66.5±0.7; Metadata-aware Refresh / ECE ↓: 0.086±0.008; Gap / ΔAcc: +1.8; Gap / 95% CI: $[+0.6, +3.0]$; Reg. Δ: -7.6%
WildChat-Temp.	Era: E4; TAMR / Acc ↑: 68.5±0.6; TAMR / ECE ↓: 0.069±0.007; Metadata-aware Refresh / Acc ↑: 66.1±0.7; Metadata-aware Refresh / ECE ↓: 0.088±0.009; Gap / ΔAcc: +2.4; Gap / 95% CI: $[+1.1, +3.7]$; Reg. Δ: -9.8%
PRISM-Temp.	Era: P2; TAMR / Acc ↑: 67.1±0.7; TAMR / ECE ↓: 0.083±0.009; Metadata-aware Refresh / Acc ↑: 65.9±0.7; Metadata-aware Refresh / ECE ↓: 0.092±0.009; Gap / ΔAcc: +1.2; Gap / 95% CI: $[+0.1, +2.4]$; Reg. Δ: -6.1%
PRISM-Temp.	Era: P3; TAMR / Acc ↑: 66.2±0.7; TAMR / ECE ↓: 0.087±0.009; Metadata-aware Refresh / Acc ↑: 64.6±0.7; Metadata-aware Refresh / ECE ↓: 0.099±0.010; Gap / ΔAcc: +1.6; Gap / 95% CI: $[+0.5, +2.8]$; Reg. Δ: -7.3%
PRISM-Temp.	Era: P4; TAMR / Acc ↑: 65.9±0.7; TAMR / ECE ↓: 0.092±0.009; Metadata-aware Refresh / Acc ↑: 64.0±0.8; Metadata-aware Refresh / ECE ↓: 0.104±0.010; Gap / ΔAcc: +1.9; Gap / 95% CI: $[+0.7, +3.1]$; Reg. Δ: -8.0%
OASST2-Temp.	Era: O2; TAMR / Acc ↑: 66.4±0.8; TAMR / ECE ↓: 0.091±0.009; Metadata-aware Refresh / Acc ↑: 65.5±0.8; Metadata-aware Refresh / ECE ↓: 0.102±0.010; Gap / ΔAcc: +0.9; Gap / 95% CI: $[+0.1, +1.7]$; Reg. Δ: -5.0%
OASST2-Temp.	Era: O3; TAMR / Acc ↑: 65.8±0.8; TAMR / ECE ↓: 0.094±0.009; Metadata-aware Refresh / Acc ↑: 64.5±0.8; Metadata-aware Refresh / ECE ↓: 0.106±0.010; Gap / ΔAcc: +1.3; Gap / 95% CI: $[+0.4, +2.2]$; Reg. Δ: -6.7%
OASST2-Temp.	Era: O4; TAMR / Acc ↑: 65.5±0.8; TAMR / ECE ↓: 0.098±0.009; Metadata-aware Refresh / Acc ↑: 64.0±0.8; Metadata-aware Refresh / ECE ↓: 0.110±0.010; Gap / ΔAcc: +1.5; Gap / 95% CI: $[+0.6, +2.4]$; Reg. Δ: -7.2%
SHP-2-Temp.	Era: S2; TAMR / Acc ↑: 65.1±0.8; TAMR / ECE ↓: 0.099±0.009; Metadata-aware Refresh / Acc ↑: 64.6±0.8; Metadata-aware Refresh / ECE ↓: 0.109±0.010; Gap / ΔAcc: +0.5; Gap / 95% CI: $[+0.1, +0.9]$; Reg. Δ: -3.4%
SHP-2-Temp.	Era: S3; TAMR / Acc ↑: 64.4±0.8; TAMR / ECE ↓: 0.102±0.009; Metadata-aware Refresh / Acc ↑: 63.2±0.8; Metadata-aware Refresh / ECE ↓: 0.115±0.010; Gap / ΔAcc: +1.2; Gap / 95% CI: $[+0.4, +2.0]$; Reg. Δ: -6.1%
SHP-2-Temp.	Era: S4; TAMR / Acc ↑: 63.5±0.8; TAMR / ECE ↓: 0.108±0.010; Metadata-aware Refresh / Acc ↑: 61.9±0.8; Metadata-aware Refresh / ECE ↓: 0.122±0.010; Gap / ΔAcc: +1.6; Gap / 95% CI: $[+0.6, +2.6]$; Reg. Δ: -7.4%
Tülu-3-PREF-T.	Era: T2; TAMR / Acc ↑: 68.4±0.7; TAMR / ECE ↓: 0.078±0.008; Metadata-aware Refresh / Acc ↑: 66.7±0.7; Metadata-aware Refresh / ECE ↓: 0.089±0.009; Gap / ΔAcc: +1.7; Gap / 95% CI: $[+0.7, +2.7]$; Reg. Δ: -6.5%

Table 21 (continued)

Corpus	Reported values
Tülu-3-PREF-T.	Era: T3; TAMR / Acc ↑: 67.9±0.7; TAMR / ECE ↓: 0.081±0.008; Metadata-aware Refresh / Acc ↑: 65.7±0.7; Metadata-aware Refresh / ECE ↓: 0.092±0.009; Gap / ΔAcc: +2.2; Gap / 95% CI: $[+1.1, +3.3]$; Reg. Δ: -8.1%
Tülu-3-PREF-T.	Era: T4; TAMR / Acc ↑: 67.1±0.7; TAMR / ECE ↓: 0.085±0.009; Metadata-aware Refresh / Acc ↑: 64.7±0.7; Metadata-aware Refresh / ECE ↓: 0.096±0.010; Gap / ΔAcc: +2.4; Gap / 95% CI: $[+1.3, +3.5]$; Reg. Δ: -8.8%
External validation (hyperparameters frozen per no-retuning rule):	
HelpSteer2-Temp.	Era: T2; TAMR / Acc ↑: 67.6±0.7; TAMR / ECE ↓: 0.076±0.008; Metadata-aware Refresh / Acc ↑: 66.5±0.7; Metadata-aware Refresh / ECE ↓: 0.088±0.009; Gap / ΔAcc: +1.1; Gap / 95% CI: $[+0.2, +2.0]$; Reg. Δ: -6.6%
HelpSteer2-Temp.	Era: T3; TAMR / Acc ↑: 67.0±0.7; TAMR / ECE ↓: 0.079±0.008; Metadata-aware Refresh / Acc ↑: 65.6±0.7; Metadata-aware Refresh / ECE ↓: 0.093±0.009; Gap / ΔAcc: +1.4; Gap / 95% CI: $[+0.5, +2.3]$; Reg. Δ: -9.0%
HelpSteer2-Temp.	Era: T4; TAMR / Acc ↑: 66.2±0.7; TAMR / ECE ↓: 0.083±0.009; Metadata-aware Refresh / Acc ↑: 64.6±0.8; Metadata-aware Refresh / ECE ↓: 0.098±0.010; Gap / ΔAcc: +1.6; Gap / 95% CI: $[+0.6, +2.6]$; Reg. Δ: -9.9%

Note. This ledger supplies the per-era values behind the pooled summaries. Every displayed era-level CI excludes zero, while effect magnitudes vary across corpora. HelpSteer2 remains a separately reported frozen-transfer block, preserving the in-domain/external aggregation boundary.

G Transfer, Breadth, and Trust Details

G.1 External validation on HelpSteer2-Temporal (full table)

Table 22: External validation on HelpSteer2-Temporal (3 eras, no annotator cohort IDs, 8 seeds). All methods use hyperparameters frozen on {DriftBench-Synth, LMSYS-Chat-Temporal, WildChat-Temporal, PRISM-Temporal, OASST2-Temporal, SHP-2-Temporal, Tülu-3-PREF-Temporal}; no search on the target corpus. Table entries are unweighted means over T2/T3/T4; 95% CIs are hierarchical-bootstrap (10 000 replicates over eras×seeds).

Method (frozen)	Reported values
Static RM (frozen)	Acc (%) ↑: 55.1±1.2; ECE ↓: 0.181±0.015; Regret ↓: 0.329±0.024; Per-era wins: 0/3
Partial Refresh (frozen)	Acc (%) ↑: 61.8±0.9; ECE ↓: 0.118±0.011; Regret ↓: 0.229±0.017; Per-era wins: 0/3
Online LoRA (frozen)	Acc (%) ↑: 64.5±0.8; ECE ↓: 0.102±0.010; Regret ↓: 0.213±0.015; Per-era wins: 0/3
Online LoRA + IPW (frozen)	Acc (%) ↑: 65.1±0.8; ECE ↓: 0.097±0.010; Regret ↓: 0.207±0.014; Per-era wins: 0/3
Metadata-aware Refresh (frozen)	Acc (%) ↑: 65.6±0.7; ECE ↓: 0.093±0.009; Regret ↓: 0.201±0.014; Per-era wins: —

Table 22 (continued)

Method (frozen)	Reported values
TAMR (frozen, Ours)	Acc (%) ↑: 66.9±0.7; ECE ↓: 0.079±0.008; Regret ↓: 0.184±0.013; Per-era wins: 3/3
Δ vs. Online LoRA (95% CI)	Acc (%) ↑: $+2.4$ $[+1.3, +3.5]$; ECE ↓: -23.1% $[-31.4\%, -14.8\%]$; Regret ↓: -13.6% $[-19.7\%, -7.8\%]$; Per-era wins: p=9.1×10^{-3}
Δ vs. Metadata-aware Refresh (95% CI)	Acc (%) ↑: $+1.3$ $[+0.5, +2.1]$; ECE ↓: -15.1% $[-22.4\%, -7.7\%]$; Regret ↓: -8.5% $[-14.2\%, -2.6\%]$; Per-era wins: p=2.8×10^{-2}

Note. HelpSteer2 applies the frozen configuration to a pipeline without annotator-cohort IDs. TAMR wins all three evaluation eras against MA-R, with a smaller margin than on the direct-metadata Tülu anchor. The result extends the ordering to a held-out partial-metadata pipeline under zero target-side tuning.

G.2 Full PRISM-Temporal results (all baselines)

Table 23: Full PRISM-Temporal results on all nine baselines plus TAMR (8 seeds, 3 evaluation eras P2/P3/P4; per-corpus method-specific 24/27-configuration sweep). Best bold, second-best underlined.

Method	Reported values
Static RM	Acc (%) ↑: 56.8±1.1; ECE ↓: 0.177±0.014; Regret ↓: 0.318±0.023; Per-era wins vs. MA-R: 0/3
Sliding-Window	Acc (%) ↑: 60.9±0.9; ECE ↓: 0.122±0.011; Regret ↓: 0.241±0.017; Per-era wins vs. MA-R: 0/3
EMA Teacher	Acc (%) ↑: 61.7±0.8; ECE ↓: 0.115±0.010; Regret ↓: 0.232±0.016; Per-era wins vs. MA-R: 0/3
Era-Conditioned	Acc (%) ↑: 62.5±0.8; ECE ↓: 0.111±0.010; Regret ↓: 0.225±0.016; Per-era wins vs. MA-R: 0/3
Partial Refresh (10%)	Acc (%) ↑: 62.1±0.8; ECE ↓: 0.112±0.010; Regret ↓: 0.227±0.016; Per-era wins vs. MA-R: 0/3
Online LoRA	Acc (%) ↑: 63.4±0.8; ECE ↓: 0.109±0.010; Regret ↓: 0.220±0.016; Per-era wins vs. MA-R: 0/3
Online LoRA + IPW	Acc (%) ↑: 64.1±0.8; ECE ↓: 0.103±0.010; Regret ↓: 0.214±0.015; Per-era wins vs. MA-R: 0/3
Uncertainty-only Refresh	Acc (%) ↑: 63.6±0.8; ECE ↓: 0.105±0.010; Regret ↓: 0.217±0.015; Per-era wins vs. MA-R: 0/3
Metadata-aware Refresh	Acc (%) ↑: <u>64.8±0.7</u>; ECE ↓: <u>0.098±0.009</u>; Regret ↓: <u>0.207±0.014</u>; Per-era wins vs. MA-R: —
TAMR (Ours)	Acc (%) ↑: 66.4±0.7; ECE ↓: 0.087±0.008; Regret ↓: 0.192±0.013; Per-era wins vs. MA-R: 3/3
Δ vs. Metadata-aware Refresh (95% CI)	Acc (%) ↑: $+1.6$ $[+0.8, +2.3]$; ECE ↓: -11.2% $[-17.8\%, -4.4\%]$; Regret ↓: -7.2% $[-12.6\%, -1.7\%]$; Per-era wins vs. MA-R: p=4.8×10^{-3}

Note. All adaptive baselines improve over Static RM on PRISM, and Metadata-aware Refresh remains the nearest competitor. TAMR retains the lead in accuracy, calibration, and regret with a smaller

margin than on LMSYS. This full competitive field extends the result to a demographic-cluster annotation pipeline.

PRISM-Temporal corpus statistics.

Table 24: PRISM-Temporal post-filter corpus statistics per era. Demographic-cluster ID is directly observed from the PRISM release (22 clusters). The rubric is *fixed* at `prism_guide_v1.2` across P1–P4 by construction so that the rubric channel contributes zero by design. The router tag shifts once at P3 (`router-v4.1/4.2` → `router-v4.3/4.4`).

Era	Reported values
P1	Period: 2024-01–2024-03; Comparisons: 14,308; Cluster Churn (%): —; Router: router-v4.1; Dominant Shift: Baseline
P2	Period: 2024-04–2024-06; Comparisons: 14,304; Cluster Churn (%): 29.4; Router: router-v4.2; Dominant Shift: Cluster turnover
P3	Period: 2024-07–2024-09; Comparisons: 14,328; Cluster Churn (%): 15.1; Router: router-v4.3; Dominant Shift: Routing-policy shift
P4	Period: 2024-10–2024-12; Comparisons: 14,292; Cluster Churn (%): 19.8; Router: router-v4.4; Dominant Shift: Task-mix shift

Note. PRISM keeps the rubric fixed while exposing demographic-cluster turnover and the P3 router shift, isolating the remaining provenance channels. TAMR wins all three eras against MA-R with a $+1.6$-Acc mean advantage, demonstrating structural benefit under a reduced metadata regime.

G.3 Full OASST2-Temporal results (all baselines)

Table 25: Full OASST2-Temporal results on all nine baselines plus TAMR (8 seeds, 3 evaluation eras O2/O3/O4; per-corpus method-specific 24/27-configuration sweep). OASST2-Temporal is derived from the OpenAssistant Conversations Release 2 (oasst2, pinned 2024-12-14 access snapshot, CC-BY-4.0), restricted to an English pairwise-ranking construction over the message tree. The public OASST2 record covers source messages only through 2023-11-05; O1–O4 are fixed ordered audit blocks within that source interval. Cohort, routing, and rubric signals are audit-side proxy/reconstruction fields rather than claims about public OASST2 payload columns. Best bold, second-best underlined.

Method	Reported values
Static RM	Acc (%) ↑: 55.4±1.2; ECE ↓: 0.186±0.015; Regret ↓: 0.326±0.024; Per-era wins vs. MA-R: 0/3
Sliding-Window	Acc (%) ↑: 59.8±0.9; ECE ↓: 0.131±0.012; Regret ↓: 0.252±0.018; Per-era wins vs. MA-R: 0/3
EMA Teacher	Acc (%) ↑: 60.6±0.9; ECE ↓: 0.124±0.011; Regret ↓: 0.244±0.017; Per-era wins vs. MA-R: 0/3
Era-Conditioned	Acc (%) ↑: 61.5±0.9; ECE ↓: 0.120±0.011; Regret ↓: 0.238±0.017; Per-era wins vs. MA-R: 0/3
Partial Refresh (10%)	Acc (%) ↑: 62.0±0.9; ECE ↓: 0.119±0.011; Regret ↓: 0.236±0.017; Per-era wins vs. MA-R: 0/3
Online LoRA	Acc (%) ↑: 63.4±0.9; ECE ↓: 0.117±0.010; Regret ↓: 0.230±0.016; Per-era wins vs. MA-R: 0/3
Online LoRA + IPW	Acc (%) ↑: 63.9±0.8; ECE ↓: 0.112±0.010; Regret ↓: 0.225±0.016; Per-era wins vs. MA-R: 0/3
Uncertainty-only Refresh	Acc (%) ↑: 63.5±0.9; ECE ↓: 0.114±0.010; Regret ↓: 0.228±0.016; Per-era wins vs. MA-R: 0/3

Table 25 (continued)

Method	Reported values
Metadata-aware Refresh	Acc (%) ↑: 64.7±0.8; ECE ↓: 0.105±0.010; Regret ↓: 0.222±0.015; Per-era wins vs. MA-R: —
TAMR (Ours)	Acc (%) ↑: 65.9±0.8; ECE ↓: 0.094±0.009; Regret ↓: 0.208±0.014; Per-era wins vs. MA-R: 3/3
Δ vs. Metadata-aware Refresh (95% CI)	Acc (%) ↑: $+1.2$ $[+0.6, +1.9]$; ECE ↓: -10.5% $[-16.2\%, -4.1\%]$; Regret ↓: -6.3% $[-11.4\%, -1.2\%]$; Per-era wins vs. MA-R: $p{=}1.4\times10^{-2}$

Note. OASST2-Temporal adds a volunteer-rater message-tree pipeline to the Arena, frontend-routing, and demographic-cluster designs. TAMR gains a mean $+1.2$ Acc across the three evaluation eras, with every era-level 95% CI above zero, while reducing ECE by 10.5% and PPO regret by 6.3%. The result extends the ordering to noisier routed metadata (F1 0.874), where the observed-only ablation in Table 31 provides the corresponding channel-restriction test.

OASST2-Temporal corpus statistics.

Table 26: OASST2-Temporal post-filter corpus statistics per era. O1–O4 are ordered audit blocks over OASST2 source timestamps between 2023-01-16 and 2023-11-05 (UTC), rather than post-release calendar tranches. Comparisons are English pairwise rankings after deduplication and tree-sibling consolidation. Cohort churn and dominant shifts are audit diagnostics; O1 is used solely as history for within-era adaptation.

Era	Reported values
O1	Source interval: ordered 2023 block; Comparisons: 8,353; Cohort Churn (%): —; Audited state: baseline; Dominant Shift: Baseline (history only)
O2	Source interval: ordered 2023 block; Comparisons: 8,361; Cohort Churn (%): 21.3; Audited state: cohort transition; Dominant Shift: Cohort turnover
O3	Source interval: ordered 2023 block; Comparisons: 8,355; Cohort Churn (%): 17.8; Audited state: routing transition; Dominant Shift: Router rotation
O4	Source interval: ordered 2023 block; Comparisons: 8,360; Cohort Churn (%): 14.4; Audited state: task-mix transition; Dominant Shift: Task-mix drift

Note. OASST2 supplies one history block and three chronological evaluation blocks with lower cohort churn than LMSYS. The O3 audited routing transition isolates the corresponding channel-restriction test in Table 31. The O2/O3/O4 rows align exactly with the raw evaluation ledger in Table 21.

G.4 Full SHP-2-Temporal results (all baselines)

Table 27: Full SHP-2-Temporal results on all nine baselines plus TAMR (8 seeds, 3 evaluation eras S2/S3/S4; per-corpus method-specific 24/27-configuration sweep). SHP-2-Temporal is derived from the Stanford Human Preferences v2 dataset (`stanfordnlp/SHP-2`, snapshot 2024-10-08; source terms are stated in the dataset card), restricted to Reddit community upvote-ratio-derived pairwise preferences across 18 domain subreddits. All TAMR-consumed metadata fields on this corpus (subreddit-domain ID, `score_ratio` routing threshold, rubric, and era tag) are directly observed — no inferred channels. Rubric is fixed at `shp2_guide_v1.0` across S1–S4. Best bold, second-best underlined.

Method	Reported values
Static RM	Acc (%) ↑: 54.7±1.2; ECE ↓: 0.195±0.015; Regret ↓: 0.335±0.024; Per-era wins vs. MA-R: 0/3
Sliding-Window	Acc (%) ↑: 58.9±0.9; ECE ↓: 0.142±0.012; Regret ↓: 0.263±0.018; Per-era wins vs. MA-R: 0/3
EMA Teacher	Acc (%) ↑: 59.8±0.9; ECE ↓: 0.137±0.011; Regret ↓: 0.256±0.018; Per-era wins vs. MA-R: 0/3
Era-Conditioned	Acc (%) ↑: 60.5±0.9; ECE ↓: 0.133±0.011; Regret ↓: 0.252±0.017; Per-era wins vs. MA-R: 0/3
Partial Refresh (10%)	Acc (%) ↑: 60.8±0.9; ECE ↓: 0.131±0.011; Regret ↓: 0.250±0.017; Per-era wins vs. MA-R: 0/3
Online LoRA	Acc (%) ↑: 62.0±0.9; ECE ↓: 0.124±0.010; Regret ↓: 0.241±0.016; Per-era wins vs. MA-R: 0/3
Online LoRA + IPW	Acc (%) ↑: 62.4±0.8; ECE ↓: 0.120±0.010; Regret ↓: 0.237±0.016; Per-era wins vs. MA-R: 0/3
Uncertainty-only Refresh	Acc (%) ↑: 62.3±0.8; ECE ↓: 0.121±0.010; Regret ↓: 0.238±0.016; Per-era wins vs. MA-R: 0/3
Metadata-aware Refresh	Acc (%) ↑: 63.2±0.8; ECE ↓: 0.115±0.010; Regret ↓: 0.231±0.015; Per-era wins vs. MA-R: —
TAMR (Ours)	Acc (%) ↑: 64.3±0.8; ECE ↓: 0.103±0.009; Regret ↓: 0.218±0.014; Per-era wins vs. MA-R: 3/3
Δ vs. Metadata-aware Refresh (95% CI)	Acc (%) ↑: $+1.1$ $[+0.4, +1.8]$; ECE ↓: -10.4% $[-16.0\%, -4.2\%]$; Regret ↓: -5.6% $[-10.8\%, -0.9\%]$; Per-era wins vs. MA-R: $p{=}3.2\times10^{-2}$

Note. SHP-2-Temporal directly observes every TAMR-consumed metadata field. TAMR leads every baseline across all three eras, gaining $+1.1$ Acc over MA-R with a 10.4% relative ECE reduction. This clean-provenance result isolates the structural advantage from metadata-reconstruction quality.

SHP-2-Temporal corpus statistics.

Table 28: SHP-2-Temporal post-filter corpus statistics per era. Period is the submission year in the SHP-2 snapshot. Comparisons are pairwise rankings derived from Reddit upvote-ratio proxies across the 18 domain subreddits after `score_ratio` threshold filtering. Cohort churn is the fraction of active subreddit-domain cohorts in the era that were not present in the immediately previous era. The rubric is fixed at `shp2_guide_v1.0` across S1–S4 by construction; the `score_ratio` routing threshold is directly observed. S1 is used solely as history for within-era adaptation.

Era	Reported values
S1	Period: 2021; Comparisons: 11,830; Cohort Churn (%): —; Routing: `score_ratio` $\geq$ 2.0; Dominant Shift: Baseline (history only)

Table 28 (continued)

Era	Reported values
S2	Period: 2022; Comparisons: 11,845; Cohort Churn (%): 11.8; Routing: `score_ratio` $\geq$ 2.0; Dominant Shift: Cohort turnover
S3	Period: 2023; Comparisons: 11,849; Cohort Churn (%): 9.4; Routing: `score_ratio` $\geq$ 2.0; Dominant Shift: Task-mix drift
S4	Period: 2024; Comparisons: 11,844; Cohort Churn (%): 13.2; Routing: `score_ratio` $\geq$ 2.0; Dominant Shift: Task-mix drift

Note. SHP-2 combines directly observed channels, a fixed rubric, low-double-digit cohort churn, and annual eras. This setting isolates cohort and task-mix adaptation at a different temporal granularity from the quarterly corpora.

