The Partial Testimony of Logs: Evaluation of Language Model Generation under Confounded Model Choice

Anonymous Author(s)
Affiliation
Address
email

Abstract

Offline evaluation of language models from usage logs is biased when model choice is confounded: the same user-side factors that drive which model is used also shape how its output is judged. Randomization removes this bias, yet production randomization remains scarce. We study a three-source design that combines a large confounded observational log (OBS) for scale, a small randomized experiment (EXP) for unconfounded scoring, and an offline simulator (SIM) that replays candidate models on cached contexts. An identification theorem shows that EXP together with SIM recovers causal model values, with OBS contributing afterward to reduce estimation variance. Six estimator families are evaluated in a controlled semi-synthetic validation and in two real-task cached benchmarks for summarization and coding. The winning family shifts with the EXP budget and with how closely the OBS signal predicts the target reward, giving a practical recipe for choosing among hybrid OBS/EXP estimators.

1 Introduction

Offline evaluation of language models increasingly relies on deployment logs, preference data, and judge annotations collected in the wild [Zheng et al., 2024, Zhao et al., 2024, Zheng et al., 2023, Chiang et al., 2024]. Once model choice is user-driven, those logs are confounded: the same unobserved factors that influence which model is used can also influence how its output is judged [Kallus et al., 2018, Rosenman et al., 2023, Cheng and Cai, 2021]. Raw comparisons of logged outcomes therefore mix self-selected subpopulations and cannot recover a common target. In summarization, a reader who prefers concise prose may select a model known for that style and rate it more favorably; in coding assistance, repository difficulty or developer familiarity may affect both which assistant is invoked and whether the resulting patch succeeds.

A randomized experiment breaks this link by overriding model choice. In practice, such experiments are costly and disruptive, so they remain small. Prior work studies how to combine a large confounded observational sample with a small randomized one [Kallus et al., 2018, Rosenman et al., 2023, Cheng and Cai, 2021, Lin et al., 2025, Yang et al., 2025, Colnet et al., 2024].

Causal evaluation under confounded model choice has two parts. *Outcome generation* asks which output a model would have produced on a given context; in many deployments an offline simulator (SIM) handles this by replaying models on cached contexts. *Outcome scoring* asks how the realized output should be evaluated without inheriting the bias in logged model choice. Separating the two motivates a three-source design: a large observational log (OBS) with self-selected model choices, a small randomized experiment (EXP) with unconfounded outcome labels, and SIM outputs under alternative model choices. The goal is to estimate how each generative model would perform when

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

logged model choice is confounded, randomized outcome labels are limited, and offline replay can regenerate candidate outputs.

Contributions. The paper makes three contributions.

- **(C1) Identification (Theorem 1).** Under the structural assumptions stated in Section 2, the simulator and the randomized experiment alone identify causal model values, even when the observational log is fully confounded.
- **(C2) Estimator role of OBS (post-identification only).** Identification and estimation are separated, and six post-identification estimator families are compared based on how they exploit OBS. A theory-to-evidence map links each family to the data regime in which it helps.
- **(C3) Empirical evaluation.** A controlled semi-synthetic validation and two real-task cached benchmarks (summarization, coding) compare the families on held-out recommendation regret. Two practical drivers explain the winning family: the size of the EXP budget, and how well OBS-derived structure matches the target reward.

Section 2 defines the causal graph, data sources, target estimands, and assumptions. Section 3 states and proves the SIM-plus-EXP identification result. Section 4 introduces the post-identification estimator families and the theory-to-evidence map. Section 5 describes the controlled validation and cached benchmarks; Sections 5.4 and 5.5 report the cached summarization and cached coding probes. Sections 6, 7, and 8 cover related work, scope and limitations, and conclusions.

2 Problem setting

OBS provides scale and auxiliary supervision from self-selected usage, EXP provides unbiased outcome labels, and SIM provides counterfactual outputs. Figure 1 shows the causal structure; Appendix A.1 gives the full three-source pipeline.

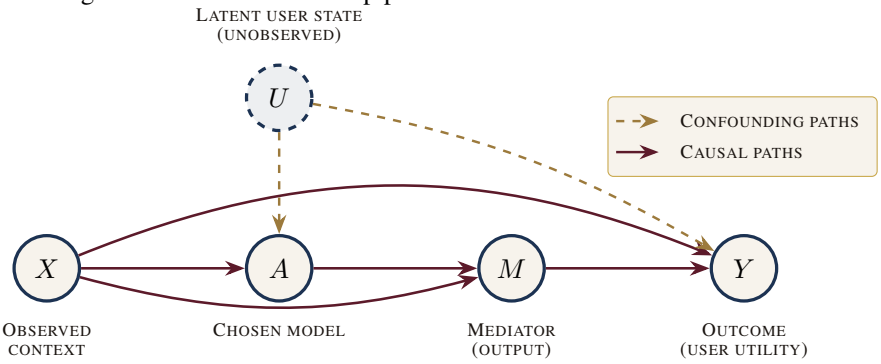

Figure 1: Causal graph with confounded model choice. The latent state U affects both model choice A and outcome Y, so OBS comparisons are biased. EXP randomization breaks the $U \to A$ link, and no latent mediator confounding blocks $U \to M$ once (X, A) is fixed.

We study (X, U, A, M, Y), where X is the observed context, U is an unobserved user state, $A \in \mathcal{A}$ indexes the chosen generative model, M is its output, and $Y \in [0, 1]$ is the outcome. The graph adds two restrictions to ordinary confounding: conditional on (X, A), U does not directly affect M, and A affects Y only through M.

Three sources. OBS is a large logged sample $D_{\mathrm{OBS}} = \{(X_i, A_i, M_i, Y_i, Z_i)\}_{i=1}^{n_{\mathrm{OBS}}}$ in which A_i may depend on U_i and auxiliary labels Z_i can train proxy representations. EXP is a smaller randomized sample $D_{\mathrm{EXP}} = \{(X_j, A_j, M_j, Y_j)\}_{j=1}^{n_{\mathrm{EXP}}}$ in which A is randomly assigned over the experimental action set $\mathcal{A}_{\mathrm{EXP}} \subseteq \mathcal{A}$, so $A \perp U \mid X$. SIM reruns model a on context x and samples $M \sim p_{\mathrm{sim}}(\cdot \mid x, a)$. EXP therefore supplies causal scores, SIM supplies counterfactual outputs, and OBS contributes auxiliary supervision used only after identification (Figure 2).

Targets. The marginal and context-conditional causal values are

$$\mu(a) := \mathbb{E}[Y(\mathrm{do}(A = a))], \qquad q(x, a) := \mathbb{E}[Y(\mathrm{do}(A = a)) \mid X = x].$$

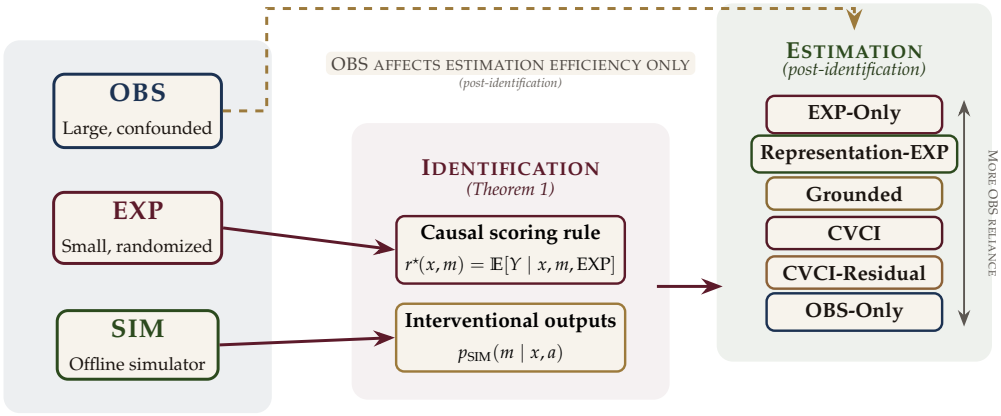

Figure 2: Three-source design. SIM and EXP are sufficient for identification of $r^\star$, q, and μ on shared support; OBS affects estimation efficiency only. Solid arrows mark the identification flow; dashed gold arrows mark post-identification statistical signal.

For finite cached benchmarks, write $r^\star(x, m) := \mathbb{E}[Y \mid X = x, M = m, \text{EXP}]$ for the EXP-side outcome regression and $m_{\text{cache}}(x, a)$ for the stored output associated with context x and model a. The cached evaluation target is

$$q_{\text{cache}}(x, a) := r^\star(x, m_{\text{cache}}(x, a)),$$

which replaces the SIM mediator average with a single cached draw and is therefore a property of the realized cache.

Assumptions. Identification uses five design assumptions, each isolating one source of bias.

Assumption A1 (Randomization in EXP)**.** In EXP, A is randomized over $\mathcal{A}_{\text{EXP}}$ so that $A \perp U$ and $A \perp U \mid X$. Without this, no source identifies the causal scoring rule.

Assumption A2 (SIM validity and shared support)**.** The simulator matches the interventional mediator law, $p_{\text{sim}}(m \mid x, a) = \mathbb{P}(M = m \mid X = x, \text{do}(A = a))$, with $\operatorname{supp} p_{\text{sim}}(\cdot \mid x, a) \subseteq \operatorname{supp} \mathbb{P}(M \mid X = x, A = a, \text{EXP})$ for every x in the EXP context support and every $a \in \mathcal{A}_{\text{EXP}}$.

Assumption A3 (Outcome consistency under intervention)**.** The conditional outcome law given (X, M, U) is invariant across OBS, EXP, and intervention: $\mathbb{P}(Y \mid X, M, U, \text{OBS}) = \mathbb{P}(Y \mid X, M, U, \text{EXP}) = \mathbb{P}(Y \mid X, M, U, \text{do}(A = a))$ on the relevant support.

Assumption A4 (No latent mediator confounding; single-round generation)**.** $U \perp M \mid X, A$, so the generative model has no edge $U \to M$ once (X, A) is fixed. This restricts the framework to single-round, stateless generation.

Assumption A5 (Context distribution alignment)**.** The marginal target distribution coincides with EXP, $P_X^{\text{tgt}} = P_X^{\text{EXP}}$. When this fails, only $q(x, a)$ on the shared support is identified; reweighting requires standard transportability arguments [Cole and Stuart, 2010, Pearl and Bareinboim, 2011, Bareinboim and Pearl, 2016].

Throughout the paper, "A1"–"A5" refer to Assumptions A1–A5; Appendix A.1 expands their motivation, support conditions, and transportability caveat.

3 Identification

EXP provides the unbiased scoring rule $r^\star(x, m) = \mathbb{E}[Y \mid X = x, M = m, \text{EXP}]$ defined in Section 2, and SIM provides the interventional mediator distribution $p_{\text{sim}}(\cdot \mid x, a)$; together they identify causal value. Throughout, the analysis is restricted to actions in the experimental action set $\mathcal{A}_{\text{EXP}}$ and to mediator values on the shared support of the EXP law of $M \mid X = x, A = a$ and the SIM law $p_{\text{sim}}(\cdot \mid x, a)$.

Theorem 1 (SIM plus EXP identifies causal value)**.** *Under the causal graph in Figure 1 and Assumptions A1–A5, for every x in the EXP context support and every $a \in \mathcal{A}_{\mathrm{EXP}}$,*

$$q(x,a) = \mathbb{E}\big[Y\big(\mathrm{do}(A=a)\big) \mid X=x\big] = \mathbb{E}_{M\sim p_{\mathrm{sim}}(\cdot\mid x,a)}[r^\star(x,M)]\,, \tag{1}$$

and therefore, with $\mu(a)$ defined under the target context distribution $P_X^{\mathrm{tgt}} = P_X^{\mathrm{EXP}}$ of Assumption A5,

$$\mu(a) = \mathbb{E}_{X\sim P_X^{\mathrm{tgt}}}[q(X,a)] = \mathbb{E}_{X\sim P_X^{\mathrm{tgt}}}\left[\mathbb{E}_{M\sim p_{\mathrm{sim}}(\cdot\mid X,a)}[r^\star(X,M)]\right]. \tag{2}$$

Hence both $q(x,a)$ and $\mu(a)$ are identified from EXP and SIM on the shared EXP/SIM support; identification places no unconfoundedness requirement on OBS.

Proof sketch. The proof uses each assumption exactly once. A4 strips M of any residual information about U given (X,A); A1 then removes the remaining $U \to A$ path inside EXP, so the conditional law of Y given (X,M) in EXP matches the law under $\mathrm{do}(A=a)$ on the shared support; A3 carries this equality from EXP to the post-intervention world, identifying the EXP regression $r^\star(x,m)$ as the causal scoring rule; A2 plugs in the SIM mediator law to obtain $q(x,a)$; and A5 averages over the target context distribution to recover $\mu(a)$. The full proof is in Appendix C.1; Appendix A.2 explains why conditioning on (X,M) in OBS still leaves residual U-confounding once A1 is dropped.

From identification to the experimental design. The controlled validation probes the theorem-level target with a known latent reward generator. The two real cached benchmarks report regret against q_{cache} with one cached output per context–action pair, diagnosing the post-identification estimator families. Across all settings, Section 4 uses OBS only after identification, to reduce variance or improve function approximation.

4 Estimators

After SIM and EXP identify the target, the remaining question is how OBS can reduce estimation error. The six estimator families differ in how they use OBS after identification: some ignore OBS outcomes, some use OBS only to learn a proxy representation, and some use OBS outcomes through a baseline or pooled fit. Proxy-learning details, value aggregation, and the theory-to-evidence map are expanded in Appendices A.3, A.6, and A.4.

Notation. Throughout, $z = (x,m)$ denotes a context–mediator pair, $\varphi(z) \in \mathbb{R}^d$ a fixed high-dimensional text-feature map, and $\psi(z) \in \mathbb{R}^k$ ($k \ll d$) a lower-dimensional proxy representation learned from OBS auxiliary labels in the spirit of embeddings adapted for causal adjustment [Veitch et al., 2020]. Write $\mathcal{F}_\varphi = \{r(z) = \mathrm{clip}_{[0,1]}(w^\top \varphi(z) + b) : (w,b) \in \mathbb{R}^{d+1}\}$ and $\mathcal{F}_\psi$ analogously on ψ for the clipped-linear function classes used as the reward-model hypothesis spaces; the proxy changes estimation only. Two families use OBS outcomes as well: f_{OBS} is a reward predictor pre-fit by ridge on D_{OBS}, used either as a baseline that EXP corrects (Grounded) or as part of a pooled OBS+EXP loss (CVCI variants).

Table 1: How the six reward estimators use OBS and EXP. Darker shading indicates heavier reliance on observational signal.

Estimator	OBS role	EXP role	Bias–variance intuition
EXP-Only	None	Fits the reward model	No confounding bias, but high variance when EXP is small.
OBS-Only	Fits the reward model on logged outcomes	None	Low variance, but inherits bias from confounded model choice.
Representation-EXP	Learns the proxy representation ψ from auxiliary labels	Fits the reward head on ψ	Effective when the target varies along directions retained by ψ.
Grounded	Provides the baseline predictor f_{OBS} and the proxy space ψ	Estimates and tunes a correction to f_{OBS}	Uses OBS for efficiency and EXP to remove low-dimensional bias.
CVCI	Contributes logged-outcome loss directly	Chooses the pooling weight by causal cross-validation and anchors the fit to randomized outcomes	Trades OBS bias against finite-EXP estimation error through direct pooling.
CVCI-Residual	Provides the baseline and contributes pooled residual loss in ψ	Chooses the pooling and shrinkage of the residual correction	Uses OBS for the coarse fit and EXP for the residual mismatch.

Table 2: Theory-to-evidence map for the six estimator families. Tier 1: formal main-paper support. Tier 2: formal appendix support. Tier 3: empirical diagnostic.

Mechanism	Tier	Formal support	Empirical reading
SIM + EXP identify the target	1	Theorem 1	All benchmarks separate scoring from replay/cache
Grounded enlarges the proxy correction class	2	Appendix Theorem on class expansion	Basis-expanded Grounded improves over single-linear in summarization
Oracle vs. finite-sample correction	2	Appendix Grounded-vs-OBS theorem	Grounded competitive but not uniformly best
CVCI-Residual residual simplicity	2	Appendix residual-vs-pooling theorem	Wins selected cells, can underperform on real summarization
Proxy/reward alignment for Representation-EXP	3	— (empirical)	Coding fix-success crossover near $\alpha_{\text{fix}} \approx 0.25$

Reward-fit objectives. All six reward models minimize a penalized squared loss with predictions clipped to $[0, 1]$. With shorthand $L_S(r) := \sum_{(x,m,y)\in D_S}(r(x, m) - y)^2$ for $S \in \{\text{OBS}, \text{EXP}\}$ and ridge penalty $\Omega(\cdot)$, the families differ in the data and feature space they use:

$$\widehat{r}_{\text{EXP}} = \arg\min_{r\in\mathcal{F}_\varphi} L_{\text{EXP}}(r) + \Omega(r), \tag{3}$$

$$\widehat{r}_{\text{OBS}} = \arg\min_{r\in\mathcal{F}_\varphi} L_{\text{OBS}}(r) + \Omega(r), \tag{4}$$

$$\widehat{r}_{\text{Rep}} = \arg\min_{r\in\mathcal{F}_\psi} L_{\text{EXP}}(r) + \Omega(r), \tag{5}$$

$$\widehat{r}_{\text{G}}(z) = \text{clip}_{[0,1]}\left[f_{\text{OBS}}(z) + \widehat{\theta}^\top\psi(z) + \widehat{b}\right], \tag{6}$$

$$\widehat{r}_{\text{CVCI}}(\lambda) = \arg\min_{r\in\mathcal{F}_\varphi} (1-\lambda)L_{\text{OBS}}(r) + \lambda L_{\text{EXP}}(r) + \Omega(r), \tag{7}$$

$$\widehat{r}_{\text{CVCIRes}}(\lambda) = f_{\text{OBS}} + \arg\min_{g\in\mathcal{F}_\psi} (1-\lambda)L^{\text{res}}_{\text{OBS}}(g) + \lambda L^{\text{res}}_{\text{EXP}}(g) + \Omega(g), \tag{8}$$

where $L^{\text{res}}_S(g) := \sum_{(x,m,y)\in D_S}(g(x, m) - (y - f_{\text{OBS}}(x, m)))^2$ is the residual loss against the OBS anchor, the Grounded correction $(\widehat{\theta}, \widehat{b})$ in (6) is fit on EXP residuals $Y - f_{\text{OBS}}(z)$, and the CVCI pooling weight $\widehat{\lambda} \in [0, 1]$ in (7)–(8) is chosen by EXP cross-validation. The unbiased–biased endpoints in (3)–(4) make EXP-Only the variance-bound baseline and OBS-Only the bias-bound endpoint; the four hybrids interpolate between them through three orthogonal levers (proxy features ψ, an OBS anchor f_{OBS}, and the pooling weight λ).

