CEDRA: DOUBLY ROBUST GRAPH AGGREGATION UNDER SELECTIVE MESSAGE EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern graph learning increasingly operates in regimes where a receiver has many plausible senders, but evaluating every exact pairwise message is expensive. In this setting, the messages that are actually computed might not represent the candidate population the graph operator is meant to summarize: selective evaluation can turn aggregation into a property of the acquisition rule rather than the declared graph target. We introduce Candidate-Evaluation Doubly Robust Aggregation (CEDRA), a principled aggregation framework that declares the candidate support and reference law before evaluation, predicts each candidate's contribution from pre-evaluation information, and then corrects the candidate-wide baseline using inverse-probability-weighted residuals from the evaluated messages. For the unnormalized first-layer pre-transform aggregate, we give an exact bias decomposition showing that CEDRA-HT recovers the declared conditional mean when either the message model or the evaluation model is pointwise correct, under positivity, stable moments, separated nuisance estimation, and mean ignorability. We also derive CEDRA-RA, a design-known risk-adaptive path that minimizes mean-squared error under an exchangeable residual model. In controlled recovery studies, CEDRA-HT reduces RMSE by 59.1% relative to evaluated-message averaging and by 38.1% relative to matched inverse-propensity aggregation. Across 720 frozen cells on six public graph supports, CEDRA-RA lowers RMSE by 22.10% relative to model-only aggregation and wins all 144 prespecified groups. CEDRA turns selective graph computation into a stated estimand, with a doubly robust correction, a risk-adaptive variant, and a matched empirical protocol to measure the gains when exact message evaluation is scarce. Code will be made available upon acceptance.

1 INTRODUCTION

A graph system may know which senders could contribute to a receiver before it can afford to compute what each sender contributes. Consider a retrieval pipeline that ranks candidate pairs using inexpensive metadata but computes an exact pairwise message only for a selected subset. A budget for these evaluations limits which messages reach the aggregation step. If the selection rule favors candidates with systematically different message values, averaging the available messages changes what the receiver represents. The problem is therefore not only missing computation: selective evaluation can change the population summarized by the graph aggregate.

This distinction is easiest to see by separating the intended population from the acquisition rule. Suppose the target gives equal weight to a fixed candidate pool, while candidates with larger conditional-mean messages are more likely to be evaluated. The evaluated-message distribution then overrepresents those candidates. More generally, the intended aggregate is specified by a candidate support $\mathcal{C}_i$ and reference weights q_{ij}, fixed before evaluation. The weights describe whose contributions matter and by how much; evaluation probabilities describe which exact contributions are obtained. These two roles should not be conflated. We study how to recover the declared aggregate when only selectively evaluated messages are available.

Candidate-Evaluation Doubly Robust Aggregation (CEDRA) combines two sources of information. A message model predicts a contribution for every candidate from information available before exact evaluation. These predictions form a candidate-wide baseline under the reference weights. Exact messages then correct that baseline through inverse-probability-weighted prediction residuals. The

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

construction uses the classical augmented inverse-probability principle (Robins et al., 1994; Bang & Robins, 2005). Its graph-specific formulation makes the receiver-level target explicit, places the baseline over the entire candidate support, and applies correction before nonlinear graph computation. Baseline placement is essential: restricting even a correct message model to evaluated candidates can retain the same selection distortion that the estimator is meant to remove.

The resulting CEDRA-HT operator provides two routes to the target. Under positivity, stable conditional moments, separated nuisance estimation, and mean ignorability, its conditional bias vanishes if either the message model or the evaluation-probability model is pointwise correct. Mean ignorability requires each candidate's evaluated message to have the same conditional mean as its potential message, given the declared and nuisance-training information. Our exact bias decomposition separates a product of the two nuisance-model errors from a residual-selection covariance term. It describes both the double-robust case and how bias can remain when evaluation and message residuals covary after adjustment. The identity concerns the unnormalized first-layer pre-transform aggregate.

Conditional unbiasedness and accurate point estimation are different objectives. When evaluation probabilities are small, inverse weighting can magnify residual noise even when the candidate-wide prediction is useful. We derive CEDRA-RA, which interpolates between model-only aggregation and full HT correction. Its correction strength is fixed before evaluation from the reference weights and known design probabilities. Under an exchangeable residual model, it minimizes mean-squared error for the realized candidate aggregate along this scalar path. HT thus addresses conditional bias, while RA supplies a design-known bias–variance tradeoff under a separate risk model. Neither criterion alone determines the behavior of a learned nonlinear predictor.

Our evaluation follows these distinct questions. Q1 tests aggregate recovery with controlled generators, Optdigits, and ten public feature tables. Q2 tests risk adaptation across a frozen 720-cell matrix on six public graph supports: RA lowers RMSE by 22.10% relative to model-only aggregation. Q3 asks whether correction helps prediction when candidate information, exact-message budgets, and the backbone recipe are matched. The study covers SAGE- and GIN-style candidate-graph predictors, with a balanced head fixed entirely from training and validation for the class-imbalanced Minesweeper and Tolokers tasks. Final Tolokers scores span 67.84–68.67% across arms, while Minesweeper remains near chance. Candidate-size, reference-law, budget, depth, and residual-selection sweeps connect these findings to the conditions the method requires. Earlier conventional-GNN comparisons remain separate because their information sets differ.

The paper makes three contributions:

- A candidate-target formulation of selective graph-message evaluation, with an explicit separation between reference weights and acquisition probabilities.
- A first-layer HT bias decomposition and double-robust identity, together with a design-known, path-optimal RA correction under its stated residual model.
- A controlled recovery and matched-prediction evaluation that separates aggregation effects from information, backbone, and downstream-head decisions, retaining complete results across the declared experimental settings.

2 RELATION TO PRIOR WORK

Propensity-aware aggregation. Neighbor Aggregation via Inverse Propensity (Navip) uses ratio-normalized inverse selection scores in recommender GNNs (Kim et al., 2022); APDA, CAGED, and DPAA address popularity or interaction bias through propagation (Zhou et al., 2023; Que et al., 2025; Aminul Islam et al., 2026). CEDRA instead targets a candidate population declared before exact evaluation, adding a candidate-wide baseline and residual correction. Adversarially robust aggregation concerns a different distortion and target (Hou et al., 2024).

Doubly robust estimation. CEDRA uses the classical augmented inverse-probability principle (Robins et al., 1994; Bang & Robins, 2005). Related estimators address recommendation risk and exposure bias (Wang et al., 2019; Li et al., 2022b;a; 2024) or network causal effects (Liu et al., 2025). Our object is a finite first-layer candidate-message aggregate: the analysis identifies the necessary baseline placement, residual correction, and operations that preserve its conditional identity.

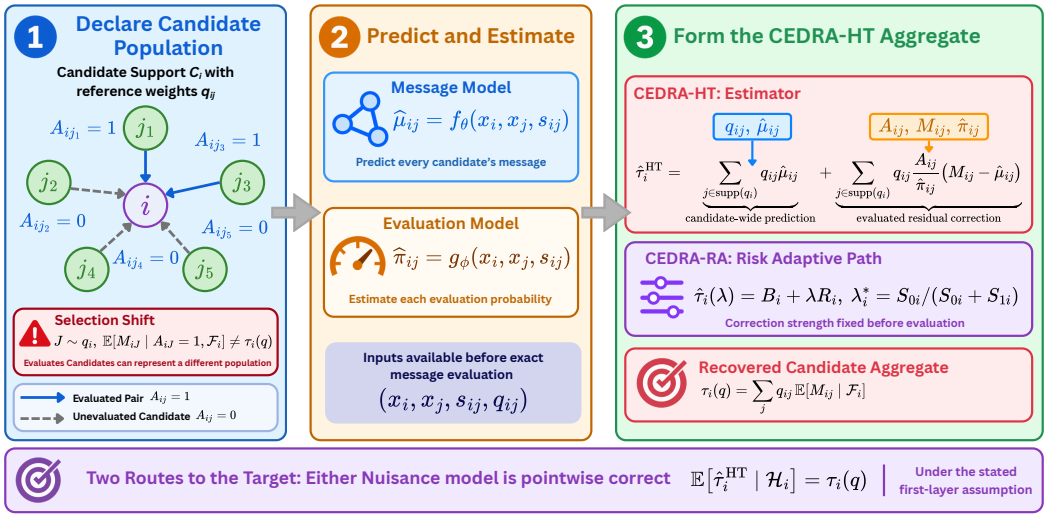


Figure 1: CEDRA declares a target population, predicts every candidate message, and corrects evaluated residuals. HT's double-robust identity applies to the unnormalized first-layer pre-transform aggregate under the stated conditions; RA optimizes a separate model-based risk.