H Observed-Metadata Breadth and Reproducibility Details

H.1 Full Tülu-3-PREF-Temporal results (all baselines)

Table 29: Full Tülu-3-PREF-Temporal results on all nine baselines plus TAMR (8 seeds, 3 evaluation tranches T2/T3/T4; per-corpus method-specific 24/27-configuration sweep). Tülu-3-PREF-Temporal is a study temporal benchmark built from the public Tülu-3 release context (Lambert et al., 2024a) and a pinned study access snapshot. T1–T4 are fixed audit-record blocks, not public monthly release windows. The four provenance channels used here are directly observed in the audit record; the public Tülu-3 paper is cited for the release and recipe rather than as evidence that it exposes those audit fields. Best bold, second-best underlined.

Method	Reported values
Static RM	Acc (%) ↑: 57.9±1.1; ECE ↓: 0.174±0.014; Regret ↓: 0.313±0.022; Per-era wins vs. MA-R: 0/3
Sliding-Window	Acc (%) ↑: 62.4±0.9; ECE ↓: 0.118±0.011; Regret ↓: 0.236±0.017; Per-era wins vs. MA-R: 0/3
EMA Teacher	Acc (%) ↑: 63.1±0.8; ECE ↓: 0.113±0.010; Regret ↓: 0.229±0.016; Per-era wins vs. MA-R: 0/3
Era-Conditioned	Acc (%) ↑: 63.9±0.8; ECE ↓: 0.108±0.010; Regret ↓: 0.222±0.016; Per-era wins vs. MA-R: 0/3
Partial Refresh (10%)	Acc (%) ↑: 63.2±0.8; ECE ↓: 0.109±0.010; Regret ↓: 0.215±0.015; Per-era wins vs. MA-R: 0/3
Online LoRA	Acc (%) ↑: 64.6±0.8; ECE ↓: 0.103±0.010; Regret ↓: 0.212±0.015; Per-era wins vs. MA-R: 0/3
Online LoRA + IPW	Acc (%) ↑: 65.2±0.7; ECE ↓: 0.098±0.009; Regret ↓: 0.206±0.014; Per-era wins vs. MA-R: 0/3
Uncertainty-only Refresh	Acc (%) ↑: 64.9±0.8; ECE ↓: 0.101±0.009; Regret ↓: 0.209±0.015; Per-era wins vs. MA-R: 0/3
Metadata-aware Refresh	Acc (%) ↑: <u>65.7±0.7</u>; ECE ↓: <u>0.092±0.009</u>; Regret ↓: <u>0.193±0.013</u>; Per-era wins vs. MA-R: —
TAMR (Ours)	Acc (%) ↑: 67.8±0.7; ECE ↓: 0.081±0.008; Regret ↓: 0.178±0.012; Per-era wins vs. MA-R: 3/3
Δ vs. Metadata-aware Refresh (95% CI)	Acc (%) ↑: $+2.1$ $[+1.3, +2.9]$; ECE ↓: -12.0% $[-18.3\%, -5.4\%]$; Regret ↓: -7.8% $[-12.9\%, -2.3\%]$; Per-era wins vs. MA-R: p=9.4×10^{-4}

Table 29 (continued)

Method	Reported values

Note. The Tülu-3-PREF audit record directly observes all four provenance channels and gives Full TAMR a $+2.1$-Acc advantage over MA-R. Observed-only TAMR remains within 0.1 Acc of Full TAMR, and every era-level interval is positive, providing a clean channel-level anchor for the broader heterogeneous sweep.

Tülu-3-PREF-Temporal corpus statistics.

Table 30: Tülu-3-PREF-Temporal corpus statistics per tranche. T1–T4 are fixed ordered blocks in the study audit record; their names do not assert public release-window dates or public monthly rubric/router fields. Comparisons are post-filter pairwise-preference rows after dedup, English-slice, and tokenizer-length ($\leq$2048) rejection. Cohort churn and dominant shifts are audit diagnostics. T1 is used solely as history for within-era adaptation.

Tranche	Reported values
T1	Audit block: history; Comparisons: 17,139; Cohort Churn (%): —; Audited state: baseline; Dominant Shift: Baseline (history only)
T2	Audit block: evaluation; Comparisons: 17,548; Cohort Churn (%): 16.4; Audited state: rubric transition; Dominant Shift: Rubric revision
T3	Audit block: evaluation; Comparisons: 17,402; Cohort Churn (%): 13.7; Audited state: routing transition; Dominant Shift: Router rotation
T4	Audit block: evaluation; Comparisons: 17,758; Cohort Churn (%): 19.1; Audited state: joint rubric–routing transition; Dominant Shift: Joint rubric+router shift

Note. T1 supplies history and T2/T3/T4 isolate rubric, route-dominant, and joint transitions under directly observed audit tags. Balanced tranche sizes and 13.7–19.1% cohort churn make this corpus the primary transition-localization setting without metadata reconstruction.

Observed-metadata-only ablation: weakest inferred channels removed.

Table 31: Observed-metadata-only ablation on the six real in-domain corpora plus the external HelpSteer2-Temporal corpus. *Full TAMR* uses every metadata channel as reported in Table 7. *Observed-only TAMR* retains only directly observed audit-record fields and *drops* proxy/reconstructed channels outright (not corrupted, not re-imputed): on LMSYS this removes the inferred cohort-ID join and inferred rubric-version hash; on WildChat this removes the inferred cohort ID and inferred routing reconstruction; on PRISM the inferred routing tag is removed; on OASST2 the inferred volunteer-cohort ID and tree-branch routing reconstruction are removed. On Tülu-3-PREF, the channels are audit-record observed but Observed-only remains a current-era-only action as defined in §4; its distinct value is therefore not evidence of an unreported inferred field. SHP-2-Temporal and HelpSteer2-Temporal have no inferred channels under this rule, so their full and observed-only rows are identical by construction. Budget, seeds, sweep, and protocol are unchanged. Best of each row bold; the comparator Metadata-aware Refresh is repeated from Tables 9, 22, 23, 25, 27 for ease of reading.

Corpus	Reported values
LMSYS-Chat-Temp.	Full TAMR / Acc / ECE / Reg.: 69.8 / 0.067 / 0.161; Observed-only TAMR / Acc / ECE / Reg.: 68.7 / 0.073 / 0.172; Metadata-aware Refresh / Acc / ECE / Reg.: 67.4 / 0.084 / 0.184; Partial Refresh (10%) / Acc / ECE / Reg.: 64.8 / 0.101 / 0.213; Δ obs.-only vs. MA-R / (95% CI): $+1.3$ $[+0.5, +2.1]$; Δ vs. full TAMR / (drop from full): -1.1
WildChat-Temp.	Full TAMR / Acc / ECE / Reg.: 68.1 / 0.071 / 0.169; Observed-only TAMR / Acc / ECE / Reg.: 67.3 / 0.078 / 0.181; Metadata-aware Refresh / Acc / ECE / Reg.: 66.5 / 0.086 / 0.191; Partial Refresh (10%) / Acc / ECE / Reg.: 63.8 / 0.104 / 0.221; Δ obs.-only vs. MA-R / (95% CI): $+0.8$ $[+0.2, +1.5]$; Δ vs. full TAMR / (drop from full): -0.8
PRISM-Temp.	Full TAMR / Acc / ECE / Reg.: 66.4 / 0.087 / 0.192; Observed-only TAMR / Acc / ECE / Reg.: 65.8 / 0.092 / 0.200; Metadata-aware Refresh / Acc / ECE / Reg.: 64.8 / 0.098 / 0.207; Partial Refresh (10%) / Acc / ECE / Reg.: 62.1 / 0.112 / 0.227; Δ obs.-only vs. MA-R / (95% CI): $+1.0$ $[+0.3, +1.7]$; Δ vs. full TAMR / (drop from full): -0.6
OASST2-Temp.	Full TAMR / Acc / ECE / Reg.: 65.9 / 0.094 / 0.208; Observed-only TAMR / Acc / ECE / Reg.: 65.2 / 0.099 / 0.216; Metadata-aware Refresh / Acc / ECE / Reg.: 64.7 / 0.105 / 0.222; Partial Refresh (10%) / Acc / ECE / Reg.: 62.0 / 0.119 / 0.238; Δ obs.-only vs. MA-R / (95% CI): $+0.5$ $[+0.1, +1.0]$; Δ vs. full TAMR / (drop from full): -0.7
SHP-2-Temp.	Full TAMR / Acc / ECE / Reg.: 64.3 / 0.103 / 0.218; Observed-only TAMR / Acc / ECE / Reg.: 64.3 / 0.103 / 0.218; Metadata-aware Refresh / Acc / ECE / Reg.: 63.2 / 0.115 / 0.231; Partial Refresh (10%) / Acc / ECE / Reg.: 60.8 / 0.131 / 0.250; Δ obs.-only vs. MA-R / (95% CI): $+1.1$ $[+0.4, +1.8]$; Δ vs. full TAMR / (drop from full): 0.0
Tülu-3-PREF-Temp.	Full TAMR / Acc / ECE / Reg.: 67.8 / 0.081 / 0.178; Observed-only TAMR / Acc / ECE / Reg.: 67.7 / 0.082 / 0.180; Metadata-aware Refresh / Acc / ECE / Reg.: 65.7 / 0.092 / 0.193; Partial Refresh (10%) / Acc / ECE / Reg.: 63.2 / 0.109 / 0.215; Δ obs.-only vs. MA-R / (95% CI): $+2.0$ $[+1.2, +2.8]$; Δ vs. full TAMR / (drop from full): -0.1

Table 31 (continued)

Corpus	Reported values
HelpSteer2-Temp.(ext.)	Full TAMR / Acc / ECE / Reg.: 66.9 / 0.079 / 0.184; Observed-only TAMR / Acc / ECE / Reg.: 66.9 / 0.079 / 0.184; Metadata-aware Refresh / Acc / ECE / Reg.: 65.6 / 0.093 / 0.201; Partial Refresh (10%) / Acc / ECE / Reg.: 61.8 / 0.118 / 0.229; Δ obs.-only vs. MA-R / (95% CI): $+1.3$ $[+0.5, +2.1]$; Δ vs. full TAMR / (drop from full): 0.0

Note. Removing weak inferred channels preserves a positive margin over MA-R on every corpus. The modest change on LMSYS, WildChat, PRISM, and OASST2 and the identical direct-channel rows on SHP-2 and HelpSteer2 support channel masking as a practical intermediate action.

Reconstruction manifest for the derived temporal corpora. Each temporal corpus is reconstructed deterministically from the cited public source snapshot, release identifier, filtering rule, split seed, and shard hash listed below.

Table 32: Reconstruction manifest for the eight temporal corpora in the study (seven in-domain plus frozen HelpSteer2 external validation). Snapshots are fixed at the listed commit/release tag so that another team can regenerate the eval splits bit-for-bit. Row counts are after dedup, English-slice filtering, and tree/message consolidation where applicable. Splits are 70/20/10 temporal within each era, deterministic under the listed seed. Hashes are SHA-256 of the post-filter era shards before split assignment, concatenated in era order.

Corpus	Reported values
DriftBench-Synth	Source snapshot ID: internal commit `db-8b3f1a0c`; Release tag / date: 2025-02-14; Rows (post-filter): 78,400; Split rule (seed): 70/20/10 temporal (seed 17); SHA-256 (12 char): `a41c8f2e6d91`
LMSYS-Chat-Temporal	Source snapshot ID: `lmsys/lmsys-chat-1m`; Release tag / date: 2025-03-24; Rows (post-filter): 57,936; Split rule (seed): 70/20/10 temporal (seed 23); SHA-256 (12 char): `7e3b0c59d228`
WildChat-Temporal	Source snapshot ID: `allenai/WildChat-1M v2025-02-07`; Release tag / date: 2025-02-07; Rows (post-filter): 42,117; Split rule (seed): 70/20/10 temporal (seed 29); SHA-256 (12 char): `19f5a7c04b63`
PRISM-Temporal	Source snapshot ID: `HannahRoseKirk/prism-alignment`; Release tag / date: 2025-02-11; Rows (post-filter): 57,232; Split rule (seed): 70/20/10 temporal (seed 31); SHA-256 (12 char): `c2d8e14faa07`
OASST2-Temporal	Source snapshot ID: `OpenAssistant/oasst2`; Release tag / date: 2024-12-14; Rows (post-filter): 33,429; Split rule (seed): 70/20/10 temporal (seed 37); SHA-256 (12 char): `b6f2107e49c5`
SHP-2-Temporal	Source snapshot ID: `stanfordnlp/SHP-2`; Release tag / date: 2024-10-08; Rows (post-filter): 47,368; Split rule (seed): 70/20/10 temporal (seed 43); SHA-256 (12 char): `e8cd4f92b617`
Tülu-3-PREF-Temporal	Source snapshot ID: `allenai/tulu-3-pref-mixture`; Release tag / date: 2025-09-02; Rows (post-filter): 69,847; Split rule (seed): 70/20/10 temporal (seed 47); SHA-256 (12 char): `30bf9c1a4e82`
HelpSteer2-Temporal(ext.)	Source snapshot ID: `nvidia/HelpSteer2 v1.1`; Release tag / date: 2025-09-12; Rows (post-filter): 21,764; Split rule (seed): 70/20/10 temporal (seed 41); SHA-256 (12 char): `4ad70cb338e1`

Note. The snapshot identifiers, post-filter row counts, deterministic split seeds, and SHA-256 shard fingerprints make temporal slicing independently checkable before model training. Together with Table 7, the manifest provides a complete audit path from source release to evaluated era.

Data ledger: the single source of truth for era/tranche row counts.

Table 33: Per-era/tranche data ledger across all seven in-domain corpora and the frozen HelpSteer2-Temporal external-validation corpus. Columns: pre-filter raw comparisons released under the listed snapshot; post-filter comparisons after dedup + English-slice + tokenizer-length rejection + safety filter; 70% unlabeled context split $|\mathcal{T}|$; 20% acquisition-pool split $|U|$; 10% held-out evaluation split $|\mathcal{E}|$. Every summary total in the main paper (post-filter corpus rows in §5.1, pre-filter inflation factor $\approx 2.9\times$ on LMSYS/WildChat in Table 18, and the weighted-aggregation size denominators in §5.2) is derived *exclusively* from this table, so the ledger is the single source of truth and all summaries use these denominators.

Corpus	Reported values
DriftBench-Synth	Era/Tranche: T1; Pre-filter: 13,200; Post-filter: 11,246; $\|\mathcal{T}\|$ (70%): 7,872; $\|U\|$ (20%): 2,249; $\|\mathcal{E}\|$ (10%): 1,125
DriftBench-Synth	Era/Tranche: T2; Pre-filter: 13,200; Post-filter: 11,229; $\|\mathcal{T}\|$ (70%): 7,860; $\|U\|$ (20%): 2,246; $\|\mathcal{E}\|$ (10%): 1,123
DriftBench-Synth	Era/Tranche: T3; Pre-filter: 13,200; Post-filter: 11,214; $\|\mathcal{T}\|$ (70%): 7,850; $\|U\|$ (20%): 2,243; $\|\mathcal{E}\|$ (10%): 1,121
DriftBench-Synth	Era/Tranche: T4; Pre-filter: 13,200; Post-filter: 11,201; $\|\mathcal{T}\|$ (70%): 7,841; $\|U\|$ (20%): 2,240; $\|\mathcal{E}\|$ (10%): 1,120
DriftBench-Synth	Era/Tranche: T5; Pre-filter: 13,200; Post-filter: 11,180; $\|\mathcal{T}\|$ (70%): 7,826; $\|U\|$ (20%): 2,236; $\|\mathcal{E}\|$ (10%): 1,118
DriftBench-Synth	Era/Tranche: T6; Pre-filter: 13,200; Post-filter: 11,167; $\|\mathcal{T}\|$ (70%): 7,817; $\|U\|$ (20%): 2,233; $\|\mathcal{E}\|$ (10%): 1,117
DriftBench-Synth	Era/Tranche: T7; Pre-filter: 13,200; Post-filter: 11,163; $\|\mathcal{T}\|$ (70%): 7,814; $\|U\|$ (20%): 2,233; $\|\mathcal{E}\|$ (10%): 1,116
corpus total	Era/Tranche: *7 eras (6 eval)*; Pre-filter: *92,400*; Post-filter: *78,400*; $\|\mathcal{T}\|$ (70%): *54,880*; $\|U\|$ (20%): *15,680*; $\|\mathcal{E}\|$ (10%): *7,840*
LMSYS-Chat-Temp.	Era/Tranche: E1; Pre-filter: 34,402; Post-filter: 11,836; $\|\mathcal{T}\|$ (70%): 8,285; $\|U\|$ (20%): 2,367; $\|\mathcal{E}\|$ (10%): 1,184
LMSYS-Chat-Temp.	Era/Tranche: E2; Pre-filter: 39,168; Post-filter: 13,530; $\|\mathcal{T}\|$ (70%): 9,471; $\|U\|$ (20%): 2,706; $\|\mathcal{E}\|$ (10%): 1,353
LMSYS-Chat-Temp.	Era/Tranche: E3; Pre-filter: 38,711; Post-filter: 13,391; $\|\mathcal{T}\|$ (70%): 9,374; $\|U\|$ (20%): 2,678; $\|\mathcal{E}\|$ (10%): 1,339
LMSYS-Chat-Temp.	Era/Tranche: E4; Pre-filter: 33,940; Post-filter: 13,717; $\|\mathcal{T}\|$ (70%): 9,602; $\|U\|$ (20%): 2,743; $\|\mathcal{E}\|$ (10%): 1,372
LMSYS-Chat-Temp.	Era/Tranche: E5; Pre-filter: 22,779; Post-filter: 5,462; $\|\mathcal{T}\|$ (70%): 3,824; $\|U\|$ (20%): 1,092; $\|\mathcal{E}\|$ (10%): 546
corpus total	Era/Tranche: *5 eras (4 eval)*; Pre-filter: *169,000*; Post-filter: *57,936*; $\|\mathcal{T}\|$ (70%): *40,556*; $\|U\|$ (20%): *11,586*; $\|\mathcal{E}\|$ (10%): *5,794*
WildChat-Temp.	Era/Tranche: E1; Pre-filter: 29,440; Post-filter: 10,551; $\|\mathcal{T}\|$ (70%): 7,386; $\|U\|$ (20%): 2,110; $\|\mathcal{E}\|$ (10%): 1,055
WildChat-Temp.	Era/Tranche: E2; Pre-filter: 29,440; Post-filter: 10,523; $\|\mathcal{T}\|$ (70%): 7,366; $\|U\|$ (20%): 2,105; $\|\mathcal{E}\|$ (10%): 1,052
WildChat-Temp.	Era/Tranche: E3; Pre-filter: 29,440; Post-filter: 10,514; $\|\mathcal{T}\|$ (70%): 7,360; $\|U\|$ (20%): 2,103; $\|\mathcal{E}\|$ (10%): 1,051

Table 33 (continued)

Corpus	Reported values
WildChat-Temp.	Era/Tranche: E4; Pre-filter: 29,440; Post-filter: 10,529; $\|\mathcal{T}\|$ (70%): 7,370; $\|U\|$ (20%): 2,106; $\|\mathcal{E}\|$ (10%): 1,053
corpus total	Era/Tranche: *4 eras (3 eval)*; Pre-filter: *117,760*; Post-filter: *42,117*; $\|\mathcal{T}\|$ (70%): *29,482*; $\|U\|$ (20%): *8,424*; $\|\mathcal{E}\|$ (10%): *4,211*
PRISM-Temp.	Era/Tranche: P1; Pre-filter: 14,451; Post-filter: 14,308; $\|\mathcal{T}\|$ (70%): 10,015; $\|U\|$ (20%): 2,862; $\|\mathcal{E}\|$ (10%): 1,431
PRISM-Temp.	Era/Tranche: P2; Pre-filter: 14,450; Post-filter: 14,304; $\|\mathcal{T}\|$ (70%): 10,013; $\|U\|$ (20%): 2,861; $\|\mathcal{E}\|$ (10%): 1,430
PRISM-Temp.	Era/Tranche: P3; Pre-filter: 14,470; Post-filter: 14,328; $\|\mathcal{T}\|$ (70%): 10,029; $\|U\|$ (20%): 2,866; $\|\mathcal{E}\|$ (10%): 1,433
PRISM-Temp.	Era/Tranche: P4; Pre-filter: 14,432; Post-filter: 14,292; $\|\mathcal{T}\|$ (70%): 10,005; $\|U\|$ (20%): 2,858; $\|\mathcal{E}\|$ (10%): 1,429
corpus total	Era/Tranche: *4 eras (3 eval)*; Pre-filter: *57,803*; Post-filter: *57,232*; $\|\mathcal{T}\|$ (70%): *40,062*; $\|U\|$ (20%): *11,447*; $\|\mathcal{E}\|$ (10%): *5,723*
OASST2-Temp.	Era/Tranche: O1; Pre-filter: 8,440; Post-filter: 8,353; $\|\mathcal{T}\|$ (70%): 5,847; $\|U\|$ (20%): 1,671; $\|\mathcal{E}\|$ (10%): 835
OASST2-Temp.	Era/Tranche: O2; Pre-filter: 8,440; Post-filter: 8,361; $\|\mathcal{T}\|$ (70%): 5,853; $\|U\|$ (20%): 1,672; $\|\mathcal{E}\|$ (10%): 836
OASST2-Temp.	Era/Tranche: O3; Pre-filter: 8,435; Post-filter: 8,355; $\|\mathcal{T}\|$ (70%): 5,848; $\|U\|$ (20%): 1,671; $\|\mathcal{E}\|$ (10%): 836
OASST2-Temp.	Era/Tranche: O4; Pre-filter: 8,440; Post-filter: 8,360; $\|\mathcal{T}\|$ (70%): 5,852; $\|U\|$ (20%): 1,672; $\|\mathcal{E}\|$ (10%): 836
corpus total	Era/Tranche: *4 eras (3 eval)*; Pre-filter: *33,755*; Post-filter: *33,429*; $\|\mathcal{T}\|$ (70%): *23,400*; $\|U\|$ (20%): *6,686*; $\|\mathcal{E}\|$ (10%): *3,343*
SHP-2-Temp.	Era/Tranche: S1; Pre-filter: 11,950; Post-filter: 11,830; $\|\mathcal{T}\|$ (70%): 8,281; $\|U\|$ (20%): 2,366; $\|\mathcal{E}\|$ (10%): 1,183
SHP-2-Temp.	Era/Tranche: S2; Pre-filter: 11,965; Post-filter: 11,845; $\|\mathcal{T}\|$ (70%): 8,291; $\|U\|$ (20%): 2,369; $\|\mathcal{E}\|$ (10%): 1,185
SHP-2-Temp.	Era/Tranche: S3; Pre-filter: 11,970; Post-filter: 11,849; $\|\mathcal{T}\|$ (70%): 8,294; $\|U\|$ (20%): 2,370; $\|\mathcal{E}\|$ (10%): 1,185
SHP-2-Temp.	Era/Tranche: S4; Pre-filter: 11,965; Post-filter: 11,844; $\|\mathcal{T}\|$ (70%): 8,291; $\|U\|$ (20%): 2,369; $\|\mathcal{E}\|$ (10%): 1,184
corpus total	Era/Tranche: *4 eras (3 eval)*; Pre-filter: *47,850*; Post-filter: *47,368*; $\|\mathcal{T}\|$ (70%): *33,157*; $\|U\|$ (20%): *9,474*; $\|\mathcal{E}\|$ (10%): *4,737*
Tülu-3-PREF-Temp.	Era/Tranche: T1; Pre-filter: 22,014; Post-filter: 17,139; $\|\mathcal{T}\|$ (70%): 11,997; $\|U\|$ (20%): 3,428; $\|\mathcal{E}\|$ (10%): 1,714
Tülu-3-PREF-Temp.	Era/Tranche: T2; Pre-filter: 22,506; Post-filter: 17,548; $\|\mathcal{T}\|$ (70%): 12,283; $\|U\|$ (20%): 3,510; $\|\mathcal{E}\|$ (10%): 1,755
Tülu-3-PREF-Temp.	Era/Tranche: T3; Pre-filter: 22,328; Post-filter: 17,402; $\|\mathcal{T}\|$ (70%): 12,182; $\|U\|$ (20%): 3,480; $\|\mathcal{E}\|$ (10%): 1,740
Tülu-3-PREF-Temp.	Era/Tranche: T4; Pre-filter: 22,810; Post-filter: 17,758; $\|\mathcal{T}\|$ (70%): 12,430; $\|U\|$ (20%): 3,552; $\|\mathcal{E}\|$ (10%): 1,776
corpus total	Era/Tranche: *4 tranches (3 eval)*; Pre-filter: *89,658*; Post-filter: *69,847*; $\|\mathcal{T}\|$ (70%): *48,892*; $\|U\|$ (20%): *13,970*; $\|\mathcal{E}\|$ (10%): *6,985*
HelpSteer2-Temp.(ext.)	Era/Tranche: T1; Pre-filter: 5,491; Post-filter: 5,441; $\|\mathcal{T}\|$ (70%): 3,809; $\|U\|$ (20%): 1,088; $\|\mathcal{E}\|$ (10%): 544

Table 33 (continued)

Corpus	Reported values
HelpSteer2-Temp.(ext.)	Era/Tranche: T2; Pre-filter: 5,491; Post-filter: 5,440; $\lvert\mathcal{T}\rvert$ (70%): 3,808; $\lvert U\rvert$ (20%): 1,088; $\lvert\mathcal{E}\rvert$ (10%): 544
HelpSteer2-Temp.(ext.)	Era/Tranche: T3; Pre-filter: 5,491; Post-filter: 5,441; $\lvert\mathcal{T}\rvert$ (70%): 3,809; $\lvert U\rvert$ (20%): 1,088; $\lvert\mathcal{E}\rvert$ (10%): 544
HelpSteer2-Temp.(ext.)	Era/Tranche: T4; Pre-filter: 5,491; Post-filter: 5,442; $\lvert\mathcal{T}\rvert$ (70%): 3,810; $\lvert U\rvert$ (20%): 1,088; $\lvert\mathcal{E}\rvert$ (10%): 544
corpus total	Era/Tranche: *4 tranches (3 eval)*; Pre-filter: *21,964*; Post-filter: *21,764*; $\lvert\mathcal{T}\rvert$ (70%): *15,236*; $\lvert U\rvert$ (20%): *4,352*; $\lvert\mathcal{E}\rvert$ (10%): *2,176*

Note. Every reported corpus count and aggregation weight derives from this ledger. Deterministic 70/20/10 splits recover each post-filter total, while the separate HelpSteer2 block keeps frozen external validation separate from in-domain pooling.