Theory-to-evidence map. Table 2 groups the six families by the kind of evidence that supports each comparison. Tier-1 statements (identification) are formal and main-paper; Tier-2 statements (oracle correction, residual simplicity, function-class enlargement) are formal but appendix-only; Tier-3 patterns (proxy alignment, reward geometry) are empirical diagnostics. The benchmarks in Sections 5–5.5 probe Tier 3, and Appendix A.4 expands the map.

From reward fits to model values. With a fitted reward $\widehat{r}$, denote by $\mathcal{X}_{\text{eval}}$ a held-out evaluation context set drawn from the target distribution $P^{\text{tgt}}_X = P^{\text{EXP}}_X$ and by $n_{\text{EXP},a}$ the number of EXP samples with $A_j = a$. The direct-method (DM) and doubly robust (DR) value estimators are

$$\widehat{\mu}^{\text{DM}}(a) = \frac{1}{|\mathcal{X}_{\text{eval}}|}\sum_{x\in\mathcal{X}_{\text{eval}}} \mathbb{E}_{M\sim p_{\text{sim}}(\cdot|x,a)}\big[\widehat{r}(x, M)\big], \tag{9}$$

$$\widehat{\mu}^{\text{DR}}(a) = \widehat{\mu}^{\text{DM}}(a) + \frac{1}{n_{\text{EXP},a}}\sum_{j:\, A_j=a}\big(Y_j - \widehat{r}(X_j, M_j)\big). \tag{10}$$

On cached benchmarks the SIM expectation collapses to a single cached draw and the DM term reduces to $\widehat{r}(x, m_{\text{cache}}(x, a))$. Empirically, on the synthetic $(\beta, n_{\text{OBS}}, n_{\text{EXP}})$ grid DR does not improve agent-level value RMSE over DM on average (Appendix Table 9: DR adds $+0.013$ to $+0.019$ RMSE for EXP-Only, OBS-Only, CVCI, and Representation-EXP), while in a separate strong-self-selection regime with intentional reward-model underfit, DR halves agent-level value RMSE (Appendix Figure 4). DR is therefore a model-level diagnostic, not a default. Sections 5–5.5 diagnose these benchmark-specific family instantiations; Appendix A.4 separates formal estimator support from empirical target-alignment diagnostics.

Table 3: Regret-optimal estimator on the semi-synthetic summarization benchmark across confounding strength β and budgets $(n_{\mathrm{OBS}}, n_{\mathrm{EXP}})$ (mean regret over 30 seeds; appendix table includes regret gaps).

Budget cell	$\beta = 0$	$\beta = 0.2$	$\beta = 0.5$	$\beta = 0.8$	$\beta = 0.9$	$\beta = 0.99$
(2,000, 20)	CVCI-Residual	CVCI	CVCI	OBS-Only	Grounded	Grounded
(20,000, 20)	CVCI	CVCI	CVCI	CVCI	CVCI	CVCI
(2,000, 100)	CVCI	CVCI	CVCI	OBS-Only	OBS-Only	OBS-Only
(20,000, 100)	CVCI	EXP-Only	EXP-Only	EXP-Only	EXP-Only	EXP-Only

5 Empirical benchmarks

5.1 Benchmark setup

Across all benchmarks, the task is to estimate a reward surface, recommend the best action on each held-out context, and measure held-out recommendation regret. The OBS choice law is governed by a confounding strength parameter $\beta \geq 0$ (with $\beta = 0$ yielding uniform OBS sampling and larger β inducing stronger user-side selection on A); EXP randomizes uniformly over $\mathcal{A}_{\mathrm{EXP}}$; and the two real cached benchmarks replace the latent target q with q_{cache}. Each fitted reward $\widehat{r}$ induces a per-context estimated value $\widehat{q}(x,a) = \mathbb{E}_{M \sim p_{\mathrm{sim}}(\cdot|x,a)}[\widehat{r}(x,M)]$ on the controlled benchmark, or $\widehat{q}(x,a) = \widehat{r}(x, m_{\mathrm{cache}}(x,a))$ on cached benchmarks, and a recommendation $\widehat{\pi}(x) = \arg\max_{a \in \mathcal{A}} \widehat{q}(x,a)$. We report

$$\mathrm{Regret}_{\mathrm{test}} = \frac{1}{|\mathcal{X}_{\mathrm{test}}|} \sum_{x \in \mathcal{X}_{\mathrm{test}}} \Big[\max_{a \in \mathcal{A}} q(x,a) - q(x, \widehat{\pi}(x)) \Big],$$

with q replaced by q_{cache} on held-out cached grids. Appendix A.12 gives the full benchmark laws and the benchmark snapshot table.

5.2 Controlled validation as a theory bridge

The controlled semi-synthetic validation replaces judged rewards in the summarization task family with a known latent reward generator, so identification-side error and estimation-side error can be measured separately. Sweeping the confounding strength β over six values and the budgets $(n_{\mathrm{OBS}}, n_{\mathrm{EXP}})$ over four settings yields a 24-cell regret map (Table 3). The pattern is regime-structured: at the smallest $(n_{\mathrm{OBS}}, n_{\mathrm{EXP}}) = (2{,}000, 20)$ row, four different families win across the six β values; at $(20{,}000, 20)$, CVCI wins all six β cells; at $(2{,}000, 100)$, the winner shifts from CVCI at low β to OBS-Only at high β; at $(20{,}000, 100)$, EXP-Only wins five of the six cells. CVCI is the most stable single winner, with most regret gaps below 0.002 in absolute terms (Appendix Table 15 reproduces the full grid with regret-gap annotations).

The cached benchmarks below retain the same families. Their winner pattern differs from this synthetic one (Section 5.4), so reward definition and auxiliary-feature alignment, not just the EXP budget, drive estimator choice in the real setting.

5.3 Main-text benchmarks

The main paper reports two real-task cached benchmarks with synthetic OBS/EXP resampling: CNN/DailyMail summarization [Hermann et al., 2015, Nallapati et al., 2016, See et al., 2017] and SWE-bench Verified coding with BouncerBench patches [Jimenez et al., 2024, OpenAI, 2024, Mathews and Nagappan, 2025]. Both use real tasks, real candidate outputs, and real judged or program-test rewards, while OBS and EXP are benchmark constructions over the cached pool. Appendix Table 14 lists the source data, cached objects, and targets.

5.4 Real-task cached summarization benchmark

The cached summarization benchmark uses 48 difficult CNN/DailyMail articles, 20 candidate summarization systems, and one judged summary per article–model pair. It evaluates smooth and sharpened user-segment reward maps on the same cache; OBS and EXP samples are obtained by resampling the cached pool under segment-dependent routing and uniform randomization. Appendix A.10 gives the

Method	$n_{\text{OBS}} = 2{,}000$		$n_{\text{OBS}} = 20{,}000$	
	$n_{\text{EXP}} = 20$	$n_{\text{EXP}} = 200$	$n_{\text{EXP}} = 20$	$n_{\text{EXP}} = 200$
CVCI	0.0128	0.0128	0.0120	0.0120
OBS-Only	0.0128	0.0128	0.0123	0.0123
Grounded	0.0200	0.0133	0.0133	0.0135
EXP-Only	0.0170	0.0121	0.0140	0.0118
Representation-EXP	0.0330	0.0184	0.0365	0.0124
Best OBS-based method	0.0128	0.0128	0.0120	0.0120
EXP-Only − best OBS-based	$+0.0042$	-0.0007	$+0.0020$	-0.0002

Table 4: Recommendation regret under the smooth aggregate reward (lower is better; mean over 30 seeds). Positive values in the last row mean an OBS-assisted family beats EXP-Only at that budget.

Table 5: Regret-optimal estimator on the real summarization benchmark across β and budgets, parallel to Table 3 (mean over 30 seeds; appendix Table 16 has regret gaps).

Budget cell	$\beta = 0$	$\beta = 0.2$	$\beta = 0.5$	$\beta = 0.8$	$\beta = 0.9$	$\beta = 0.99$
(2,000, 20)	Grounded	Grounded	Grounded	OBS-Only	OBS-Only	Grounded
(20,000, 20)	Grounded	Grounded	EXP-Only	EXP-Only	EXP-Only	EXP-Only
(2,000, 100)	OBS-Only	OBS-Only	OBS-Only	CVCI	CVCI	CVCI
(20,000, 100)	CVCI-Residual	CVCI	CVCI	CVCI	CVCI	CVCI

exact slate construction, reward maps, budget grid, and standard errors; Appendix Table 11 reports the full aggregate ranking summary.

Supervision scarcity. Table 4 traces a budget-driven crossover. At $n_{\text{EXP}} = 20$ the best OBS-assisted family improves on EXP-Only, and at $n_{\text{EXP}} = 200$ EXP-Only catches up or slightly wins. This matches the post-identification view in Table 2: once EXP and SIM identify the target, additional randomized labels improve the EXP-side reward fit. In the aggregate smooth/sharpened comparison, CVCI has the best average rank and excess regret, with a small gap to OBS-Only relative to cell-level uncertainty (Appendix Tables 11 and 13). Reward-shape and R^2 diagnostics in Appendix Table 12 show that varying the target reward while fixing the cached outputs changes alignment with the rubric-linear and proxy classes.

Real-vs-synthetic winner-map contrast. On the same six-β four-budget grid, the regret winner pattern on the cached real benchmark differs from the synthetic surface (Table 5). Grounded wins five of six low-budget cells (n_{OBS}=2,000, n_{EXP}=20) where CVCI and CVCI-Residual dominate the synthetic version, and CVCI reclaims the high-OBS large-EXP row (20,000, 100). Ranking diagnostics tell yet another story: Grounded wins top-1 and top-3 in all 24 real cells while winning regret in only six (Appendix A.14). Estimator choice is therefore a function of (data regime, deployment loss, target reward) jointly, and a single "best" family does not transfer between the synthetic and the real benchmark even when the budget grid is identical.

5.5 Real-task cached coding benchmark

The cached coding benchmark uses SWE-bench Verified issues [Jimenez et al., 2024, OpenAI, 2024] and one BouncerBench candidate patch per issue–agent pair [Mathews and Nagappan, 2025]. It fixes $n_{\text{OBS}} = 2{,}000$, $n_{\text{EXP}} = 100$, and $\beta = 0.5$, then varies the reward definition. Each patch has true fix success c_1 and patch-quality components c_2, c_3, c_4. The fix-success weight α_{fix} shifts target mass toward c_1, while ω_{weak} changes the non-success components from additive to weakest-link aggregation. Appendix A.15 gives the full construction, R^2 diagnostics, and pairwise heatmaps.

Reward-alignment crossover. For user type u, let $w_{u1} \in [0, 1]$ denote the fix-success mass for that type and $\bar{w}_{uk} \geq 0$ ($k = 2, 3, 4$, $\sum_k \bar{w}_{uk} = 1$) the normalized weights on the patch-quality components. The benchmark utility is then

$$Y_u = w_{u1}\alpha_{\text{fix}}c_1 + (1 - w_{u1})\left((1 - \omega_{\text{weak}})\sum_{k=2}^{4} \bar{w}_{uk}c_k + \omega_{\text{weak}} \min_{k=2,3,4} c_k\right),$$

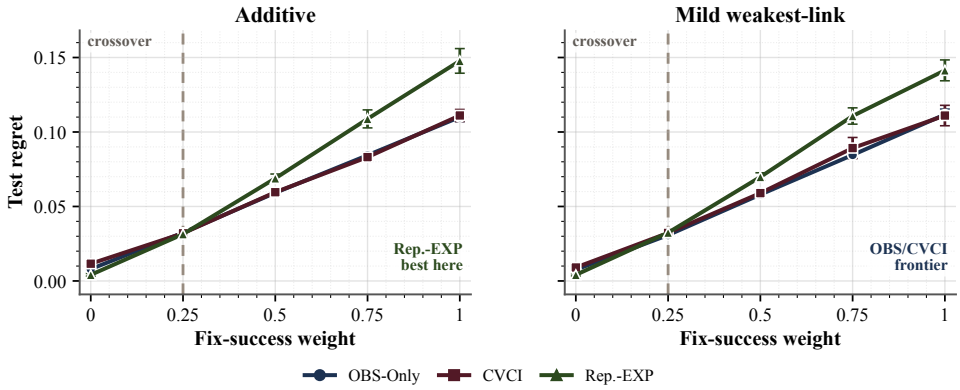

Figure 3: Mean regret in the cached coding benchmark as the fix-success weight α_{fix} increases (30 seeds; lower is better). At low α_{fix}, the target follows auxiliary patch-quality components and Representation-EXP is competitive; above the crossover near 0.25, OBS-based families overtake it.

and q_{cache} averages Y_u over benchmark user types. Figure 3 shows a crossover near $\alpha_{\text{fix}} \approx 0.25$: Representation-EXP is among the lowest-regret families when the target follows patch quality, while OBS-based families win once true fix success dominates. On the additive slice, the regret gap between Representation-EXP and the better of OBS-Only and CVCI moves from -0.0039 at $\alpha_{\text{fix}} = 0$ to $+0.0003$ at 0.25 and $+0.0415$ at 1. The mild weakest-link slice ($\omega_{\text{weak}} = 0.25$) preserves the same crossover and ordering, isolating the effect to which component of the reward dominates. Tilting toward c_1 moves the target away from auxiliary patch-quality signal that Representation-EXP can extract from OBS labels and toward the binary fix-success signal that pooled OBS+EXP fits absorb directly. The full 5×5 pairwise delta heatmaps in Appendix Figure 5 show the same target-alignment diagnostic across both aggregation modes.

6 Related work

The paper combines off-policy evaluation, hybrid observational–randomized estimation, and LLM evaluation. Standard Direct Method with Replay (DM) and Doubly Robust (DR) estimators come from contextual-bandit, policy-evaluation, and missing-data work [Robins et al., 1994, Bang and Robins, 2005, Dudík et al., 2011, 2014, Swaminathan and Joachims, 2015, Jiang and Li, 2016, Thomas and Brunskill, 2016], where logged outcomes are typically assumed unconfounded given context. Our setting differs because user-side factors can confound both model choice and outcome, so SIM plus EXP first identifies the reward surface and OBS enters only afterward.

The estimator families are closest to experimental grounding, data fusion, trial generalizability, and recent methods for combining large biased samples with smaller randomized samples [Kallus et al., 2018, Cole and Stuart, 2010, Pearl and Bareinboim, 2011, Bareinboim and Pearl, 2016, Rosenman et al., 2023, Cheng and Cai, 2021, Lin et al., 2025, Yang et al., 2025, Colnet et al., 2024]. The experiments draw on summarization and LLM-evaluation work, including CNN/DailyMail [Hermann et al., 2015, Nallapati et al., 2016, See et al., 2017, Stiennon et al., 2020], benchmark and judge reliability studies [Liang et al., 2023, Zheng et al., 2023, Chiang et al., 2024], real-world logs [Zheng et al., 2024, Zhao et al., 2024], evaluator-bias analyses [Wang et al., 2023, Dubois et al., 2024, Panickssery et al., 2024, Verga et al., 2024], preference-learning pipelines [Christiano et al., 2017, Ouyang et al., 2022, Rafailov et al., 2023], and causal NLP views of text as treatment, outcome, mediator, or proxy [Feder et al., 2022]. Appendix A.16 preserves the fuller positioning.

7 Discussion and Future Work

Each benchmark probes a different driver of estimator choice. The controlled validation isolates the EXP-budget effect with a known reward generator; the summarization benchmark stresses supervision scarcity at fixed reward geometry; and the coding benchmark stresses reward-target geometry at fixed supervision. The pattern across the three sweeps yields concrete deployment guidance.