Sampling and graph construction. GraphSAINT corrects subgraph-sampling bias (Zeng et al., 2019), and TOP reconstructs messages omitted by mini-batch boundaries (Shi et al., 2025). These sample an existing graph; CEDRA starts from a declared candidate law. Structure learning and rewiring instead alter the graph or objective (Franceschi et al., 2019; Chen et al., 2020; Dong & Kluger, 2023). Because graph conclusions depend on datasets, splits, and tuning (Platonov et al., 2023; Tori et al., 2025; Zhu et al., 2020; Lim et al., 2021), we distinguish identified recovery, matched aggregation effects, and contextual whole-system comparisons.

3 Problem: Selective Message Evaluation

3.1 Candidate Law and Estimand

For receiver i, fix a finite candidate support $\mathcal{C}_i$ and nonnegative reference weights q_{ij} summing to one before evaluation indicators A_{ij} are realized. The support defines who could contribute, q_i defines how they contribute, and evaluation determines which exact messages are obtained. Uniform or locality-weighted laws are admissible (Appendix A). Let $M_{ij} \in \mathbb{R}^d$ be the potential first-layer message, available only if $A_{ij} = 1$, and let W_{ij} be cheap pre-evaluation information available for every candidate. With $\mathcal{F}_i$ generated by the candidate law and declared adjustment information, define

$$\pi_{ij}^\star := \Pr(A_{ij} = 1 \mid \mathcal{F}_i), \qquad m_{ij} := \mathbb{E}[M_{ij} \mid \mathcal{F}_i], \qquad \tau_i(q) := \sum_{j \in \mathrm{supp}(q_i)} q_{ij} m_{ij}. \tag{1}$$

Here $\mathrm{supp}(q_i) = \{j \in \mathcal{C}_i : q_{ij} > 0\}$. The theorem targets the conditional mean $\tau_i(q)$; full evaluation returns the realized aggregate $T_i(q) = \sum_j q_{ij} M_{ij}$, whose conditional expectation is $\tau_i(q)$. These are descriptive targets, not causal effects. Selection changes the population represented by evaluated messages when candidate means covary with evaluation probabilities; Proposition 2 gives the exact selection-law identity. Finite-sample ratios additionally have random-denominator effects.

4 CEDRA ESTIMATOR AND GUARANTEES

4.1 CEDRA-HT: PREDICT ALL, CORRECT EVALUATED RESIDUALS

Restricting the regression baseline to evaluated messages retains selection bias even with a correct message model (Appendix F). CEDRA instead predicts all candidates and corrects only evaluated residuals, combining Horvitz–Thompson inverse-probability weighting with augmentation (Horvitz & Thompson, 1952; Robins et al., 1994; Bang & Robins, 2005; Dudík et al., 2011):

$$\widehat{\tau}_i^{\mathrm{HT}} = \underbrace{\sum_{j \in \mathrm{supp}(q_i)} q_{ij} \widehat{\mu}_{ij}}_{\text{candidate-wide baseline}} + \underbrace{\sum_{j \in \mathrm{supp}(q_i)} q_{ij} \frac{A_{ij}}{\widehat{\pi}_{ij}} (M_{ij} - \widehat{\mu}_{ij})}_{\text{evaluated-message HT residual}}. \tag{2}$$

The graph-specific contract is the declared receiver-level population, the candidate-wide baseline, and correction before graph-dependent nonlinear computation. Figure 1 shows the two paths.

4.2 CONDITIONAL BIAS AND DOUBLE ROBUSTNESS

Let $\mathcal{H}_i = \mathcal{F}_i \vee \mathcal{G}$ include nuisance training information. We require a fixed pre-evaluation target; stable conditional moments $\mathbb{E}[A_{ij} \mid \mathcal{H}_i] = \pi^\star_{ij}$ and $\mathbb{E}[M_{ij} \mid \mathcal{H}_i] = m_{ij}$; true and estimated positivity on the target support; $\mathcal{H}_i$-measurable separated nuisances; and integrable first-layer potential messages defined by a fixed or separately trained map. Assumptions 1–5 give the full conditions. Cross-fitting alone does not establish moment stability under arbitrary graph dependence. Define

$$c_{ij} := \mathrm{Cov}(A_{ij}, M_{ij} \mid \mathcal{H}_i) = \mathbb{E}[A_{ij}(M_{ij} - m_{ij}) \mid \mathcal{H}_i], \tag{3}$$

where moment stability gives the equality.

Theorem 1 (Conditional bias decomposition for CEDRA-HT)**.** *Under Assumptions 1–5, conditionally on $\mathcal{H}_i$, only the jointly distributed evaluation/message array remains random; no independence across candidates is required. Then*

$$\mathbb{E}\big[\widehat{\tau}_i^{\mathrm{HT}} - \tau_i(q) \mid \mathcal{H}_i\big] = \sum_{j \in \mathrm{supp}(q_i)} q_{ij} \left\{ \frac{\pi^\star_{ij} - \widehat{\pi}_{ij}}{\widehat{\pi}_{ij}} (m_{ij} - \widehat{\mu}_{ij}) + \frac{c_{ij}}{\widehat{\pi}_{ij}} \right\}. \tag{4}$$

Corollary 1 (Mean-ignorable product remainder and double robustness)**.** *If, in addition, $c_{ij} = 0$ for every $j \in \mathrm{supp}(q_i)$, equivalently $\mathbb{E}[A_{ij}(M_{ij} - m_{ij}) \mid \mathcal{H}_i] = 0$, then*

$$\mathbb{E}[\widehat{\tau}_i^{\mathrm{HT}} - \tau_i(q) \mid \mathcal{H}_i] = \sum_{j \in \mathrm{supp}(q_i)} q_{ij} \frac{\pi^\star_{ij} - \widehat{\pi}_{ij}}{\widehat{\pi}_{ij}} (m_{ij} - \widehat{\mu}_{ij}). \tag{5}$$

A sufficient condition is $A_{ij} \perp\!\!\!\perp M_{ij} \mid \mathcal{H}_i$. The conditional bias is zero if, on $\mathrm{supp}(q_i)$, either nuisance is pointwise correct.

Correctness of either nuisance removes the product term, while c_{ij} measures residual selection dependence. Substituting the conditional moments into each baseline-plus-residual term and summing proves the result (Appendix G). With correct design probabilities independent of realized messages, HT also targets realized $T_i(q)$ (Corollary 2); fixed affine maps preserve the identity (Corollary 3). Neither result extends automatically to nonlinear prediction, and concentration requires additional dependence and moment conditions.

4.3 RISK-ADAPTIVE CORRECTION PATH

CEDRA-HT protects the design-unbiased branch. When its candidate-wide prediction is already accurate, however, a smaller residual correction can improve point estimation by avoiding unnecessary inverse-weight variance. Write $B_i = \sum_j q_{ij} \widehat{\mu}_{ij}$, $r_{ij} = M_{ij} - \widehat{\mu}_{ij}$, and $R_i = \sum_j q_{ij} A_{ij} r_{ij} / \pi_{ij}$. The path $\widehat{T}_i(\lambda) = B_i + \lambda R_i$ connects model-only aggregation at $\lambda = 0$ to full CEDRA-HT at $\lambda = 1$.

Proposition 1 (Path-optimal design-known shrinkage)**.** *For one scalar message coordinate, condition on the fixed design and fitted baseline. Assume conditionally independent $A_{ij} \sim \mathrm{Bernoulli}(\pi_{ij})$ and*

residuals that are independent across candidates, independent of the evaluation design, mean zero, and exchangeable with variance σ^2. Define $S_{0i} = \sum_j q_{ij}^2$ and $S_{1i} = \sum_j q_{ij}^2(1-\pi_{ij})/\pi_{ij}$. Then the mean-squared error for estimating the realized candidate aggregate $T_i(q)$ is

$$\mathrm{MSE}\{\widehat{T}_i(\lambda); T_i(q)\} = \sigma^2\{S_{0i}(1-\lambda)^2 + S_{1i}\lambda^2\}, \qquad \lambda_i^\star = \frac{S_{0i}}{S_{0i}+S_{1i}}. \tag{6}$$

The minimized risk is $\sigma^2 S_{0i}S_{1i}/(S_{0i}+S_{1i})$, which is no greater than the risk of either endpoint. If all $\pi_{ij} = p_i$, then $\lambda_i^\star = p_i$.

Interpretation. We call $\widehat{T}_i(\lambda_i^\star)$ CEDRA-RA. Its strength is fixed from q_i and known probabilities before evaluation, not tuned to outcomes. At a common probability p_i, $\lambda_i^\star = p_i$: stronger subsampling relies more on the baseline. RA has the stated model-based risk guarantee; HT retains the double-robust and design-unbiased branches. Neither implies predictive dominance.