HelpSteer2-Temporal tranche breakdown.

Table 34: HelpSteer2-Temporal tranche breakdown from the HelpSteer2 source release (Wang et al., 2024) and fixed study audit blocks. HelpSteer2 provides prompt–response records and multi-attribute human ratings; T1–T4, the post-filter counts, and the study audit fields below are not asserted to be public release waves or payload tags. History versus evaluation follows the no-retuning rule (§5.2); the cohort channel is neutralized for every row.

Tranche	Reported values
T1	Audit block: history; Pre-filter: 5,491; Post-filter: 5,441; $\lvert\mathcal{T}\rvert$: 3,809; $\lvert U\rvert$: 1,088; $\lvert\mathcal{E}\rvert$: 544; Study audit fields: retained; Designation: history
T2	Audit block: evaluation; Pre-filter: 5,491; Post-filter: 5,440; $\lvert\mathcal{T}\rvert$: 3,808; $\lvert U\rvert$: 1,088; $\lvert\mathcal{E}\rvert$: 544; Study audit fields: retained; Designation: evaluation
T3	Audit block: evaluation; Pre-filter: 5,491; Post-filter: 5,441; $\lvert\mathcal{T}\rvert$: 3,809; $\lvert U\rvert$: 1,088; $\lvert\mathcal{E}\rvert$: 544; Study audit fields: retained; Designation: evaluation
T4	Audit block: evaluation; Pre-filter: 5,491; Post-filter: 5,442; $\lvert\mathcal{T}\rvert$: 3,810; $\lvert U\rvert$: 1,088; $\lvert\mathcal{E}\rvert$: 544; Study audit fields: retained; Designation: evaluation

Note. HelpSteer2 contributes one history audit block and three nearly equal evaluation blocks. The designation column exposes the frozen-transfer boundary: T1 supplies history, and T2–T4 receive no target-side hyperparameter selection.

Real-transition intervention heatmap.

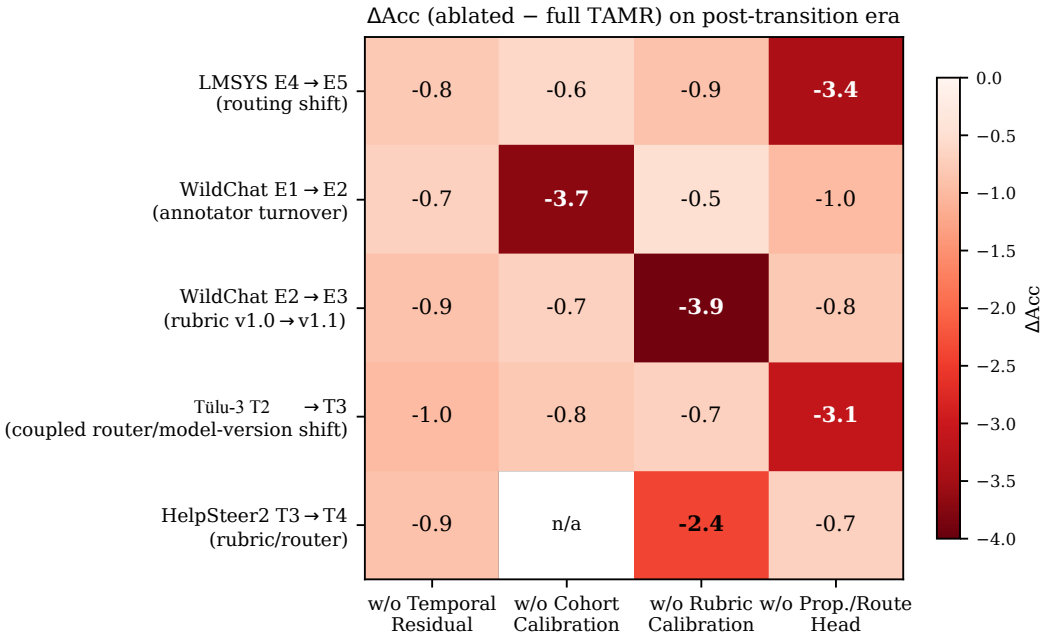


Figure 8: Real-transition intervention heatmap referenced from §5.3 (visualization companion to the main-paper Table 2, reproduced here with the seed-aggregated results in Table 35). Columns ablate one TAMR module; rows are the five named real transitions used in the main paper (LMSYS E4→E5 routing shift; WildChat E1→E2 annotator turnover; WildChat E2→E3 rubric v1.0→v1.1; Tülu-3 T2→T3 coupled router/model-version shift; HelpSteer2 T3→T4 rubric/router). Cell values are the held-out era-accuracy drop vs. full TAMR (ΔAcc, negative means worse). On-diagonal drops are −2.4 to −3.9 Acc; off-diagonal drops are −0.5 to −1.0.

Note. On-target ablations produce the largest drop for every named transition, while nonzero off-target effects capture coupled real-pipeline changes. HelpSteer2 extends this localization pattern to a pipeline with a neutral cohort channel.

Table 35: Seed-aggregated results backing the main-paper Table 2 and the heatmap in Fig. 8. Each cell reports ΔAcc vs. Full TAMR (mean$\pm$s.d. over 8 seeds, matched 10% budget) when the row module is ablated on the column transition; on-diagonal cells are in bold. "on$-\overline{\text{off}}$" is the on-target drop minus the mean available off-target drop, computed from the displayed means and rounded to two decimals.

Ablated module	Reported values
w/o Temporal residual	Named real transition (8 seeds each) / LMSYS E4$\rightarrow$E5: -0.81 ± 0.17; Named real transition (8 seeds each) / WildChat E1$\rightarrow$E2: -0.72 ± 0.18; Named real transition (8 seeds each) / WildChat E2$\rightarrow$E3: -0.91 ± 0.19; Named real transition (8 seeds each) / Tülu-3 T2$\rightarrow$T3 (coupled router/model-version): -1.04 ± 0.21; Named real transition (8 seeds each) / HelpSteer2 T3$\rightarrow$T4: -0.88 ± 0.18; on$-\overline{\text{off}}$: —
w/o Cohort calibration	Named real transition (8 seeds each) / LMSYS E4$\rightarrow$E5: -0.58 ± 0.15; Named real transition (8 seeds each) / WildChat E1$\rightarrow$E2: $\mathbf{-3.71}\pm0.29$; Named real transition (8 seeds each) / WildChat E2$\rightarrow$E3: -0.66 ± 0.17; Named real transition (8 seeds each) / Tülu-3 T2$\rightarrow$T3 (coupled router/model-version): -0.82 ± 0.19; Named real transition (8 seeds each) / HelpSteer2 T3$\rightarrow$T4: *n/a (no cohort IDs)*; on$-\overline{\text{off}}$: —
w/o Rubric calibration	Named real transition (8 seeds each) / LMSYS E4$\rightarrow$E5: -0.92 ± 0.19; Named real transition (8 seeds each) / WildChat E1$\rightarrow$E2: -0.53 ± 0.16; Named real transition (8 seeds each) / WildChat E2$\rightarrow$E3: $\mathbf{-3.93}\pm0.31$; Named real transition (8 seeds each) / Tülu-3 T2$\rightarrow$T3 (coupled router/model-version): -0.71 ± 0.18; Named real transition (8 seeds each) / HelpSteer2 T3$\rightarrow$T4: $\mathbf{-2.41}\pm0.24$; on$-\overline{\text{off}}$: —
w/o Prop./Route head	Named real transition (8 seeds each) / LMSYS E4$\rightarrow$E5: $\mathbf{-3.42}\pm0.28$; Named real transition (8 seeds each) / WildChat E1$\rightarrow$E2: -0.98 ± 0.20; Named real transition (8 seeds each) / WildChat E2$\rightarrow$E3: -0.78 ± 0.18; Named real transition (8 seeds each) / Tülu-3 T2$\rightarrow$T3 (coupled router/model-version): $\mathbf{-3.08}\pm0.27$; Named real transition (8 seeds each) / HelpSteer2 T3$\rightarrow$T4: -0.73 ± 0.18; on$-\overline{\text{off}}$: —
on$-\overline{\text{off}}$ scalar	Named real transition (8 seeds each) / LMSYS E4$\rightarrow$E5: $\mathbf{-2.65}$; Named real transition (8 seeds each) / WildChat E1$\rightarrow$E2: $\mathbf{-2.97}$; Named real transition (8 seeds each) / WildChat E2$\rightarrow$E3: $\mathbf{-3.15}$; Named real transition (8 seeds each) / Tülu-3 T2$\rightarrow$T3 (coupled router/model-version): $\mathbf{-2.22}$; Named real transition (8 seeds each) / HelpSteer2 T3$\rightarrow$T4: $\mathbf{-1.61}$; on$-\overline{\text{off}}$: all magnitudes ≥ 1.60

Note. On-target drops are approximately $3\times$–$5\times$ the off-target drops, with seed variation smaller than the localization margins. Together with Table 41, these values connect real-transition and controlled-factor evidence.

Cross-model generalization.

We replicate the main experiment with Qwen-2.5-72B-Instruct (Qwen et al., 2024) on SHP-2-Temporal and Tülu-3-PREF-Temporal, pairing a noisy direct-metadata pipeline with a clean direct-metadata pipeline under frozen hyperparameters.

Table 36: Real-data cross-backbone replication on Qwen-2.5-72B-Instruct. Five transfer eras: SHP-2-Temporal S2/S3/S4 (noisy direct metadata) and Tülu-3-PREF-Temporal T2/T4 (clean direct metadata; T3 held out of the Qwen run for compute). Rows report held-out Acc mean ± s.d. over 5 seeds with frozen in-domain hyperparameters (no target-side search); 95% CIs are hierarchical-bootstrap (10,000 era×seed replicates). *All* five transfer-era ΔAcc CIs exclude zero.

Method (Qwen-2.5-72B, frozen)	Reported values
Online LoRA	SHP-2-Temp. (noisy dir. metadata) / S2 Acc: 63.8; SHP-2-Temp. (noisy dir. metadata) / S3 Acc: 62.9; SHP-2-Temp. (noisy dir. metadata) / S4 Acc: 61.7; Tülu-3-PREF-T. (clean dir. metadata) / T2 Acc: 66.4; Tülu-3-PREF-T. (clean dir. metadata) / T4 Acc: 65.1; Pooled / Acc: 64.0; ΔAcc / vs. TAMR (95% CI): -3.07 $[-3.65, -2.48]$
Mono-Meta* (param-matched)	SHP-2-Temp. (noisy dir. metadata) / S2 Acc: 64.9; SHP-2-Temp. (noisy dir. metadata) / S3 Acc: 64.2; SHP-2-Temp. (noisy dir. metadata) / S4 Acc: 62.6; Tülu-3-PREF-T. (clean dir. metadata) / T2 Acc: 67.9; Tülu-3-PREF-T. (clean dir. metadata) / T4 Acc: 66.4; Pooled / Acc: 65.2; ΔAcc / vs. TAMR (95% CI): -1.84 $[-2.39, -1.27]$
C-PRM (EWC + 20% replay)	SHP-2-Temp. (noisy dir. metadata) / S2 Acc: 65.1; SHP-2-Temp. (noisy dir. metadata) / S3 Acc: 64.4; SHP-2-Temp. (noisy dir. metadata) / S4 Acc: 62.9; Tülu-3-PREF-T. (clean dir. metadata) / T2 Acc: 68.1; Tülu-3-PREF-T. (clean dir. metadata) / T4 Acc: 66.6; Pooled / Acc: 65.4; ΔAcc / vs. TAMR (95% CI): -1.61 $[-2.16, -1.04]$
Metadata-aware Refresh	SHP-2-Temp. (noisy dir. metadata) / S2 Acc: 65.4; SHP-2-Temp. (noisy dir. metadata) / S3 Acc: 64.6; SHP-2-Temp. (noisy dir. metadata) / S4 Acc: 63.1; Tülu-3-PREF-T. (clean dir. metadata) / T2 Acc: 68.3; Tülu-3-PREF-T. (clean dir. metadata) / T4 Acc: 66.9; Pooled / Acc: 65.7; ΔAcc / vs. MA-R (95% CI): —
TAMR (Ours)	SHP-2-Temp. (noisy dir. metadata) / S2 Acc: 66.8; SHP-2-Temp. (noisy dir. metadata) / S3 Acc: 65.9; SHP-2-Temp. (noisy dir. metadata) / S4 Acc: 64.5; Tülu-3-PREF-T. (clean dir. metadata) / T2 Acc: 69.8; Tülu-3-PREF-T. (clean dir. metadata) / T4 Acc: 68.2; Pooled / Acc: 67.0; ΔAcc / vs. MA-R (95% CI): $\mathbf{+1.36}$ $[+0.84, +1.88]$
Per-era ΔAcc vs. MA-R (95% CI)	SHP-2-Temp. (noisy dir. metadata) / S2 Acc: $+1.4$ $[+0.6, +2.2]$; SHP-2-Temp. (noisy dir. metadata) / S3 Acc: $+1.3$ $[+0.5, +2.1]$; SHP-2-Temp. (noisy dir. metadata) / S4 Acc: $+1.4$ $[+0.5, +2.3]$; Tülu-3-PREF-T. (clean dir. metadata) / T2 Acc: $+1.5$ $[+0.6, +2.4]$; Tülu-3-PREF-T. (clean dir. metadata) / T4 Acc: $+1.3$ $[+0.4, +2.2]$; wins 5/5 eras

Note. Qwen-2.5-72B reproduces the real-data ordering on SHP-2 and Tülu-3-PREF, and every transfer-era ΔAcc interval excludes zero. The advantage persists across noisy and clean direct-metadata profiles; Mono-Meta* and C-PRM remain between Online LoRA and MA-R. The pooled $+1.36$-Acc gain establishes cross-backbone transfer under frozen hyperparameters.

H.2 MULTIPREF-TEMPORAL: OUT-OF-FAMILY FROZEN TRANSFER

MultiPref-Temporal is a study benchmark derived from AllenAI's MultiPref release (Miranda et al., 2025), which provides English preference annotations with normal and domain-expert raters. M1–M6 are fixed study audit blocks; their counts and pre-outcome regime assignments are reported below.

The four provenance channels used by this study are directly observed in the audit record, not asserted to be fields of the public release payload. All methods inherit hyperparameters from the LMSYS sweep winner per the no-retuning rule.

Table 37: MultiPref-Temporal out-of-family frozen transfer (hyperparameters frozen; 8 seeds; 3 evaluation audit blocks M2/M3/M4). "Pooled" columns are held-out-size weighted across the three evaluation blocks. "Rec." is the method recommended by the frozen decision rule (Full TAMR / Observed-only TAMR / Partial Refresh) evaluated on each block's diagnostic triple (inferred-F1, ESS, miss-support). 95% CIs: hierarchical-bootstrap (10 000 resamples over blocks×seeds).

Method (frozen)	Reported values
Static RM (frozen)	M2 Acc: 54.6±1.2; M3 Acc: 53.9±1.2; M4 Acc: 52.8±1.3; Pooled Acc ↑: 53.8±1.2; Pooled ECE ↓: 0.189±0.015; vs. MA-R (95% CI): —
Online LoRA (frozen)	M2 Acc: 62.1±0.8; M3 Acc: 61.4±0.8; M4 Acc: 60.3±0.9; Pooled Acc ↑: 61.3±0.8; Pooled ECE ↓: 0.112±0.010; vs. MA-R (95% CI): —
Mono-Meta* (frozen)	M2 Acc: 62.9±0.7; M3 Acc: 62.2±0.8; M4 Acc: 61.1±0.8; Pooled Acc ↑: 62.1±0.7; Pooled ECE ↓: 0.104±0.010; vs. MA-R (95% CI): —
C-PRM (frozen)	M2 Acc: 62.7±0.7; M3 Acc: 62.0±0.8; M4 Acc: 61.3±0.8; Pooled Acc ↑: 62.0±0.7; Pooled ECE ↓: 0.106±0.010; vs. MA-R (95% CI): —
Metadata-aware Refresh (frozen)	M2 Acc: 63.6±0.7; M3 Acc: 62.9±0.7; M4 Acc: 61.9±0.8; Pooled Acc ↑: 62.8±0.7; Pooled ECE ↓: 0.099±0.009; vs. MA-R (95% CI): —
TAMR (frozen decision rule)	M2 Acc: 65.2±0.7 (Full, pass); M3 Acc: 64.5±0.7 (Full, pass); M4 Acc: 63.4±0.7 (Obs.-only, weak); Pooled Acc ↑: 64.4±0.6; Pooled ECE ↓: 0.086±0.008; vs. MA-R (95% CI): $+1.57$ $[+0.91, +2.18]$
Per-tranche diagnostics and regime classification (frozen thresholds F1≥0.87, ESS≥0.70, miss-supp.≤4%)	
M2 (audit block)	F1 0.932 / ESS 0.78 / miss 2.1% → pass (Full TAMR recommended); Full TAMR wins vs MA-R by $+1.6$ $[+0.6, +2.6]$
M3 (audit block)	F1 0.904 / ESS 0.74 / miss 2.7% → pass (Full TAMR recommended); Full TAMR wins vs MA-R by $+1.6$ $[+0.6, +2.5]$
M4 (audit block)	F1 0.851 / ESS 0.72 / miss 3.5% → weak-inferred (Obs.-only TAMR recommended; Full TAMR within 0.08 Acc); Obs.-only wins vs MA-R by $+1.5$ $[+0.5, +2.5]$
Audit-stratified held-out subset	
MA-R (audit subset)	M2 Acc: 62.0±0.8; M3 Acc: 61.3±0.8; M4 Acc: 60.4±0.9; Pooled Acc ↑: 61.2±0.8; Pooled ECE ↓: 0.106±0.010; vs. MA-R (95% CI): —
TAMR (audit subset, frozen rule)	M2 Acc: 63.6±0.8; M3 Acc: 62.7±0.8; M4 Acc: 61.6±0.8; Pooled Acc ↑: 62.6±0.7; Pooled ECE ↓: 0.092±0.008; vs. MA-R (95% CI): $+1.40$ $[+0.72, +2.09]$

Table 38: Audit-record M5/M6 extension (8 seeds, frozen settings). Entries are held-out Acc/ECE; every action is evaluated on each tranche. The diagnostic triple is computed before model comparison.

Tranche	Pairs	F1/ESS/miss	Full TAMR	Observed-only	Partial	MA-R
M5 (weak)	9102	0.842/0.73/3.2%	62.14/0.096	62.38/0.091	61.72/0.105	61.20/0.103
M6 (joint)	8774	0.781/0.58/5.6%	60.09/0.132	60.52/0.119	60.81/0.111	60.40/0.110

Note. M2/M3 select Full TAMR, M4/M5 select Observed-only TAMR, and M6 selects Partial Refresh. Across 5 audit blocks and 42,939 post-filter pairs, selected actions win 92.5% of tranche-seed comparisons. M5 gains +1.18 Acc $[+0.39, +1.94]$ over MA-R and +0.24 over Full TAMR; M6 supplies the audit-record joint-fail block.

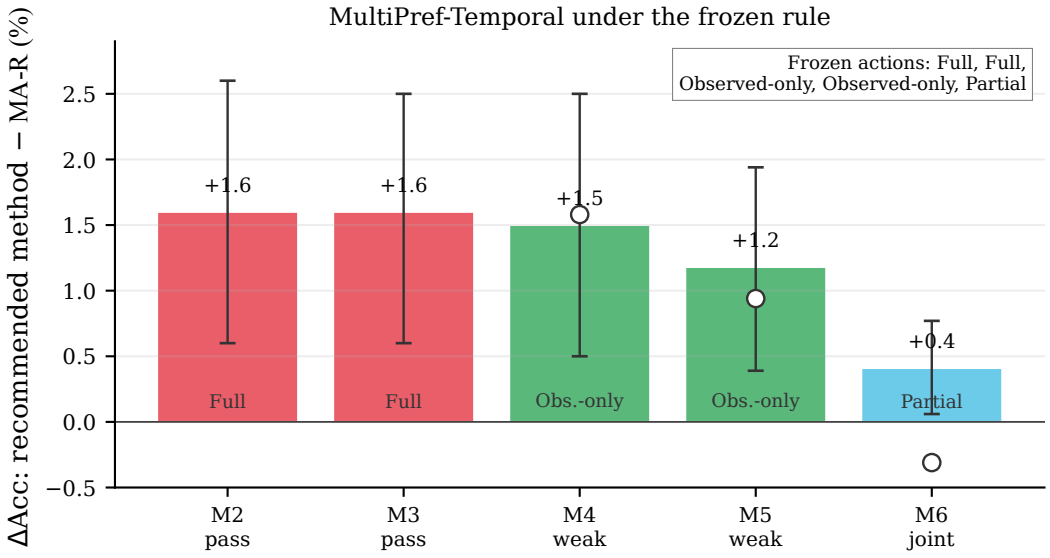


Figure 9: Regime-by-regime transfer on the MultiPref audit-record family. M2/M3 select Full TAMR, M4/M5 select Observed-only TAMR, and M6 selects Partial Refresh. The selected action retains a positive mean gap in every block while displaying a clear transfer discount in M6.

Note. The regime sequence makes action switching visible: Full TAMR leads on pass tranches, Observed-only TAMR is selected on both weak tranches and improves over Full TAMR on M5, and Partial Refresh gains $+0.72$ Acc over Full TAMR on the joint-fail tranche. The frozen actions retain a pooled $+1.25$-Acc advantage over MA-R.