A practical recipe. Tie estimator choice to three observable quantities. (i) The deployment loss: regret and top-rank metrics rank families differently, with Grounded winning all 24 top-1 and top-3 ranking cells in the cached summarization benchmark while winning regret in only six. (ii) The EXP budget n_{EXP} relative to the OBS budget n_{OBS}: OBS-assisted families gain most when EXP is scarce, and EXP-Only catches up as EXP grows. (iii) How well the auxiliary representation ψ or the OBS predictor f_{OBS} tracks the held-out target: both Representation-EXP and OBS-anchored families inherit the alignment of the OBS signal with the deployment reward.

When OBS information helps and when it does not. Whenever the deployment reward is well predicted by the same features that drive OBS outcomes, the pooled fits in (7) reduce variance without paying a meaningful bias cost; CVCI and Grounded are then strong defaults. When the deployment reward depends on a signal that OBS labels do not contain — the binary fix-success component in coding, or a sharply re-weighted user segment in summarization — the same OBS pull becomes a bias liability and EXP-Only or Representation-EXP becomes preferable. The coding crossover near $\alpha_{\mathrm{fix}} \approx 0.25$ in Figure 3 is a concrete instance of this transition.

Scope and extensions. The real-task benchmarks use real candidate outputs and real judged or program-test rewards; OBS and EXP mechanisms are controlled constructions over cached pools, which is what makes the supervision-scarcity and reward-geometry sweeps cleanly interpretable. Two extensions follow directly: broader tasks with naturally observed deployment logs paired with live randomized trials, and multi-round human–agent interaction, which relaxes Assumption A4 to allow stateful $U \to M$ paths through dialogue history. Appendix A.16 expands on scope, limitations, and alternative scoring rules.

8 Conclusion

Offline evaluation of language models separates cleanly into identification and post-identification estimation. SIM and EXP together identify causal model values on shared support, while OBS contributes after identification through auxiliary signal or biased outcomes whenever they predict the causal target. Across the controlled validation and two cached real-task probes, the best estimator tracks the EXP budget and the alignment between OBS-derived structure and the target reward; live randomization remains the prerequisite for any deployment-grade comparison.

References

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. doi: 10.1073/pnas.1510507113. URL https://www.pnas.org/doi/10.1073/pnas.1510507113.

David Cheng and Tianxi Cai. Adaptive combination of randomized and observational data, 2021. URL https://arxiv.org/abs/2111.15012.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03741.

Stephen R. Cole and Elizabeth A. Stuart. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, 172(1):107–115, 2010. doi: 10.1093/aje/kwq084.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: A review. *Statistical Science*, 39(1):165–191, 2024. doi: 10.1214/23-STS889. URL https://doi.org/10.1214/23-STS889.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, 2011. URL https://arxiv.org/abs/1103.4601.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014. doi: 10.1214/14-STS500. URL https://doi.org/10.1214/14-STS500.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022. doi: 10.1162/tacl_a_00511.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1506.03340.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 652–661, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06770.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, volume 31, pages 10911–10920, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/566f0ea4f6c2e947f36795c8f58ba901-Abstract.html.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2211.09110.

Xi Lin, Jens Magelund Tarp, and Robin J. Evans. Combining experimental and observational data through a power likelihood. *Biometrics*, 81(1):ujaf008, 2025. doi: 10.1093/biomtc/ujaf008. URL https://academic.oup.com/biometrics/article/81/1/ujaf008/8016472.

Noble Saji Mathews and Meiyappan Nagappan. Is your automated software engineer trustworthy?, 2025. URL https://arxiv.org/abs/2506.17812.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, 2016. URL https://arxiv.org/abs/1602.06023.

OpenAI. Introducing SWE-bench Verified. OpenAI Blog, 2024. URL https://openai.com/index/introducing-swe-bench-verified/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations, 2024. URL https://arxiv.org/abs/2404.13076.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011. doi: 10.1609/aaai.v25i1.7861.

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.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994. doi: 10.1080/01621459.1994.10476818.

Evan T. R. Rosenman, Guillaume Basse, Art B. Owen, and Mike Baiocchi. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 79(4):2961–2973, 2023. doi: 10.1111/biom.13827. URL https://pubmed.ncbi.nlm.nih.gov/36629736/.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1073–1083, 2017. URL https://arxiv.org/abs/1704.04368.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2009.01325.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 814–823, 2015. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2139–2148, 2016. URL https://proceedings.mlr.press/v48/thomasa16.html.

Victor Veitch, Dhanya Sridhar, and David M. Blei. Adapting text embeddings for causal inference. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 919–928, 2020. URL https://proceedings.mlr.press/v124/veitch20a.html.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models, 2024. URL https://arxiv.org/abs/2404.18796.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators, 2023. URL https://arxiv.org/abs/2305.17926.

Xuelin Yang, Licong Lin, Susan Athey, Michael I. Jordan, and Guido W. Imbens. Cross-validated causal inference: a modern method to combine experimental and observational data, 2025. URL https://arxiv.org/abs/2511.00727.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1M ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2405.01470.

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://arxiv.org/abs/2306.05685.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BOfDKxfwt0.

A Additional details

A.1 Extended problem setting and assumptions

This appendix preserves the full three-source design and assumption motivation behind Section 2.

Accessible data sources. Three data sources are available, each contributing something the others cannot provide alone. OBS is a large observational dataset

$$D_{\mathrm{OBS}} = \{(X_i, A_i, M_i, Y_i, Z_i)\}_{i=1}^{n_{\mathrm{OBS}}},$$

where A_i may depend on the corresponding latent state U_i and $Z_i \in [0,1]^K$ are auxiliary labels that may or may not be available. The auxiliary labels provide extra supervision for learning a compact proxy representation of the context–output pair; user feedback ratings and pairwise preferences are natural sources and can often be collected at scale [Christiano et al., 2017, Ouyang et al., 2022, Chiang et al., 2024].

EXP is a much smaller dataset

$$D_{\mathrm{EXP}} = \{(X_j, A_j, M_j, Y_j)\}_{j=1}^{n_{\mathrm{EXP}}},$$

in which model choice is randomized and therefore unconfounded. Since A_j is drawn from the experimental action set $\mathcal{A}_{\mathrm{EXP}}$ according to a fixed distribution, $A \perp U$ and $A \perp U \mid X$, so EXP outcomes are causally interpretable. Randomizing model choice disrupts user experience, which is why these experiments remain small in practice.

The simulator reruns any generative model on any context and returns a mediator draw from $M \sim p_{\mathrm{sim}}(\cdot \mid X = x, A = a)$. Operationally, SIM reruns generative model a on held-out context x and records the generated output. In generative systems, SIM is especially natural because prompts, retrieved documents, tool traces, and other environment inputs can often be stored as deterministic inputs, while the remaining randomness comes from the generative model itself. Figure 2 (main text) shows the three-source pipeline.

Target estimand details. The marginal model value asks how well each model would perform on average if used by the full target population:

$$\mu(a) := \mathbb{E}\big[Y\big(\mathrm{do}(A = a)\big)\big].$$

The context-conditional model value asks what outcome each model would deliver for a particular observed context:

$$q(x, a) := \mathbb{E}\big[Y\big(\mathrm{do}(A = a)\big) \mid X = x\big].$$

Section 3 shows how both targets are recovered: SIM supplies the model-specific mediator distribution and EXP provides unconfounded scores for realized outputs. In real cached benchmarks, the finite cached target q_{cache} approximates $q(x, a)$ while the superpopulation estimand remains latent.

Design assumptions for SIM-based identification. **A1 (Randomization in EXP).** In EXP, A is randomized over $\mathcal{A}_{\mathrm{EXP}}$ so that $A \perp U$ and $A \perp U \mid X$ hold. Randomization is the sole source of unconfounded outcome information; without it, identification is impossible regardless of OBS size.

A2 (SIM validity and support). The simulator matches the interventional mediator distribution, $p_{\mathrm{sim}}(m \mid x, a) = \mathbb{P}(M = m \mid X = x, \mathrm{do}(A = a))$, and is supported within the EXP mediator support: for every x in the EXP context support and every $a \in \mathcal{A}_{\mathrm{EXP}}$,

$$\operatorname{supp} p_{\mathrm{sim}}(\cdot \mid x, a) \;\subseteq\; \operatorname{supp} \mathbb{P}(M \mid X = x, A = a, \mathrm{EXP}).$$

Throughout the paper, "the shared EXP/SIM support" refers to this support condition. SIM validity is satisfied whenever any randomness in the generated output comes from the generative model itself and all deployment-time inputs that affect the generative model's behavior are included in X.

A3 (Outcome consistency under intervention). OBS and EXP measure the same outcome Y, and the conditional outcome law given (X, M, U) is invariant across OBS, EXP, and the post-intervention world: for every (x, m, u) on the relevant support,

$$\begin{aligned}\mathbb{P}(Y \mid X = x, M = m, U = u, \mathrm{OBS}) \;&=\; \mathbb{P}(Y \mid X = x, M = m, U = u, \mathrm{EXP}) \\ &=\; \mathbb{P}(Y \mid X = x, M = m, U = u, \mathrm{do}(A = a)).\end{aligned}$$

Only the model-choice mechanism differs across OBS, EXP, and intervention. Without this, EXP supervision could not transport either to the intervention or to OBS-trained reward models.

A4 (No latent mediator confounding; single-round generation). Conditional on (X, A), the latent state U is independent of the mediator M in OBS, EXP, and under intervention:

$$U \perp M \mid X, A.$$

Equivalently, the SCM has no edge $U \to M$ once (X, A) is fixed, and any deployment-time inputs that drive the generative model are absorbed into X. A4 restricts the framework to single-round, stateless generation. In repeated, adaptive, or multi-round interactions the residual dependence of M on U given (X, A) is generally nonzero, so A4 fails.

A5 (Context distribution alignment). The marginal model value $\mu(a)$ is defined with respect to a target context distribution P_X^{tgt} that coincides with the EXP context distribution: $P_X^{\mathrm{tgt}} = P_X^{\mathrm{EXP}}$. Equivalently, the EXP sample is drawn from the same context population over which $\mu(a)$ is averaged. When this fails, only $q(x, a)$ on the shared support is identified, and reweighting to a different P_X^{tgt} requires standard transportability and trial-generalizability arguments [Cole and Stuart, 2010, Pearl and Bareinboim, 2011, Bareinboim and Pearl, 2016].

A.2 Identification remarks

SIM resolves generation; EXP resolves scoring. SIM regenerates the mediator M under $\mathrm{do}(A = a)$, while EXP supplies the outcome labels needed to score that mediator. So SIM handles counterfactual generation and EXP identifies causal scoring. The randomized sample is what identifies the scoring rule $r^\star(x, m)$ for a realized context–output pair.

Why observational outcomes target a different object. In deployment, A depends on the latent variable U, and U also affects Y directly. Conditioning on (X, M) therefore leaves self-selection bias in place, so in general

$$\mathbb{E}[Y \mid X = x, M = m, \mathrm{OBS}] \neq r^\star(x, m).$$

This is why the additional assumptions in Section 4 are estimator-side assumptions for using OBS, while identification itself comes from SIM and EXP.

A.3 Estimator details and concrete implementations

Exact reward-model objectives. This subsection records the exact fitting problems and benchmark-specific implementation details deferred from Section 4. EXP-Only fits

$$(\widehat{w}_{\mathrm{EXP}}, \widehat{b}_{\mathrm{EXP}}) \in \arg\min_{w,b} \sum_{j \in \mathrm{EXP}} \big(Y_j - (w^\top \varphi(z_j) + b)\big)^2 + \alpha \|w\|_2^2,$$

with prediction rule $\widehat{r}_{\mathrm{EXP}}(z) = \mathrm{clip}_{[0,1]}(\widehat{w}_{\mathrm{EXP}}^\top \varphi(z) + \widehat{b}_{\mathrm{EXP}})$.

Representation-EXP fits the same ridge head on the proxy representation:

$$(\widehat{w}_{\psi}, \widehat{b}_{\psi}) \in \arg\min_{w,b} \sum_{j \in \mathrm{EXP}} \big(Y_j - (w^\top \psi(z_j) + b)\big)^2 + \alpha \|w\|_2^2,$$

with $\widehat{r}_{\psi}(z) = \mathrm{clip}_{[0,1]}(\widehat{w}_{\psi}^\top \psi(z) + \widehat{b}_{\psi})$.

OBS-Only fits the raw-feature ridge model on OBS outcomes,

$$(\widehat{w}_{\mathrm{OBS}}, \widehat{b}_{\mathrm{OBS}}) \in \arg\min_{w,b} \sum_{i \in \mathrm{OBS}} \big(Y_i - (w^\top \varphi(z_i) + b)\big)^2 + \alpha \|w\|_2^2,$$

and predicts $f_{\mathrm{OBS}}(z) = \mathrm{clip}_{[0,1]}(\widehat{w}_{\mathrm{OBS}}^\top \varphi(z) + \widehat{b}_{\mathrm{OBS}})$.

For Grounded, define EXP residual targets $\Delta_j = f_{\mathrm{OBS}}(z_j) - Y_j$ and fit

$$(\widehat{\theta}, \widehat{c}) \in \arg\min_{\theta,c} \sum_{j \in \mathrm{EXP}} \big(\Delta_j - (\theta^\top \psi(z_j) + c)\big)^2 + \lambda_\theta \|\theta\|_2^2.$$

Table 6: How each estimator family is implemented in the appendix validation and the real benchmarks. Family names are shared across settings, but the proxy channel and EXP-side tuning differ by benchmark.

Paper family	Semi-synthetic instantiation	Real-benchmark instantiation	EXP-side tuning
EXP-Only	Raw-feature ridge fit on EXP only	Same family on cached real outputs	Fixed defaults
OBS-Only	Raw-feature ridge fit on OBS outcomes only	Same family on cached real outputs	None
Representation-EXP	EXP-only reward head on a proxy learned from OBS auxiliary labels	EXP-only reward head on a heuristic proxy learned from cached trajectory metadata	Fixed defaults
Grounded	OBS baseline plus low-dimensional proxy correction	OBS baseline plus richer proxy-basis correction on the heuristic auxiliary proxy	Small EXP-only tuning
CVCI	Direct pooled OBS/EXP fit in raw text features	Same family with model-level EXP cross-validation for the pooling weight	Benchmark-specific EXP tuning
CVCI-Residual	OBS baseline plus pooled residual fit in proxy space	Same family with the heuristic auxiliary proxy and model-level EXP cross-validation	Benchmark-specific EXP tuning

The resulting predictor subtracts the fitted proxy-side correction from the OBS baseline and clips the result to $[0, 1]$. In the real benchmarks, the grounded family keeps the same outer form but replaces the single linear proxy correction with a small library of proxy-basis corrections built on the heuristic auxiliary proxy. Appendix A.5 compares this richer correction class with a single-linear baseline and with a weak pooled-anchor variant.

CVCI pools the OBS and EXP regression objectives:

$$(\widehat{w}_\lambda, \widehat{b}_\lambda) \in \arg\min_{w,b} \Bigg[(1-\lambda) \frac{1}{n_{\mathrm{EXP}}} \sum_{j \in \mathrm{EXP}} \big(Y_j - (w^\top \varphi(z_j) + b)\big)^2 + \lambda \frac{1}{n_{\mathrm{OBS}}} \sum_{i \in \mathrm{OBS}} \big(Y_i - (w^\top \varphi(z_i) + b)\big)^2 \Bigg] + \alpha \|w\|_2^2, \tag{11}$$

with $\widehat{r}_\lambda(z) = \mathrm{clip}_{[0,1]}(\widehat{w}_\lambda^\top \varphi(z) + \widehat{b}_\lambda)$. Here $\lambda = 1$ is the OBS-only endpoint and $\lambda = 0$ is the EXP-only endpoint. The final predictor refits (11) on the full OBS and EXP data at the value of λ selected by Appendix A.11.

CVCI-Residual residualizes around f_{OBS} using the residual target $Y_j - f_{\mathrm{OBS}}(z_j)$ and pools only that residual fit in proxy space:

$$(\widehat{\theta}_\lambda, \widehat{c}_\lambda) \in \arg\min_{\theta,c} \Bigg[(1-\lambda) \frac{1}{n_{\mathrm{EXP}}} \sum_{j \in \mathrm{EXP}} \big(Y_j - f_{\mathrm{OBS}}(z_j) - (\theta^\top \psi(z_j) + c)\big)^2 + \lambda \frac{1}{n_{\mathrm{OBS}}} \sum_{i \in \mathrm{OBS}} \big(Y_i - f_{\mathrm{OBS}}(z_i) - (\theta^\top \psi(z_i) + c)\big)^2 \Bigg] + \alpha_\psi \|\theta\|_2^2,$$

with $\widehat{r}_{\mathrm{res},\lambda}(z) = \mathrm{clip}_{[0,1]}\big(f_{\mathrm{OBS}}(z) + \widehat{\theta}_\lambda^\top \psi(z) + \widehat{c}_\lambda\big)$. In the experiments, both the pooling weight λ and the residual ridge penalty α_ψ are selected by Appendix A.11; when residual-CVCI candidates are tied within numerical tolerance, the implementation resolves the tie toward the more regularized candidate.

How each estimator family is implemented. Table 6 shows how each estimator family is implemented in the appendix validation and in the two real benchmarks. The family names are the same across settings, but the proxy construction and EXP-side tuning differ by benchmark.

A.4 Theory-to-evidence map for estimator comparisons

The estimator comparison is guided by three evidence tiers. The identification theorem is part of the main paper. The estimator-comparison results in Appendix C provide formal support for several oracle and approximation statements. Target-alignment patterns in the benchmarks diagnose whether a target reward aligns with a proxy, baseline, or correction class. They provide empirical rather than formal evidence. Table 7 summarizes this map.