5 FROM IDENTIFICATION TO IMPLEMENTATION

Message models use receiver-separated information, never their target message or post-evaluation variables as inputs. Probabilities are learned in the learned-nuisance studies and design-known in matched prediction. HT carries the exact identity; SN self-normalizes its residual and is generally biased; RA optimizes the stated risk path. For matched prediction, a frozen task-trained pair encoder and cross-fitted outcome predictions are shared across arms. Each aggregate plus cheap receiver features feeds a SAGE- or GIN-style candidate-graph predictor with identical initialization and training recipe, not identical fitted weights. Correction remains at the first-layer interface, including in the depth sweep. This is not joint end-to-end encoder/nuisance training. Appendix K.1 specifies the contract.

Calibration and prediction use different acquisition stages. The pair encoder is trained with full messages on calibration receivers and then frozen; a separate evaluated-message pilot trains the outcome model on disjoint receivers. These shared training costs are accounted separately from test queries. At test time, feasible arms obtain exact messages only through the counted selective interface, while complete candidate-message arrays remain evaluator inputs. The reported query fraction therefore describes selective prediction after calibration, rather than the acquisition cost of learning the shared representations.

6 EMPIRICAL VALIDATION

Q1: Does CEDRA recover the declared aggregate? Controlled populations, Optdigits, and ten public feature tables test known randomized evaluation. **Q2: Does risk adaptation improve recovery across regimes?** The frozen 720-cell matrix crosses two teachers, two candidate laws, two evaluation mechanisms, three budgets, and six public graph supports, with five masks per design group. **Q3: How does aggregation affect prediction at a fixed exact-message budget?** Six graphs, three seeds, two candidate-graph backbones, and 15 additive settings yield 540 configurations and 4,860 arm-output slots. Within a backbone, feasible arms share candidates, cheap inputs, splits, masks/probabilities, encoder, applicable nuisances, initialization, and training recipe. Evaluation is imposed, not naturally occurring. Appendix K gives complete protocols, retained outcomes, and the separate mean-ignorability stress.

7 RESULTS

Table 1 follows the evidence hierarchy: identified recovery, risk-adaptive recovery, and matched prediction. Table 3 gives every matched primary score; the historical system-level GNN comparisons remain separate in Section 7.4.

7.1 Q1: CEDRA RECOVERS THE DECLARED TARGET

The controlled studies validate both routes in the double-robust construction. In the complete 72×20 oracle factorial, CEDRA-HT lowers RMSE from 0.3118 for the evaluated-message mean to 0.1274,

Table 1: Evidence map for the three empirical questions. Panels A and B report recovery of declared aggregates under randomized evaluation; Panel C gives the fixed HT-versus-model-only contrasts from the separate matched-Q3 study, whose complete controls appear in Table 3. RMSE effects are relative reductions; prediction effects and their paired-seed 95% intervals are in balanced-accuracy percentage points (pp). Public-feature family envelopes are descriptive, use no source-task labels, and are not model-selection estimates. Positive effects favor CEDRA; bold marks the lower-RMSE entry. The three protocols are not pooled. BA denotes balanced accuracy and n counts complete artifacts or paired dataset-seed cells.

Study and dataset	Task / metric	n	Method	Comparator	CEDRA	Comp.	Effect
A. Identified target recovery under randomized evaluation							
Controlled oracle	Synthetic / RMSE	1,440	CEDRA-HT	Eval. mean	**0.1274**	0.3118	+59.1%
Controlled learned	Synthetic / RMSE	180	CEDRA-HT	Navip	**0.1893**	0.3058	+38.1%
Optdigits, near primary	Feature agg. / RMSE	20	CEDRA-HT	Eval.-q	**0.0388**	0.0403	+3.6%
Optdigits, constant desc.	Feature agg. / RMSE	20	CEDRA-SN	Eval.-q	**0.0443**	0.0504	+12.0%
Optdigits, far desc.	Feature agg. / RMSE	20	CEDRA-SN	Eval. mean	0.0635	**0.0586**	−8.3%
Abalone	Feature agg. / RMSE	20	Best CEDRA	Best non-CEDRA	**0.06627**	0.06984	+5.1%
Airfoil	Feature agg. / RMSE	20	Best CEDRA	Best non-CEDRA	**0.18482**	0.19028	+2.9%
Wine Quality	Feature agg. / RMSE	20	Best CEDRA	Best non-CEDRA	**0.19596**	0.20147	+2.7%
MAGIC	Feature agg. / RMSE	20	Best CEDRA	Best non-CEDRA	**0.18088**	0.18695	+3.2%
Rice	Feature agg. / RMSE	20	Best CEDRA	Best non-CEDRA	**0.01782**	0.02937	+39.3%
Optdigits	Feature agg. / RMSE	20	Best CEDRA	Best non-CEDRA	**0.15549**	0.16920	+8.1%
Dry Bean	Feature agg. / RMSE	20	Best CEDRA	Best non-CEDRA	**0.08811**	0.09374	+6.0%
Letter	Feature agg. / RMSE	20	Best CEDRA	Best non-CEDRA	**0.14164**	0.14439	+1.9%
PenDigits	Feature agg. / RMSE	20	Best CEDRA	Best non-CEDRA	0.12110	**0.12050**	−0.5%
Landsat	Feature agg. / RMSE	20	Best CEDRA	Best non-CEDRA	**0.05796**	0.06627	+12.5%
B. Frozen risk-adaptive robustness matrix							
Frozen 720-cell matrix	32-D msg. / RMSE	720	CEDRA-RA	Model-only	**0.09107**	0.11690	+22.10%
C. Final balanced-head prediction at expected evaluation fraction 0.3							
Tolokers, SAGE-style	Prediction / BA	3	CEDRA-HT	Matched arms	$68.06 \pm 0.54\%$		
Tolokers, GIN-style	Prediction / BA	3	CEDRA-HT	Matched arms	$68.28 \pm 0.65\%$		

a 59.1% reduction. Within this factorial, RMSE is 0.045 when only the message model is correct. When only the evaluation model is correct, empirical mean bias remains near zero and RMSE is 0.509. With both models misspecified, bias is 0.199 and RMSE is 0.535, consistent with the requirement that at least one nuisance model be correct. In the separate learned-nuisance grid, cross-fitted CEDRA-HT reduces RMSE from 0.3058 for matched Navip to 0.1893, a 38.1% reduction, while the stabilized CEDRA-SN reaches 0.159.

The frozen Optdigits study extends target recovery to real covariates. In the primary contrast, CEDRA-HT lowers near-selection RMSE from 0.04027 for the observed-q ratio to 0.03882, a 3.6% reduction, and wins all 20 paired seeds with Holm-adjusted $p = 1.53 \times 10^{-5}$. The stabilized CEDRA-SN attains the lowest near-selection RMSE among feasible non-oracle estimators at 0.03281, 12.2% below Navip, and improves by 12.0% under constant selection.

Across ten public feature tables totaling 90,066 source rows, evaluated on held-out test receivers over 20 randomized seeds per dataset under the near-selection mechanism, a CEDRA family member has the lowest descriptive RMSE on nine datasets, with gains from 1.9% to 39.3%. PenDigits is the sole exception, favoring the comparator by 0.5%. These complete, label-free recovery results show that the estimator applies across varied real covariates while preserving the candidate-message target.

The far-selection stress test marks a regime in which stabilized correction does not improve RMSE: CEDRA-SN has 8.3% higher RMSE than the evaluated-message mean. This result motivates the separately frozen risk-adaptation study in Q2.

7.2 Q2: Risk adaptation improves recovery across regimes

When the candidate-wide prediction is strong, full inverse correction can add more variance than the remaining residual requires. CEDRA-RA resolves this regime with a correction weight fixed

Table 2: Risk-adaptive recovery over all 720 cells. Values are mean RMSE; lower is better. Across 144 design groups with five fixed masks each, CEDRA-RA wins 144/144 comparisons with model-only and IPW, 140/144 with the evaluated-message mean, and 111/144 with full CEDRA-HT.

Teacher	Candidate law	Eval. mean	Model	IPW	Full CEDRA-HT	CEDRA-RA
Additive tanh	Original edges	0.29028	0.14106	0.54816	0.22099	**0.11411**
Additive tanh	Projected neighbors	0.17605	0.11362	0.24997	0.09333	**0.08360**
Gated ReLU	Original edges	0.19680	0.12607	0.39874	0.19797	**0.10204**
Gated ReLU	Projected neighbors	0.12155	0.08684	0.17329	0.07799	**0.06453**
Overall	All four families	0.19617	0.11690	0.34254	0.14757	**0.09107**

before evaluation from q_i and the known evaluation probabilities, without outcome-selected clipping or normalization.