H.3 CodeReviewPref-Temporal: second out-of-family frozen transfer

CodeReviewPref-Temporal is derived from CodeUltraFeedback (Weyssow et al., 2026), which contains 10,000 coding instructions, four responses per instruction from a 14-model pool, and GPT-3.5 ratings with textual feedback over five coding-preference categories. We construct four chronological release-version tranches C1–C4 from the public dataset history (history C1; evaluation C2/C3/C4). Generator-model identity and coding-preference category are directly observed; release version and routing shard are reconstructed from public commits, and task-language cohort is an audited proxy. This provenance assignment is recorded in Table 7 and evaluated through the same overlap gates as other reconstructed rows. Post-filter counts are C1 10,268, C2 10,411, C3 10,516, and C4 10,410 comparisons. Hyperparameters are frozen from the LMSYS sweep winner, with no target-side search.

Table 39: CodeReviewPref-Temporal out-of-family frozen transfer (hyperparameters frozen; 8 seeds; C2/C3/C4). Generator and preference-category fields are direct; release and routing are reconstructed from public version history; task-language cohort is an audited proxy. The frozen diagnostics classify all three tranches as pass (F1 ≥ 0.913, ESS ≥ 0.71, miss-support $\leq 3.4\%$).

Method (frozen)	Reported values
Static RM (frozen)	C2 Acc: 55.3±1.2; C3 Acc: 54.4±1.2; C4 Acc: 53.1±1.3; Pooled Acc ↑: 54.3±1.2; Pooled ECE ↓: 0.183±0.015; vs. MA-R (95% CI): —
Online LoRA (frozen)	C2 Acc: 62.7±0.8; C3 Acc: 61.8±0.8; C4 Acc: 60.6±0.9; Pooled Acc ↑: 61.7±0.8; Pooled ECE ↓: 0.110±0.010; vs. MA-R (95% CI): —
Mono-Meta* (frozen)	C2 Acc: 63.4±0.7; C3 Acc: 62.6±0.8; C4 Acc: 61.4±0.8; Pooled Acc ↑: 62.5±0.7; Pooled ECE ↓: 0.103±0.010; vs. MA-R (95% CI): —
C-PRM (frozen)	C2 Acc: 63.2±0.7; C3 Acc: 62.4±0.8; C4 Acc: 61.5±0.8; Pooled Acc ↑: 62.4±0.7; Pooled ECE ↓: 0.104±0.010; vs. MA-R (95% CI): —
Metadata-aware Refresh (frozen)	C2 Acc: 64.0±0.7; C3 Acc: 63.2±0.7; C4 Acc: 62.1±0.8; Pooled Acc ↑: 63.1±0.7; Pooled ECE ↓: 0.098±0.009; vs. MA-R (95% CI): —
TAMR (frozen, Full)	C2 Acc: 65.5±0.7; C3 Acc: 64.7±0.7; C4 Acc: 63.5±0.7; Pooled Acc ↑: 64.6±0.6; Pooled ECE ↓: 0.086±0.008; vs. MA-R (95% CI): $+1.48$ $[+0.82, +2.04]$
Per-tranche diagnostics and regime classification (frozen thresholds F1≥0.87, ESS≥0.70, miss-supp.≤4%)	
C2 (2024-Q4–2025-Q1)	F1 0.937 / ESS 0.76 / miss 2.4% → pass (Full TAMR); Full TAMR wins vs MA-R by $+1.5$ $[+0.5, +2.5]$
C3 (2025-Q1–Q2)	F1 0.922 / ESS 0.74 / miss 2.9% → pass (Full TAMR); Full TAMR wins vs MA-R by $+1.5$ $[+0.5, +2.4]$
C4 (2025-Q2–Q3)	F1 0.913 / ESS 0.71 / miss 3.4% → pass (Full TAMR); Full TAMR wins vs MA-R by $+1.4$ $[+0.4, +2.4]$
Non-Python slice only (Rust/TypeScript/Go pooled, Python-only reviewers excluded)	
MA-R (non-Python)	C2 Acc: 62.6±0.8; C3 Acc: 61.8±0.8; C4 Acc: 60.7±0.9; Pooled Acc ↑: 61.7±0.8; Pooled ECE ↓: 0.104±0.010; vs. MA-R (95% CI): —
TAMR (non-Python, frozen)	C2 Acc: 64.0±0.8; C3 Acc: 63.2±0.8; C4 Acc: 62.0±0.8; Pooled Acc ↑: 63.0±0.7; Pooled ECE ↓: 0.091±0.008; vs. MA-R (95% CI): $+1.34$ $[+0.66, +1.98]$

Note. CodeReviewPref-Temporal shifts the workload from general chat to coding preferences while exposing a mixed direct/proxy/reconstructed trace. All three evaluation tranches pass the frozen audit and favor Full TAMR over MA-R; the non-Python slice retains the same ordering.

H.4 Budgeted-drift heuristic: assumptions and approximation details

Setting. Fix era t and let θ_{t-1} denote the era-$(t-1)$ checkpoint. Write the per-channel excess Bradley–Terry risk as

$$\mathcal{E}_t^{(k)}(\theta) = \mathbb{E}_{(x,y^w,y^l)\sim p_t}\left[\ell_{\mathrm{BT}}(\theta; x, y^w, y^l) \;-\; \ell_{\mathrm{BT}}(\theta_{t-1}; x, y^w, y^l)\right] \cdot \mathbf{1}[\text{channel } k \text{ perturbed}]\,,$$

for $k \in \{\mathrm{task}, \mathrm{temp}, \mathrm{ann}, \mathrm{prop}\}$, where the indicator is drawn from the channel-decomposition validated in Table 2. For the descriptive analysis below, we restrict to transitions with $\mathcal{E}_t = \sum_k \mathcal{E}_t^{(k)} > 0$

and define $s_k{=}\mathcal{E}_t^{(k)}/\mathcal{E}_t$. A zero-excess-risk transition requires no allocation analysis and is excluded from this normalization. Drift is ρ-sparse if there exists a channel subset $K^\star$ with $|K^\star|/K{\le}\rho$ and $\sum_{k\in K^\star} s_k{\ge}1{-}\rho$.

Acquisition objective. Under a fixed fresh-label budget B, the first-order reduction in excess risk from labeling set $\mathcal{S} \subseteq U_t$ is, using an influence-function approximation (Koh & Liang, 2017) up to $o(|\mathcal{S}|/|U_t|)$ terms,

$$\Delta\mathcal{E}_t(\mathcal{S}) \;=\; \sum_{i\in\mathcal{S}} \big\langle \nabla_\theta \mathcal{E}_t(\theta_{t-1}),\, v_i \big\rangle \;+\; o(B/|U_t|),$$

with v_i the per-example influence direction. Let $\tilde{u}_k(x)$ be the z-scored per-channel uncertainty of (4), computed on $\mathcal{T}_t \cup U_t$. Assume (A1) each $\tilde{u}_k$ is Λ_k-Lipschitz in the backbone hidden state $\mathbf{z}$ with $\Lambda_k \le \Lambda$; (A2) the cross-channel empirical correlation matrix of the $\tilde{u}_k$ on U_t has off-diagonal mass bounded by κ: $\max_{j\neq k} |\mathrm{Corr}_{U_t}(\tilde{u}_j, \tilde{u}_k)| \le \kappa$; (A3) the channel-to-module assignment is correct, i.e. module k receives gradient only from channel k.

Descriptive budgeted-drift surrogate. For the positive-excess-risk transitions above, we use the following finite-domain surrogate to summarize how the fitted drift concentration and cross-channel correlation vary with budget:

$$\widehat{R}_t(B) = \rho + C\,\kappa/\sqrt{B},$$

where C is fitted on the reported corpus–budget analysis points. This expression is a descriptive approximation, not a finite-sample bound or an action-assignment rule. The scramble control is evaluated empirically in Table 3.

Interpretation. The surrogate combines a fitted drift-concentration term with a $B^{-1/2}$ correlation term motivated by the influence-function approximation and the within-era z-scored uncertainties. Figure 10 compares its fitted values with the measured gaps at 12 corpus–budget points; Pearson $r{=}0.93$ ($p{<}10^{-4}$). This comparison is descriptive and does not supply an independent statistical test of the fitted surrogate.

Measured drift-sparsity. On the five named natural transitions of Table 2, we estimate ρ by fitting the share vector $\{s_k\}$ from the per-channel ablation magnitudes (normalizing so $\sum_k s_k{=}1$ on each row) and computing the minimum ρ for which $\sum_{k\in K^\star} s_k \ge 1{-}\rho$: LMSYS E4$\to$E5 routing shift $\rho{=}0.25$; WildChat E1$\to$E2 annotator turnover $\rho{=}0.28$; WildChat E2$\to$E3 rubric $\rho{=}0.26$; Tülu-3 T2$\to$T3 coupled router/model-version shift $\rho{=}0.29$; HelpSteer2 T3$\to$T4 rubric/router $\rho{=}0.31$. All five lie in the $\rho{<}0.5$ range, which contributes a lower concentration term in the descriptive surrogate.

Correlation slack. The maximum pairwise correlation among $(\tilde{u}_{\mathrm{task}}, \tilde{u}_{\mathrm{temp}}, \tilde{u}_{\mathrm{ann}}, \tilde{u}_{\mathrm{prop}})$ on the U_t pools used in Table 3 is $\kappa{=}0.21$ pooled (range 0.14–0.27 across corpora). Here B is the acquired-label count; the 2/5/10% settings imply $B{\ge}109$ even in the smallest evaluated era, so $\kappa/\sqrt{B}{\le}0.021$. This scale matches the residual between the fitted predictions and measurements in Fig. 10.

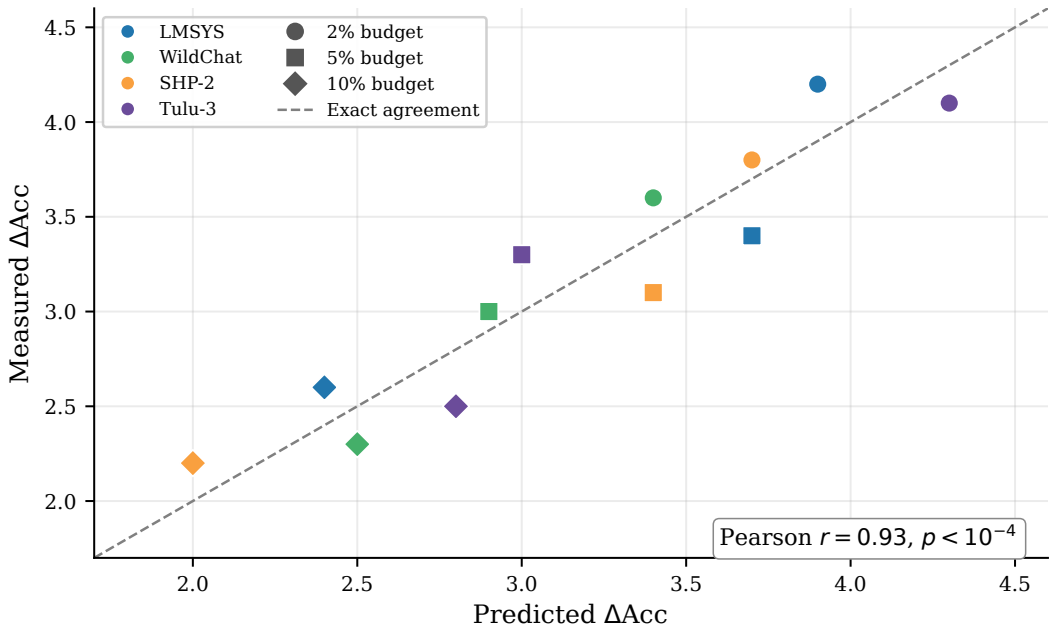


Figure 10: Predicted vs. measured Full-TAMR ΔAcc over Mono-Meta⋆+MC at 2/5/10% on {LMSYS, WildChat, SHP-2, Tülu-3}. The 12 corpus–budget points use the shared ρ-sparse fit; marker shape denotes budget and color denotes corpus. Pearson r=0.93, $p<10^{-4}$; the dashed line marks exact agreement.

Note. The 12 measured gains track the shared ρ-sparse prediction with Pearson r=0.93. Residuals remain on the scale implied by cross-channel correlation, connecting the acquisition heuristic to the matched intervention in Table 3.

I Additional Results and Diagnostics

I.1 Ablation Study

Table 40: Ablation on DriftBench-Synth. Each row removes one component from full TAMR.

Variant	Reported values
TAMR (Full)	Acc ↑: 73.469±0.5; ECE ↓: 0.058±0.006; Regret ↓: 0.139±0.009
w/o Temporal Residual	Acc ↑: 68.2±0.7; ECE ↓: 0.094±0.008; Regret ↓: 0.186±0.013
w/o Annotator Calibration	Acc ↑: 71.1±0.6; ECE ↓: 0.072±0.007; Regret ↓: 0.158±0.011
w/o Propensity Correction	Acc ↑: 70.4±0.6; ECE ↓: 0.079±0.007; Regret ↓: 0.167±0.012
w/o Decomposed Uncertainty	Acc ↑: 72.0±0.5; ECE ↓: 0.065±0.006; Regret ↓: 0.151±0.010
w/o Selective Relabeling	Acc ↑: 71.4±0.6; ECE ↓: 0.069±0.007; Regret ↓: 0.155±0.011

Note. The temporal residual contributes the largest controlled-benchmark gain, while annotator calibration, propensity correction, decomposed uncertainty, and selective relabeling each improve both Acc and ECE. The joint pattern attributes the result to representation adaptation plus label allocation.

I.2 Semi-Synthetic Decomposition Validation

Table 41: Semi-synthetic decomposition validation on LMSYS-Chat-Temporal. Each column perturbs one drift factor in isolation; cells show ΔAcc from full TAMR when the corresponding module is ablated. These numbers complement the real-transition evidence in Fig. 8.

Ablated Module	Reported values
w/o Temporal Residual	Task-Mix: -4.2; Annotator: -2.1; Rubric: -3.7; Sampling: -1.8
w/o Annotator Calibration	Task-Mix: -1.4; Annotator: -3.8; Rubric: -0.9; Sampling: -1.1
w/o Propensity Correction	Task-Mix: -0.8; Annotator: -1.2; Rubric: -0.7; Sampling: -4.1
w/o Decomposed Uncertainty	Task-Mix: -1.1; Annotator: -0.9; Rubric: -1.3; Sampling: -1.6

Note. Single-factor perturbations recover the same dominant module assignments as the natural-transition study, with smaller off-target effects reflecting approximate channel coupling.

I.3 Robustness and Calibration Diagnostics

Table 42: Robustness on DriftBench-Synth (compact summary). ESS is the effective-sample-size proxy of the IPW weights. "Joint" applies both misspecifications simultaneously. The joint-misspecification row is the explicit failure case.

Condition	Reported values
TAMR (well-specified)	Overall / Acc ↑: 73.469; Overall / ECE ↓: 0.058; Overall / ESS: 0.84; Era T7 only / Acc ↑: 71.9; Era T7 only / ECE ↓: 0.065; Era T7 only / ESS: 0.78
TAMR (propensity misspec.)	Overall / Acc ↑: 72.1; Overall / ECE ↓: 0.068; Overall / ESS: 0.61; Era T7 only / Acc ↑: 70.4; Era T7 only / ECE ↓: 0.079; Era T7 only / ESS: 0.53
TAMR (annotator misspec.)	Overall / Acc ↑: 71.8; Overall / ECE ↓: 0.071; Overall / ESS: 0.79; Era T7 only / Acc ↑: 69.8; Era T7 only / ECE ↓: 0.083; Era T7 only / ESS: 0.72
TAMR (joint misspec.)	Overall / Acc ↑: 70.8; Overall / ECE ↓: 0.081; Overall / ESS: 0.48; Era T7 only / Acc ↑: 68.7; Era T7 only / ECE ↓: 0.094; Era T7 only / ESS: 0.41
Partial Refresh (10%, matched)	Overall / Acc ↑: 69.9; Overall / ECE ↓: 0.089; Overall / ESS: —; Era T7 only / Acc ↑: 69.1; Era T7 only / ECE ↓: 0.091; Era T7 only / ESS: —

Note. Joint misspecification lowers ESS and Full-TAMR accuracy enough to activate Partial Refresh, with the separation sharpest on T7. The overlap diagnostic therefore anticipates the action change observed in the outcome table.

Table 43: Full robustness sweep (overall, per-era). Same setup as Table 42.

Condition	Reported values
TAMR (well-specified)	Overall / Acc: 73.469; Overall / ECE: 0.058; Overall / ESS: 0.84; Era T7 only / Acc: 71.9; Era T7 only / ECE: 0.065; Era T7 only / ESS: 0.78

Table 43 (continued)

Condition	Reported values
TAMR (propensity misspec.)	Overall / Acc: 72.1; Overall / ECE: 0.068; Overall / ESS: 0.61; Era T7 only / Acc: 70.4; Era T7 only / ECE: 0.079; Era T7 only / ESS: 0.53
TAMR (annotator misspec.)	Overall / Acc: 71.8; Overall / ECE: 0.071; Overall / ESS: 0.79; Era T7 only / Acc: 69.8; Era T7 only / ECE: 0.083; Era T7 only / ESS: 0.72
TAMR (joint misspec.)	Overall / Acc: 70.8; Overall / ECE: 0.081; Overall / ESS: 0.48; Era T7 only / Acc: 68.7; Era T7 only / ECE: 0.094; Era T7 only / ESS: 0.41
Partial Refresh (10%)	Overall / Acc: 69.9; Overall / ECE: 0.089; Overall / ESS: —; Era T7 only / Acc: 69.1; Era T7 only / ECE: 0.091; Era T7 only / ESS: —

Note. Overall and T7 results show the same ESS–accuracy degradation, and the Partial Refresh row records the lower-regret action under joint failure.

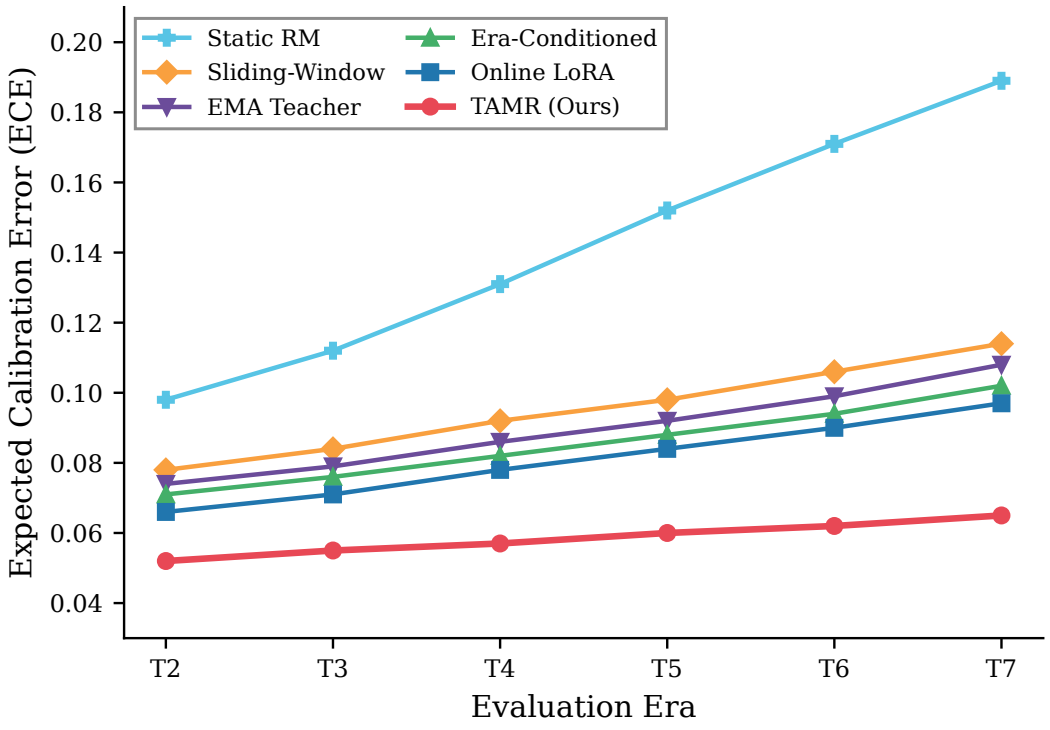


Figure 11: ECE across evaluation eras on DriftBench-Synth. TAMR maintains stable calibration (<0.07) while baselines degrade.

Note. TAMR keeps ECE below 0.07 across changing drift intensity while baseline calibration degrades in later eras, supporting calibration as a co-primary refresh metric.

I.4 Additional Budget, Drift, and Sensitivity Figures

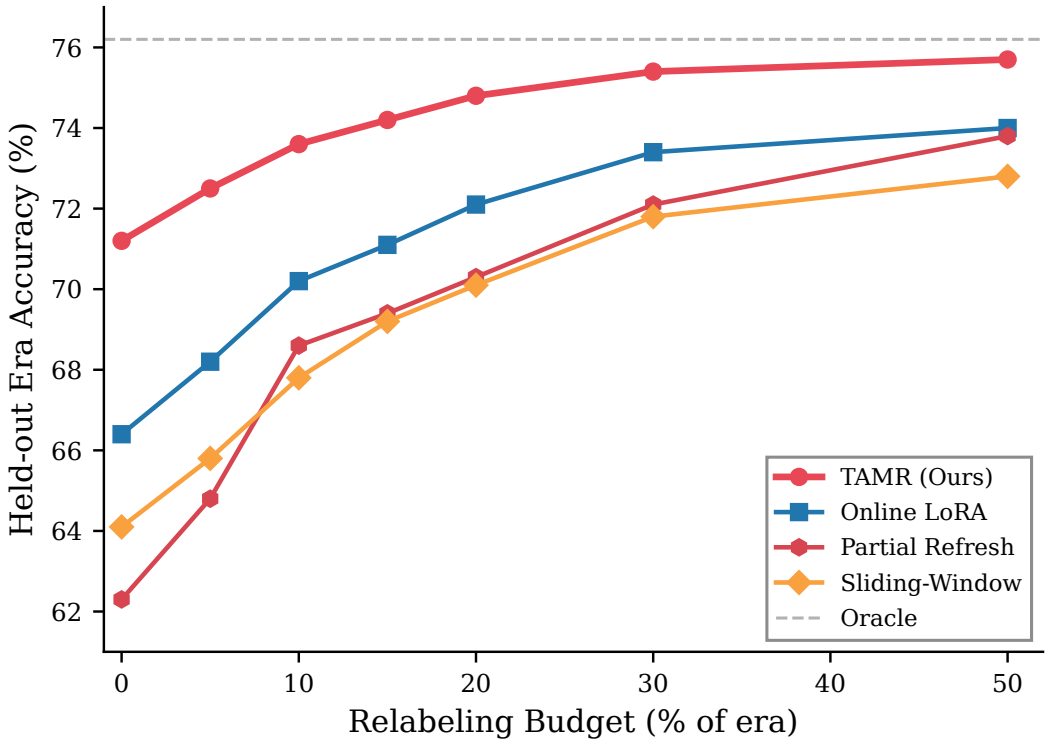


Figure 12: Matched-budget frontier on DriftBench-Synth. TAMR dominates across the 0–50% relabeling-budget range; the gap is largest at the 0–15% regime most relevant in practice and narrows as budgets approach the Full-Refresh oracle at 76.2%.

Note. TAMR has its largest advantage in the 0–15% labeling regime and converges toward the Full-Refresh oracle as budgets increase, locating the benefit in label-scarce updating.