Throughout Section 5.4, Grounded refers to the rich proxy-basis version in Table 6. Appendix A.5 compares it with the single-linear baseline and the pooled-anchor variant.

Table 7: Theory-to-evidence map for the estimator comparison. Tier 1 is formal support in the main paper; Tier 2 is formal support from the appendix; Tier 3 is empirical diagnostic evidence only.

Mechanism	Tier	Formal support	Empirical diagnostic	Interpretation rule
SIM plus EXP identify the target	1	Theorem 1	All benchmarks separate randomized scoring from replayed or cached outputs	EXP supplies unconfounded scores; SIM or cached replay supplies mediators.
Grounded can enlarge a proxy-side approximation class	2	Theorem 2	Basis-expanded Grounded improves over the single-linear grounded baseline in Appendix A.5	Treat this as benchmark evidence for richer correction bases; the theorem establishes class expansion, not finite-sample dominance.
Oracle correction versus finite-sample correction	2	Theorem 4 and (15)	Grounded is competitive but not uniformly best in the cached benchmarks	The oracle result is not a finite-sample dominance guarantee.
Residual simplicity	2	Theorem 3	CVCI-Residual wins only selected cells and often underperforms in real summarization	Read wins or losses as benchmark evidence about residual simplicity, not as direct tests of the oracle condition.
Proxy alignment and reward definition	3	Empirical diagnostic unless additional approximation-class results are formalized	Summarization R^2 diagnostics and coding fix-success sweep	Interpret as empirical diagnostics rather than formal oracle statements.

A.5 Grounded-family implementations for the real cached benchmark

A common grounded template. All real-benchmark grounded variants start from the same decomposition: fit an OBS baseline f_{OBS} from logged outcomes, then use the heuristic auxiliary proxy to estimate a correction that is subtracted from that baseline. The variants differ along two design axes:

1. the correction class used on the proxy representation, and
2. whether OBS enters the second stage only through f_{OBS} or also through a small pooled anchor after the correction direction has been fixed.

The grounded-family comparison instead organizes three principled instantiations of the same baseline-minus-correction idea.

Variant 1: single-linear grounded correction. This is the most straightforward grounded instantiation. It uses the heuristic auxiliary proxy, keeps a single linear correction family, and tunes the correction strength on EXP. For the auxiliary real summarization comparison over $\beta \in \{0, 0.2, 0.5, 0.8, 0.9, 0.99\}$, $n_{\mathrm{OBS}} \in \{2{,}000, 20{,}000\}$, and $n_{\mathrm{EXP}} \in \{20, 100\}$, we report the version of this linear class tuned by model-level EXP cross-validation, which serves as the matched linear baseline for the richer grounded variants. That single-linear class uses the predictor

$$\widehat{r}_{\mathrm{lin},\alpha}(z) = \mathrm{clip}_{[0,1]}\Big(f_{\mathrm{OBS}}(z) - \alpha\big(\widehat{\theta}^{\top}\psi_{\mathrm{aux}}(z) + \widehat{c}\big)\Big),$$

where ψ_{aux} is the learned heuristic auxiliary proxy and the correction is linear in that proxy.

Variant 2: Grounded as a rich but still small proxy-side correction class. The main-text Grounded instantiation keeps the same baseline-minus-correction outer form, but replaces the single linear correction by a small library of proxy-basis corrections. Write $\tilde{\psi}_{\mathrm{aux}}(z) \in \mathbb{R}^{16}$ for the standardized 16-dimensional SVD compression of the learned heuristic auxiliary proxy. For

$$B \in \{B_{\mathrm{id}}, B_{\mathrm{poly}}\}, \qquad B_{\mathrm{id}}(u) = u, \qquad B_{\mathrm{poly}}(u) = \big[u,\ u \odot u\big],$$

and ridge level $\tau \in \{10^{-2}, 1, 100\}$, the rich grounded direction fits

$$(\widehat{\theta}_{B,\tau}, \widehat{c}_{B,\tau}) \in \arg\min_{\theta,c} \sum_{j \in \mathrm{EXP}} \big(\Delta_j - (\theta^{\top} B(\tilde{\psi}_{\mathrm{aux}}(z_j)) + c)\big)^2 + \tau\|\theta\|_2^2, \qquad \Delta_j = f_{\mathrm{OBS}}(z_j) - Y_j,$$

and predicts

$$\widehat{r}_{\mathrm{rich},B,\alpha}(z) = \mathrm{clip}_{[0,1]}\Big(f_{\mathrm{OBS}}(z) - \alpha\big(\widehat{\theta}_{B,\tau}^{\top} B(\tilde{\psi}_{\mathrm{aux}}(z)) + \widehat{c}_{B,\tau}\big)\Big).$$

Table 8: Grounded-family implementations for the real cached benchmark. The table compares the single-linear grounded baseline, the richer proxy-basis grounded instantiation used in the main text, and a weak pooled-anchor variant built on the same rich correction direction. 'Regret wins' counts wins among the seven methods obtained by adding the pooled-anchor variant to the six main-text methods.

Variant	Correction family	Second-stage OBS use	Tuning	Regret wins	Macro regret	Mean top-1	Mean top-3
Single-linear correction baseline	Single linear correction on the heuristic auxiliary proxy	None beyond the baseline f_{OBS}	Model-level EXP cross-validation for α_{corr}	1	0.0357	0.628	0.646
Grounded rich proxy-basis correction	Rich proxy-basis correction on standardized 16-dimensional heuristic auxiliary features with $B \in \{\mathrm{id}, \mathrm{poly2}\}$ and $\tau \in \{10^{-2}, 1, 100\}$	None beyond the baseline f_{OBS}	Model-level EXP cross-validation over $(B, \tau, \alpha_{\mathrm{corr}})$	6	0.0318	0.736	0.736
Rich correction with weak pooled anchor	Same rich EXP-only correction direction as Grounded, followed by anchored coefficient pooling	OBS enters only through the pooled coefficients (b, α)	Model-level EXP cross-validation over (B, τ, λ) with closed-form anchored (b, α)	4	0.0326	0.744	0.744

The final implementation selects (B, τ, α) by agent-CV on EXP. Relative to the linear grounded class, this replaces the correction family

$$\mathcal{H}_{\mathrm{lin}} = \Big\{ u \mapsto \theta^\top u + c \Big\}$$

by

$$\mathcal{H}_{\mathrm{rich}} = \Big\{ u \mapsto \theta^\top B(u) + c : \; B \in \{B_{\mathrm{id}}, B_{\mathrm{poly}}\} \Big\},$$

so $\mathcal{H}_{\mathrm{lin}} \subseteq \mathcal{H}_{\mathrm{rich}}$. The corresponding grounded predictor class therefore expands the proxy-side approximation space while still learning the correction direction from EXP only.

Variant 3: pooled-anchor grounded. The appendix pooled-anchor variant keeps the rich EXP-only correction direction but adds a weak pooled anchor in the second stage. Let

$$\widehat{c}_{B,\tau}(z) = \widehat{\theta}_{B,\tau}^\top B(\tilde{\psi}_{\mathrm{aux}}(z)) + \widehat{c}_{B,\tau}$$

denote the grounded correction direction learned from EXP. The pooled-anchor predictor is

$$\widehat{r}_{\mathrm{anchor},\lambda}(z) = \mathrm{clip}_{[0,1]}\Big(\widehat{b}_\lambda \, f_{\mathrm{OBS}}(z) - \widehat{\alpha}_\lambda \, \widehat{c}_{B,\tau}(z) \Big),$$

where $(\widehat{b}_\lambda, \widehat{\alpha}_\lambda)$ are obtained from the anchored objective

$$\begin{aligned}(\widehat{b}_\lambda, \widehat{\alpha}_\lambda) \in \arg\min_{b,\alpha} \Bigg[\lambda \frac{1}{n_{\mathrm{OBS}}} \sum_{i \in \mathrm{OBS}} \big(Y_i - (b f^{\mathrm{cf}}_{\mathrm{OBS}}(z_i) - \alpha \widehat{c}_{B,\tau}(z_i)) \big)^2 \\ + (1-\lambda) \frac{1}{n_{\mathrm{EXP}}} \sum_{j \in \mathrm{EXP}} \big(Y_j - (b f_{\mathrm{OBS}}(z_j) - \alpha \widehat{c}_{B,\tau}(z_j)) \big)^2 \Bigg] \\ + \rho_b (b-1)^2 + \rho_\alpha (\alpha - 1)^2.\end{aligned}$$

Here $f^{\mathrm{cf}}_{\mathrm{OBS}}$ is a cross-fitted OBS baseline prediction. This variant pools only two calibration coefficients after EXP has already fixed the correction direction.

Why these implementations should behave differently. The three variants place their modeling flexibility in different parts of the decomposition. The linear grounded baseline puts the entire second stage inside the smallest class $\mathcal{H}_{\mathrm{lin}}$. So it is the most conservative option, but it can underfit whenever the discrepancy between f_{OBS} and the causal target is only approximately linear in the proxy. Because Grounded replaces $\mathcal{H}_{\mathrm{lin}}$ by the strictly larger class $\mathcal{H}_{\mathrm{rich}} \supseteq \mathcal{H}_{\mathrm{lin}}$ while still estimating the correction direction from EXP only, it should help precisely when the residual mismatch is slightly nonlinear in the proxy yet still low-dimensional. The pooled-anchor variant changes a different part of the problem. It keeps the correction direction fixed at the grounded solution and lets OBS influence only the coarse scaling coefficients (b, α). It should help when the main issue is the relative calibration of the baseline and the correction, especially once EXP is large enough to identify a stable correction direction.

Results on the $(\beta, n_{\mathrm{OBS}}, n_{\mathrm{EXP}})$ comparison. Relative to the matched single-linear baseline, Grounded lowers macro regret from 0.0357 to 0.0318, increases regret wins from 1 to 6, and has lower mean regret in 15 of the 24 $(\beta, n_{\mathrm{OBS}}, n_{\mathrm{EXP}})$ settings. The gains are concentrated in $(2000, 20)$ and $(20000, 100)$, where Grounded wins all six beta values, and in $(20000, 20)$, where it wins three of six.

The pooled-anchor variant behaves differently. It beats Grounded in 11 of the 24 settings and has the strongest average ranking metrics in this three-way comparison, but its macro regret is 0.0326 and it wins regret in only four settings, all with $n_{\mathrm{EXP}} = 100$. Those wins occur only when $n_{\mathrm{EXP}} = 100$:

$$(\beta, n_{\mathrm{OBS}}, n_{\mathrm{EXP}}) \in \{(0, 2000, 100),\ (0.5, 2000, 100),\ (0.8, 20000, 100),\ (0.9, 20000, 100)\}.$$

The selected pooling weight is usually small: across 720 fitted models, $\lambda_\star \leq 0.03$ in 565 fits and $\lambda_\star = 0$ in 264 fits. In this comparison, the pooled anchor changes calibration more than it changes the regret winner.

A.6 Value estimators and empirical benchmark metrics

The reward models in Section 4 estimate a conditional outcome surface. To obtain generative-model-level causal values, we combine those reward predictions with SIM. These DM / DR aggregators are used for the semi-synthetic generative-model values. They sit alongside the direct held-out-grid metrics in the real cached benchmark.

Let $\{x_i\}_{i=1}^{n_{\mathrm{eval}}}$ be evaluation contexts and let $m_{i,a,b} \sim p_{\mathrm{sim}}(\cdot \mid x_i, a)$ denote SIM-generated mediators. The default SIM plug-in estimator is

$$\widehat{\mu}^{\mathrm{DM}}(a) = \frac{1}{n_{\mathrm{eval}}} \sum_{i=1}^{n_{\mathrm{eval}}} \frac{1}{B} \sum_{b=1}^{B} \widehat{r}(x_i, m_{i,a,b}).$$

This simply averages predicted rewards over SIM-generated outputs.

A doubly robust variant augments the plug-in estimator with a residual term computed from EXP. Because EXP randomizes generative-model choice uniformly, $p_{\mathrm{EXP}}(a) = 1/|\mathcal{A}_{\mathrm{EXP}}|$. Define

$$\widehat{q}(x, a) = \frac{1}{B_{\mathrm{DR}}} \sum_{b=1}^{B_{\mathrm{DR}}} \widehat{r}\big(x, m_{a,b}\big), \qquad m_{a,b} \sim p_{\mathrm{sim}}(\cdot \mid x, a),$$

as the Monte Carlo approximation to the SIM-integrated reward. The estimator is

$$\widehat{\mu}^{\mathrm{DR}}(a) = \frac{1}{n_{\mathrm{EXP}}} \sum_{j=1}^{n_{\mathrm{EXP}}} \widehat{q}(X_j, a) + \frac{1}{n_{\mathrm{EXP}}} \sum_{j=1}^{n_{\mathrm{EXP}}} \frac{\mathbb{I}\{A_j = a\}}{p_{\mathrm{EXP}}(a)} \big(Y_j - \widehat{r}(X_j, M_j)\big).$$

The first term is the SIM-based plug-in estimate. The second term corrects it using the randomized residual observed in EXP. For reward models trained on EXP, we use cross-fitting: EXP is partitioned into K_{cf} folds, and the reward model used in the residual term is trained on the remaining $K_{\mathrm{cf}} - 1$ folds.

Role of DM / DR versus direct cached-benchmark metrics. In the appendix semi-synthetic validation, we center held-out recommendation regret computed against the generator's true interventional rewards. The same runs also produce model-level value diagnostics such as the agent-level RMSE of $\widehat{\mu}^{\mathrm{DM}}(a)$ or $\widehat{\mu}^{\mathrm{DR}}(a)$, which remain useful for diagnosing reward-surface fit. In the real cached benchmark, because the full held-out cached test grid is available, we additionally evaluate the fitted reward surface directly on that grid. Let $\mathcal{X}_{\mathrm{test}}$ be the held-out test contexts and $\mathcal{G}_{\mathrm{test}} = \mathcal{X}_{\mathrm{test}} \times \mathcal{A}$ the held-out cached grid. For each $(x, a) \in \mathcal{G}_{\mathrm{test}}$, let $q_{\mathrm{cache}}(x, a)$ be the judged scalar reward of the single cached held-out trajectory for context x and agent a, and let $\widehat{q}(x, a)$ be the fitted reward model evaluated on that cached trajectory. Define the benchmark-level agent aggregates

$$\bar{q}_{\mathrm{cache}}(a) = \frac{1}{|\mathcal{X}_{\mathrm{test}}|} \sum_{x \in \mathcal{X}_{\mathrm{test}}} q_{\mathrm{cache}}(x, a), \qquad \bar{q}_{\mathrm{pred}}(a) = \frac{1}{|\mathcal{X}_{\mathrm{test}}|} \sum_{x \in \mathcal{X}_{\mathrm{test}}} \widehat{q}(x, a).$$

The real-benchmark empirical metrics are then

$$\mathrm{RMSE}_{xa} = \left(\frac{1}{|\mathcal{G}_{\mathrm{test}}|} \sum_{(x,a) \in \mathcal{G}_{\mathrm{test}}} \big(\widehat{q}(x, a) - q_{\mathrm{cache}}(x, a)\big)^2 \right)^{1/2},$$

$$\mathrm{RMSE}_{\mathrm{agent}} = \left(\frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \left(\bar{q}_{\mathrm{pred}}(a) - \bar{q}_{\mathrm{cache}}(a) \right)^2 \right)^{1/2}, \qquad \widehat{\pi}(x) = \arg\max_{a \in \mathcal{A}} \widehat{q}(x, a),$$

$$\mathrm{Regret}_{\mathrm{test}} = \frac{1}{|\mathcal{X}_{\mathrm{test}}|} \sum_{x \in \mathcal{X}_{\mathrm{test}}} \left[\max_{a \in \mathcal{A}} q_{\mathrm{cache}}(x, a) - q_{\mathrm{cache}}(x, \widehat{\pi}(x)) \right].$$

These real-benchmark metrics summarize the finite held-out cached grid. The natural next extension is repeated cached replays / repeated judged trajectories per (x, a) so that the benchmark can estimate latent $q(x, a)$ more directly.

A.7 Empirical role of DM / DR on the controlled benchmark

The same semi-synthetic runs also let us compare DM and DR on model-level value estimation. These diagnostics answer a narrower question than the regret tables: when does randomized residual correction improve the SIM plug-in value estimate?

Method	DM RMSE	DR RMSE	Δ (DR−DM)
EXP-Only	0.0303	0.0432	+0.0128
OBS-Only	0.0268	0.0409	+0.0141
CVCI	0.0212	0.0399	+0.0187
Representation-EXP	0.0263	0.0424	+0.0161

Table 9: Average agent-level value RMSE across the synthetic $(\beta, n_{\mathrm{OBS}}, n_{\mathrm{EXP}})$ grid with $\beta \in \{0, 0.2, 0.5, 0.8, 0.9, 0.99\}$, $n_{\mathrm{OBS}} \in \{2{,}000, 20{,}000\}$, and $n_{\mathrm{EXP}} \in \{20, 100\}$. Positive Δ means that replacing DM by DR increases error.

Table 9 shows that, on this synthetic $(\beta, n_{\mathrm{OBS}}, n_{\mathrm{EXP}})$ grid, the SIM plug-in estimator has lower average agent-level value RMSE than DR for every representative method shown. Across these 24 settings, DR does not improve model-level value estimation on average.