Across all 720 frozen cells, CEDRA-RA achieves the lowest overall RMSE: 0.09107 versus 0.11690 for model-only aggregation, 0.14757 for full CEDRA-HT, 0.19617 for evaluated-message averaging, and 0.34254 for pure IPW. These values correspond to reductions of 22.10%, 38.29%, 53.58%, and 73.41%. Across the 144 design groups, CEDRA-RA beats model-only aggregation and IPW in all 144, the evaluated-message mean in 140, and full CEDRA-HT in 111; its mean RMSE is lower than full CEDRA-HT in every teacher-by-candidate-law family. CEDRA-RA's advantage over model-only aggregation holds across both message teachers, both candidate laws, both randomized evaluation mechanisms, all three budgets, and all six public graph supports.

7.3 Q3: MATCHED PREDICTION UNDER SELECTIVE MESSAGE EVALUATION

The reference-setting comparison fixes $K = 40$, an expected evaluation fraction of 0.3, normalized exponential-rank weights of strength 1, and depth 2. Within each dataset and backbone it changes only the aggregation rule. Evaluated-q averaging, a Navip-style inverse-propensity ratio, pure HT/IPW, model-only aggregation, and EdgeResid provide five fixed controls; all three CEDRA variants remain visible. The full-message row is a separate full-cost reference. Minesweeper and Tolokers use the balanced head fixed from training and validation. These are candidate-graph predictors with a shared frozen task-trained pair encoder, not reproductions of native-adjacency GNN baselines.

The controls distinguish the value of evaluated information from the choice of correction. Model-only asks whether evaluated residuals add predictive value beyond the shared candidate-wide predictions. Comparing CEDRA-HT with pure HT/IPW tests the contribution of augmentation, while evaluated-q and Navip-style compare selective averaging and ratio weighting. EdgeResid retains the same baseline and evaluated residuals but uses propensity scores as non-inverse gates, testing inverse correction against another residual aggregation rule. An HT improvement over model-only can therefore coexist with losses against selective controls: the former concerns the value of evaluated information in this pipeline, while the latter limits claims about how best to aggregate it.

Final matched results. Under the balanced head, all Tolokers arms fall between 67.84% and 68.67% balanced accuracy, while Minesweeper remains between 49.71% and 51.49%. CEDRA-HT reaches 68.06% with SAGE and 68.28% with GIN on Tolokers. Outcome-only and HT/IPW are also competitive, so the table supports robust selective-evaluation comparison rather than universal dominance by one aggregation rule. Appendix K.1.1 gives the balanced-head specification.

Evaluation budget. Figure 2 shows the complete budget sweep. From expected fractions 0.1 to 1, HT's equal-graph mean rises from 46.02% to 47.74% for SAGE and 45.75% to 47.34% for GIN. At fraction 0.5, it reaches 47.69% and 47.17%, compared with the corresponding full-cost references of 47.84% and 47.38%. These descriptive proximity comparisons do not establish equivalence or superiority over the other selective controls. The budget counts exact pair-message queries, not sender-feature acquisition or total runtime.

Table 3: Final matched prediction results at $K = 40$, expected evaluation fraction 0.3, rank-law strength 1, and depth 2. Test balanced accuracy is mean ± sample SD in percent over three paired seeds. Minesweeper and Tolokers use the balanced head fixed from training and validation; all methods within each dataset share the same head protocol. Within each dataset and backbone, bold and italics mark the best and second-best distinct means among the eight budget-feasible arms; ties share a rank. The separated full-message reference has full query cost and is excluded from ranking.

Method	Cora	Roman-empire	Amazon-ratings	Minesweeper	Tolokers	Questions
A. SAGE-style candidate-graph backbone						
Evaluated-q	47.38 ± 0.51	51.09 ± 0.35	$\mathbf{31.63 \pm 1.02}$	50.43 ± 1.86	68.36 ± 0.74	$\mathit{54.85 \pm 4.20}$
Navip-style	46.86 ± 1.12	$\mathit{51.27 \pm 0.40}$	30.86 ± 0.68	49.71 ± 1.96	68.37 ± 0.69	54.77 ± 4.14
HT/IPW	47.96 ± 2.75	51.07 ± 0.40	30.34 ± 0.54	$\mathit{50.94 \pm 0.70}$	68.40 ± 0.68	52.79 ± 4.72
Model-only	48.31 ± 1.34	49.03 ± 0.42	27.53 ± 0.42	$\mathbf{51.07 \pm 0.78}$	$\mathbf{68.67 \pm 0.75}$	52.72 ± 2.64
EdgeResid	$\mathit{48.46 \pm 1.00}$	51.09 ± 0.45	$\mathit{31.13 \pm 1.45}$	50.04 ± 0.29	68.16 ± 0.59	52.72 ± 4.46
CEDRA-HT	$\mathbf{48.66 \pm 1.25}$	51.10 ± 0.64	30.24 ± 0.34	50.74 ± 0.36	68.06 ± 0.54	52.54 ± 4.15
CEDRA-SN	47.35 ± 0.83	$\mathbf{51.40 \pm 0.42}$	30.62 ± 0.35	50.86 ± 0.81	$\mathit{68.42 \pm 0.84}$	$\mathbf{55.21 \pm 3.39}$
CEDRA-RA	47.15 ± 3.72	49.57 ± 0.96	29.31 ± 0.72	50.10 ± 1.03	68.09 ± 0.56	52.24 ± 3.08
Full-cost reference	48.51 ± 0.38	51.39 ± 0.58	32.10 ± 1.01	51.05 ± 0.70	68.56 ± 0.72	55.03 ± 4.36
B. GIN-style candidate-graph backbone						
Evaluated-q	48.71 ± 0.91	50.57 ± 0.31	$\mathbf{30.99 \pm 1.02}$	$\mathbf{51.49 \pm 0.05}$	67.84 ± 0.39	52.05 ± 4.88
Navip-style	$\mathbf{49.95 \pm 0.90}$	$\mathit{50.70 \pm 0.39}$	30.39 ± 0.76	51.35 ± 0.48	68.21 ± 0.58	52.17 ± 4.88
HT/IPW	48.63 ± 0.19	50.35 ± 0.87	30.36 ± 0.62	51.20 ± 0.21	$\mathbf{68.56 \pm 0.80}$	$\mathit{52.37 \pm 4.83}$
Model-only	47.61 ± 1.69	49.45 ± 0.49	27.88 ± 0.87	$\mathbf{51.49 \pm 0.47}$	$\mathit{68.53 \pm 0.79}$	51.97 ± 3.42
EdgeResid	48.46 ± 1.11	50.57 ± 0.37	$\mathit{30.94 \pm 1.17}$	$\mathit{51.38 \pm 0.61}$	68.24 ± 0.45	52.00 ± 5.15
CEDRA-HT	48.81 ± 1.62	50.69 ± 0.24	30.08 ± 0.89	50.58 ± 1.83	68.28 ± 0.65	51.81 ± 4.44
CEDRA-SN	$\mathit{49.55 \pm 0.58}$	$\mathbf{50.81 \pm 0.42}$	30.53 ± 0.48	50.87 ± 0.95	68.20 ± 0.67	52.07 ± 4.90
CEDRA-RA	49.33 ± 0.99	49.59 ± 1.09	29.17 ± 0.83	50.22 ± 1.30	68.15 ± 0.61	$\mathbf{53.29 \pm 3.17}$
Full-cost reference	49.28 ± 0.87	50.92 ± 0.32	31.56 ± 0.91	51.34 ± 0.97	68.24 ± 0.67	52.10 ± 4.89

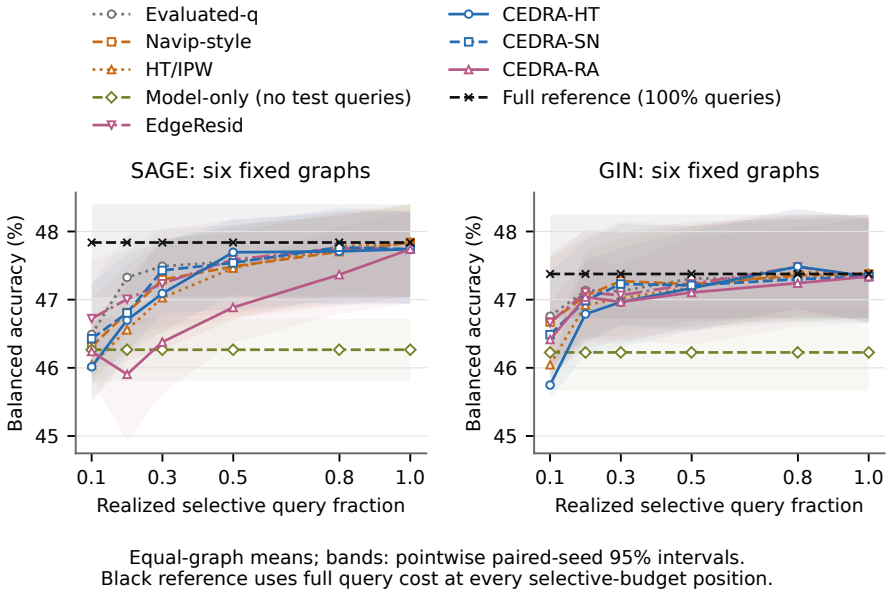


Figure 2: Matched prediction versus realized selective test-query fraction. Curves are equal-graph means; bands are pointwise paired-seed 95% intervals.