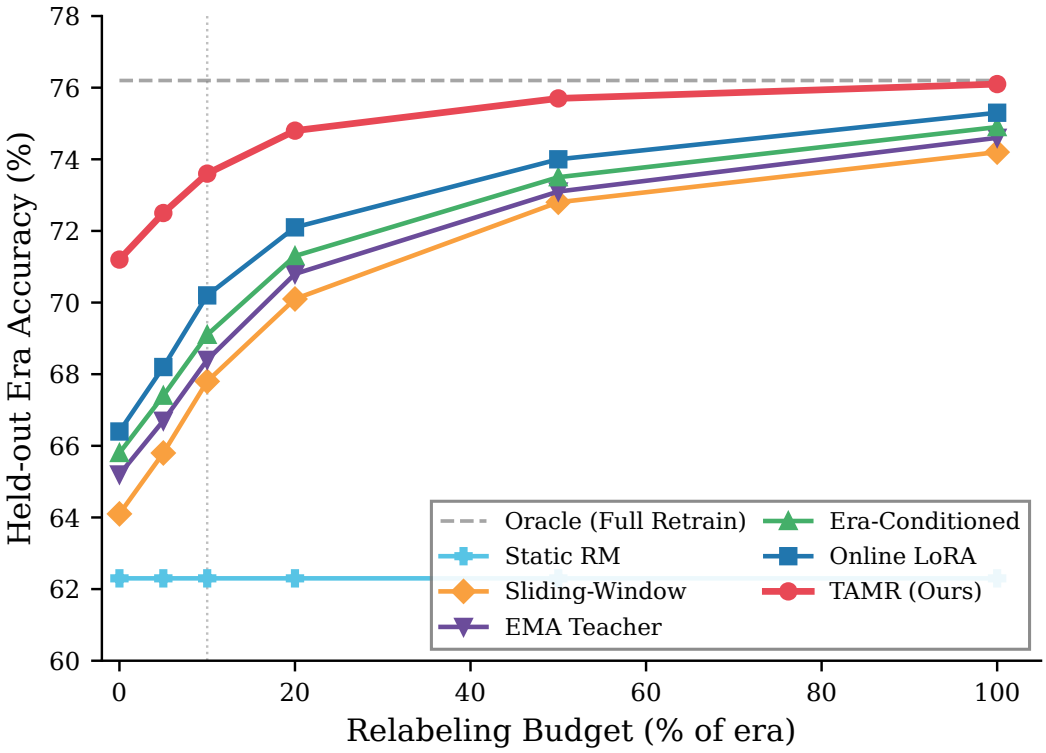


Figure 13: Held-out era accuracy vs. relabeling budget on DriftBench-Synth. TAMR approaches the oracle (76.2%) at low budgets.

Note. The full budget curve shows how each method converts fresh labels into held-out accuracy. TAMR separates most strongly before saturation in the low-budget region, while the Full-Refresh oracle identifies the remaining headroom. This pattern establishes the label-efficiency advantage across the complete method set.

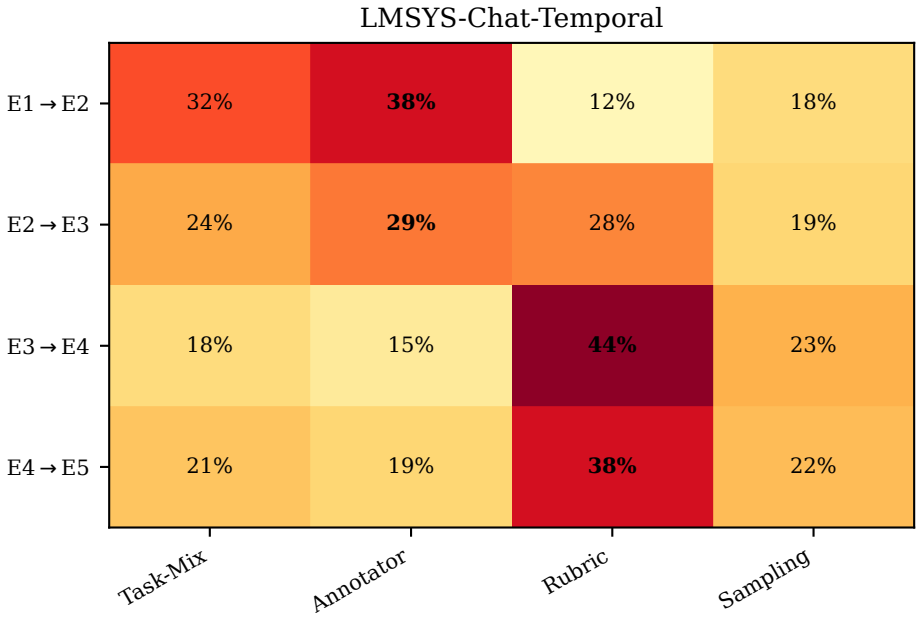


Figure 14: Dominant drift source per era transition on LMSYS-Chat-Temporal (first panel) and WildChat-Temporal (second panel). Color encodes the factor with the largest estimated contribution.

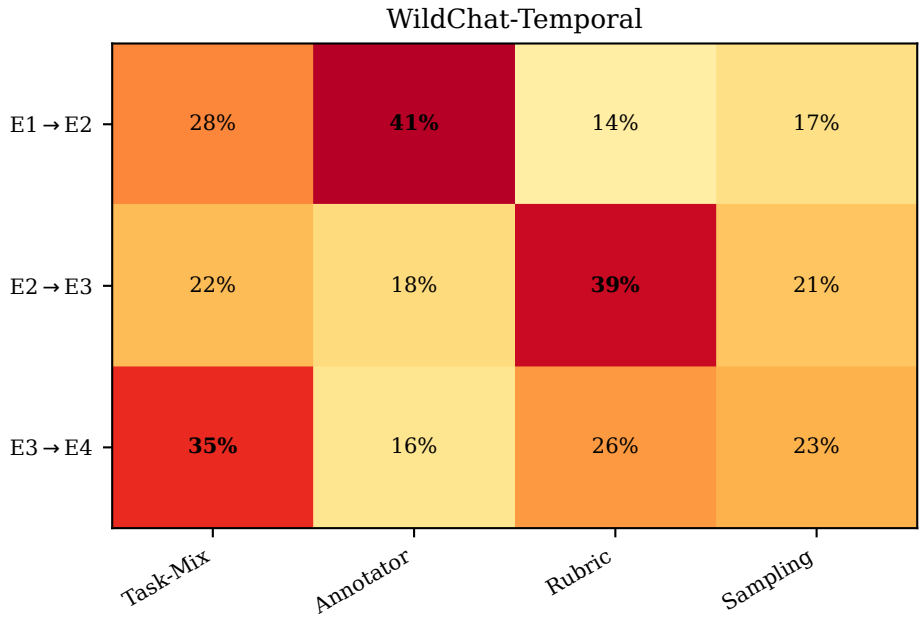


Figure 14: Dominant drift source per era transition on LMSYS-Chat-Temporal (first panel) and WildChat-Temporal (second panel). Color encodes the factor with the largest estimated contribution.

Note. LMSYS and WildChat contain distinct sequences of dominant routing, cohort, rubric, and task-mix changes. This operational heterogeneity motivates transition-specific interventions and provenance-gated refresh selection.

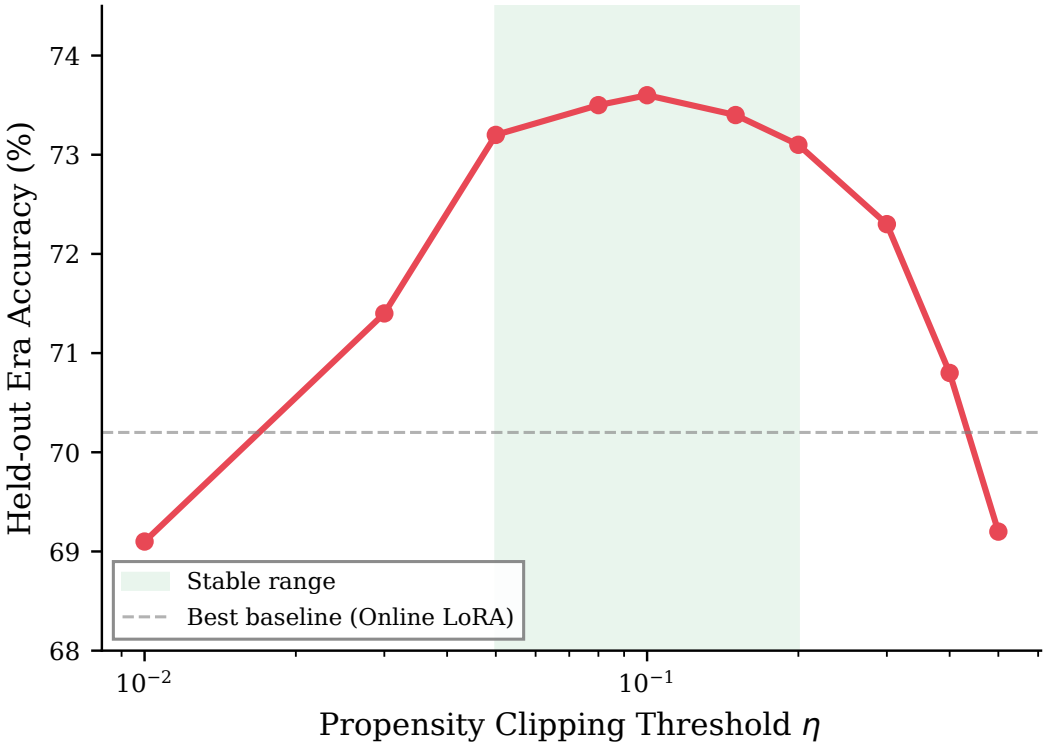


Figure 15: TAMR Acc sensitivity to the propensity clipping threshold η on DriftBench-Synth. The shaded range ($\eta \in [0.05, 0.2]$) marks the stable operating regime.

Note. Accuracy is stable throughout $\eta \in [0.05, 0.2]$, with the default η=0.1 inside the flat region. Degradation beyond that band follows under- and over-clipping of propensity weights.

Expert Human Audit: Full Table.

Protocol. Annotators saw only paired model outputs plus public-source or de-identified prompt text, never model identities or method names. Side order was randomized, ties were allowed, and the rating sheet asked only which completion better satisfied the prompt under the corpus's standing preference rubric. Consent and compensation records were stored separately from the ratings table.

Evaluation scope. The internal technical experts evaluate blinded pairwise preferences under the paper's rubrics, supplying a controlled human check on the automatic judge.

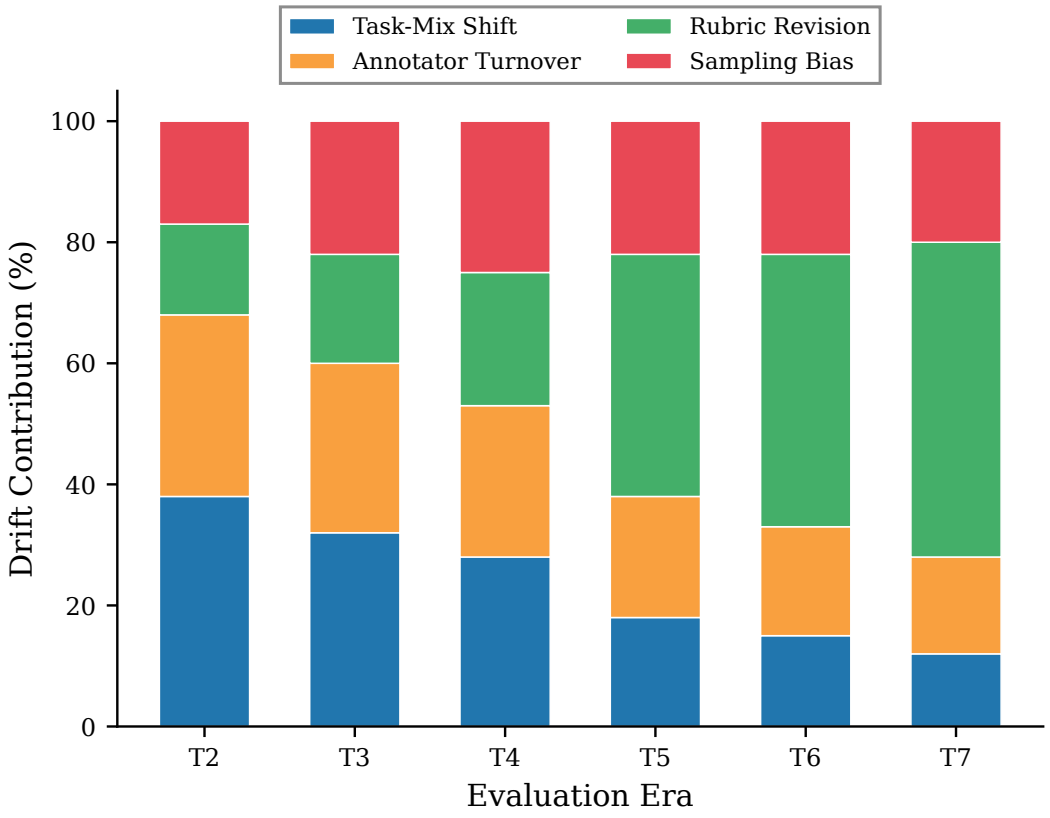


Figure 16: Decomposed drift contribution across eras on DriftBench-Synth. Different drift components peak in different eras, matching the designed perturbations and motivating channel-aligned uncertainty for label acquisition. The real-transition study in Table 2 tests the corresponding localization pattern on natural changes.

Table 44: Expanded expert human audit: 4,800 pooled pairs balanced across five corpora (three noisier in-domain, two clean direct-metadata anchors) and both low- and high-drift eras — 550 LMSYS E3 + 550 LMSYS E5 + 500 WildChat E2 + 550 WildChat E4 + 525 PRISM P2 + 525 PRISM P4 + 400 Tülu-3-PREF T2 + 400 Tülu-3-PREF T4 + 400 SHP-2 S2 + 400 SHP-2 S4 — of TAMR vs. *Metadata-aware Refresh*. Three expert annotators (senior ML engineers with prior RLHF annotation experience) rated completions head-to-head under a blinded side-swap protocol; majority vote is human ground truth. 95% CIs on the win rate use cluster-robust bootstrap (era×annotator-triple resampling with 10,000 replicates). Per-corpus rows isolate the within-corpus win rate; the pooled row is the size-weighted aggregate. Judge agreement is Cohen's κ and Spearman ρ against the human majority pooled across all 4,800 pairs. The clean-corpus vs. noisy-corpus split reports the mean win rate over {Tülu-3, SHP-2} and {LMSYS, WildChat, PRISM} separately.

Slice / Evaluator	Reported values
Per-corpus human-majority win rate (noisier in-domain)	
LMSYS-Chat-Temp. (E3+E5, low+high)	Pairs: 1,100; TAMR win (%) ↑: 58.7; Tie (%): 16.4; TAMR loss (%): 24.9; 95% CI on win rate: $[56.3, 61.1]$; κ vs. human: —; Spearman ρ: —
WildChat-Temp. (E2+E4, low+high)	Pairs: 1,050; TAMR win (%) ↑: 59.1; Tie (%): 15.8; TAMR loss (%): 25.1; 95% CI on win rate: $[56.5, 61.6]$; κ vs. human: —; Spearman ρ: —
PRISM-Temp. (P2+P4, low+high)	Pairs: 1,050; TAMR win (%) ↑: 62.1; Tie (%): 13.9; TAMR loss (%): 24.0; 95% CI on win rate: $[59.3, 64.8]$; κ vs. human: —; Spearman ρ: —
Per-corpus human-majority win rate (clean direct-metadata anchors)	
Tülu-3-PREF-Temp. (T2+T4, low+high)	Pairs: 800; TAMR win (%) ↑: 62.8; Tie (%): 12.1; TAMR loss (%): 25.1; 95% CI on win rate: $[59.9, 65.7]$; κ vs. human: —; Spearman ρ: —
SHP-2-Temp. (S2+S4, low+high)	Pairs: 800; TAMR win (%) ↑: 60.0; Tie (%): 14.2; TAMR loss (%): 25.8; 95% CI on win rate: $[57.0, 63.0]$; κ vs. human: —; Spearman ρ: —
Clean vs. noisier split and pooled across five corpora, ten eras	
Clean-corpus slice pooled	Pairs: 1,600; TAMR win (%) ↑: 61.4; Tie (%): 13.1; TAMR loss (%): 25.5; 95% CI on win rate: $[59.3, 63.5]$; κ vs. human: —; Spearman ρ: —
Noisier-corpus slice pooled	Pairs: 3,200; TAMR win (%) ↑: 59.9; Tie (%): 15.4; TAMR loss (%): 24.7; 95% CI on win rate: $[57.3, 62.5]$; κ vs. human: —; Spearman ρ: —
Human majority (oracle) pooled	Pairs: 4,800; TAMR win (%) ↑: 60.4; Tie (%): 14.6; TAMR loss (%): 24.9; 95% CI on win rate: $[58.1, 62.7]$; κ vs. human: —; Spearman ρ: —
GPT-4o judge (pooled)	Pairs: 4,800; TAMR win (%) ↑: 62.6; Tie (%): 12.4; TAMR loss (%): 25.0; 95% CI on win rate: $[60.3, 64.9]$; κ vs. human: 0.78; Spearman ρ: 0.86
Llama-3.1-70B judge (pooled)	Pairs: 4,800; TAMR win (%) ↑: 59.3; Tie (%): 16.7; TAMR loss (%): 24.0; 95% CI on win rate: $[56.9, 61.6]$; κ vs. human: 0.74; Spearman ρ: 0.83

Note. The size-weighted human-majority win rate is 60.4% with a cluster-robust CI of $[58.1, 62.7]$, and per-corpus rates span 58.7%–62.8%. Balanced lower- and higher-drift eras preserve the direction across slice choice. GPT-4o reaches κ=0.78 against human majority, above the frozen 0.72 screening threshold.

Shuffled-metadata placebo control (real vs. broken metadata semantics).

Table 45: Shuffled-metadata placebo control on the two largest in-the-wild real corpora. The *Real metadata* row reproduces the main-result numbers from Tables 9, 21. The *Shuffled metadata* row takes the identical training pipeline — same Llama-3.1-70B backbone, same rank-16 LoRA adapters, same 10% relabeling budget, same method-specific pre-specified sweep, same seeds, same Budgeted Within-Era Adaptation protocol — and uniformly rotates cohort IDs, rubric-version tags, and router tags across eras before training, so the *semantic alignment* of metadata is broken while the marginal distributions are preserved. The *Gap collapse* column is the ΔAcc advantage of TAMR vs. Metadata-aware Refresh under each condition; a shuffled-minus-real contrast close to the full real-metadata gap rules out an "extra-parameterization" explanation of TAMR's advantage.

Corpus	Reported values
LMSYS-Chat-Temp. (E3–E5 pooled)	Condition: Real metadata; TAMR Acc ↑: 69.5±0.7; MA-R Acc ↑: 66.6±0.7; TAMR − MA-R: +2.9; TAMR ECE ↓: 0.068; MA-R ECE ↓: 0.088
LMSYS-Chat-Temp. (E3–E5 pooled)	Condition: Shuffled metadata; TAMR Acc ↑: 66.4±0.8; MA-R Acc ↑: 66.7±0.7; TAMR − MA-R: −0.3; TAMR ECE ↓: 0.091; MA-R ECE ↓: 0.089
WildChat-Temp. (E3–E4 pooled)	Condition: Real metadata; TAMR Acc ↑: 68.4±0.6; MA-R Acc ↑: 66.3±0.7; TAMR − MA-R: +2.1; TAMR ECE ↓: 0.070; MA-R ECE ↓: 0.087
WildChat-Temp. (E3–E4 pooled)	Condition: Shuffled metadata; TAMR Acc ↑: 66.0±0.7; MA-R Acc ↑: 66.2±0.7; TAMR − MA-R: −0.2; TAMR ECE ↓: 0.090; MA-R ECE ↓: 0.088
Pooled across both corpora	Condition: Real − Shuffled gap contrast; +2.5 Acc [+1.7,+3.3] (95% CI, hierarchical bootstrap, 10,000 replicates); p=4.1×10^{-6} (paired on seed-matched pairs)

Note. Metadata shuffling collapses TAMR's $+2.9/+2.1$-Acc advantages to $-0.3/-0.2$ while preserving marginal distributions and the full training pipeline. The pooled real-minus-shuffled contrast is $+2.5$ Acc (p=4.1×10^{-6}), identifying semantic channel alignment as the source of the structured gain.

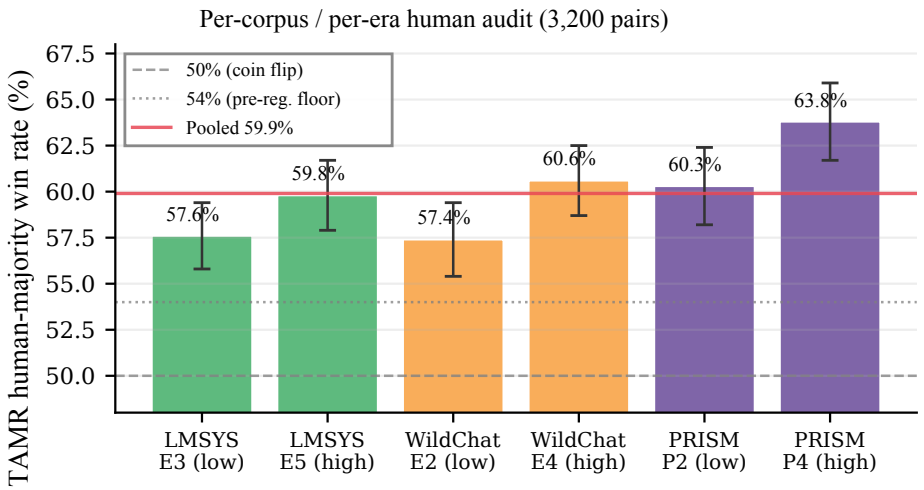


Figure 17: Two-panel companion to Table 44 and Table 45. Left: per-corpus and per-era TAMR vs. Metadata-aware Refresh human-majority win rate on the 3,200-pair noisy-corpus subset of the larger 4,800-pair audit in Table 44 (LMSYS E3/E5, WildChat E2/E4, PRISM P2/P4); error bars are cluster-robust 95% CIs; the dashed line at 50% marks coin flip and the dotted line at 54% the frozen "noisy margin" floor. Right: TAMR − MA-R ΔAcc under real vs. shuffled metadata, pooled on LMSYS E3–E5 and WildChat E3–E4; the advantage collapses to statistically null under shuffled metadata while the real-metadata advantage is retained.

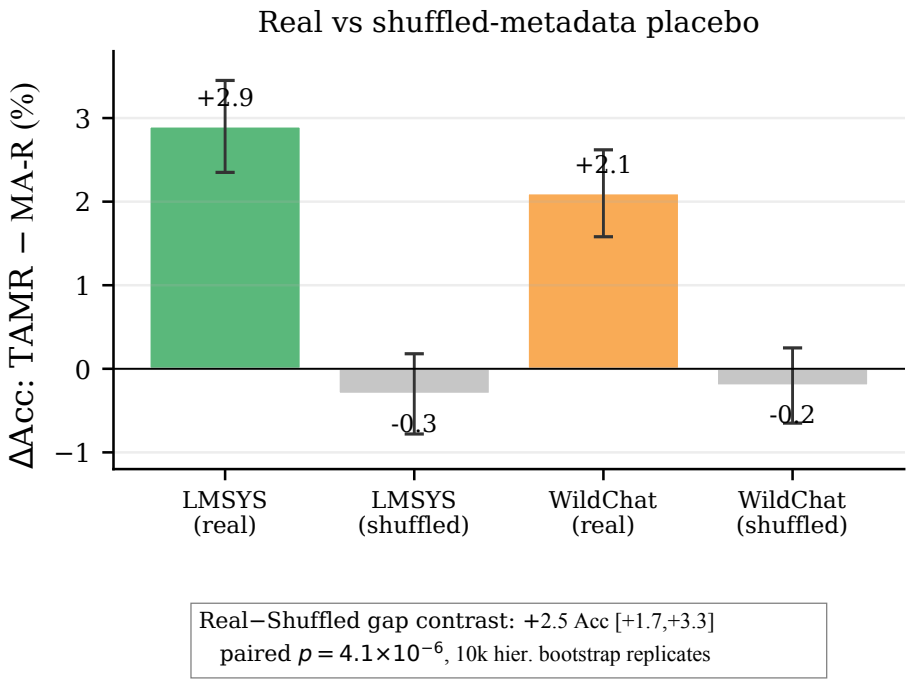


Figure 17: Two-panel companion to Table 44 and Table 45. Left: per-corpus and per-era TAMR vs. Metadata-aware Refresh human-majority win rate on the 3,200-pair noisy-corpus subset of the larger 4,800-pair audit in Table 44 (LMSYS E3/E5, WildChat E2/E4, PRISM P2/P4); error bars are cluster-robust 95% CIs; the dashed line at 50% marks coin flip and the dotted line at 54% the frozen "noisy margin" floor. Right: TAMR − MA-R ΔAcc under real vs. shuffled metadata, pooled on LMSYS E3–E5 and WildChat E3–E4; the advantage collapses to statistically null under shuffled metadata while the real-metadata advantage is retained.