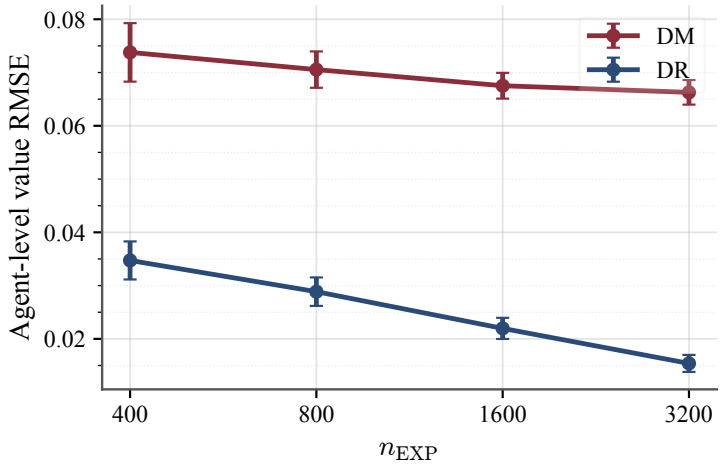

Figure 4: Targeted DM/DR diagnostic under strong self-selection, intentional reward-model underfit, and increasing EXP budget with fixed OBS budget. In this regime, DR materially reduces agent-level value RMSE.

Figure 4 shows a different case, with stronger self-selection, intentional reward-model underfit, lower outcome noise, and increasing EXP budget at fixed OBS budget. In that setting, DR lowers agent-level RMSE from 0.0738 to 0.0347 at $n_{\mathrm{EXP}} = 400$, and from 0.0663 to 0.0154 at $n_{\mathrm{EXP}} = 3200$. This is a narrower result about model-level value estimation; it does not change the main regret comparisons in the real summarization and coding benchmarks.

A.8 Empirical role of DM / DR on the real summarization benchmark

The real summarization benchmark also allows a DM/DR comparison for model-level value estimation. Here the target is the benchmark-level agent average over the held-out cached test contexts, and DR uses the known uniform EXP propensity together with cross-fitting for methods trained on EXP. This is a model-level value check built on top of the fitted reward model, not a replacement for the regret comparisons in the main text.

Method	n_{EXP}	DM RMSE	DR RMSE	Δ (DR−DM)
OBS-Only	100	0.0479	0.0459	-0.0021
OBS-Only	200	0.0479	0.0410	-0.0070
OBS-Only	800	0.0479	0.0310	-0.0169
CVCI	100	0.0479	0.0724	+0.0245
CVCI	200	0.0479	0.0572	+0.0093
CVCI	800	0.0479	0.0325	-0.0154
Representation-EXP	100	0.0226	0.0749	+0.0522
Representation-EXP	200	0.0187	0.0634	+0.0447
Representation-EXP	800	0.0190	0.0298	+0.0108
EXP-Only	100	0.1092	0.2495	+0.1403
EXP-Only	200	0.0809	0.1528	+0.0718
EXP-Only	800	0.0604	0.0798	+0.0194

Table 10: Budget-specific DM/DR agent-level value RMSE on the real summarization benchmark at $n_{\mathrm{OBS}} = 20{,}000$. Negative Δ means that replacing DM by DR reduces error.

Table 10 makes the heterogeneity explicit. OBS-Only improves in 23 of the 30 benchmark cells and becomes more favorable as n_{EXP} grows. CVCI changes sign only at $n_{\mathrm{EXP}} = 800$, where the residual term is large enough to help more than it hurts. The EXP-side methods behave differently: Representation-EXP worsens in 28 of 30 cells, and EXP-Only worsens in 27 of 30.

The natural interpretation is that DR is correcting the remaining OBS-side bias in model-level values rather than uniformly improving all fitted reward models. On this benchmark, that makes it most useful for the OBS-heavy baseline, conditionally useful for CVCI once the EXP budget is larger, and generally unhelpful for the EXP-side methods whose remaining error is not well summarized by a simple EXP residual correction.

A.9 Additional summarization diagnostics

Reward maps and routing. The summarization benchmark fixes one judged cached summary for each article–model pair and evaluates two reward maps on that same cache. Each cached summary carries a rubric vector $s(x,a) = \big(s_{\mathrm{faith}}(x,a), s_{\mathrm{cov}}(x,a), s_{\mathrm{clar}}(x,a), s_{\mathrm{conc}}(x,a)\big) \in [0,1]^4$ for faithfulness, coverage, clarity, and conciseness, together with an unsupported-claims count $u(x,a)$. Let $\mathcal{T}$ denote the four user segments and let $w_\tau \in \Delta^4$ be the segment-specific weight vector. For sharpening exponent $\gamma \geq 1$ and unsupported-claims penalty $\lambda \geq 0$, define

$$\bar{w}_{\tau k}^{(\gamma)} = \frac{w_{\tau k}^{\gamma}}{\sum_{\ell=1}^{4} w_{\tau \ell}^{\gamma}}, \qquad \ell_\tau^{(\gamma,\lambda)}(x,a) = \sum_{k=1}^{4} \bar{w}_{\tau k}^{(\gamma)} s_k(x,a) - \lambda \frac{u(x,a)}{20}.$$

The smooth map is

$$g_\tau^{\mathrm{sm}}(x,a) = \ell_\tau^{(1,0)}(x,a) = \langle w_\tau, s(x,a) \rangle, \qquad q_{\mathrm{cache}}^{\mathrm{sm}}(x,a) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} g_\tau^{\mathrm{sm}}(x,a).$$

With $\sigma(t) = (1+e^{-t})^{-1}$ and $[t]_{[0,1]} = \min\{1, \max\{0,t\}\}$, the sharpened map is

$$g_\tau^{\mathrm{sh}}(x,a) = \sigma\left(\frac{[\ell_\tau^{(16,0.1)}(x,a)]_{[0,1]} - 0.9}{0.02} \right), \qquad q_{\mathrm{cache}}^{\mathrm{sh}}(x,a) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} g_\tau^{\mathrm{sh}}(x,a).$$

OBS and EXP are sampled only from training contexts. User segments are drawn uniformly from $\mathcal{T}$, EXP randomizes uniformly over $\mathcal{A}$, and OBS uses the β-indexed router

$$p_{\mathrm{OBS}}^{\beta}(a \mid x, \tau) = (1-\varepsilon) \frac{\exp(\eta_a(x) + \beta\, \xi_a(\tau))}{\sum_{a' \in \mathcal{A}} \exp(\eta_{a'}(x) + \beta\, \xi_{a'}(\tau))} + \frac{\varepsilon}{|\mathcal{A}|},$$

Method	Avg. rank	Top-3	Excess (%)
CVCI	**2.04**	**43**	**4.8**
OBS-Only	2.83	32	7.9
EXP-Only	2.90	32	8.8
Grounded	3.21	26	7.9
CVCI-Residual	4.40	9	24.0
Representation-EXP	5.62	2	77.6

Table 11: Overall recommendation regret in the cached summarization benchmark. *Avg. rank* is the average rank across the 48 settings; *Top-3* is the count of settings in which a method ranked among the top three; *Excess (%)* is the mean percentage regret increase relative to the best method in each setting.

Average regret over the four budget settings			
Method	Smooth	Sharpened	Δ_{shape}
CVCI	0.0126	0.0423	$+0.0297$
OBS-Only	0.0125	0.0477	$+0.0352$
EXP-Only	0.0147	0.0376	$+0.0229$
Representation-EXP	0.0267	0.0773	$+0.0507$
Held-out target-fit R^2			
Quantity	Smooth	Sharpened	Δ
Proxy-bundle R^2	0.161	0.171	$+0.010$
Rubric-linear R^2	1.000	0.779	-0.221

Table 12: Recommendation regret and held-out target-fit R^2 under smooth versus sharpened reward maps, with cached outputs held fixed. Sharpening the reward inflates regret for every estimator family and drops rubric-linear R^2, while proxy-bundle R^2 stays low.

where $\eta_a(x) = \langle \theta_a^X, f_{\text{ctx}}(x) \rangle - \kappa c_a$ is an article-dependent logit and $\xi_a(\tau) = \langle \theta_a^U, e(\tau) \rangle$ is a segment-affinity logit. Thus $\beta = 0$ removes the latent-user term, and larger β makes routing more segment-specific.

In the cached summarization benchmark, CVCI has the best average rank, top-3 coverage, and excess regret across 30 seeds and 48 settings. The gap to the next family is small relative to the spread of cell-level standard errors reported in Appendix A.8, so this result reads as family-level evidence within this benchmark instantiation.

In Table 12, the two held-out R^2 values are out-of-sample linear fits of the cached target surface using the auxiliary proxy representation and the raw rubric vector $s(x, a)$, respectively. This table changes the target reward while holding the cached outputs fixed, so the regret increase and the drop in rubric-linear fit diagnose target alignment rather than a formal oracle condition.

Table 13 reports model-level value RMSE for DM and DR. In this benchmark, DR mainly helps OBS-heavy methods, while DM remains stronger for the other families. Appendix A.8 gives the budget-specific breakdown and implementation details.

Method	DM RMSE	DR RMSE	Δ (DR$-$DM)
OBS-Only	0.0479	**0.0393**	-0.0086
CVCI	0.0479	0.0540	$+0.0061$
Representation-EXP	**0.0201**	0.0560	$+0.0359$
EXP-Only	0.0835	0.1607	$+0.0772$

Table 13: Model-level value RMSE for DM and DR, averaged at $n_{\text{OBS}} = 20{,}000$ and $n_{\text{EXP}} \in \{100, 200, 800\}$ over 30 seeds. Negative Δ means DR helps over DM; positive Δ means DR hurts.

A.10 Exact real benchmark protocol

Dataset and candidate generative models. Contexts X are CNN/DailyMail articles [Hermann et al., 2015, See et al., 2017]. The candidate set contains 20 summarization systems: 16 prompt-based abstractive systems built from two Gemma 3 instruction models [Gemma Team, 2025], together with four extractive baselines. For each context–generative-model pair, the cache stores one summary together with its judged metadata.

Judged rewards and proxy features. For each cached output, the judge returns rubric-style signals that are combined into the scalar reward used throughout the real benchmark comparisons. The real benchmark instead uses a heuristic auxiliary-feature channel derived from cached trajectories and judged-output metadata. That channel is best read as an OBS proxy, distinct from the hidden representation $\phi^\star$ used in the semi-synthetic study.

Fixed benchmark protocol. The real-data study fixes a 48-context subset of relatively difficult CNN/DailyMail articles, selected once by an article-level difficulty score. Within this fixed pool, each seed performs an $80/20$ context-level train/test split, uses no separate context-level validation split, and tunes hyperparameters only on a held-out fraction of EXP when the estimator requires it. OBS and EXP are sampled only from the training contexts. The observational-choice protocol is a fixed paired-sampling design under a self-selection mechanism that routes each decision through a five-system candidate slate before the final model choice.

Budget views and aggregation. The real summarization study uses three budget views. The main-text regret comparisons use

$$\beta \in \{0, 0.2, 0.5, 0.8, 0.9, 0.99\}, \qquad n_{\mathrm{OBS}} \in \{2{,}000, 20{,}000\}, \qquad n_{\mathrm{EXP}} \in \{20, 200\},$$

with matched smooth and sharpened reward maps, giving the Section 5.4 summaries over six β values, two OBS budgets, two EXP budgets, and two reward maps. The DM/DR value diagnostics in Table 10 instead fix $n_{\mathrm{OBS}} = 20{,}000$ and use $n_{\mathrm{EXP}} \in \{100, 200, 800\}$. The winner-map table below is a separate smooth-reward $(\beta, n_{\mathrm{OBS}}, n_{\mathrm{EXP}})$ comparison with

$$\beta \in \{0, 0.2, 0.5, 0.8, 0.9, 0.99\}, \qquad n_{\mathrm{OBS}} \in \{2{,}000, 20{,}000\}, \qquad n_{\mathrm{EXP}} \in \{20, 100\},$$

reported to show the full 24-cell winner map. All reported means and standard errors are recomputed from seed-level paired runs, with 30 seeds in every reported $(\beta, n_{\mathrm{OBS}}, n_{\mathrm{EXP}}, \text{method})$ cell.

A.11 Tuning on EXP: model-level cross-validation, EXP holdout, and defaults

We use three EXP-side tuning strategies. Model-level EXP cross-validation partitions the generative models *appearing in the sampled EXP data for that seed/cell* and validates at the held-out model level over that sampled EXP model set. EXP holdout reserves a validation subset from the sampled n_{EXP} budget itself and tunes on that holdout. Fixed defaults use dataset-configured settings with no extra EXP hyperparameter sweep.

More precisely, n_{EXP} is the total EXP budget for a given seed/cell. If a method uses an explicit EXP holdout, that holdout is carved from the sampled EXP budget itself; no extra EXP rows are sampled outside budget, and no separate context-level validation split is required for this.

For Grounded-, CVCI-, and CVCI-Residual-style families tuned by model-level EXP cross-validation, let $\mathcal{A}_{\mathrm{EXP}}$ be the set of generative models appearing in the sampled EXP data and partition it into K_{cv} folds $\{\mathcal{A}^{(k)}\}_{k=1}^{K_{\mathrm{cv}}}$. The corresponding validation indices are

$$\mathcal{I}_{\mathrm{val}}^{(k)} = \{j \in \mathrm{EXP} : A_j \in \mathcal{A}^{(k)}\}, \qquad \mathcal{I}_{\mathrm{tr}}^{(k)} = \mathrm{EXP} \setminus \mathcal{I}_{\mathrm{val}}^{(k)}.$$

For a candidate tuning parameter η (for example, $\eta = \alpha_{\mathrm{corr}}$ for Grounded or $\eta = \lambda$ for CVCI), fit the corresponding estimator using all OBS data together with the EXP indices in $\mathcal{I}_{\mathrm{tr}}^{(k)}$. Then evaluate held-out generative-model means with

$$\mathcal{L}^{(k)}(\eta) = \sum_{a \in \mathcal{A}^{(k)}} \omega_{k,a} \left(\overline{\widehat{r}}_{\eta}^{(k)}(a) - \overline{Y}^{(k)}(a) \right)^2, \tag{12}$$

where $\overline{\widehat{r}}_\eta^{(k)}(a) = \frac{1}{n_{k,a}} \sum_{j \in \mathcal{I}_{\text{val}}^{(k)}:\, A_j = a} \widehat{r}_\eta^{(-k)}(z_j)$, $\overline{Y}^{(k)}(a) = \frac{1}{n_{k,a}} \sum_{j \in \mathcal{I}_{\text{val}}^{(k)}:\, A_j = a} Y_j$, and $n_{k,a}$ is the number of held-out EXP examples for generative model a in fold k. We use count weights $\omega_{k,a} \propto n_{k,a}$, normalized so that $\sum_{a \in \mathcal{A}^{(k)}} \omega_{k,a} = 1$, and select the parameter that minimizes the average fold loss. If too few generative models appear in the sampled EXP data to sustain the requested number of model-level folds, the procedure falls back to sample-level cross-validation and records the effective mode.

In the real main sweep, Grounded, CVCI, and CVCI-Residual use model-level EXP cross-validation. The simpler historical linear grounded baseline instead uses an EXP holdout through correction-α tuning on the held-out EXP subset. EXP-Only, OBS-Only, and Representation-EXP use fixed defaults.

When candidate hyperparameters are tied within numerical tolerance, the implementation resolves ties toward the more regularized / less aggressive solution. For residual CVCI-Residual this means preferring larger residual ridge penalties and, if still tied, larger OBS pooling weights. For holdout-tuned grounded corrections, ties in the correction-α grid are resolved toward smaller shrinkage factors.

Because the `news_hard top-48` real benchmark contains only 48 contexts, each $80/20$ context split leaves roughly 10 held-out test contexts per seed. Single-seed regret is therefore noisy, and all real-benchmark conclusions are aggregated across many seeds.

A.12 Benchmark setup details

Each benchmark setting specifies a training-context distribution, a held-out test set, an action set, a simulator, an experimental reward map, and an observational logging family. Within one such cell, the observational law factors as

$$p_{\text{OBS}}(x, u, a, m, y) = P_{\text{tr}}(x)\, P(u \mid x)\, p_{\text{OBS}}^{\beta}(a \mid x, u)\, p_{\text{sim}}(m \mid x, a)\, P_Y(y \mid x, m, u),$$

and the experimental law replaces the choice model by an unconfounded randomization rule $\pi_{\text{EXP}}(a \mid x)$:

$$p_{\text{EXP}}(x, u, a, m, y) = P_{\text{tr}}(x)\, P(u \mid x)\, \pi_{\text{EXP}}(a \mid x)\, p_{\text{sim}}(m \mid x, a)\, P_Y(y \mid x, m, u).$$

Here P_Y is the benchmark-specific outcome law. In the main-text benchmark settings, $\pi_{\text{EXP}}(a \mid x) = 1/|\mathcal{A}|$, and $r^\star(x, m) = \mathbb{E}[Y \mid X = x, M = m, \text{EXP}]$ integrates out $U \mid X = x$ under p_{EXP}. The identified target remains

$$q(x, a) = \mathbb{E}_{M \sim p_{\text{sim}}(\cdot \mid x, a)}[r^\star(x, M)].$$

Each estimator returns $\widehat{q}$, inducing $\widehat{\pi}(x) = \arg\max_{a \in \mathcal{A}} \widehat{q}(x, a)$, and the main metric is held-out recommendation regret. The β-indexed observational choice law is benchmark-specific. In the appendix controlled validation, $p_{\text{OBS}}^{\beta}(a \mid x, u) = (1 - \beta)\pi_X(a \mid x) + \beta\, \pi_U(a \mid u)$, where π_X is a context-driven policy and π_U is a latent-user policy. Thus $\beta = 0$ removes latent-user routing, whereas $\beta = 1$ uses only the latent-user component. Section 5.4 instead uses a softmax router in which β scales the latent-user logit contribution.