Candidate law, size, and depth. The complete additive sweeps test $K = 10, 20, 40, 80, 160$, uniform and three exponential-rank laws, and depths $1, 2, 3$ (Appendix K.2). Larger pools do not monotonically improve prediction: HT's SAGE mean is 46.28%, 47.09%, and 46.97% at $K = 10, 40, 160$. Reference-law changes have modest predictive effects but change the target itself. Equal-

graph HT and full-reference performance both decline with depth in these matched predictors. Depth changes propagation after the corrected input, not the first-layer guarantee.

When mean ignorability fails. A separate signed residual-selection stress crosses five dependence strengths with four nuisance regimes. At zero dependence, HT's analytical bias vanishes when either nuisance is correct; away from zero, its measured signed bias tracks the covariance remainder even when both nuisances are correct (Appendix K.3). The experiment thus verifies both branches of the identity and measures the consequence of violating its identifying assumption.

7.4 Downstream performance context on observed graphs

Historical conventional GNNs receive the original features and observed graph, while candidate-aware systems use a candidate population and split feature view. Shared splits/seeds do not remove this information difference, so these are whole-system comparisons, not isolated correction effects. The 54-run CEDRA-Guard—a task-aligned first-layer adapter that gates a clipped CEDRA residual into a frozen downstream GraphSAGE model (Appendix K)—is a successor to the earlier CEDRA-Fuse adapter and gives a pooled GraphSAGE contrast of -0.22 balanced-accuracy points (95% interval $[-0.47, 0.03]$), with three positive dataset means and a validation-selected zero path in 22 runs. It improves on Fuse by 0.091 points but does not establish predictive superiority. In the ten-dataset nonlinear-head study, CEDRA leads on four means and non-CEDRA on six; none of the four positive intervals excludes zero. Separate Optdigits and earlier 18-arm graph studies also do not beat their selected comparators. Appendix K preserves all scores, per-dataset effects, and protocols.

8 Scope and outlook

The guarantee requires a declared pre-evaluation target, positivity, stable moments, separated nuisances, and mean ignorability, and is confined to unnormalized first-layer pre-transform HT aggregation; Appendix K.3 stress-tests its covariance remainder, and Appendix J gives the full scope, including the untested naturally selective log, deployment scale, and protocol.

Using the operator. The reference law should encode the population of interest rather than reproduce the acquisition mechanism: changing q_i changes the estimand, whereas changing the evaluation budget changes how much information is obtained about that estimand. HT is the relevant operator when preserving the conditional target is the objective. RA addresses a different choice, trading correction against a useful baseline under its residual risk model. In a predictive application, these estimator properties motivate candidate operators but do not select the best one. The matched Q3 design provides the corresponding evaluation template: hold the information and acquisition process fixed, compare corrections, and account for calibration separately from selective test queries.

Deployment recipe. A CEDRA layer records four objects: candidate support, reference law, acquisition probabilities, and message model. Declare $\mathcal{C}_i$ and q_i before evaluation; estimate π_{ij} only from pre-evaluation information; fit $\widehat{\mu}_{ij}$ on separated or cross-fitted units; and report pair-query and sender-feature costs. Use HT for target preservation and RA when its residual risk model is credible, while treating self-normalization and clipping as distinct estimators. Under one fixed backbone and head, compare against model-only, evaluated-q, pure IPW, and a separated full-message reference.

9 Conclusion

CEDRA treats selective message evaluation as estimation of an explicit receiver-level target: it predicts across a declared candidate population and corrects the candidate-wide baseline with evaluated residuals. The HT decomposition identifies two routes to conditional unbiasedness and the residual-selection term that can defeat them; RA supplies a design-known correction strength under its residual model. The matched prediction study separates aggregation from information and backbone differences, with complete budget, candidate-law, size, and depth diagnostics. Its final balanced-head results reach 68.06% and 68.28% for CEDRA-HT on Tolokers, while Minesweeper remains near chance and competing aggregation rules remain strong. Together, they support a disciplined recipe: specify the aggregate before acquisition, and compare corrections under matched budgets.

AI Disclosure Section

In this work, we used generative AI tools to draft, edit, and polish text. We have not used generative AI tools for research ideation, theoretical development, proof construction, algorithm or method design; all such content is the authors' original work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

E. Alpaydin and Fevzi. Alimoglu. Pen-Based Recognition of Handwritten Digits. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5MG6K.

E. Alpaydin and C. Kaynak. Optical Recognition of Handwritten Digits. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50P49.

Md Aminul Islam, Ahmed Sayeed Faruk, Sourav Medya, and Elena Zheleva. Debiasing message passing to mitigate popularity bias in gnn-based collaborative filtering. *arXiv e-prints*, pp. arXiv–2605, 2026.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

R. Bock. MAGIC Gamma Telescope. UCI Machine Learning Repository, 2004. DOI: https://doi.org/10.24432/C52C8B.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

Thomas Brooks, D. Pope, and Michael Marcolini. Airfoil Self-Noise. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C5VW2C.

Gabriel Cardoso, Sergey Samsonov, Achille Thin, Eric Moulines, and Jimmy Olsson. Br-snis: bias reduced self-normalized importance sampling. *Advances in Neural Information Processing Systems*, 35:716–729, 2022.

Yu Chen, Lingfei Wu, and Mohammed Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. *Advances in neural information processing systems*, 33: 19314–19326, 2020.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Ilkay Cinar and Murat Koklu. Rice (Cammeo and Osmancik). UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C5MW4Z.

Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine Quality. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C56S3T.

Mingze Dong and Yuval Kluger. Towards understanding and reducing graph structural noise for gnns. In *International Conference on Machine Learning*, pp. 8202–8226. PMLR, 2023.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In *International conference on machine learning*, pp. 1972–1982. PMLR, 2019.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Zhichao Hou, Ruiqi Feng, Tyler Derr, and Xiaorui Liu. Robust graph neural networks via unbiased aggregation. *Advances in Neural Information Processing Systems*, 37:110097–110130, 2024.

Minseok Kim, Jinoh Oh, Jaeyoung Do, and Sungjin Lee. Debiasing neighbor aggregation for graph neural network in recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 4128–4132, 2022.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Murat Köklü and İlker Ali Özkan. Dry Bean. UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C50S4B.

Haoxuan Li, Yan Lyu, Chunyuan Zheng, and Peng Wu. Tdr-cl: Targeted doubly robust collaborative learning for debiased recommendations. *arXiv preprint arXiv:2203.10258*, 2022a.

Haoxuan Li, Chunyuan Zheng, and Peng Wu. Stabledr: Stabilized doubly robust learning for recommendation on data missing not at random. *arXiv preprint arXiv:2205.04701*, 2022b.

Haoxuan Li, Chunyuan Zheng, Shuyi Wang, Kunhan Wu, Eric Hao Wang, Peng Wu, Zhi Geng, Xu Chen, and Xiao-Hua Zhou. Relaxing the accurate imputation assumption in doubly robust learning for debiased collaborative filtering. In *ICML*, pp. 29448–29460, 2024.

Derek Lim, Felix Hohne, Xiuyu Li, Sijia Linda Huang, Vaishnavi Gupta, Omkar Bhalerao, and Ser Nam Lim. Large scale learning on non-homophilous graphs: New benchmarks and strong simple methods. *Advances in neural information processing systems*, 34:20887–20902, 2021.

Jizhou Liu, Dake Zhang, and Eric J Tchetgen Tchetgen. Auto-doubly robust estimation of causal effects on a network. *arXiv preprint arXiv:2506.23332*, 2025.

Warwick Nash, Tracy Sellers, Simon Talbot, Andrew Cawthorn, and Wes Ford. Abalone. UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C55C7W.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of gnns under heterophily: Are we really making progress? *arXiv preprint arXiv:2302.11640*, 2023.