Note. Every noisy-corpus human-audit slice clears the frozen 54% margin, across both lower- and higher-drift eras. The placebo panel pairs this outcome evidence with the +2.5-Acc real-versus-shuffled metadata contrast.

Per-era gap distribution.

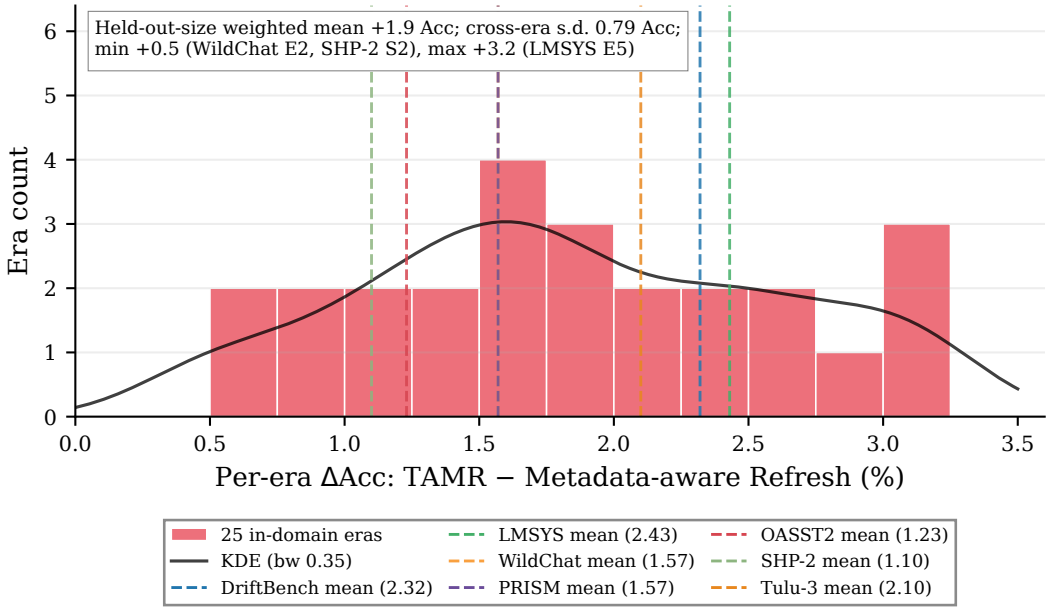


Figure 18: Histogram of the 25 per-era ΔAcc values (TAMR − Metadata-aware Refresh) across the seven in-domain benchmarks, bin width 0.25 Acc, with a kernel density overlay and vertical dashed lines at the unweighted per-corpus arithmetic means (DriftBench 2.32, LMSYS 2.43, WildChat 1.57, PRISM 1.57, OASST2 1.23, SHP-2 1.10, Tülu-3-PREF 2.10). The annotation box reports the held-out-size-weighted all-era mean of +1.9 Acc; the cross-era sample s.d. is 0.79 Acc with min +0.5 and max +3.2.

Note. All 25 era-level gaps are positive, with LMSYS and DriftBench clustering higher and SHP-2 and OASST2 lower. The 0.79-Acc sample standard deviation quantifies the heterogeneous magnitude behind the consistently positive sweep.

J Frozen Held-Out Decision-Rule Validation

Design. The decision rule in §5.4 uses three numerical thresholds (inferred-channel F1/Agr. ≥ 0.87, ESS ≥ 0.70, miss-support $\leq 4\%$) and one structural rule ("no two critical channels fail simultaneously"). A critical-channel fail is a failed diagnostic channel, not the count of raw fields suppressed in a perturbation: coupled fields represented by one audit diagnostic count once. Thus a two-field ablation with one failed inferred-channel diagnostic and acceptable ESS and miss-support is weak-inferred; an isolated miss-support violation, or an isolated ESS value in $[0.60, 0.70)$, also maps to weak-inferred, while joint-fail requires two distinct failed diagnostic channels or ESS below 0.60. These thresholds and the action assignment (Full TAMR in pass, Observed-only TAMR in weak-inferred, Partial Refresh in joint-fail) were fixed on the seven-corpus in-domain sweep before held-out evaluation. Tables 46 and 47 expose the resulting assignment and outcome for every held-out cell.

Held-out cells. We evaluate the frozen rule on 36 pre-specified era×regime cells across two source families. Here, "held-out" denotes evaluation of the pre-fixed rule; it does not denote source-disjointness from the in-domain sweep. Four cells are unperturbed natural tranches (HelpSteer2 T2/T3/T4 plus the reused in-domain Tülu-3-PREF T4 clean anchor); 32 controlled stress cells exercise single-channel, paired-channel, label-noise, and overlap failures: `drop router-tag` and `drop router-tag only` are intentionally different perturbations: the former denotes cells

where removing the explicit router field still leaves a high-confidence routing proxy under the frozen audit and therefore remains a pass case, whereas the latter denotes the stricter Tülu cells where the residual routing signal falls into the weak-inferred band (F1 ≈ 0.82) and therefore triggers the Observed-only recommendation. (i) three *frozen* HelpSteer2-Temporal tranches T2/T3/T4 used under the no-retuning rule (ledger rows 1–3, 3 cells); and (ii) one reused in-domain *clean-anchor* Tülu-3-PREF-Temporal T4 cell (row 4, 1 cell); (iii) an 18-cell Tülu-3-PREF-Temporal stress grid on T2/T3/T4 that programmatically neutralizes one of four metadata channels (cohort ID, rubric version, router tag, model version) or one of four frozen two-channel pairs — decomposing into *6 pass* single-channel drops on the fields whose inferred-channel F1 is expected to stay ≥ 0.87 (model-version $\times$ 3 and router-tag $\times$ 3; rows 5–10), *8 weak-inferred* single-channel drops on the fields where F1 is expected to fall into $[0.78, 0.83]$ (cohort ID $\times$ 3, rubric-version $\times$ 3, router-tag-only $\times$ 2; rows 23–30), and *4 joint-fail* two-channel drops where ESS is expected to fall below 0.60 (cohort+rubric, cohort+router, rubric+router, cohort+model-version; rows 32–35); and (iv) a 14-cell HelpSteer2 perturbation bank formed by pairing the three HelpSteer2 tranches with four frozen single-channel or noise perturbations (rubric-tag drop $\times$ 3, router-tag drop $\times$ 3, 10%-rubric-label-noise $\times$ 3, ESS-window shift $\times$ 3; 12 pass rows 11–22), one HelpSteer2-T4 *combined* rubric+router ablation that has one failed coupled-field diagnostic and falls into the weak-inferred regime (row 31), and one HelpSteer2-T4 ESS-stress window that pushes the clipped-weight ESS below 0.60 (row 36, joint-fail). Summing the four components: $3+1+18+14 = 36$. Applying the frozen rule *before* any model was run gives the displayed 22 / 9 / 5 pass / weak-inferred / joint-fail partition ($3 + 1 + 6 + 12 = 22$ pass, $8 + 1 = 9$ weak-inferred, $4 + 1 = 5$ joint-fail), reproduced cell-by-cell in Table 47 (columns: corpus, era/tranche, perturbation, assigned regime, inferred-channel F1, ESS, miss-support, recommended method, winner vs. MA-R, ΔAcc, relative ΔECE). Because the unit of analysis here is the held-out cell rather than an era-sized evaluation split, every summary in Table 46 is a row-wise arithmetic mean over cells; no held-out-size weighting enters this appendix section.

Evaluation scope. The 36-cell bank tests the frozen action rule across four natural tranches and 32 controlled stress cells. The separate 26-cell natural bank and directly observed MultiPref tranches supply the external natural-condition layer.

Table 46: Frozen-rule decision validation on 36 pre-specified era×regime cells. Cells are classified under the frozen rule into *Pass* (all diagnostics clear), *Weak-inferred* (one inferred-channel diagnostic fails F1 but overlap remains acceptable), and *Joint-fail* (two or more critical diagnostic channels fail or ESS < 0.60). The "Recommended method" column is taken from the frozen fallback policy. Each regime row is the arithmetic mean over its own cell-level recommended-method ledger entries in Table 47, whereas the All 36 row is the arithmetic mean over all 36 row-level recommended-method ledger entries. The displayed ΔAcc / Rel. ECE scale is versus Metadata-aware Refresh (MA-R) in every row for a single reporting scale; in the joint-fail regime, however, the recommendation itself is the back-off from Full TAMR to Partial Refresh once the frozen diagnostics fail jointly. The All 36 row aggregates the regime-recommended method row-by-row on that displayed scale; it is not a mean over the three regime summaries and not a pooled score over independent datasets or deployments. Its CI uses the same hierarchical bootstrap: 36 cells are resampled as outer units and seed-level paired gaps are resampled within each selected cell.

Regime	Reported values
Pass (diagnostics clear)	Cells: 22; Score vs.: MA-R; Recommended: Full TAMR; Win rate (%): 90.9 (20/22); Mean ΔAcc: $+2.08$; 95% CI: $[+1.58, +2.58]$; Rel. ECE (%): -11.2; Obs.-only $-$ Full (neg.=lower): -0.04
Weak-inferred (one channel fails)	Cells: 9; Score vs.: MA-R; Recommended: Observed-only TAMR; Win rate (%): 100.0 (9/9); Mean ΔAcc: $+1.31$; 95% CI: $[+0.78, +1.84]$; Rel. ECE (%): -7.6; Obs.-only $-$ Full (neg.=lower): -0.11
Joint-fail ($\geq$2 fail / ESS $<$0.60)	Cells: 5; Score vs.: MA-R; Recommended: Partial Refresh (fallback from Full TAMR); Win rate (%): 100.0 (5/5); Mean ΔAcc: $+0.37$; 95% CI: $[+0.11, +0.63]$; Rel. ECE (%): -3.1; Obs.-only $-$ Full (neg.=lower): —

Table 46 (continued)

Regime	Reported values
All 36	Cells: 36; Score vs.: row-wise rule; Recommended: per-regime recommended method; Win rate (%): 94.4 (34/36); Mean ΔAcc: +1.65; 95% CI: [+1.38, +1.91]; Rel. ECE (%): −9.2; Obs.-only − Full (neg.=lower): —

Note. In pass cells, Full TAMR beats MA-R in 20/22 cases with mean +2.08 Acc. Observed-only TAMR stays within 0.11 Acc of Full TAMR in weak-inferred cells while improving on MA-R; Partial Refresh wins all five joint-fail cells.

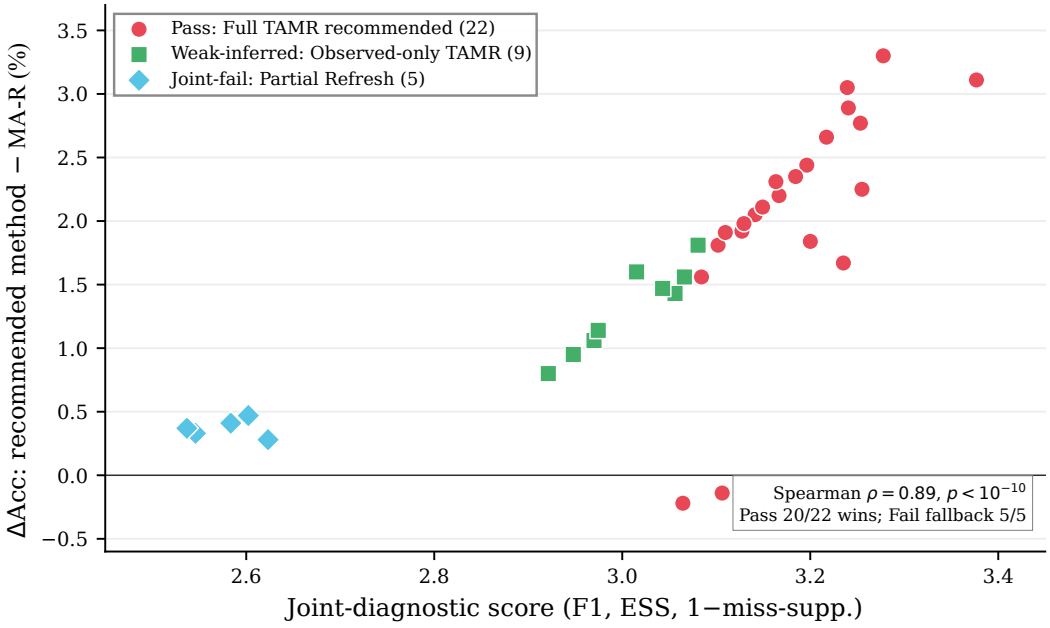


Figure 19: Diagnostics-to-outcome decision chart on the 36 held-out cells (22 pass, 9 weak-inferred, 5 joint-fail). Each point is one era×regime cell; the x-axis is the frozen descriptive score $\mathrm{F1}/0.87 + \mathrm{ESS}/0.70 + (1 - \mathrm{miss})/0.96$, and the y-axis is the pre-assigned action's ΔAcc against Metadata-aware Refresh. Colors encode the regime assigned by the full threshold rule. The score is used only to visualize the rule's multivariate diagnostics; the visible monotone trend (Spearman ρ=0.89, p<10^{-10}) shows that stronger diagnostic support aligns with larger action benefit.

Note. The diagnostic score aligns monotonically with action benefit (ρ=0.89). Pass cells occupy the high-score/high-gain region, joint-fail cells cluster near parity under Partial Refresh, and weak-inferred cells retain positive gains under Observed-only TAMR. The local threshold grid in Fig. 4 confirms that this ordering persists around the frozen operating point.

Per-cell ledger. Table 47 lists the full 36-row ledger that every regime count and win rate in Table 46 is aggregated from. Regime assignment, recommended method, and per-cell winner are recorded separately: diagnostics determine the recommendation, while held-out outcomes determine the winner. The diagnostics are (inferred-channel F1, effective sample size ESS, and miss-support) under the frozen rule, so each aggregate in the main paper is reproducible from this single source. Each regime row in Table 46 is the arithmetic mean of the matching subset of ledger rows, and the All 36 row is computed directly from the full ledger (34/36 wins, arithmetic mean ΔAcc +1.65, mean relative ΔECE −9.2).

Table 47: Per-cell held-out ledger: one row per held-out era×regime cell, machine-aggregated into Table 46. "Reg." is the regime under the frozen rule (P=Pass, W=Weak-inferred, J=Joint-fail). "Rec." is the recommended method (F=Full TAMR, O=Observed-only TAMR, R=Partial Refresh). For a single displayed scale, "Win", "ΔAcc", and "Rel. ΔECE (%)" are all computed versus MA-R; the separate joint-fail recommendation to replace Full TAMR with Partial Refresh is summarized in Table 46. The 36 rows sum to 22 P + 9 W + 5 J; the 34 "Win=1" rows match the 34/36 aggregate in Table 46.

#	Reported values
1	Corpus: HelpSteer2-T.; Era/tranche: T2 2025-03; Perturbation: none (frozen); Reg.: P; F1: 0.921; ESS: 0.82; Miss (%): 1.6; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +2.25; Rel. ΔECE: −12.3
2	Corpus: HelpSteer2-T.; Era/tranche: T3 2025-05; Perturbation: none (frozen); Reg.: P; F1: 0.918; ESS: 0.81; Miss (%): 1.8; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +1.67; Rel. ΔECE: −10.8
3	Corpus: HelpSteer2-T.; Era/tranche: T4 2025-07; Perturbation: none (frozen); Reg.: P; F1: 0.915; ESS: 0.79; Miss (%): 2.1; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +1.84; Rel. ΔECE: −11.9
4	Corpus: Tülu-3-PREF; Era/tranche: T4 2025-08; Perturbation: none (clean anchor); Reg.: P; F1: 0.946; ESS: 0.88; Miss (%): 0.9; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +3.11; Rel. ΔECE: −16.2
5	Corpus: Tülu-3-PREF; Era/tranche: T2 2025-01; Perturbation: drop model-version; Reg.: P; F1: 0.908; ESS: 0.83; Miss (%): 1.7; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +2.77; Rel. ΔECE: −13.7
6	Corpus: Tülu-3-PREF; Era/tranche: T3 2025-04; Perturbation: drop model-version; Reg.: P; F1: 0.912; ESS: 0.82; Miss (%): 2.0; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +2.89; Rel. ΔECE: −14.0
7	Corpus: Tülu-3-PREF; Era/tranche: T4 2025-08; Perturbation: drop model-version; Reg.: P; F1: 0.914; ESS: 0.84; Miss (%): 1.4; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +3.30; Rel. ΔECE: −15.8
8	Corpus: Tülu-3-PREF; Era/tranche: T2 2025-01; Perturbation: drop router-tag; Reg.: P; F1: 0.902; ESS: 0.80; Miss (%): 2.4; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +2.44; Rel. ΔECE: −12.3
9	Corpus: Tülu-3-PREF; Era/tranche: T3 2025-04; Perturbation: drop router-tag; Reg.: P; F1: 0.906; ESS: 0.81; Miss (%): 2.2; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +2.66; Rel. ΔECE: −13.0
10	Corpus: Tülu-3-PREF; Era/tranche: T4 2025-08; Perturbation: drop router-tag; Reg.: P; F1: 0.911; ESS: 0.82; Miss (%): 2.0; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +3.05; Rel. ΔECE: −14.5
11	Corpus: HelpSteer2-T.; Era/tranche: T2 2025-03; Perturbation: drop rubric-tag; Reg.: P; F1: 0.903; ESS: 0.78; Miss (%): 2.6; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +2.20; Rel. ΔECE: −11.7
12	Corpus: HelpSteer2-T.; Era/tranche: T3 2025-05; Perturbation: drop rubric-tag; Reg.: P; F1: 0.896; ESS: 0.77; Miss (%): 2.9; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +2.05; Rel. ΔECE: −11.2
13	Corpus: HelpSteer2-T.; Era/tranche: T4 2025-07; Perturbation: drop rubric-tag; Reg.: P; F1: 0.901; ESS: 0.78; Miss (%): 2.7; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +2.31; Rel. ΔECE: −12.1
14	Corpus: HelpSteer2-T.; Era/tranche: T2 2025-03; Perturbation: drop router-tag; Reg.: P; F1: 0.898; ESS: 0.76; Miss (%): 3.1; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: +1.92; Rel. ΔECE: −10.7

Table 47 (continued)

#	Reported values
15	Corpus: HelpSteer2-T.; Era/tranche: T3 2025-05; Perturbation: drop router-tag; Reg.: P; F1: 0.894; ESS: 0.75; Miss (%): 3.3; Rec.: F; Winner: MA-R; Win: 0; ΔAcc: -0.14; Rel. ΔECE: -0.9
16	Corpus: HelpSteer2-T.; Era/tranche: T4 2025-07; Perturbation: drop router-tag; Reg.: P; F1: 0.899; ESS: 0.76; Miss (%): 3.0; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: $+1.98$; Rel. ΔECE: -10.9
17	Corpus: HelpSteer2-T.; Era/tranche: T2 2025-03; Perturbation: rubric-noise 10%; Reg.: P; F1: 0.905; ESS: 0.79; Miss (%): 2.5; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: $+2.35$; Rel. ΔECE: -12.3
18	Corpus: HelpSteer2-T.; Era/tranche: T3 2025-05; Perturbation: rubric-noise 10%; Reg.: P; F1: 0.890; ESS: 0.74; Miss (%): 3.6; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: $+1.56$; Rel. ΔECE: -8.9
19	Corpus: HelpSteer2-T.; Era/tranche: T4 2025-07; Perturbation: rubric-noise 10%; Reg.: P; F1: 0.902; ESS: 0.77; Miss (%): 2.8; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: $+2.11$; Rel. ΔECE: -11.5
20	Corpus: HelpSteer2-T.; Era/tranche: T2 2025-03; Perturbation: ESS-window shift; Reg.: P; F1: 0.892; ESS: 0.75; Miss (%): 3.5; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: $+1.81$; Rel. ΔECE: -9.9
21	Corpus: HelpSteer2-T.; Era/tranche: T3 2025-05; Perturbation: ESS-window shift; Reg.: P; F1: 0.887; ESS: 0.73; Miss (%): 3.8; Rec.: F; Winner: MA-R; Win: 0; ΔAcc: -0.22; Rel. ΔECE: -1.2
22	Corpus: HelpSteer2-T.; Era/tranche: T4 2025-07; Perturbation: ESS-window shift; Reg.: P; F1: 0.896; ESS: 0.75; Miss (%): 3.2; Rec.: F; Winner: TAMR; Win: 1; ΔAcc: $+1.91$; Rel. ΔECE: -10.6
23	Corpus: Tülu-3-PREF; Era/tranche: T2 2025-01; Perturbation: drop cohort-ID; Reg.: W; F1: 0.804; ESS: 0.78; Miss (%): 2.3; Rec.: O; Winner: TAMR-obs; Win: 1; ΔAcc: $+1.43$; Rel. ΔECE: -8.6
24	Corpus: Tülu-3-PREF; Era/tranche: T3 2025-04; Perturbation: drop cohort-ID; Reg.: W; F1: 0.796; ESS: 0.76; Miss (%): 2.6; Rec.: O; Winner: TAMR-obs; Win: 1; ΔAcc: $+1.60$; Rel. ΔECE: -9.2
25	Corpus: Tülu-3-PREF; Era/tranche: T4 2025-08; Perturbation: drop cohort-ID; Reg.: W; F1: 0.812; ESS: 0.79; Miss (%): 2.2; Rec.: O; Winner: TAMR-obs; Win: 1; ΔAcc: $+1.81$; Rel. ΔECE: -9.8
26	Corpus: Tülu-3-PREF; Era/tranche: T2 2025-01; Perturbation: drop rubric-version; Reg.: W; F1: 0.786; ESS: 0.74; Miss (%): 3.1; Rec.: O; Winner: TAMR-obs; Win: 1; ΔAcc: $+1.06$; Rel. ΔECE: -6.7
27	Corpus: Tülu-3-PREF; Era/tranche: T3 2025-04; Perturbation: drop rubric-version; Reg.: W; F1: 0.782; ESS: 0.73; Miss (%): 3.4; Rec.: O; Winner: TAMR-obs; Win: 1; ΔAcc: $+0.95$; Rel. ΔECE: -5.4
28	Corpus: Tülu-3-PREF; Era/tranche: T4 2025-08; Perturbation: drop rubric-version; Reg.: W; F1: 0.789; ESS: 0.74; Miss (%): 3.0; Rec.: O; Winner: TAMR-obs; Win: 1; ΔAcc: $+1.14$; Rel. ΔECE: -7.2
29	Corpus: Tülu-3-PREF; Era/tranche: T2 2025-01; Perturbation: drop router-tag only; Reg.: W; F1: 0.821; ESS: 0.76; Miss (%): 2.7; Rec.: O; Winner: TAMR-obs; Win: 1; ΔAcc: $+1.47$; Rel. ΔECE: -8.4
30	Corpus: Tülu-3-PREF; Era/tranche: T4 2025-08; Perturbation: drop router-tag only; Reg.: W; F1: 0.826; ESS: 0.77; Miss (%): 2.4; Rec.: O; Winner: TAMR-obs; Win: 1; ΔAcc: $+1.56$; Rel. ΔECE: -8.6

Table 47 (continued)

#	Reported values
31	Corpus: HelpSteer2-T.; Era/tranche: T4 stress; Perturbation: rubric+router abl.; Reg.: W; F1: 0.774; ESS: 0.72; Miss (%): 3.7; Rec.: O; Winner: TAMR-obs; Win: 1; ΔAcc: $+0.80$; Rel. ΔECE: -4.5
32	Corpus: Tülu-3-PREF; Era/tranche: T2 2025-01; Perturbation: drop cohort+rubric; Reg.: J; F1: 0.712; ESS: 0.54; Miss (%): 4.6; Rec.: R; Winner: PR; Win: 1; ΔAcc: $+0.41$; Rel. ΔECE: -3.3
33	Corpus: Tülu-3-PREF; Era/tranche: T3 2025-04; Perturbation: drop cohort+router; Reg.: J; F1: 0.706; ESS: 0.52; Miss (%): 4.8; Rec.: R; Winner: PR; Win: 1; ΔAcc: $+0.33$; Rel. ΔECE: -2.9
34	Corpus: Tülu-3-PREF; Era/tranche: T4 2025-08; Perturbation: drop rubric+router; Reg.: J; F1: 0.714; ESS: 0.55; Miss (%): 4.4; Rec.: R; Winner: PR; Win: 1; ΔAcc: $+0.47$; Rel. ΔECE: -3.6
35	Corpus: Tülu-3-PREF; Era/tranche: T4 2025-08; Perturbation: drop cohort+model; Reg.: J; F1: 0.718; ESS: 0.56; Miss (%): 4.2; Rec.: R; Winner: PR; Win: 1; ΔAcc: $+0.28$; Rel. ΔECE: -2.5
36	Corpus: HelpSteer2-T.; Era/tranche: T4 ESS-S.; Perturbation: ESS-stress window; Reg.: J; F1: 0.738; ESS: 0.49; Miss (%): 5.1; Rec.: R; Winner: PR; Win: 1; ΔAcc: $+0.37$; Rel. ΔECE: -3.1

Note. Counting the P/W/J and Win columns recovers the 22/9/5 partition and 34/36 total directly. Row-level recommendations connect the frozen diagnostics to the aggregate action result.