The real cached benchmarks fix one realized mediator $m_{\text{cache}}(x, a)$ for each context-action pair and report the finite cached target $q_{\text{cache}}(x, a) = r^\star(x, m_{\text{cache}}(x, a))$. Their reported recommendation regret therefore replaces q by q_{cache} on the held-out cache. Randomness then comes from OBS/EXP sampling and the benchmark's latent draws, not from regenerating mediators at evaluation time.

A.13 Semi-synthetic controlled validation

The semi-synthetic validation keeps the summarization task family but replaces judged rewards by a known latent generator, so held-out regret can be evaluated directly against the true interventional surface. Each context–mediator pair is mapped to an unobserved representation

$$\phi^\star(x, m) = H\varphi(x, m) \in \mathbb{R}^{d_\phi},$$

where H is a fixed signed-hash projection of sparse hashed text features $\varphi(X, M) \in \mathbb{R}^{2^{18}}$. The benchmark-specific outcome law is

$$P_Y(\cdot \mid x, m, u) = \mathcal{L}\big(\text{sigmoid}\big((w^\star)^\top \phi^\star(x, m) + \lambda^\top u + \varepsilon\big)\big), \qquad \varepsilon \sim \mathcal{N}(0, \sigma_Y^2),$$

Table 14: Benchmarks used in the main text. The two cached benchmarks evaluate finite cached targets q_{cache} rather than the latent superpopulation target q; their OBS and EXP samples are constructed by resampling and reweighting the cached pool under specified routing and randomization rules.

Benchmark	Source data	Cached object	Evaluation target
Controlled semi-synthetic	CNN/DailyMail article family with a known latent reward generator	SIM-generated summaries from the generator pool	True interventional reward surface q (theorem-level target).
Real-task cached summarization	48 difficult CNN/DailyMail articles; 20 candidate models	One judged summary per article and model	Finite cached target $q_{\mathrm{cache}}^{\mathrm{sm}}$ (or $q_{\mathrm{cache}}^{\mathrm{sh}}$): segment-average rubric score over four user types under the smooth (or sharpened) reward map.
Real-task cached coding	SWE-bench Verified issues with a BouncerBench patch pool	One cached candidate patch per issue and agent	Finite cached target q_{cache}: user-average utility over fix success (c_1) and patch-quality components (c_2, c_3, c_4).

SO

$$r^\star(x,m) = \mathbb{E}[Y \mid X=x, M=m, \mathrm{EXP}] = \mathbb{E}_{U\sim P(\cdot|x),\,\varepsilon}\Big[\,\mathrm{sigmoid}\big((w^\star)^\top \phi^\star(x,m) + \lambda^\top U + \varepsilon\big)\Big].$$

OBS and EXP share this same P_Y, so the only confounding channel is model choice through $p^{\beta}_{\mathrm{OBS}}(a \mid x,u)$, while EXP randomizes uniformly over the candidate agent set. This appendix validation varies only β, n_{OBS}, and n_{EXP}, so changes in the regret winner can be attributed to confounding strength and data budget.

Implementation details. The semi-synthetic generator uses sparse hashed text features $\varphi(X,M) \in \mathbb{R}^{2^{18}}$ for all reward models. The hidden representation has dimension $d_\phi = 8$ and is constructed by a seeded signed-hash projection; it is fixed across replicates and is not observed by any estimator. When hybrid estimators use the proxy map $\psi(z)$ from Section 4, the PCA dimension is $d_\psi = 20$. The context-only embedding $\psi_X(X)$ that appears in the observational choice model is a separate object. It is fit on raw article text alone, with no context–output pairs. Operationally, we sample 5,000 contexts from the observational split, fit a TF–IDF vectorizer with English stop-word removal, unigram–bigram features, and a vocabulary cap of 50,000 terms, and then apply rank-50 truncated SVD. Thus $\psi_X(X) \in \mathbb{R}^{50}$ is a low-dimensional context representation used only for model choice and for the $U \mid X$ model in the semi-synthetic generator. This is distinct from the proxy representation $\psi(X,M)$ in Section 4, which is learned from OBS auxiliary labels and is used by the hybrid reward estimators. Reference values and held-out regret metrics are computed on $n_{\mathrm{true}} = 100$ and $n_{\mathrm{eval}} = 100$ held-out contexts respectively, using $B = 5$ simulator draws per context–agent pair.

Fixed protocol in the appendix controlled validation. The appendix semi-synthetic validation fixes one confounding protocol and varies only the self-selection weight β, the observational budget n_{OBS}, and the experimental budget n_{EXP}. EXP randomizes uniformly over the candidate agent set. OBS mixes a context-only softmax policy with a latent-user shortlist policy, where β is the probability of drawing from the confounded latent-user component. This validation uses

$$\beta \in \{0, 0.2, 0.5, 0.8, 0.9, 0.99\}, \qquad n_{\mathrm{OBS}} \in \{2{,}000, 20{,}000\}, \qquad n_{\mathrm{EXP}} \in \{20, 100\},$$

and reports 30 paired seeds in every available cell. The direct latent effect on the reward score and the auxiliary-label noise level are both fixed throughout this validation grid.

Budget-axis repetition for pure baselines. OBS-Only ignores EXP labels, so its reported value is constant across the two EXP columns for a fixed $(\beta, n_{\mathrm{OBS}})$ pair. EXP-Only ignores OBS outcomes, so its reported value is constant across the two OBS rows for a fixed $(\beta, n_{\mathrm{EXP}})$ pair. This repetition is carried through consistently in the synthetic regime table and summary table.

A.14 Appendix results for the summarization benchmark

Synthetic summarization results. Table 15 shows the regret winner in each synthetic summarization setting over six β values and four $(n_{\mathrm{OBS}}, n_{\mathrm{EXP}})$ budget settings.

The smallest-budget row $(2{,}000, 20)$ is the most mixed: CVCI-Residual wins once, CVCI wins twice, OBS-Only wins once, and Grounded wins twice. At $(20{,}000, 20)$, CVCI wins all six beta values. At

Budget cell	$\beta = 0$	$\beta = 0.2$	$\beta = 0.5$	$\beta = 0.8$	$\beta = 0.9$	$\beta = 0.99$
$n_{\mathrm{OBS}} = 2{,}000$, $n_{\mathrm{EXP}} = 20$	CVCI-Residual +0.0004	CVCI +0.0002	CVCI +0.0009	OBS-Only +0.0004	Grounded +0.0010	Grounded +0.0004
$n_{\mathrm{OBS}} = 20{,}000$, $n_{\mathrm{EXP}} = 20$	CVCI +0.0005	CVCI +0.0015	CVCI +0.0015	CVCI +0.0021	CVCI +0.0018	CVCI +0.0017
$n_{\mathrm{OBS}} = 2{,}000$, $n_{\mathrm{EXP}} = 100$	CVCI +0.0008	CVCI +0.0017	CVCI +0.0012	OBS-Only +0.0006	OBS-Only +0.0008	OBS-Only +0.0005
$n_{\mathrm{OBS}} = 20{,}000$, $n_{\mathrm{EXP}} = 100$	CVCI +0.0002	EXP-Only +0.0012	EXP-Only +0.0016	EXP-Only +0.0013	EXP-Only +0.0012	EXP-Only +0.0016

Table 15: Regret-optimal estimator on the semi-synthetic summarization benchmark. The number beneath each winner is the regret gap to the runner-up.

Budget cell	$\beta = 0$	$\beta = 0.2$	$\beta = 0.5$	$\beta = 0.8$	$\beta = 0.9$	$\beta = 0.99$
$n_{\mathrm{OBS}} = 2{,}000$, $n_{\mathrm{EXP}} = 20$	Grounded +0.0002	Grounded +0.0011	Grounded +0.0036	OBS-Only +0.0000	OBS-Only +0.0008	Grounded +0.0006
$n_{\mathrm{OBS}} = 20{,}000$, $n_{\mathrm{EXP}} = 20$	Grounded +0.0016	Grounded +0.0049	EXP-Only +0.0013	EXP-Only +0.0062	EXP-Only +0.0019	EXP-Only +0.0019
$n_{\mathrm{OBS}} = 2{,}000$, $n_{\mathrm{EXP}} = 100$	OBS-Only +0.0007	OBS-Only +0.0002	OBS-Only +0.0002	CVCI +0.0000	CVCI +0.0000	CVCI +0.0008
$n_{\mathrm{OBS}} = 20{,}000$, $n_{\mathrm{EXP}} = 100$	CVCI-Residual +0.0048	CVCI +0.0009	CVCI +0.0052	CVCI +0.0012	CVCI +0.0013	CVCI +0.0046

Table 16: Regret-optimal estimator on the real summarization benchmark. The number beneath each winner is the regret gap to the runner-up.

$(2{,}000, 100)$, the winner shifts from CVCI at lower β to OBS-Only at higher β. At $(20{,}000, 100)$, EXP-Only wins all but the $\beta = 0$ cell.

Across the 24 settings, CVCI is the most stable overall winner, but no single method wins everywhere. Most regret gaps are small, so Table 15 is best read as a map of where the winner changes.

Additional synthetic aggregate interpretation. Across the same 24 settings, CVCI is also the aggregate regret leader. EXP-Only and OBS-Only form the next tier, while Grounded and CVCI-Residual win only a few cells. The ranking diagnostics are flatter than regret: no single method leads top-1 or top-3 across the grid.

Real summarization results on the smooth auxiliary comparison. Table 16 shows the regret winner in each real summarization setting for the auxiliary smooth-reward comparison over six β values and four $(n_{\mathrm{OBS}}, n_{\mathrm{EXP}})$ budget settings.

At $(2{,}000, 20)$, Grounded wins four beta values and OBS-Only wins two; the $\beta = 0.8$ cell is essentially a tie. At $(20{,}000, 20)$, Grounded wins at low β, while EXP-Only wins at higher β. Once $n_{\mathrm{EXP}} = 100$, the high-OBS row $(20{,}000, 100)$ is dominated by CVCI, with a single CVCI-Residual win at $\beta = 0$, while the low-OBS row $(2{,}000, 100)$ shifts from OBS-Only at lower β to CVCI at higher β.

Additional real ranking–regret comparison. Across these 24 settings, CVCI is the aggregate regret leader and Grounded is second. Ranking tells a different story: Grounded wins top-1 and top-3 in all 24 settings but wins regret in only 6.

A.15 Public cached coding benchmark

Fixed protocol. The coding benchmark uses a public cached grid of issue descriptions, patch mediators, and execution-derived rewards. The coding comparison in Section 5.4 fixes

$$\beta = 0.5, \qquad n_{\mathrm{OBS}} = 2{,}000, \qquad n_{\mathrm{EXP}} = 100,$$

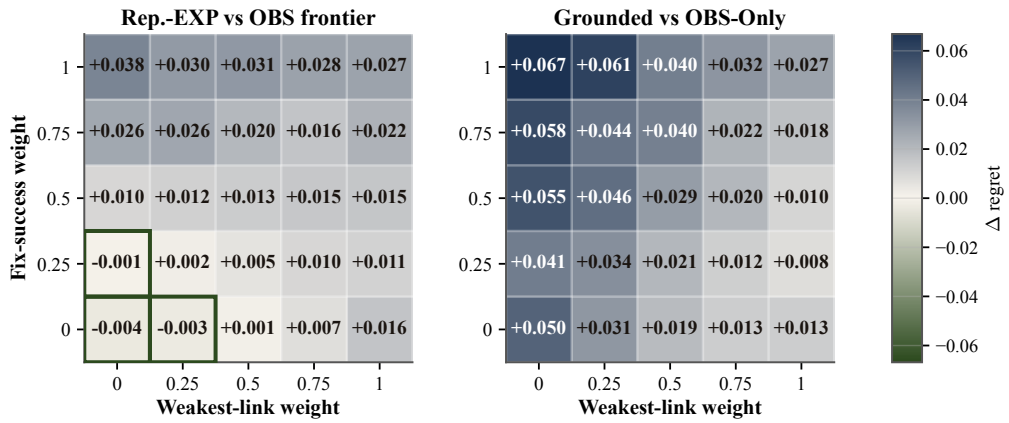

Figure 5: Pairwise regret differences $\Delta\,\text{regret} = \text{regret}(\text{family A}) - \text{regret}(\text{family B})$ across the 5×5 grid of $(\alpha_{\text{fix}}, \omega_{\text{weak}})$ in the cached coding benchmark instantiation; negative values mean family A has lower regret than family B in that cell. Left panel: Representation-EXP minus the best OBS-based family. Right panel: Grounded minus OBS-Only.

uses outcome-precision 64, and evaluates only the feature-based implementations of the estimator families. It then varies two reward-design parameters:

$$\text{fix-success weight} \in \{0, 0.25, 0.5, 0.75, 1\}, \qquad \text{weakest-link weight} \in \{0, 0.25, 0.5, 0.75, 1\},$$

with 10 paired seeds in every cell. The first parameter rescales the fix-success component of the reward, while the second moves the non-fix patch-quality score from additive aggregation toward a weakest-link rule.

Reward construction and target-alignment diagnostics. Each cached patch is scored by four components c_1, c_2, c_3, c_4, where c_1 records true fix success and c_2, c_3, c_4 summarize patch size, locality, and issue focus. These patch-quality components can disagree with true fix success, which is why changing the reward definition matters here.

For user type u with weights $w_u = (w_{u1}, w_{u2}, w_{u3}, w_{u4})$, normalized non-success weights are $\bar{w}_{uk} = w_{uk}/(w_{u2} + w_{u3} + w_{u4})$ for $k \in \{2, 3, 4\}$, and

$$g_u(c_2, c_3, c_4) = (1 - \omega_{\text{weak}}) \sum_{k=2}^{4} \bar{w}_{uk} c_k + \omega_{\text{weak}} \min\{c_2, c_3, c_4\}.$$

The benchmark utility for user type u is

$$Y_u(x, a) = w_{u1}\, \alpha_{\text{fix}}\, c_1(x, a) + (1 - w_{u1})\, g_u(c_2(x, a), c_3(x, a), c_4(x, a)),$$

and the evaluation target is the cached user-average surface $q_{\text{cache}}(x, a) = \mathbb{E}[Y_U(x, a) \mid X = x]$. Representation-EXP maps each issue–patch pair into an auxiliary representation learned from observational patch statistics and fits EXP rewards on that representation.

With the data budget fixed, rankings change because the reward puts more or less weight on true fix success. At $\omega_{\text{weak}} = 0.25$, the regret gap between Representation-EXP and the better of OBS-Only and CVCI moves from -0.0034 at $\alpha_{\text{fix}} = 0$, to $+0.0024$ at $\alpha_{\text{fix}} = 0.25$, to $+0.0375$ at $\alpha_{\text{fix}} = 1$. This sweep tests whether auxiliary patch statistics retain the target signal as the target moves from patch quality toward true fix success.

Held-out R^2 comes from linear fits of the cached target using either auxiliary patch statistics or c_1 alone. On the additive slice, the auxiliary-feature fit drops from $R^2 = 0.347$ at $\alpha_{\text{fix}} = 0$ to 0.081 at $\alpha_{\text{fix}} = 1$, while the c_1-only fit rises from essentially zero to $R^2 = 0.967$. The R^2 shift matches the change in method ranking: Representation-EXP helps when the target follows patch-quality features, but it loses that advantage once the reward is driven mainly by fix success.

Changing ω_{weak} changes only how the non-success components are combined. Its largest effect is on Grounded: at zero fix-success weight, the grounded-minus-OBS regret gap shrinks from about

Fix-success weight\weakest-link weight	0	0.25	0.5	0.75	1.0
0.00	Representation-EXP +0.0039	Representation-EXP +0.0033	CVCI-Residual +0.0001	OBS-Only +0.0001	OBS-Only +0.0002
0.25	Representation-EXP +0.0006	OBS-Only +0.0004	CVCI +0.0004	CVCI +0.0013	CVCI +0.0009
0.50	OBS-Only +0.0004	OBS-Only +0.0010	CVCI +0.0003	CVCI +0.0008	CVCI +0.0004
0.75	CVCI +0.0013	OBS-Only +0.0029	OBS-Only +0.0004	OBS-Only +0.0003	OBS-Only +0.0013
1.00	OBS-Only +0.0011	CVCI +0.0007	CVCI +0.0001	OBS-Only +0.0068	OBS-Only +0.0056

Table 17: Regret-optimal estimator on the coding benchmark at fixed $n_{\mathrm{OBS}} = 2{,}000$, $n_{\mathrm{EXP}} = 100$, $\beta = 0.5$, and reward-noise precision 64. Columns give the weakest-link weight in the non-fix patch-quality score, and rows give the weight on true fix success. Most top-two gaps are small, so the table mainly shows where the regret winner changes.

$+0.0495$ under additive scoring to about $+0.0126$ under the pure weakest-link rule. Figure 5 (right) shows the same effect. Even so, Grounded remains worse than the best OBS-based method throughout this fixed-budget setting.