Yue Que, Yingyi Zhang, Xiangyu Zhao, and Chen Ma. Causality-aware graph aggregation weight estimator for popularity debiasing in top-k recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 2471–2481, 2025.

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.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pp. 1670–1679. PMLR, 2016.

Zhihao Shi, Jie Wang, Zhiwei Zhuang, Xize Liang, Bin Li, and Feng Wu. Accurate and scalable graph neural networks via message invariance. In *International Conference on Learning Representations*, volume 2025, pp. 19920–19951, 2025.

David Slate. Letter Recognition. UCI Machine Learning Repository, 1991. DOI: https://doi.org/10.24432/C5ZP40.

Ashwin Srinivasan. Statlog (Landsat Satellite). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C55887.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015.

Floriano Tori, Vincent Holst, and Vincent Ginis. The effectiveness of curvature-based rewiring and the role of hyperparameters in gnns revisited. In *International Conference on Learning Representations*, volume 2025, pp. 93981–94002, 2025.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*, pp. 6638–6647. PMLR, 2019.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. Graphsaint: Graph sampling based inductive learning method. *arXiv preprint arXiv:1907.04931*, 2019.

Huachi Zhou, Hao Chen, Junnan Dong, Daochen Zha, Chuang Zhou, and Xiao Huang. Adaptive popularity debiasing aggregator for graph collaborative filtering. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pp. 7–17, 2023.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. *Advances in neural information processing systems*, 33:7793–7804, 2020.

A Candidate construction, full assumptions, and auxiliary results

A.1 A locality-preserving reference law

For a declared support $\mathcal{C}_i$, one locality-preserving reference law is

$$q_{ij} = \frac{\kappa(\rho(S_i, S_j)/h_i)\,\mathbf{1}\{j \in \mathcal{C}_i\}}{\sum_{k\in\mathcal{C}_i} \kappa(\rho(S_i, S_k)/h_i)}, \tag{7}$$

where $\kappa \geq 0$ is non-increasing, $0 < h_i < \infty$, and the distance ρ, bandwidth h_i, covariates S, and pool are chosen by design or from training-only information. The denominator is positive. Uniform weights are valid inside a meaningful local pool; uniform mass over every graph node need not preserve an application's locality. This construction is an example, not an empirically tuned kernel-bandwidth result.

A.2 How selection changes the target

Let $J \sim q_i$ be auxiliary and conditionally independent of the evaluation/message array. Given $\mathcal{F}_i$, $\Pr(J = j) = q_{ij}$; conditional on $\mathcal{F}_i$ and $J = j$, define $(A_{iJ}, M_{iJ}) = (A_{ij}, M_{ij})$. For Proposition 2 only, assume for every supported candidate that

$$\mathbb{E}[A_{ij}M_{ij} \mid \mathcal{F}_i] = \pi^\star_{ij} m_{ij}, \tag{8}$$

as follows from $A_{ij} \perp\!\!\!\perp M_{ij} \mid \mathcal{F}_i$. Theorem 1 does not impose this identity: it conditions on the larger $\mathcal{H}_i$ and retains the covariance term c_{ij}.

Proposition 2 (Candidate-law selection identity)**.** *If Equation 8 holds and* $\sum_j q_{ij}\pi^\star_{ij} > 0$*, then, componentwise,*

$$\mathbb{E}[M_{iJ} \mid A_{iJ} = 1, \mathcal{F}_i] = \frac{\sum_{j\in\mathrm{supp}(q_i)} q_{ij}\mathbb{E}[A_{ij}M_{ij} \mid \mathcal{F}_i]}{\sum_{j\in\mathrm{supp}(q_i)} q_{ij}\Pr(A_{ij} = 1 \mid \mathcal{F}_i)}, \tag{9}$$

$$\mathbb{E}[M_{iJ} \mid A_{iJ} = 1, \mathcal{F}_i] - \tau_i(q) = \frac{\mathrm{Cov}_{J\sim q_i}(\pi^\star_{iJ}, m_{iJ} \mid \mathcal{F}_i)}{\sum_{j\in\mathrm{supp}(q_i)} q_{ij}\pi^\star_{ij}}. \tag{10}$$

The selected-message mean shifts when candidates with different conditional messages also have different evaluation probabilities, and equals the target when this covariance is zero. Finite-sample ratios additionally introduce a random denominator. Appendix E proves the identity.

A.3 Full first-layer conditions

Let $\mathcal{G}$ denote nuisance-training information and $\mathcal{H}_i = \mathcal{F}_i \vee \mathcal{G}$. The following conditions separate target construction, evaluation randomness, nuisance training, and downstream computation.

Assumption 1 (Fixed target)**.** *$\mathcal{C}_i$, q_i, and their construction variables are fixed before $\{A_{ij}\}$ is realized and are constructed without those indicators or labels reserved for evaluation.*

Assumption 2 (Conditional moment stability)**.** *For every $j \in \mathrm{supp}(q_i)$, $\mathbb{E}[A_{ij} \mid \mathcal{H}_i] = \pi^\star_{ij}$, $\mathbb{E}[M_{ij} \mid \mathcal{H}_i] = m_{ij}$, and $\mathbb{E}[\|A_{ij}M_{ij}\| \mid \mathcal{H}_i] < \infty$. Thus nuisance training preserves these conditional first moments; this does not impose mean ignorability between evaluation and message.*

Assumption 3 (Target-support positivity)**.** *There is an $\eta > 0$ such that $\pi^\star_{ij} \geq \eta$ whenever $q_{ij} > 0$. The evaluation-probability estimates also satisfy $\widehat{\pi}_{ij} > 0$ on this support.*

Assumption 4 (Nuisance separation)**.** *The evaluated nuisance estimates are fixed oracle values or $\mathcal{H}_i$-measurable. Cross-fitting is sufficient only when its split and dependence structure imply Assumption 2; withholding A_{ij} alone is not sufficient under arbitrary graph dependence.*

Assumption 5 (Exact object)**.** *M_{ij} has a finite first moment and is a first-layer pre-transform potential message whose map is fixed or trained outside the evaluated unit. The theorem concerns Equation 2 before activation, normalization, dropout, residual updates, pooling, or prediction.*

From identity to estimation. Theorem 1 is a conditional identity. Identification requires pre-evaluation information sufficient for its moment and positivity conditions; estimation additionally requires suitable nuisance models and separation. The randomized studies establish identification by design and evaluate estimation empirically. If $\|c_{ij}\| \leq \Gamma_{ij}$, Equation 4 bounds the covariance contribution to conditional bias by $\sum_j q_{ij}\Gamma_{ij}/\widehat{\pi}_{ij}$.

A.4 Realized-target and affine-map corollaries

Corollary 2 (Design-unbiased realized-candidate target)**.** *Let $\mathcal{K}_i = \mathcal{H}_i \vee \sigma(\{M_{ij} : j \in \mathrm{supp}(q_i)\})$. If $\Pr(A_{ij} = 1 \mid \mathcal{K}_i) = \pi^\star_{ij}$, $\widehat{\pi}_{ij} = \pi^\star_{ij}$, and the evaluated nuisances do not use realized evaluation indicators, then*

$$\mathbb{E}[\widehat{\tau}_i^{\mathrm{HT}} \mid \mathcal{K}_i] = T_i(q). \tag{11}$$

No independence among candidate-evaluation indicators is needed for this expectation identity.

Corollary 3 (Fixed affine-map preservation)**.** *Let $L_i(x) = G_i x + b_i$, where G_i, b_i are $\mathcal{H}_i$-measurable and do not use realized evaluation indicators. Then*

$$\mathbb{E}[L_i(\widehat{\tau}_i^{\mathrm{HT}}) - L_i(\tau_i(q)) \mid \mathcal{H}_i] = G_i\,\mathbb{E}[\widehat{\tau}_i^{\mathrm{HT}} - \tau_i(q) \mid \mathcal{H}_i]. \tag{12}$$

Thus an immediate fixed affine map preserves conditional unbiasedness. This includes fixed linear pooling when pooled aggregates share a conditioning information set and pooling precedes nonlinear or data-dependent computation.

Appendix G gives the proofs. The expectation identities do not supply variance or concentration bounds without further dependence and moment conditions.

B ESTIMATOR AND PRACTICAL-MODEL DETAILS

Table 4: Estimator map. The final column states each method's precise aggregate-level scope; Section 7 evaluates predictive transfer separately. Navip is included as a related inverse-propensity neighbor aggregator in its published recommender setting (Kim et al., 2022).