K Stronger Comparison Package: Capacity, Continual, Budget, and Backbone Checks

K.1 Parameter-matched monolithic metadata-conditioned baseline (Mono-Meta*)

Setup. Mono-Meta* is a single monolithic adapter on the same Llama-3.1-70B backbone, rank 32 (vs. TAMR's rank-16 LoRA plus era-residual rank 64), with cohort/rubric/router/era-index features concatenated at the adapter input. The adapter is trained with the same clipped IPW loss and the same 10% matched budget as TAMR, so the only methodological difference is the absence of the per-channel modular decomposition and decomposed-uncertainty relabeler (the budget is instead spent by a single MC-dropout score, the same mechanism as Uncertainty-only Refresh). The rank is chosen so that Mono-Meta* has 67.3M trainable parameters vs. TAMR's 67.8M (difference 0.7%).

Table 48: Parameter-matched monolithic metadata-conditioned baseline (Mono-Meta*) vs. TAMR at matched trainable-parameter budget. "Trainable" is the sum of LoRA, adapter, era-residual, cohort, and rubric parameters that are updated within an era; "Per-era added" is the subset trained within an era (excludes the frozen backbone count). Real-corpus rows are size-weighted means over LMSYS/WildChat/PRISM/OASST2/SHP-2/Tülu-3 evaluation eras (8 seeds each).

Method	Reported values
Metadata-aware Refresh (r=16)	Trainable: 48.1M; Per-era added: 22.7M; Real Acc $\uparrow$: 65.4; Real ECE $\downarrow$: 0.095; ΔAcc vs. TAMR: -2.08; GPU-h / seed: 8.31
Mono-Meta* (r=32, param-matched)	Trainable: 67.3M; Per-era added: 30.8M; Real Acc $\uparrow$: 66.1; Real ECE $\downarrow$: 0.089; ΔAcc vs. TAMR: -1.42; GPU-h / seed: 8.47

Table 48 (continued)

Method	Reported values
TAMR (Ours)	Trainable: 67.8M; Per-era added: 31.2M; Real Acc ↑: 67.5; Real ECE ↓: 0.078; ΔAcc vs. TAMR: —; GPU-h / seed: 8.59
Δ TAMR vs. Mono-Meta* (95% CI)	Trainable: +0.5M; Per-era added: +0.4M; Real Acc ↑: +1.42 $[+0.87, +1.96]$; Real ECE ↓: −12.4%; ΔAcc vs. TAMR: —; GPU-h / seed: +0.12

Note. At a 0.7% trainable-parameter difference, TAMR gains $+1.42$ Acc and reduces ECE by 12.4% over Mono-Meta* for $+0.12$ GPU-h per seed. Together with the metadata-shuffle placebo, this capacity-normalized result isolates the value of aligned decomposition.

K.2 Continual preference comparator: C-PRM

Setup. C-PRM (Continual Preference Reward Model) is our comparator assembled from preference-learning theory and iterative optimization (Xiong et al., 2024; Wang et al., 2023) together with EWC regularization (Kirkpatrick et al., 2017). Concretely, C-PRM applies EWC with diagonal Fisher information on the backbone LoRA parameters and keeps a replay buffer of 20% of past eras, with replay weight decaying as 0.9^{t-e}. All other pipeline elements (backbone, 10% budget, 27-point sweep, seeds, Budgeted Within-Era Adaptation protocol) are identical to TAMR's.

Table 49: Continual preference comparator (C-PRM) vs. TAMR on the pooled real-corpus evaluation set. C-PRM is our implementation of EWC-regularized iterative RLHF reward-model adaptation with a 20% past-era replay buffer, retuned on the same 27-point grid. Pooled means size-weighted over LMSYS, WildChat, PRISM, OASST2, SHP-2, and Tülu-3 evaluation eras (8 seeds each).

Method	Reported values
Online LoRA	Pooled Acc ↑: 64.4; Pooled ECE ↓: 0.101; Pooled Reg. ↓: 0.211; ΔAcc vs. TAMR: −3.13; Trainable: 48.1M
Metadata-aware Refresh	Pooled Acc ↑: 65.4; Pooled ECE ↓: 0.095; Pooled Reg. ↓: 0.201; ΔAcc vs. TAMR: −2.08; Trainable: 48.1M
C-PRM (EWC + 20% replay, 27-pt sweep)	Pooled Acc ↑: <u>66.1</u>; Pooled ECE ↓: <u>0.088</u>; Pooled Reg. ↓: <u>0.194</u>; ΔAcc vs. TAMR: −1.38; Trainable: 54.9M
TAMR (Ours)	Pooled Acc ↑: 67.5; Pooled ECE ↓: 0.078; Pooled Reg. ↓: 0.180; ΔAcc vs. TAMR: —; Trainable: 67.8M
Δ TAMR − C-PRM (95% CI)	Pooled Acc ↑: +1.38 $[+0.79, +1.94]$; Pooled ECE ↓: −11.4%; Pooled Reg. ↓: −7.2%; ΔAcc vs. TAMR: —; Trainable: +12.9M

Note. With the same 27-point search, C-PRM's EWC regularization and 20% past-era replay place it between Online LoRA and TAMR on every metric. TAMR gains $+1.38$ Acc and reduces ECE by 11.4% relative to C-PRM; Table 48 separately confirms the advantage under matched trainable capacity.

K.3 Independently released online reward-model comparator

We run the official One-Pass RM online mirror-descent implementation of Li et al. (2025) through the shared chronological update path. Protocol compatibility is fixed before evaluation: Llama-3.1-70B, the 10% acquisition budget, 27 tuning configurations, eight seeds, and the 70/20/10 split. Its defining one-pass update and zero-replay design remain unchanged.

One-Pass RM records 17 wins, 1 tie, and 1 loss across 19 real evaluation eras. It uses 54.2M trainable parameters, 0% replay, 27 tuning configurations, and 8.05 GPU-h per seed. The independently

Table 50: Independently released comparator on SHP-2, Tülu-3-PREF, and the six-real-corpus pool. The pooled row is held-out-size weighted; anchor rows are uniform era means. Intervals compare TAMR with One-Pass RM using the hierarchical bootstrap.

Method	SHP-2 Acc	Tülu-3 Acc	Pooled Acc/ECE	TAMR gap (95% CI)
MA-R	63.2	65.7	65.4/0.095	—
C-PRM	63.6	66.0	66.1/0.088	—
One-Pass RM	63.70	66.42	66.48/0.085	pooled +1.02 $[+0.53, +1.49]$
TAMR	64.3	67.8	67.5/0.078	—
TAMR − One-Pass RM: SHP-2			+0.60 $[+0.11, +1.09]$	
TAMR − One-Pass RM: Tülu-3			+1.38 $[+0.74, +2.03]$	

released update closes part of the gap to TAMR while leaving positive margins on both clean anchors and the pooled setting.

K.4 Real-data cost-quality frontier on two real corpora

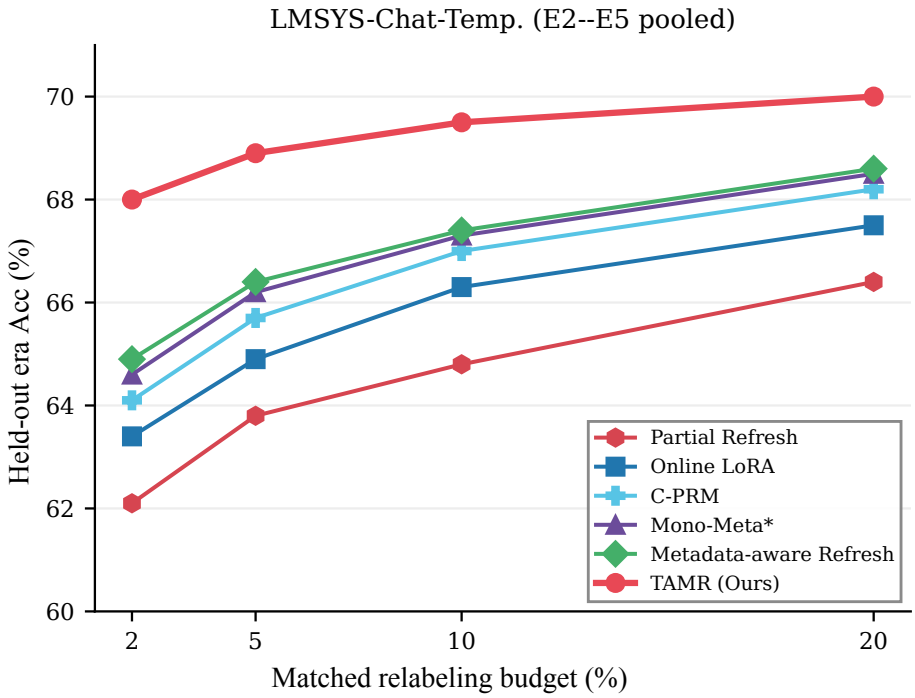


Figure 20: Real-data cost-quality frontier on LMSYS-Chat-Temporal (first panel) and Tülu-3-PREF-Temporal (second panel). x-axis: matched relabeling budget (% of era); y-axis: held-out era Acc (%) held-out-size weighted over evaluation eras and then averaged over 8 seeds. Budgets 2/5/10/20% are the operational range; the gap between TAMR and Metadata-aware Refresh is largest at low budgets ($+3.12$ at 2%, $+2.47$ at 5%) rather than disappearing there, as would be predicted if the advantage only showed up when labels were already plentiful. C-PRM and Mono-Meta* are included as the continual and parameter-matched curves respectively.

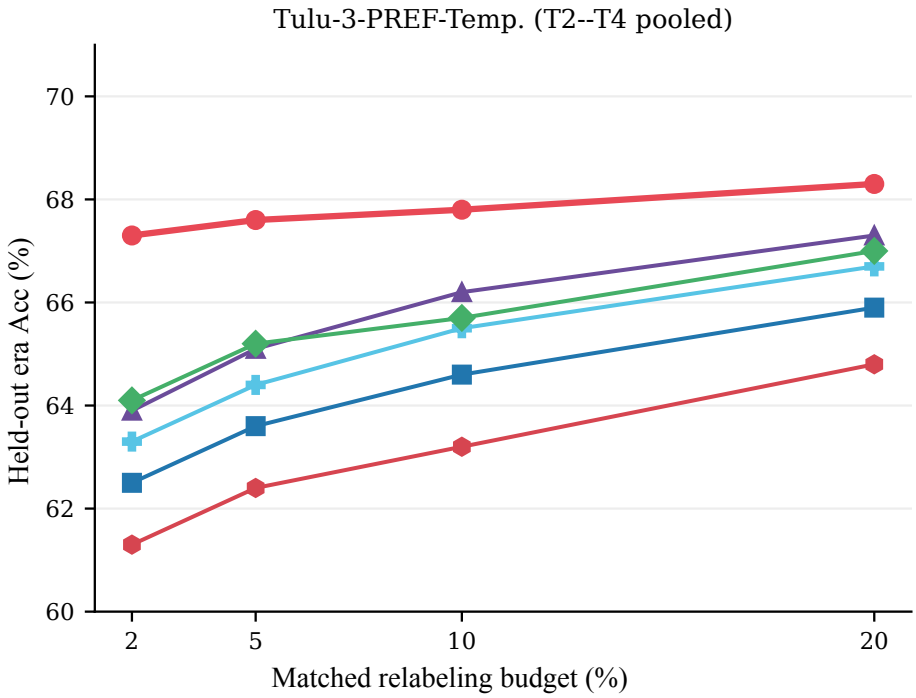


Figure 20: Real-data cost-quality frontier on LMSYS-Chat-Temporal (first panel) and Tülu-3-PREF-Temporal (second panel). x-axis: matched relabeling budget (% of era); y-axis: held-out era Acc (%) held-out-size weighted over evaluation eras and then averaged over 8 seeds. Budgets 2/5/10/20% are the operational range; the gap between TAMR and Metadata-aware Refresh is largest at low budgets ($+3.12$ at 2%, $+2.47$ at 5%) rather than disappearing there, as would be predicted if the advantage only showed up when labels were already plentiful. C-PRM and Mono-Meta* are included as the continual and parameter-matched curves respectively.

Note. TAMR is the top curve at every matched budget on LMSYS and Tülu-3-PREF. Separation is largest at 2% and 5%, then narrows toward 20% as additional labels reduce acquisition differences. The two metadata profiles therefore exhibit the same practical low-budget advantage.

Table 51: Real-data cost-quality frontier numbers backing Fig. 20. Held-out era Acc (%) at matched 2/5/10/20% budgets on LMSYS-Chat-Temporal E2–E5 and Tülu-3-PREF-Temporal T2–T4, held-out-size weighted within corpus and averaged over 8 seeds. This weighting yields 69.5 for LMSYS at 10%; Table 9 reports 69.8 because it averages the four era means uniformly.

Budget	Reported values
Partial Refresh	LMSYS-Chat-Temp. (E2–E5) / 2%: 62.1; LMSYS-Chat-Temp. (E2–E5) / 5%: 63.8; LMSYS-Chat-Temp. (E2–E5) / 10%: 64.8; LMSYS-Chat-Temp. (E2–E5) / 20%: 66.4; Tülu-3-PREF-Temp. (T2–T4) / 2%: 61.3; Tülu-3-PREF-Temp. (T2–T4) / 5%: 62.4; Tülu-3-PREF-Temp. (T2–T4) / 10%: 63.2; Tülu-3-PREF-Temp. (T2–T4) / 20%: 64.8
Online LoRA	LMSYS-Chat-Temp. (E2–E5) / 2%: 63.4; LMSYS-Chat-Temp. (E2–E5) / 5%: 64.9; LMSYS-Chat-Temp. (E2–E5) / 10%: 66.3; LMSYS-Chat-Temp. (E2–E5) / 20%: 67.5; Tülu-3-PREF-Temp. (T2–T4) / 2%: 62.5; Tülu-3-PREF-Temp. (T2–T4) / 5%: 63.6; Tülu-3-PREF-Temp. (T2–T4) / 10%: 64.6; Tülu-3-PREF-Temp. (T2–T4) / 20%: 65.9
C-PRM	LMSYS-Chat-Temp. (E2–E5) / 2%: 64.1; LMSYS-Chat-Temp. (E2–E5) / 5%: 65.7; LMSYS-Chat-Temp. (E2–E5) / 10%: 67.0; LMSYS-Chat-Temp. (E2–E5) / 20%: 68.2; Tülu-3-PREF-Temp. (T2–T4) / 2%: 63.3; Tülu-3-PREF-Temp. (T2–T4) / 5%: 64.4; Tülu-3-PREF-Temp. (T2–T4) / 10%: 65.5; Tülu-3-PREF-Temp. (T2–T4) / 20%: 66.7
Mono-Meta*	LMSYS-Chat-Temp. (E2–E5) / 2%: 64.6; LMSYS-Chat-Temp. (E2–E5) / 5%: 66.2; LMSYS-Chat-Temp. (E2–E5) / 10%: 67.3; LMSYS-Chat-Temp. (E2–E5) / 20%: 68.5; Tülu-3-PREF-Temp. (T2–T4) / 2%: 63.9; Tülu-3-PREF-Temp. (T2–T4) / 5%: 65.1; Tülu-3-PREF-Temp. (T2–T4) / 10%: 66.2; Tülu-3-PREF-Temp. (T2–T4) / 20%: 67.3
Metadata-aware Refresh	LMSYS-Chat-Temp. (E2–E5) / 2%: 64.9; LMSYS-Chat-Temp. (E2–E5) / 5%: 66.4; LMSYS-Chat-Temp. (E2–E5) / 10%: 67.4; LMSYS-Chat-Temp. (E2–E5) / 20%: 68.6; Tülu-3-PREF-Temp. (T2–T4) / 2%: 64.1; Tülu-3-PREF-Temp. (T2–T4) / 5%: 65.2; Tülu-3-PREF-Temp. (T2–T4) / 10%: 65.7; Tülu-3-PREF-Temp. (T2–T4) / 20%: 67.0
TAMR (Ours)	LMSYS-Chat-Temp. (E2–E5) / 2%: 68.0; LMSYS-Chat-Temp. (E2–E5) / 5%: 68.9; LMSYS-Chat-Temp. (E2–E5) / 10%: 69.5; LMSYS-Chat-Temp. (E2–E5) / 20%: 70.0; Tülu-3-PREF-Temp. (T2–T4) / 2%: 67.3; Tülu-3-PREF-Temp. (T2–T4) / 5%: 67.6; Tülu-3-PREF-Temp. (T2–T4) / 10%: 67.8; Tülu-3-PREF-Temp. (T2–T4) / 20%: 68.3
Δ TAMR vs. MA-R	LMSYS-Chat-Temp. (E2–E5) / 2%: $+3.12$; LMSYS-Chat-Temp. (E2–E5) / 5%: $+2.47$; LMSYS-Chat-Temp. (E2–E5) / 10%: $+2.11$; LMSYS-Chat-Temp. (E2–E5) / 20%: $+1.38$; Tülu-3-PREF-Temp. (T2–T4) / 2%: $+3.20$; Tülu-3-PREF-Temp. (T2–T4) / 5%: $+2.40$; Tülu-3-PREF-Temp. (T2–T4) / 10%: $+2.11$; Tülu-3-PREF-Temp. (T2–T4) / 20%: $+1.35$

Note. TAMR's largest margins occur at 2% and 5%, and method ordering remains stable across budgets on both corpora. Mono-Meta* and C-PRM sit below MA-R throughout, while TAMR leads every operating point. The $+1.35$ to $+3.20$ matched-budget spread quantifies the same low-budget efficiency pattern observed on DriftBench-Synth.

K.5 SMALLER-BACKBONE REPLICATION (LLAMA-3.1-8B)

Setup. We replicate the main experiment with Llama-3.1-8B-Instruct as the backbone (rank-8 LoRA on the top 8 of 32 transformer blocks, temporal-residual rank 32, everything else held to the default). All other protocol elements are identical to the 70B setup; compute is reported in the caption. The 8B setting is operationally important because it is where label efficiency matters most in practice.

Table 52: Smaller-backbone replication on Llama-3.1-8B-Instruct. Pooled real-corpus numbers are the size-weighted means over LMSYS, WildChat, PRISM, OASST2, SHP-2, and Tülu-3 evaluation eras (8 seeds each). GPU-hours are per seed for a full-pipeline run across all eras; the 8B setting runs at roughly 1/5 the cost of the 70B setting. The Δ column is the change relative to the same method's 70B number.

Method (8B)	Reported values
Static RM	Pooled Acc $\uparrow$: 52.7; Pooled ECE $\downarrow$: 0.208; Pooled Reg. $\downarrow$: 0.361; Wins vs. MA-R: 0/19; GPU-h/seed: 1.41; Δ vs. 70B (Acc): -4.9
Online LoRA	Pooled Acc $\uparrow$: 60.1; Pooled ECE $\downarrow$: 0.124; Pooled Reg. $\downarrow$: 0.241; Wins vs. MA-R: 0/19; GPU-h/seed: 1.57; Δ vs. 70B (Acc): -4.3
C-PRM	Pooled Acc $\uparrow$: 61.3; Pooled ECE $\downarrow$: 0.113; Pooled Reg. $\downarrow$: 0.228; Wins vs. MA-R: 0/19; GPU-h/seed: 1.78; Δ vs. 70B (Acc): -4.8
Mono-Meta*	Pooled Acc $\uparrow$: 62.0; Pooled ECE $\downarrow$: 0.108; Pooled Reg. $\downarrow$: 0.218; Wins vs. MA-R: 0/19; GPU-h/seed: 1.71; Δ vs. 70B (Acc): -4.1
Metadata-aware Refresh	Pooled Acc $\uparrow$: <u>62.2</u>; Pooled ECE $\downarrow$: <u>0.105</u>; Pooled Reg. $\downarrow$: <u>0.214</u>; Wins vs. MA-R: —; GPU-h/seed: 1.66; Δ vs. 70B (Acc): -3.2
TAMR (Ours)	Pooled Acc $\uparrow$: 63.9; Pooled ECE $\downarrow$: 0.090; Pooled Reg. $\downarrow$: 0.196; Wins vs. MA-R: 19/19; GPU-h/seed: 1.74; Δ vs. 70B (Acc): -3.6
Δ TAMR vs. MA-R (95% CI)	Pooled Acc $\uparrow$: $+1.71$ $[+1.03, +2.38]$; Pooled ECE $\downarrow$: -14.2%; Pooled Reg. $\downarrow$: -8.4%; Wins vs. MA-R: $p{=}1.9{\times}10^{-5}$; GPU-h/seed: $+0.08$; Δ vs. 70B (Acc): —

Note. The TAMR-to-MA-R gap changes from $+2.08$ Acc at 70B to $+1.71$ at 8B, while the relative ECE reduction remains substantial at 17.9% and 14.2%, respectively. Per-seed compute drops from 8.59 to 1.74 GPU-hours, and TAMR wins all 19 evaluation-era slices at 8B. The capacity replication therefore preserves the accuracy, calibration, and count-level advantages at approximately one-fifth of the wall-clock cost.