Appendix coding results. In this coding appendix comparison, we treat a winner change as substantive when the top-two regret gap exceeds 3×10^{-3}; smaller gaps are near-ties. Table 17 reports the regret winner in each cell. Every top-two gap is below 0.01, and 20 of the 25 cells are below 1.5×10^{-3}. Most flips between OBS-Only and CVCI therefore reflect near-ties in the high-fix-success part of the grid. The clearest differences appear at low fix-success weight under additive scoring, where Representation-EXP outperforms the OBS-based methods, and in the steady narrowing of the grounded-versus-OBS gap as weakest-link scoring becomes stronger.

Secondary axes omitted from the main text. Additional coding comparisons over confounding strength and outcome noise were qualitatively weaker. They changed absolute difficulty, but they did not produce ranking shifts as clear as those induced by the fix-success weight and weakest-link weight in this fixed-budget probe. For that reason, the main text centers this reward-design comparison and leaves the other axes to this appendix discussion.

A.16 Extended positioning, scope, and limitations

Off-policy evaluation. Direct Method with Replay (DM) and Doubly Robust (DR) are standard estimators in contextual-bandit and policy-evaluation settings [Dudík et al., 2011, 2014, Swaminathan and Joachims, 2015, Jiang and Li, 2016, Thomas and Brunskill, 2016], and they descend from the foundational doubly robust and missing-data lineage of Robins et al. [1994] and Bang and Robins [2005]. The OPE setting these papers analyze starts from logged action–outcome pairs in which outcomes are assumed unconfounded given context: the central technical issue is action coverage under the logging policy, and the recorded outcomes faithfully represent the causal target. In our setting, even for observed actions, the recorded outcomes are confounded because user-side factors influence both who selects which model and how its output is judged. Because SIM can regenerate outputs for any model on any held-out context, DM and DR enter only after the reward surface itself has been estimated. We use SIM and EXP to identify the causal scoring rule and then use OBS to reduce variance.

Experimental grounding and hybrid estimation. The closest methodological antecedents are experimental grounding [Kallus et al., 2018], data-fusion and transportability formalisms [Pearl and Bareinboim, 2011, Bareinboim and Pearl, 2016], and trial-to-target generalizability [Cole and Stuart, 2010]. A related line of work studies how to combine large biased samples with smaller randomized ones [Rosenman et al., 2023, Cheng and Cai, 2021, Lin et al., 2025, Yang et al., 2025, Colnet et al., 2024]. That literature asks when observational signal can be used to reduce variance without giving up the causal credibility of the randomized sample; in those papers, the unit-level outcome is observed on both samples and the modeling work is done on a fixed outcome label. The estimator families in Section 4 are concrete instantiations of the same bias–variance question. Our setting differs in one specific way: the target outcome is not directly observed even on the randomized sample because the

unit "output" must first be regenerated by SIM and then scored. Identification in our setting depends on SIM validity (Assumption A2) together with EXP randomization.

Summarization and LLM-based evaluation. Our experiments use CNN/DailyMail contexts [Hermann et al., 2015, Nallapati et al., 2016, See et al., 2017, Stiennon et al., 2020] and rely on LLM-judged rubric scores in the pool-based study. Recent LLM-evaluation work covers benchmark design [Liang et al., 2023], judge reliability and pairwise preference platforms [Zheng et al., 2023, Chiang et al., 2024], and real-world usage logs [Zheng et al., 2024, Zhao et al., 2024]. It also documents biases in automatic evaluators, including positional unfairness [Wang et al., 2023], length bias [Dubois et al., 2024], and self-preference [Panickssery et al., 2024], and studies panel-based mitigations [Verga et al., 2024]. Preference data is central in alignment pipelines, both with explicit learned reward models [Christiano et al., 2017, Stiennon et al., 2020, Ouyang et al., 2022] and with implicit reward formulations such as direct preference optimization [Rafailov et al., 2023]. The LLM-evaluation work cited above primarily treats judged scores as outcome data and studies how reliable those scores are and how to average them. We use cached outputs and judged scores as inputs to a causal evaluation problem; they become the causal target only after EXP-identified scoring.

A complementary line of work in causal NLP studies how text can serve as treatment, outcome, mediator, or proxy for unobserved confounders [Feder et al., 2022]; OBS-derived auxiliary labels in our setting play exactly the proxy role identified there. Taken together, the setting sits across the three literatures rather than fitting cleanly inside any one of them.

Scope and future work. Ranking diagnostics can diverge sharply from recommendation regret. In the cached summarization benchmark, Grounded wins every top-1 and top-3 ranking setting but wins regret in only 6 settings. Regret is driven by a small number of costly recommendation mistakes that average ranking metrics do not isolate. The benchmarks therefore center $\mathrm{Regret}_{\mathrm{test}}$, with context-action RMSE and model-level RMSE as secondary diagnostics.

The two real-task benchmarks use real candidate model outputs and real LLM-judged rubric or program-test scores on those outputs. Their OBS and EXP mechanisms are benchmark constructions over the cached pool. The empirical comparisons therefore diagnose how the estimator families behave under controlled synthetic OBS/EXP resampling on real tasks. The empirical scope is deliberately narrow: cached summarization isolates supervision scarcity, and cached coding isolates reward geometry. Family-level wins in these benchmarks are benchmark-specific diagnostics.

Estimator choice depends on what limits performance in the benchmark at hand. In practice, the key checks are whether more EXP labels materially improve EXP-based methods and whether a structured representation predicts the held-out target better than a simple success-only baseline. Deployment decisions should still be confirmed by live randomized evaluation. These conclusions are specific to the task families and benchmark constructions studied here. Future work includes broader tasks, repeated mediator draws, naturally observed deployment logs paired with in-deployment randomized trials, and multi-round human–agent interaction, which would require relaxing or replacing Assumption A4.

A.17 Implementation notes

The synthetic validation uses only cached data generation; no external API calls occur after the generator pool is constructed. DR uses EXP propensities from uniform randomization together with cross-fitting for reward models trained on EXP. Additional synthetic diagnostics and older experimental variants are omitted from the present main-text comparisons.

B Rubric subscores for CNN/DailyMail summaries

Given an article X and candidate summary M, the judge outputs four integer subscores $\tilde{s}(X, M) \in \{0, 1, 2, 3, 4, 5\}^4$: *faithfulness*, *coverage*, *clarity*, and *conciseness*. Scores use the anchors below; intermediate values interpolate between adjacent anchors.

Faithfulness (0–5). Factual consistency with the article. Unsupported claims include invented entities, numbers, events, causal attributions, or quotes.

- **5:** No unsupported claims; all salient statements are grounded in X.
- **4:** Minor unsupported detail(s) that do not affect the main facts.
- **3:** Multiple unsupported or weakly supported statements; main story remains recognizable.
- **2:** Several unsupported claims affecting key details (e.g., actor/action/outcome/number).
- **1:** Many unsupported claims; substantial contradiction or fabrication.
- **0:** Predominantly hallucinatory or contradictory relative to X.

Coverage (0–5). Inclusion of the article's major points (headline facts, primary actors, core events, outcomes, and key qualifiers).

- **5:** Covers all major points; omissions are limited to minor details.
- **4:** Covers most major points; at most one major point missing.
- **3:** Covers some major points; several major omissions.
- **2:** Covers few major points; summary reflects only a small slice of the article.
- **1:** Minimal coverage; largely misses the article's main content.
- **0:** Misses most major points or is largely off-topic relative to X.

Clarity (0–5). Readability and organization (coherence, grammaticality, referential clarity, and logical flow).

- **5:** Clear and well-structured; unambiguous references and fluent phrasing.
- **4:** Generally clear; minor awkwardness or minor ambiguity.
- **3:** Understandable with effort; noticeable disfluency, repetition, or unclear references.
- **2:** Hard to follow; frequent grammatical issues or unclear structure.
- **1:** Mostly unclear; major coherence failures.
- **0:** Unreadable or incoherent.

Conciseness (0–5). Information density at an appropriate length; redundancy and irrelevant detail reduce the score; excessive brevity that omits necessary content also reduces the score.

- **5:** Efficient summary with minimal redundancy and no filler.
- **4:** Slight redundancy or mild over/under-length without major impact.
- **3:** Noticeable redundancy or length mismatch; still usable as a summary.
- **2:** Clearly too verbose or too terse; substantial inefficiency or truncation.
- **1:** Extremely verbose or extremely short; poor summary form.
- **0:** Pathological length or pervasive redundancy; not a usable summary.

C Full statements and proofs for identification and population estimator comparisons

This appendix collects the proof of the main identification result and the full statements and proofs for the population estimator comparisons referenced in Section 4. It also explains how these approximation-class results motivate the empirical target-alignment diagnostics used in the benchmarks; those diagnostics are not additional formal results. Throughout, $Z = (X, M)$, P_E denotes the EXP law of Z, and for a fixed target outcome Y we write

$$\mathcal{R}_E(f) = \sigma_E^2 + \|f - r^\star\|_E^2, \qquad \|g\|_E^2 := \mathbb{E}_{Z\sim P_E}[g(Z)^2], \qquad \sigma_E^2 := \mathbb{E}\big[(Y - r^\star(Z))^2 \mid \mathrm{EXP}\big], \tag{13}$$

where $r^\star(z) := \mathbb{E}[Y \mid Z = z, \mathrm{EXP}]$. For readability, we ignore the final clipping to $[0, 1]$ and absorb the intercept into both ψ and φ by augmenting them with a constant feature.

C.1 Proof of Theorem 1

Proof of Theorem 1. By Assumption A1, $A \perp U \mid X$ in EXP. Assumption A4 states the structural restriction $U \perp M \mid X, A$ in OBS, in EXP, and under $\mathrm{do}(A = a)$, so that conditional on (X, A) the mediator M carries no residual information about U. Combining A1 and A4 with the chain rule for conditional independence gives

$$U \perp (A, M) \mid X \qquad \text{in EXP},$$

so on the shared EXP/SIM support the scoring rule $r^\star(x,m) = \mathbb{E}[Y \mid X = x, M = m, \mathrm{EXP}]$ is identified from EXP and is free of self-selection bias. Assumption A2 gives $p_{\mathrm{sim}}(m \mid x, a) = \mathbb{P}(M = m \mid X = x, \mathrm{do}(A = a))$, so for any x in the EXP support,

$$\begin{aligned} q(x,a) &= \mathbb{E}[Y(\mathrm{do}(A=a)) \mid X = x] \\ &= \mathbb{E}_{M \sim p_{\mathrm{sim}}(\cdot \mid x,a)}[\mathbb{E}[Y \mid X = x, M = M, \mathrm{do}(A = a)]] \\ &= \mathbb{E}_{M \sim p_{\mathrm{sim}}(\cdot \mid x,a)}[r^\star(x,M)]\,, \end{aligned}$$

where the last equality uses Assumption A3 (the conditional outcome law given (X, M, U) is the same in OBS, EXP, and under $\mathrm{do}(A = a)$, and $r^\star$ integrates U out under the EXP $U \mid X$ distribution, which by A1 equals the post-intervention $U \mid X$ distribution). This proves (1). Averaging $q(X, a)$ under the EXP context distribution and invoking Assumption A5 ($P_X^{\mathrm{tgt}} = P_X^{\mathrm{EXP}}$) gives (2). Outside Assumption A5, only $q(x, a)$ on the shared support is identified; reweighting to a different target context distribution requires a standard covariate-shift adjustment. □

Define the linear classes

$$\mathcal{H}_\psi := \{z \mapsto \theta^\top \psi(z) : \theta \in \mathbb{R}^{d_\psi}\}, \qquad \mathcal{F}_\varphi := \{z \mapsto w^\top \varphi(z) : w \in \mathbb{R}^{d_\varphi}\},$$

and the grounded / residual affine class

$$\mathcal{G}(f_{\mathrm{OBS}}, \psi) := \{f_{\mathrm{OBS}} - h : \ h \in \mathcal{H}_\psi\} = f_{\mathrm{OBS}} + \mathcal{H}_\psi,$$

where the equality uses that $\mathcal{H}_\psi$ is a linear space.

C.2 Summary of the fixed-target comparisons and empirical alignment diagnostics

The main text uses the population comparisons as conditional guidance for reading the experiments, not as a guarantee that one estimator dominates. We separate the formal results included below from empirical diagnostics that are motivated by the same approximation-gap viewpoint.

Formal results included below.

- **At the oracle level, correcting OBS cannot hurt.** Projecting OBS bias onto the correction space weakly improves EXP risk relative to OBS-Only; shrinkage matters only because the estimated correction can be noisy in finite samples. See Theorem 4.
- **Grounded expands the proxy-side approximation class.** If the OBS baseline is not already linear in ψ, then correcting an expressive baseline by a low-dimensional proxy residual can represent targets that a pure ψ-linear EXP fit cannot. See Theorem 2.
- **CVCI-Residual is favored when the hard part is already in the OBS baseline.** Residualization beats direct pooling when f_{OBS} absorbs the difficult part of the reward surface and the remaining discrepancy is simpler in proxy space than the full target is in raw text features. See Theorem 3.
- **In the shared linear proxy special case, Grounded is centered ridge.** When both the OBS baseline and the EXP target are linear in the same proxy representation, Grounded becomes ridge centered at the OBS coefficient, and it beats EXP-Only exactly when that OBS center is a better shrinkage target than zero. See Theorem 5 and Corollary 6.

Empirical diagnostics motivated by the formal results. The experiments do not test the oracle conditions directly. They report proxy-alignment and reward-shape diagnostics that are related to the approximation classes in the formal results. In summarization, the smooth-to-sharpened reward comparison changes the target while keeping the cached outputs fixed. In coding, the fix-success sweep moves the target away from auxiliary patch-quality features and toward true fix success. These diagnostics are empirical probes of estimator–target alignment, not additional theorems.

- **Target–proxy alignment is measured empirically.** The representation-based estimators are most useful when the target reward varies along directions retained by the auxiliary proxy. The summarization and coding experiments report held-out target-fit R^2 as a proxy for this alignment.
- **Reward definitions can change estimator rankings through approximation-class alignment.** Changing the reward definition can move the target toward or away from the approximation class used by an estimator. The smooth/sharpened summarization comparison and the coding fix-success sweep are diagnostics consistent with this mechanism.

- **Regret is the primary downstream metric.** The experiments report recommendation regret because the evaluation objective is action selection; RMSE and ranking metrics are secondary diagnostics for reward-surface fit and model-level value fit.

C.3 What representation correction buys

Theorem 2 (What representation correction buys)**.** *For any fixed baseline f_{OBS}, the grounded class $\mathcal{G}(f_{\mathrm{OBS}}, \psi)$ satisfies the following three statements.*

(a) If there exists $\theta_\star \in \mathbb{R}^{d_\psi}$ such that

$$r^\star(z) = f_{\mathrm{OBS}}(z) - \theta_\star^\top \psi(z) \qquad \text{for all } z,$$

then $r^\star \in \mathcal{G}(f_{\mathrm{OBS}}, \psi)$. Conditional on the fitted baseline f_{OBS}, EXP only needs to estimate the d_ψ-dimensional residual coefficient $\theta_\star$.

(b) If $f_{\mathrm{OBS}} \notin \mathcal{H}_\psi$, then

$$\mathcal{G}(f_{\mathrm{OBS}}, \psi) \neq \mathcal{H}_\psi, \qquad f_{\mathrm{OBS}} \in \mathcal{G}(f_{\mathrm{OBS}}, \psi) \setminus \mathcal{H}_\psi.$$

So grounded correction can represent targets that a pure ψ-linear EXP fit cannot.

(c) If $f_{\mathrm{OBS}} \in \mathcal{H}_\psi$, then

$$\mathcal{G}(f_{\mathrm{OBS}}, \psi) = \mathcal{H}_\psi.$$

In that special case, representation correction does not enlarge the function class.

Proof. We prove the three claims in order.

For part (a), the displayed identity

$$r^\star(z) = f_{\mathrm{OBS}}(z) - \theta_\star^\top \psi(z)$$

is exactly the statement that $r^\star \in \mathcal{G}(f_{\mathrm{OBS}}, \psi)$. Once f_{OBS} is treated as fixed, the only EXP-side unknown is the d_ψ-dimensional coefficient vector $\theta_\star$.

For part (b), note that $0 \in \mathcal{H}_\psi$, hence

$$f_{\mathrm{OBS}} = f_{\mathrm{OBS}} - 0 \in \mathcal{G}(f_{\mathrm{OBS}}, \psi).$$

If $\mathcal{G}(f_{\mathrm{OBS}}, \psi) = \mathcal{H}_\psi$, then this would imply $f_{\mathrm{OBS}} \in \mathcal{H}_\psi$, contradicting the assumption. Therefore $\mathcal{G}(f_{\mathrm{OBS}}, \psi) \neq \mathcal{H}_\psi$, and f_{OBS} is a member of $\mathcal{G}(f_{\mathrm{OBS}}, \psi)$ that does not belong to $\mathcal{H}_\psi$.