Method	Intended object and baseline	Weight or normalization	Aggregate-level scope
Observed mean	Evaluated-message ratio; no candidate baseline	Degree ratio	No candidate-target claim
Candidate mean	Realized $T_i(q)$; all candidate messages	Fixed q_i	Exact full-access reference; conditionally unbiased for $\tau_i(q)$
Navip	Propensity-adjusted neighbor mean; no candidate baseline	Inverse, ratio-normalized	Published recommender setting
Pure HT	$\tau_i(q)$; no candidate baseline	Inverse, unnormalized	Correct evaluation-probability model under mean ignorability
Model only (outcome regression)	$\tau_i(q)$; candidate-wide regression baseline	None	Correct message regression
CEDRA-HT	$\tau_i(q)$; candidate-wide regression baseline	Inverse residual, unnormalized	Either nuisance correct under Corollary 1
CEDRA-RA	Realized candidate aggregate under a residual model	Design-known shrunken HT residual	Path-MSE-optimal under Proposition 1; generally biased
CEDRA-SN	Intended $\tau_i(q)$; candidate-wide regression baseline	Hájek-type residual ratio	Practical; no propensity-branch exactness
Structure learning	Learned graph or objective; method-specific	Architecture-specific	Different target

B.1 NUISANCE ESTIMATION AND INFORMATION CONTROL

We fit an evaluation-probability model $\widehat{\pi}(W_{ij})$ and a message regression $\widehat{\mu}(W_{ij})$ on training folds separated from evaluated receivers whenever the data structure permits. The split unit is the receiver: all candidate pairs for an evaluated receiver are withheld from both nuisance fits, while a node may still recur as a sender for receivers in different folds.

In the controlled generators, receiver units and evaluation draws are generated under the declared mechanism, so this separation preserves the evaluated first moments. On a single observational graph, shared senders, overlapping neighborhoods, or latent graph-wide variables can couple the folds; receiver-level cross-fitting alone therefore does not prove Assumption 2. This cross-fitting separation follows the standard orthogonal-learning use of sample splitting (Chernozhukov et al., 2018), but its validity remains a property of the data-generating dependence rather than of the split name.

The learned-nuisance controlled and historical public-graph studies fit evaluation-probability logits by proper binary cross-entropy over all declared candidates outside the evaluated receiver folds; they do not use focal loss or a fixed negative-to-positive subsample. A candidate subset held out from fitting is used for logistic calibration. Probability conversion, any case-control prior correction, held-out calibration, and clipping remain separate operations; prior correction requires known class-specific sampling probabilities. In public `observed_union`, the resulting values are candidate-set edge-classification scores, not identified pre-evaluation probabilities. The matched-Q3 study instead supplies every propensity-based arm with the same known randomized design probabilities and does not evaluate learned propensity estimation.

The message-regression heads receive only the declared construction features; they never receive the target message vector that they are trained to predict. The message model may not use A_{ij}

or post-evaluation quantities as shortcuts. Evaluation-probability diagnostics include held-out log loss, Brier score, calibration error, weight quantiles, and effective sample size; message-model diagnostics include held-out error and the relative norm of the baseline and residual components. These diagnostics assess nuisance quality but do not validate conditional moment stability or mean ignorability on observational graphs.

If the message regression is correct, a residual component whose conditional mean is near zero is an intended outcome rather than evidence that the augmentation failed. That same behavior does not rule out an overly broad q_i making the candidate-wide baseline dominate local signal, removing useful locality, or contributing to representation collapse. We therefore retain the outcome-only and full-candidate-mean controls and report baseline/residual shares, pairwise-distance, variance, effective-rank, and normalized-Dirichlet diagnostics; these are empirical stress tests, not a non-collapse guarantee.

Exact candidate-wide evaluation costs $O(|\mathcal{C}_i|d)$ arithmetic and $O(|\mathcal{C}_i|d)$ transient message storage per receiver for message width d, compared with observed-neighbor aggregation over $O(d_i d)$ entries for realized degree d_i. Pure HT avoids candidate-wide regression calls but still needs candidate metadata and evaluated-message weights; direct candidate aggregation has the same candidate-message order as CEDRA when every message is available. Batching can reduce peak memory without changing total candidate evaluation. Candidate subsampling introduces another sampling design and its own weights; a sampled candidate baseline is an approximation unless that design is analyzed explicitly. The new matched protocol evaluates $K = 10, 20, 40, 80, 160$ at a fixed expected fraction, with actual pair-query counts and head time/memory recorded; this does not hold absolute query count fixed or establish scalability beyond 160 candidates. Shared full-message encoder calibration and the nuisance pilot have separate acquisition costs. Exact pair-query savings need not save expensive sender-feature acquisition because a sender can be queried by many receivers, and the head timer omits candidate construction, nuisance prediction, and aggregation. Separate nuisance-fit duration is not recorded, so these measurements do not yield a complete pipeline timing decomposition.

B.2 Self-normalization and clipping

For numerical stability, one may use the residual-ratio variant

$$\widehat{\tau}_i^{\mathrm{SN}} = \sum_{j \in \mathrm{supp}(q_i)} q_{ij} \widehat{\mu}_{ij} + \frac{\sum_{j \in \mathrm{supp}(q_i)} q_{ij} A_{ij} \widehat{\pi}_{ij}^{-1} (M_{ij} - \widehat{\mu}_{ij})}{\sum_{j \in \mathrm{supp}(q_i)} q_{ij} A_{ij} \widehat{\pi}_{ij}^{-1}}. \tag{13}$$

We call this CEDRA-SN and, in the implementation, fall back to the candidate baseline when its denominator is at most the declared numerical floor $\delta = 10^{-12}$. Equation 13 displays the positive-denominator case. It is a Hájek-type random ratio (Swaminathan & Joachims, 2015; Cardoso et al., 2022) and is generally finite-sample biased. A related self-normalized inverse-propensity-scoring ratio has been studied for recommendation evaluation with selectively observed ratings (Schnabel et al., 2016). In particular, the ratio does not inherit the exact propensity-correct branch of Corollary 1; any claim of lower variance must be empirical and accompanied by denominator and effective-sample-size diagnostics.

Appendix H gives an exact positive-denominator ratio identity. Writing N_i and D_i for the residual numerator and inverse-weight denominator, the identity separates the ratio expectation into the ratio of expectations, a $-\operatorname{Cov}(N_i, D_i)/\mathbb{E}[D_i]^2$ term, and an exact higher-order remainder. Its second-order expansion adds $\mathbb{E}[N_i]\operatorname{Var}(D_i)/\mathbb{E}[D_i]^3$. Thus SN can reduce RMSE by damping inverse-weight variation while changing finite-sample bias through numerator-denominator dependence; neither a denominator mean near one nor correct propensities restore the HT branch.

For a declared $0 < \epsilon < 1/2$, replacing $\widehat{\pi}_{ij}$ by its symmetric implementation clip $\min\{1 - \epsilon, \max(\widehat{\pi}_{ij}, \epsilon)\}$ changes the propensity-error factor when clipping is active and can trade variance for bias when the message regression is imperfect. The upper clip is used only for numerical symmetry; inverse-probability weighting requires the lower bound. Clipping does not repair a violation of true positivity, and true overlap must be varied separately from the clipping threshold. Clipped HT still satisfies the conditional bias decomposition with the clipped score substituted for $\widehat{\pi}$, including the covariance term. Under mean ignorability, it normally forfeits the propensity-correct branch wherever clipping changes a correct raw score. Clipped HT and clipped SN are therefore labeled practical stabilizations; SN additionally has random-ratio bias.

Table 5: Experiment-to-estimand map. RMSE in the aggregate studies is computed against the displayed reference object. Prediction studies use label metrics and do not test either aggregate theorem.