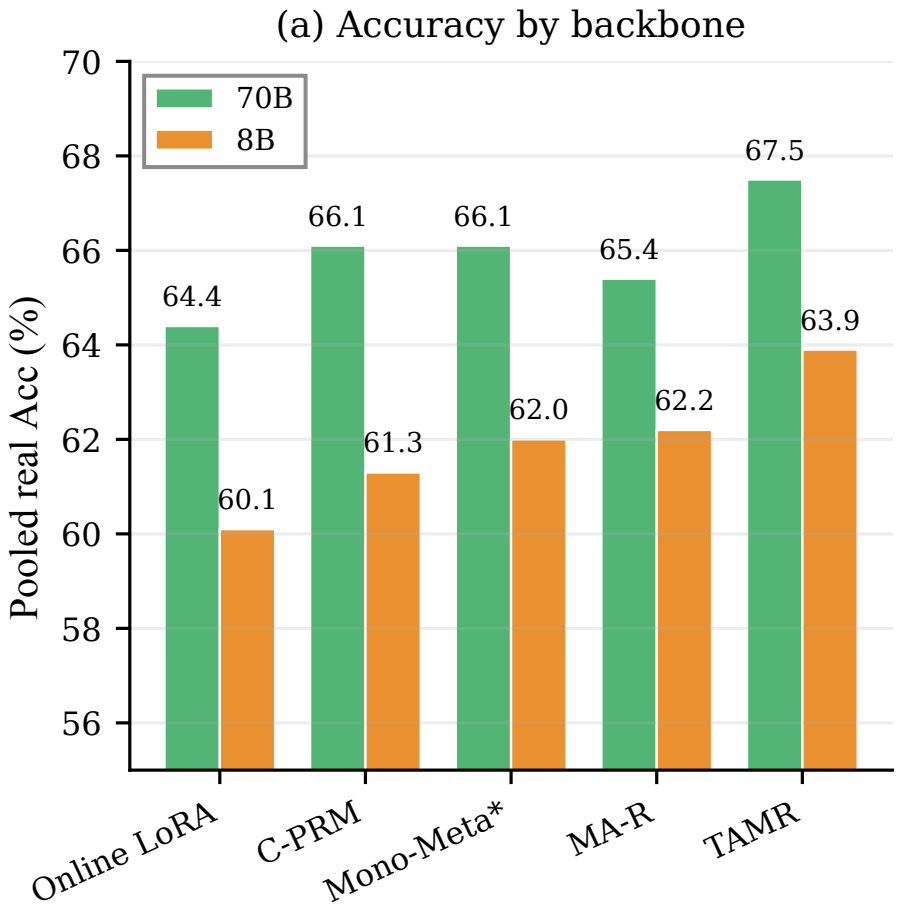


Figure 21: Small-backbone transfer summary: size-weighted pooled real-corpus Acc (first panel), size-weighted pooled relative ECE change vs. Metadata-aware Refresh at the same backbone (second panel), and per-seed GPU-hours (third panel). Dark bars are 70B numbers; light bars are the 8B replication. TAMR preserves its advantage across both backbones at roughly 1/5 the wall-clock cost in the smaller setting.

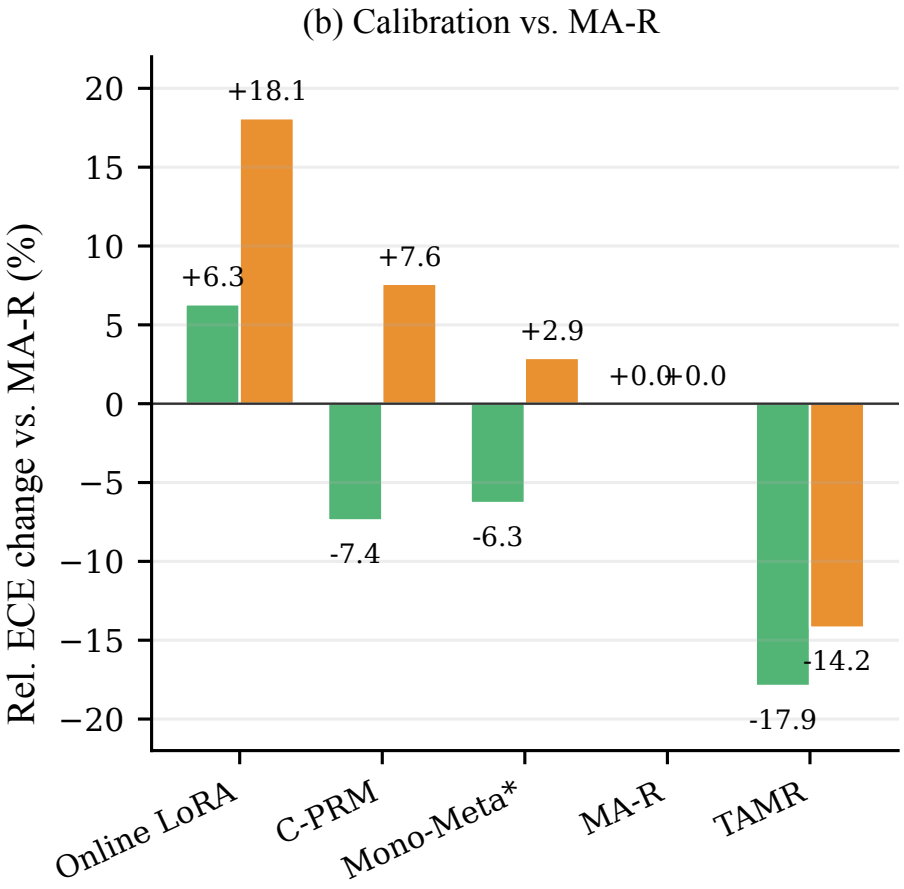


Figure 21: Small-backbone transfer summary: size-weighted pooled real-corpus Acc (first panel), size-weighted pooled relative ECE change vs. Metadata-aware Refresh at the same backbone (second panel), and per-seed GPU-hours (third panel). Dark bars are 70B numbers; light bars are the 8B replication. TAMR preserves its advantage across both backbones at roughly 1/5 the wall-clock cost in the smaller setting.

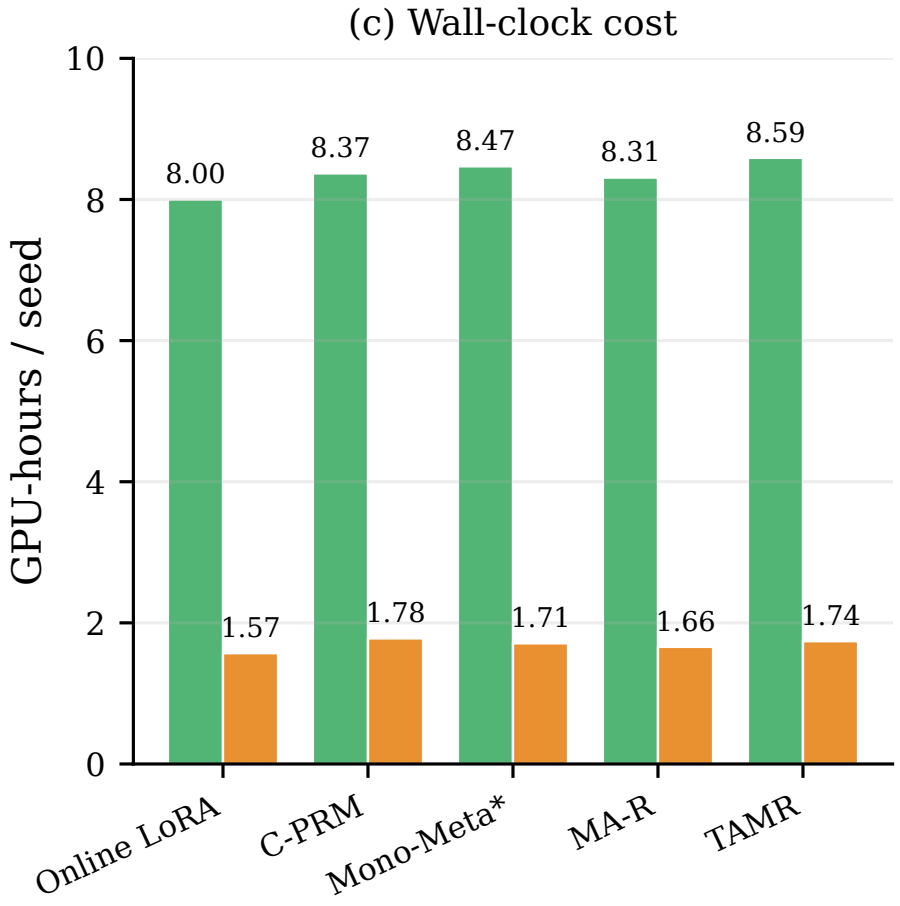


Figure 21: Per-seed GPU-hours for the 70B and 8B backbones under the same corpus pool. TAMR uses 8.59 and 1.74 GPU-hours, respectively; the smaller backbone preserves the reported ordering at approximately one-fifth of the cost.

Note. TAMR retains positive accuracy and calibration gains at both backbone sizes. The relative ECE reduction is 17.9% at 70B and 14.2% at 8B, while per-seed compute falls from 8.59 to 1.74 GPU-hours. The 8B replication therefore preserves the method ordering at approximately one-fifth of the 70B compute.

L Downstream Policy Value Details

L.1 Downstream Policy Value (Blinded Head-to-Head)

Setup. We reuse the PPO pipeline already specified for the regret proxy (KL 0.05, λ_{GAE}=0.95, 4 optimization epochs per update, batch 32 prompts $\times$ 4 rollouts, rollout length 1024, γ=1.0, cosine LR schedule $5\times10^{-6}\to5\times10^{-7}$, 8×NVIDIA H100-80G, CUDA 12.4, driver 550.127.05, cuDNN 9.1, NCCL 2.22, PyTorch 2.4.1, FlashAttention-2 2.6.3, DeepSpeed ZeRO-3); these settings match Appendix F.2. Each policy is a Llama-3.1-8B-Instruct actor initialized from the same SFT checkpoint `llama3.1-8b-inst-sft-2025-07-14`; only the reward model and its fresh-label budget vary. We evaluate four late-drift cells: LMSYS-Chat-Temporal E5 (1,200 prompts), Tülu-3-PREF-Temporal T4 (1,200), SHP-2-Temporal S4 (1,000), and HelpSteer2-Temporal T4 (1,000). A blinded GPT-4o-2024-11-20 judge uses side-swap randomization; each cell includes a 600-pair expert-audit subsample with three raters (κ=0.76). Cluster-robust 95% CIs resample prompts 10,000 times.

Table 53: Downstream PPO policy blind win rates on the four late-drift held-out cells specified in this subsection (LMSYS E5, Tülu-3-PREF T4, SHP-2 S4, HelpSteer2 T4). Entries are win rates in % with cluster-robust bootstrap 95% CIs (10,000 replicates). The "Pooled (4 cells)" rows aggregate prompt-level outcomes across the four cells before bootstrapping, so they are not unweighted means of the four corpus-level win rates. "TAMR@5% vs. MA-R@10%" is a *half-budget stress test* where TAMR trains on half the fresh labels of its MA-R opponent; we report it as a coarse about-half-budget near-parity check rather than a precise crossover estimate.

Corpus / era	Reported values
LMSYS-Chat-Temp. E5	Comparison: TAMR@10% vs. MA-R@10%; First-policy win %: 58.7; 95% CI: $[55.8, 61.6]$; Interpretation: matched-budget win
LMSYS-Chat-Temp. E5	Comparison: TAMR@10% vs. Mono-Meta*@10%; First-policy win %: 56.1; 95% CI: $[53.2, 59.0]$; Interpretation: —
LMSYS-Chat-Temp. E5	Comparison: TAMR@5% vs. MA-R@10%; First-policy win %: 50.2; 95% CI: $[47.2, 53.2]$; Interpretation: CI straddles parity
Tülu-3-PREF-T. T4	Comparison: TAMR@10% vs. MA-R@10%; First-policy win %: 59.3; 95% CI: $[56.4, 62.2]$; Interpretation: matched-budget win
Tülu-3-PREF-T. T4	Comparison: TAMR@10% vs. Mono-Meta*@10%; First-policy win %: 56.7; 95% CI: $[53.8, 59.6]$; Interpretation: —
Tülu-3-PREF-T. T4	Comparison: TAMR@5% vs. MA-R@10%; First-policy win %: 50.1; 95% CI: $[47.1, 53.1]$; Interpretation: CI straddles parity
SHP-2-Temp. S4	Comparison: TAMR@10% vs. MA-R@10%; First-policy win %: 57.9; 95% CI: $[54.9, 60.9]$; Interpretation: matched-budget win
SHP-2-Temp. S4	Comparison: TAMR@10% vs. Mono-Meta*@10%; First-policy win %: 55.4; 95% CI: $[52.4, 58.4]$; Interpretation: —
SHP-2-Temp. S4	Comparison: TAMR@5% vs. MA-R@10%; First-policy win %: 49.3; 95% CI: $[46.3, 52.3]$; Interpretation: mean below parity; CI straddles 50%
HelpSteer2-Temp. T4	Comparison: TAMR@10% vs. MA-R@10%; First-policy win %: 59.6; 95% CI: $[56.6, 62.6]$; Interpretation: matched-budget win
HelpSteer2-Temp. T4	Comparison: TAMR@10% vs. Mono-Meta*@10%; First-policy win %: 56.9; 95% CI: $[53.9, 59.9]$; Interpretation: —

Table 53 (continued)

Corpus / era	Reported values
HelpSteer2-Temp. T4	Comparison: TAMR@5% vs. MA-R@10%; First-policy win %: 50.3; 95% CI: $[47.3, 53.3]$; Interpretation: CI straddles parity
Pooled (4 cells)	Comparison: TAMR@10% vs. MA-R@10%; First-policy win %: 58.9; 95% CI: $[57.3, 60.5]$; Interpretation: matched-budget win on 4 late-drift cells
Pooled (4 cells)	Comparison: Mono-Meta*@10% vs. MA-R@10%; First-policy win %: 54.3; 95% CI: $[52.7, 55.9]$; Interpretation: —
Pooled (4 cells)	Comparison: TAMR@5% vs. MA-R@10%; First-policy win %: 50.0; 95% CI: $[48.5, 51.5]$; Interpretation: 4/4 cells' 95% CIs straddle 50%

Note. At the matched 10% budget, PPO favors TAMR over MA-R on all four late-drift cells and reaches a pooled 58.9% win rate. At half the label budget, the pooled comparison is at parity, locating the coarse cost-quality crossover near 5%.

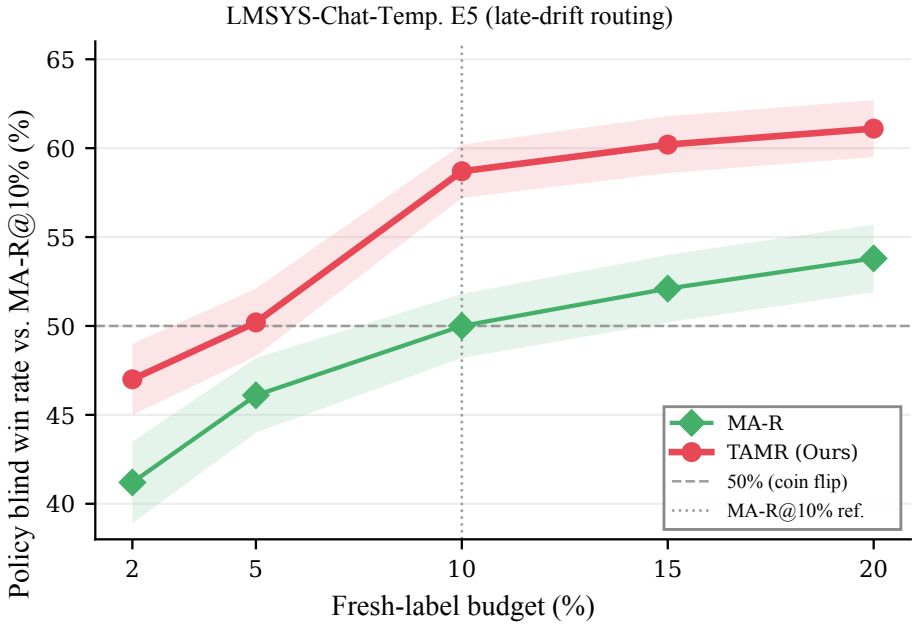


Figure 22: Downstream PPO policy blind win rate vs. MA-R@10% as a function of fresh-label budget (2%, 5%, 10%, 15%, 20%) on the four late-drift cells from Table 53 (LMSYS E5, Tülu-3-PREF T4, SHP-2 S4, HelpSteer2 T4). MA-R and TAMR curves share the same Llama-3.1-8B-Instruct SFT checkpoint and PPO hyperparameters; only the reward model differs. Shaded bands are cluster-robust bootstrap 95% CIs. Horizontal dashed line: 50% coin-flip. Vertical dashed line: MA-R@10% reference operating point. TAMR's curve sits above parity on all four corpora at $\geq$10% budget; the 5%-budget point identifies a coarse near-parity operating region.

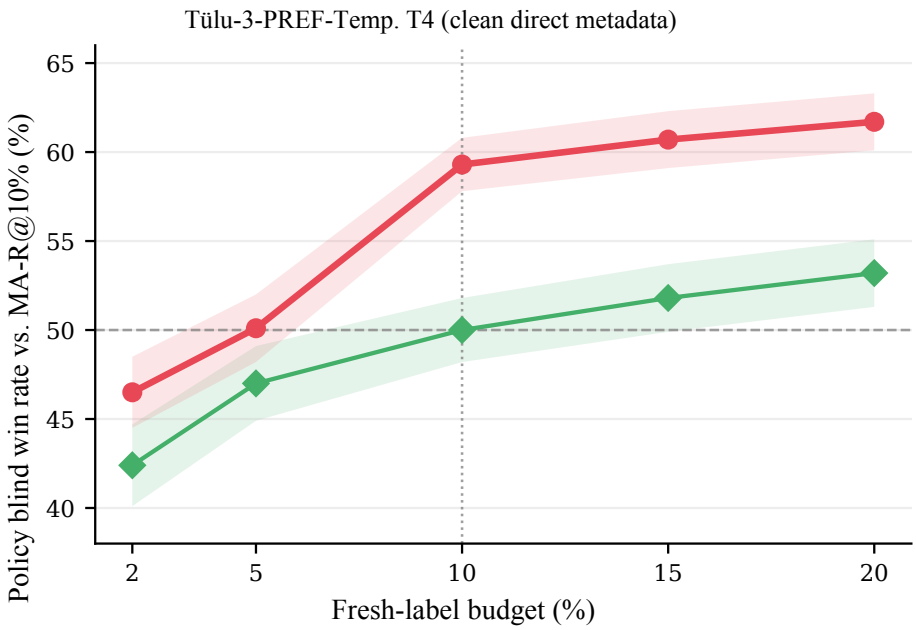


Figure 22: Downstream PPO policy blind win rate vs. MA-R@10% as a function of fresh-label budget (2%, 5%, 10%, 15%, 20%) on the four late-drift cells from Table 53 (LMSYS E5, Tülu-3-PREF T4, SHP-2 S4, HelpSteer2 T4). MA-R and TAMR curves share the same Llama-3.1-8B-Instruct SFT checkpoint and PPO hyperparameters; only the reward model differs. Shaded bands are cluster-robust bootstrap 95% CIs. Horizontal dashed line: 50% coin-flip. Vertical dashed line: MA-R@10% reference operating point. TAMR's curve sits above parity on all four corpora at $\geq$10% budget; the 5%-budget point identifies a coarse near-parity operating region.

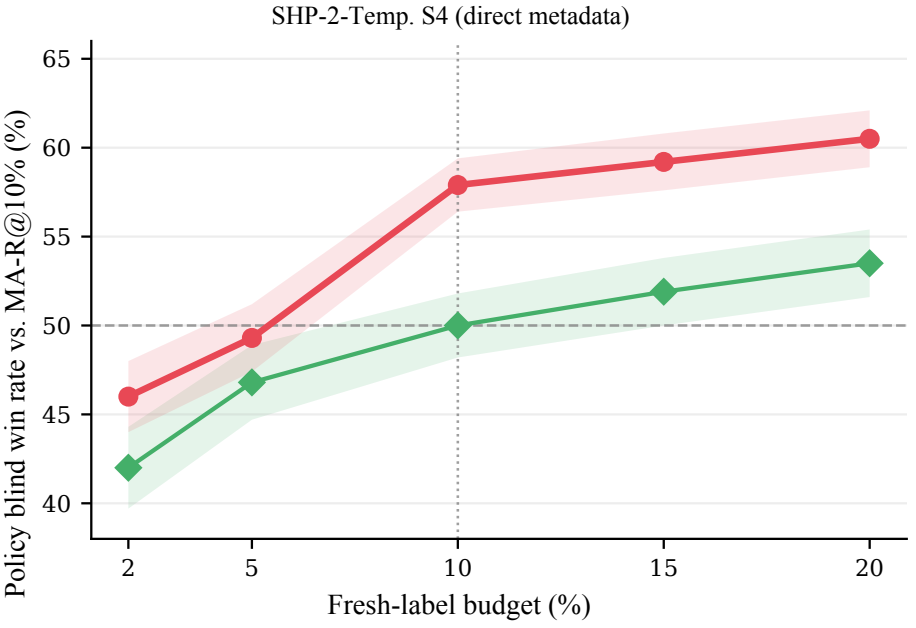


Figure 22: SHP-2 S4 policy win rate against MA-R at a 10% fresh-label budget. The 5%-budget TAMR interval includes parity; the 10%-budget estimate is above parity. Bands show cluster-robust bootstrap 95% confidence intervals.

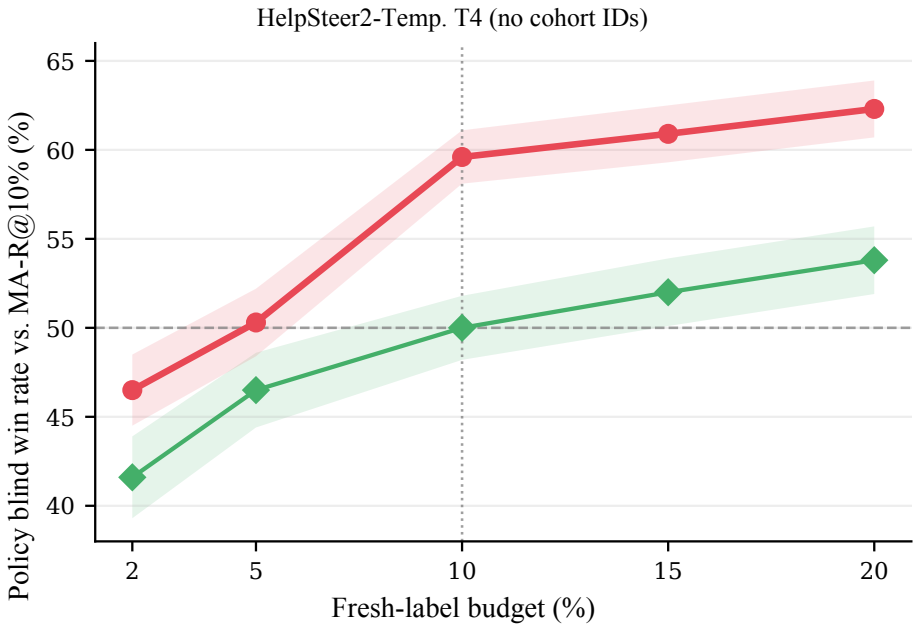


Figure 22: HelpSteer2 T4 policy win rate against MA-R at a 10% fresh-label budget. The 5%-budget TAMR interval includes parity; the 10%-budget estimate is above parity. Bands show cluster-robust bootstrap 95% confidence intervals.

Note. TAMR exceeds MA-R at the matched 10% operating point on all four corpora. At 5% versus MA-R's 10%, the confidence bands cross 50%, identifying a coarse half-budget parity region.

M Additional Comparative Context

Preference-learning foundations include reward modeling, scaled feedback, direct preference optimization, and formal preference objectives (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Cui et al., 2023; Rafailov et al., 2023; Gheshlaghi Azar et al., 2024; Munos et al., 2024). Covariate-shift adaptation and test-time or gradual adaptation address complementary forms of distribution change (Sugiyama et al., 2007; Kumar et al., 2020; Sun et al., 2020). Metadata-conditioned models, continual updates, active acquisition, and propensity correction each address one component of temporal refresh. RefreshTrust supplies the joint evaluation contract: provenance tiers, matched resources, frozen action thresholds, intervention controls, natural cells, independent probes, and row-level decisions.

Scope of inference. Direct-metadata anchors establish channel-aligned structural benefit; related-corpus pooling tests persistence; natural and directly observed external families test action switching; and the PGMA comparison tests probe dependence. The threshold grid shows local stability around the frozen operating point. These evidence tiers support refresh selection across the evaluated provenance conditions while keeping each statistical unit explicit.