For part (c), suppose $f_{\mathrm{OBS}}(z) = \beta_{\mathrm{OBS}}^\top \psi(z)$ for some $\beta_{\mathrm{OBS}} \in \mathbb{R}^{d_\psi}$. Every $g \in \mathcal{G}(f_{\mathrm{OBS}}, \psi)$ has the form

$$g(z) = \beta_{\mathrm{OBS}}^\top \psi(z) - \theta^\top \psi(z) = (\beta_{\mathrm{OBS}} - \theta)^\top \psi(z) \in \mathcal{H}_\psi,$$

so $\mathcal{G}(f_{\mathrm{OBS}}, \psi) \subseteq \mathcal{H}_\psi$. Conversely, for any $g(z) = \beta^\top \psi(z) \in \mathcal{H}_\psi$, choosing $\theta = \beta_{\mathrm{OBS}} - \beta$ gives

$$g(z) = f_{\mathrm{OBS}}(z) - \theta^\top \psi(z) \in \mathcal{G}(f_{\mathrm{OBS}}, \psi).$$

Thus $\mathcal{H}_\psi \subseteq \mathcal{G}(f_{\mathrm{OBS}}, \psi)$, proving equality. □

C.4 Residualization versus direct pooling

Theorem 3 (Residualization versus direct pooling)**.** *Let f_{OBS} be the observational baseline used by Grounded and CVCI-Residual, and assume $f_{\mathrm{OBS}}, r^\star \in L^2(P_E)$ together with finite-dimensional linear classes $\mathcal{F}_\varphi$ and $\mathcal{H}_\psi$, so that both classes are closed in $L^2(P_E)$ and the projections in question exist (when* arg min *is non-unique, any minimizer suffices, since the displayed risks depend only on the projections). Define the oracle direct and residual predictors by*

$$f_{\mathrm{cvci}}^\dagger \in \arg\min_{f \in \mathcal{F}_\varphi} \mathcal{R}_E(f), \qquad h_{\mathrm{res}}^\dagger \in \arg\min_{h \in \mathcal{H}_\psi} \mathcal{R}_E(f_{\mathrm{OBS}} + h).$$

Then

$$\mathcal{R}_E(f_{\mathrm{OBS}} + h_{\mathrm{res}}^\dagger) \leq \mathcal{R}_E(f_{\mathrm{cvci}}^\dagger)$$

if and only if

$$\mathrm{dist}_E(r^\star, f_{\mathrm{OBS}} + \mathcal{H}_\psi) \leq \mathrm{dist}_E(r^\star, \mathcal{F}_\varphi),$$

where $\mathrm{dist}_E(r, \mathcal{A}) := \inf_{g \in \mathcal{A}} \|r - g\|_E$. *In particular, if*

$$r^\star - f_{\mathrm{OBS}} \in \mathcal{H}_\psi \qquad \textit{but} \qquad r^\star \notin \overline{\mathcal{F}_\varphi},$$

then

$$\mathcal{R}_E(f_{\mathrm{OBS}} + h_{\mathrm{res}}^\dagger) < \mathcal{R}_E(f_{\mathrm{cvci}}^\dagger).$$

Proof. By (13),

$$\inf_{f\in\mathcal{F}_\varphi} \mathcal{R}_E(f) = \sigma_E^2 + \inf_{f\in\mathcal{F}_\varphi} \|f - r^\star\|_E^2 = \sigma_E^2 + \operatorname{dist}_E^2(r^\star, \mathcal{F}_\varphi).$$

Likewise,

$$\inf_{h\in\mathcal{H}_\psi} \mathcal{R}_E(f_{\mathrm{OBS}} + h) = \sigma_E^2 + \inf_{h\in\mathcal{H}_\psi} \|f_{\mathrm{OBS}} + h - r^\star\|_E^2 = \sigma_E^2 + \operatorname{dist}_E^2(r^\star, f_{\mathrm{OBS}} + \mathcal{H}_\psi).$$

Subtracting the common irreducible term σ_E^2 gives the equivalence.

If $r^\star - f_{\mathrm{OBS}} \in \mathcal{H}_\psi$, then

$$\operatorname{dist}_E(r^\star, f_{\mathrm{OBS}} + \mathcal{H}_\psi) = 0.$$

If also $r^\star \notin \overline{\mathcal{F}_\varphi}$, then

$$\operatorname{dist}_E(r^\star, \mathcal{F}_\varphi) > 0.$$

Hence

$$\mathcal{R}_E(f_{\mathrm{OBS}} + h^\dagger_{\mathrm{res}}) < \mathcal{R}_E(f^\dagger_{\mathrm{cvci}}).$$ □

C.5 Oracle grounding and shrinkage

Theorem 4 (Oracle grounding is no worse than OBS-only)**.** *Let $b := f_{\mathrm{OBS}} - r^\star$ denote the OBS bias function on EXP. Let $\mathcal{H}_\psi \subset L^2(P_E)$ be the resulting closed linear correction space, and let*

$$h^\dagger := \Pi_{\mathcal{H}_\psi} b$$

be the $L^2(P_E)$ projection of b onto that space. For any $\alpha \in [0,1]$, define the oracle grounded predictor

$$f_{\mathrm{G},\alpha} := f_{\mathrm{OBS}} - \alpha h^\dagger.$$

Then

$$\mathcal{R}_E(f_{\mathrm{G},\alpha}) = \mathcal{R}_E(f_{\mathrm{OBS}}) - (2\alpha - \alpha^2)\|h^\dagger\|_E^2. \tag{14}$$

Hence

$$\mathcal{R}_E(f_{\mathrm{G},\alpha}) \le \mathcal{R}_E(f_{\mathrm{OBS}}) \qquad \text{for all } \alpha \in [0,1],$$

with strict inequality whenever $\alpha > 0$ and $h^\dagger \neq 0$.

Proof. Write

$$b = h^\dagger + u, \qquad u := b - h^\dagger.$$

Because $h^\dagger$ is the orthogonal projection of b onto the closed linear subspace $\mathcal{H}_\psi$, we have $u \perp h^\dagger$ in $L^2(P_E)$.

Now

$$f_{\mathrm{G},\alpha} - r^\star = (f_{\mathrm{OBS}} - r^\star) - \alpha h^\dagger = u + (1-\alpha)h^\dagger.$$

Therefore,

$$\mathcal{R}_E(f_{\mathrm{G},\alpha}) - \sigma_E^2 = \|u + (1-\alpha)h^\dagger\|_E^2 = \|u\|_E^2 + (1-\alpha)^2\|h^\dagger\|_E^2,$$

where the cross term vanishes by orthogonality. On the other hand,

$$\mathcal{R}_E(f_{\mathrm{OBS}}) - \sigma_E^2 = \|b\|_E^2 = \|u + h^\dagger\|_E^2 = \|u\|_E^2 + \|h^\dagger\|_E^2.$$

Subtracting the two displays gives

$$\mathcal{R}_E(f_{\mathrm{G},\alpha}) - \mathcal{R}_E(f_{\mathrm{OBS}}) = \big((1-\alpha)^2 - 1\big)\|h^\dagger\|_E^2 = -(2\alpha - \alpha^2)\|h^\dagger\|_E^2,$$

which is (14). Since $2\alpha - \alpha^2 \ge 0$ on $[0,1]$, the inequality follows, and it is strict whenever $\alpha > 0$ and $h^\dagger \neq 0$. □

Derivation of the noisy-correction identity. For any $\widetilde{h} \in L^2(P_E)$,

$$\mathcal{R}_E(f_{\mathrm{OBS}} - \alpha\widetilde{h}) - \mathcal{R}_E(f_{\mathrm{OBS}}) = \alpha^2\|\widetilde{h}\|_E^2 - 2\alpha\langle b, \widetilde{h}\rangle_E. \tag{15}$$

Indeed, $f_{\mathrm{OBS}} - \alpha\widetilde{h} - r^\star = b - \alpha\widetilde{h}$, so

$$\mathcal{R}_E(f_{\mathrm{OBS}} - \alpha\widetilde{h}) - \sigma_E^2 = \|b - \alpha\widetilde{h}\|_E^2 = \|b\|_E^2 + \alpha^2\|\widetilde{h}\|_E^2 - 2\alpha\langle b, \widetilde{h}\rangle_E.$$

Subtracting $\mathcal{R}_E(f_{\mathrm{OBS}}) - \sigma_E^2 = \|b\|_E^2$ yields (15).

C.6 The shared linear proxy special case

Theorem 5 (Grounded is centered ridge in the linear proxy special case)**.** *Assume the observational baseline is already linear in the proxy representation,*

$$f_{\mathrm{OBS}}(z) = \beta_{\mathrm{OBS}}^{\top}\psi(z),$$

with the intercept absorbed into ψ. Let the EXP sample be $\{(\psi_j, Y_j)\}_{j=1}^{n}$, and write

$$\Psi = \begin{bmatrix} \psi_1^{\top} \\ \vdots \\ \psi_n^{\top} \end{bmatrix} \in \mathbb{R}^{n \times d_\psi}, \qquad Y = (Y_1, \ldots, Y_n)^{\top}.$$

Under full correction $\alpha_{\mathrm{corr}} = 1$, grounded correction solves

$$\widehat{\delta} \in \arg\min_{\delta \in \mathbb{R}^{d_\psi}} \|\Psi\beta_{\mathrm{OBS}} - Y - \Psi\delta\|_2^2 + \lambda\|\delta\|_2^2$$

and outputs

$$\widehat{\beta}_{\mathrm{G}} := \beta_{\mathrm{OBS}} - \widehat{\delta}.$$

Then $\widehat{\beta}_{\mathrm{G}}$ is exactly the centered ridge estimator

$$\widehat{\beta}_{\mathrm{G}} \in \arg\min_{\beta \in \mathbb{R}^{d_\psi}} \|Y - \Psi\beta\|_2^2 + \lambda\|\beta - \beta_{\mathrm{OBS}}\|_2^2.$$

Equivalently,

$$\widehat{\beta}_{\mathrm{G}} = (\Psi^{\top}\Psi + \lambda I)^{-1}(\Psi^{\top}Y + \lambda\beta_{\mathrm{OBS}}).$$

For general $\alpha_{\mathrm{corr}} \in [0, 1]$,

$$\widehat{\beta}_{\mathrm{G},\alpha_{\mathrm{corr}}} := \beta_{\mathrm{OBS}} - \alpha_{\mathrm{corr}}\widehat{\delta} = (1 - \alpha_{\mathrm{corr}})\beta_{\mathrm{OBS}} + \alpha_{\mathrm{corr}}\widehat{\beta}_{\mathrm{G}}.$$

Proof. Set

$$\beta = \beta_{\mathrm{OBS}} - \delta.$$

Then $\delta = \beta_{\mathrm{OBS}} - \beta$, so

$$\Psi\beta_{\mathrm{OBS}} - Y - \Psi\delta = \Psi\beta_{\mathrm{OBS}} - Y - \Psi(\beta_{\mathrm{OBS}} - \beta) = \Psi\beta - Y.$$

Therefore

$$\|\Psi\beta_{\mathrm{OBS}} - Y - \Psi\delta\|_2^2 + \lambda\|\delta\|_2^2 = \|Y - \Psi\beta\|_2^2 + \lambda\|\beta - \beta_{\mathrm{OBS}}\|_2^2.$$

So minimizing over δ is equivalent to minimizing over β, which proves the centered-ridge representation.

For the closed form, differentiate the centered-ridge objective:

$$-2\Psi^{\top}(Y - \Psi\beta) + 2\lambda(\beta - \beta_{\mathrm{OBS}}) = 0,$$

hence

$$(\Psi^{\top}\Psi + \lambda I)\beta = \Psi^{\top}Y + \lambda\beta_{\mathrm{OBS}},$$

which yields the stated formula for $\widehat{\beta}_{\mathrm{G}}$.

Finally,

$$\widehat{\beta}_{\mathrm{G},\alpha_{\mathrm{corr}}} = \beta_{\mathrm{OBS}} - \alpha_{\mathrm{corr}}\widehat{\delta} = \beta_{\mathrm{OBS}} - \alpha_{\mathrm{corr}}(\beta_{\mathrm{OBS}} - \widehat{\beta}_{\mathrm{G}}) = (1 - \alpha_{\mathrm{corr}})\beta_{\mathrm{OBS}} + \alpha_{\mathrm{corr}}\widehat{\beta}_{\mathrm{G}}. \quad \square$$

Corollary 6 (When the OBS center beats EXP-only)**.** *Assume in addition that the EXP target is linear in the same representation,*

$$Y = \Psi\beta_\star + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid \Psi] = 0, \qquad \mathrm{Var}(\varepsilon \mid \Psi) = \sigma^2 I_n.$$

Let $S := \Psi^{\top}\Psi$, $A := (S + \lambda I)^{-1}$, and define the EXP-only ridge estimator

$$\widehat{\beta}_{\mathrm{EXP}} := A\Psi^{\top}Y.$$

For any positive semidefinite matrix G, write $\|u\|_G^2 := u^{\top}Gu$. Then, conditional on Ψ,

$$\mathbb{E}\Big[\|\widehat{\beta}_{\mathrm{G}} - \beta_\star\|_G^2 \mid \Psi\Big] = \lambda^2(\beta_\star - \beta_{\mathrm{OBS}})^{\top}AGA(\beta_\star - \beta_{\mathrm{OBS}}) + \sigma^2\,\mathrm{tr}(GASA),$$

while

$$\mathbb{E}\Big[\|\widehat{\beta}_{\mathrm{EXP}} - \beta_\star\|_G^2 \mid \Psi\Big] = \lambda^2\beta_\star^{\top}AGA\,\beta_\star + \sigma^2\,\mathrm{tr}(GASA).$$

Therefore the linear grounded estimator is better than EXP-Only if and only if

$$(\beta_\star - \beta_{\mathrm{OBS}})^{\top}AGA(\beta_\star - \beta_{\mathrm{OBS}}) \le \beta_\star^{\top}AGA\,\beta_\star. \tag{16}$$

Proof. First,

$$\widehat{\beta}_{\mathrm{EXP}} - \beta_\star = A\Psi^\top(\Psi\beta_\star + \varepsilon) - \beta_\star = (AS - I)\beta_\star + A\Psi^\top\varepsilon.$$

Since $AS - I = -\lambda A$,

$$\widehat{\beta}_{\mathrm{EXP}} - \beta_\star = -\lambda A\beta_\star + A\Psi^\top\varepsilon. \tag{17}$$

For grounded, note that

$$\Psi\beta_{\mathrm{OBS}} - Y = \Psi(\beta_{\mathrm{OBS}} - \beta_\star) - \varepsilon = \Psi\delta_\star - \varepsilon.$$

Hence

$$\widehat{\beta}_{\mathrm{G}} - \beta_\star = \beta_{\mathrm{OBS}} - \beta_\star - A\Psi^\top(\Psi\delta_\star - \varepsilon) = \delta_\star - AS\delta_\star + A\Psi^\top\varepsilon.$$

Using $I - AS = \lambda A$ gives

$$\widehat{\beta}_{\mathrm{G}} - \beta_\star = \lambda A\delta_\star + A\Psi^\top\varepsilon. \tag{18}$$

Now take conditional G-risk. From (17),

$$\mathbb{E}\Big[\|\widehat{\beta}_{\mathrm{EXP}} - \beta_\star\|_G^2 \mid \Psi\Big] = \lambda^2\beta_\star^\top AGA\,\beta_\star + \mathbb{E}\big[(A\Psi^\top\varepsilon)^\top G(A\Psi^\top\varepsilon) \mid \Psi\big],$$

because the cross term vanishes by $\mathbb{E}[\varepsilon \mid \Psi] = 0$. The noise term is

$$\mathbb{E}\big[(A\Psi^\top\varepsilon)^\top G(A\Psi^\top\varepsilon) \mid \Psi\big] = \mathrm{tr}\big(GA\Psi^\top\,\mathbb{E}[\varepsilon\varepsilon^\top \mid \Psi]\,\Psi A\big) = \sigma^2\,\mathrm{tr}(GASA).$$

This gives the EXP-only expression.

The grounded formula is identical, replacing the bias vector $-\lambda A\beta_\star$ by $\lambda A\delta_\star$ from (18). Comparing the two expressions yields (16). □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and Section 1 state the identification claim, the post-identification role of OBS, and the empirical scope of the estimator comparison.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Section 7 describes the cached-benchmark scope, benchmark-specific interpretation, and future work needed for broader deployments.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

   Justification: Section 2 states Assumptions A1–A5, Section 3 states the main theorem and proof sketch, and Appendix C gives full statements and proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: Sections 5, 5.4, and 5.5, together with Appendices A.10, A.11, A.13, and A.15, specify the benchmark construction, resampling rules, budgets, tuning, seeds, and metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: The draft cites public source datasets and cached patch resources, but it does not yet provide an anonymized code or data artifact with executable reproduction commands.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Appendix A.10 gives the fixed pool, train/test split, OBS/EXP sampling, and aggregation protocol, while Appendix A.11 specifies tuning and fallback rules.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: The main-text tables report means over paired seeds, and Appendix A.10 states that means and standard errors are recomputed from seed-level paired runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No].

   Justification: The draft describes experimental budgets and cached-data construction, but it does not yet report hardware, memory, runtime, or total compute requirements.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes].

Justification: The work analyzes public/cached benchmark data and does not introduce human-subject data collection, high-risk model release, or deployed decision systems.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No].

Justification: The draft discusses methodological scope and deployment validation, but it does not yet contain a dedicated discussion of positive and negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The paper does not release a new high-risk model or scraped dataset; it evaluates methods on cached benchmark artifacts and public datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No].

Justification: The draft credits the datasets, models, and patch resources with citations and URLs, but it does not yet explicitly list licenses or terms of use for each asset.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A].

Justification: The submission draft does not introduce or release a new dataset, model, or code package.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: The paper does not conduct new crowdsourcing or human-subject experiments.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not conduct new human-subject research or collect new participant data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes].

    Justification: The paper describes LLM model outputs and LLM-judged rubric scores in Sections 5, 5.4, and Appendix A.10.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.