Study	Reported reference	Why that reference is available	Applicable result
Controlled synthetic, oracle and learned	$\tau_i(q) = T_i(q)$ pathwise	The generator centers realization noise under q_i	Theorem 1 and Corollary 1 under the declared mean-ignorable mechanism
Optdigits aggregate	Realized $T_i(q)$	The offline evaluator reads all deliberately masked candidate messages	Corollary 2 for the true-propensity arm; other arms are finite-sample estimators compared with the same realized target
Ten-dataset breadth aggregate	Realized $T_i(q)$	The offline evaluator reads all deliberately masked candidate messages	Same design-based interpretation as Optdigits aggregate, with the disclosed seed-offset deviation
Risk-adaptive public-support aggregate	Realized $T_i(q)$	A fixed nonlinear teacher defines selectively evaluated messages on real graph supports	Proposition 1 supplies the model-based risk target; full HT intervals remain separate
Matched-Q3 aggregate diagnostics	Realized $T_i(q)$ on test receivers	A separate full-message evaluator uses the same frozen task-trained pair encoder	Known-probability design-based aggregate comparison, not a predictive guarantee
Matched-Q3 prediction	Test labels	Matched candidate-graph predictors receive one of eight selective aggregates	Isolates aggregation within each backbone; prediction is outside the first-layer identity
Mean-ignorability stress	Conditional $\tau_i(q)$ and realized $T_i(q)$	Synthetic residual signs and acquisition law are fully specified	Tests the covariance remainder, including nonzero c_{ij}
Optdigits prediction	Test labels	Balanced accuracy after deep nonlinear prediction	Theorem-ineligible
Six public graphs	Test labels	Balanced accuracy, macro-F1, and ROC-AUC on fixed graphs	Violates fixed-target construction; descriptive stress test

B.3 Depth and nonlinear prediction

The historical public-graph network projects the message view to a fixed hidden width, computes one candidate score from the construction view before the layer loop, and reuses that score at every layer. A separate construction-view message-regression head is used at each layer, with output dimension equal to that layer's candidate-message width. Although the score is not recomputed from hidden states, later messages and learned maps depend on the realized graph and do not represent the potential first-layer messages in Assumption 5. Its fixed-depth representation and component-zeroing diagnostics are not a depth sweep or causal attribution. The separate matched-Q3 experiment does vary depths $1, 2, 3$, but keeps selective correction at the first-layer interface and propagates only cheap features and already accessible representations afterward. Neither design extends the theorem to hidden-layer corrections, nonlinear activations, normalization, dropout, residual connections, pooling, embeddings, losses, or final predictions.

C Evaluation design details

Controlled mechanism validation. The controlled simulation fixes $\mathcal{C}_i$ and q_i and uses bounded independent Bernoulli evaluation with known probabilities. Its factorial crosses three levels of edge-feature coupling, three levels of label-message alignment, two expected degrees, two community-mixing levels, and two overlap bounds, for $72 \times 20 = 1{,}440$ prespecified condition-by-seed artifacts. A separate $9 \times 20 = 180$-artifact grid fits two-fold cross-fitted nuisances. Its learned Navip, HT, and SN estimators have explicitly named matched unclipped and symmetrically clipped-at-0.02 arms

Table 6: Evidence hierarchy. The assumption-status column distinguishes satisfied, theorem-ineligible, and violated regimes; these labels are not empirical outcomes. A known randomized mechanism can validate the aggregate identity without making its imposed graph naturally observed.

Evidence source	Primary role	Selection-assumption status
Controlled finite-population simulation	Estimator bias/RMSE and nuisance factorial	Satisfied: fixed candidates and bounded Bernoulli evaluation
UCI Optdigits with imposed bounded Bernoulli evaluation	Aggregate-mechanism study over real public features	Satisfied for the imposed evaluation mechanism
Ten UCI feature datasets with imposed bounded Bernoulli evaluation	Aggregate-mechanism breadth extension	Satisfied for the executed evaluation mechanisms
UCI Optdigits with one imposed randomized graph per seed	Transductive held-out prediction over 18 frozen arms	Outside the theorem's scope: deep prediction with learned nuisances
Six public observed graphs (Cora plus five modern heterophily benchmarks)	Predictive stress test under shared protocols	Candidate membership uses observed adjacency, outside fixed-target validation
CEDRA-Guard on six public graphs	Supervised integration with feature-only candidates and a zero-correction path	Predictive objective with deterministic observed graph, outside the first-layer identity
Matched-Q3 on six public graphs	Prediction and held-out aggregate error at matched acquisition budgets	Imposed known evaluation probabilities; only the aggregate endpoint has design-based scope
Synthetic residual-dependent evaluation	Bias and both conditional- and realized-target RMSE over five signed dependence strengths	Mean ignorability holds at zero strength and is deliberately violated otherwise

that reuse the same nuisance fits. The oracle grid includes both-correct, propensity-correct-only, outcome-correct-only, and both-wrong regimes; the misspecified propensity and outcome nuisances are fixed candidate-constant laws set before data generation. Reports use 20,000-resample percentile bootstrap intervals, paired for contrasts, exact two-sided paired sign tests with ties excluded, and Holm adjustment within declared endpoint families. The tests concern sign balance, not a zero mean effect. Artifact validation rejects missing or duplicate runs and verifies the recorded code revision, source tree, protocol, inputs, and outputs. Primary mechanism endpoints are target bias and RMSE; ROC-AUC, calibration, effective sample size (ESS), denominator stability, and clipping are diagnostics rather than theorem endpoints.

Optdigits randomized aggregate study. The UCI Optical Recognition of Handwritten Digits dataset contains 5,620 rows with 64 integer features in official 3,823-row training and 1,797-row writer-separated test files (Alpaydin & Kaynak, 1998). We scale the features without fitted statistics and fix disjoint 32-coordinate checkerboard views: one for construction and one for messages. Using all construction views, including unlabeled test features, we form a directed $K = 40$ candidate pool per receiver and an exponentially rank-weighted q_i. We then impose three known Bernoulli evaluation mechanisms with $0.10 \leq \pi_{ij} \leq 0.50$ and use 20 shared evaluation-draw seeds. This is a controlled selective-evaluation stress test, not a claim of natural missingness: despite full public availability, every retained feasible/non-oracle aggregator deliberately masks messages with $A_{ij} = 0$; arms designated to use the external outcome nuisance use a receiver-separated, non-oracle regression. The separately labeled full-access candidate reference reads all candidate messages. This transductive design is *semi-synthetic randomized message evaluation over real public features*: the evaluation pattern is imposed, not naturally observed, and the satisfied status applies only to its declared finite

candidate population and randomized evaluation-probability law. The aggregate study contains ten prespecified controls and makes no predictive neural claim.

Ten-dataset real-feature randomized aggregate study. The additional breadth panel covers Abalone, Airfoil Self-Noise, Wine Quality white, MAGIC Gamma Telescope, Rice Cammeo and Osmancik, Optdigits, Dry Bean, Letter Recognition, PenDigits, and Landsat Satellite (Nash et al., 1994; Brooks et al., 1989; Cortez et al., 2009; Bock, 2004; Cinar & Koklu, 2019; Alpaydin & Kaynak, 1998; Köklü & Özkan, 2020; Slate, 1991; Alpaydin & Alimoglu, 1996; Srinivasan, 1993). It contains 90,066 rows and spans three regression source tasks, two binary-classification source tasks, and five multiclass-classification source tasks. Numeric inputs are standardized from the deterministic training split, alternating columns provide disjoint construction and message views, and labels do not enter candidates, evaluation, nuisances, aggregate targets, or target-recovery assessment. Each receiver has 40 directed nearest candidates under the construction view, rank-decay reference weights, and one of three imposed evaluation-probability laws with probabilities in $[0.10, 0.50]$. All seven feasible estimators, three mechanisms, ten datasets, and 20 seeds are retained. The executed random-number seed used $51000 + s$ rather than the $10000 + s$ written in the frozen prose protocol. Because that fixed nonselective offset is a protocol deviation, this panel is a complete reproducible breadth study rather than a perfectly preregistered confirmation. The separately generated ridge and logistic downstream probes are exploratory because they do not match the neural-head and GraphSAGE text in the frozen protocol; they are not used for the target-recovery claim.

Optdigits transductive prediction. A separate predictive benchmark reuses the fixed features, candidates, views, and only the prespecified "near" randomized mechanism, with one shared graph for all 18 arms in each of 20 paired seeds. It splits labels into 3,058 model-training, 765 validation, and 1,797 final-test nodes. Conventional arms receive all 64 features; the eight conventional graph baselines also receive the realized graph, whereas MLP receives no graph. Candidate-aware arms receive the disjoint construction/message views and candidate-population inputs. In Appendix Table 21, "joint" means that the outcome-only regression is optimized jointly with its predictor, whereas "external" means that CEDRA consumes a receiver-separated nuisance fit; these labels describe different training pathways, not extra pre-evaluation variables. Because those pathways and information sets are not matched, their comparison cannot isolate residual weighting or establish architectural superiority. Each seed writes its validation-selected checkpoints before a separate evaluator checks all 360 checkpoints, selects the strongest non-CEDRA comparator from validation evidence alone, and loads final-test labels once for aggregate balanced-accuracy evaluation. This is a transductive semi-synthetic predictive stress test, not evidence for the first-layer identity.

Baselines and fairness of comparison. For predictive public-graph comparisons, every method uses the same data split, seed list, early-stopping rule, validation metric, and prespecified configuration. We specified the configuration and compute budget before inspecting outcomes, and the preregistration includes no hyperparameter or model-specific search. The protocol does not provide a separate contemporaneous rationale for every chosen value, and the configuration must not be described as exhaustive or literature-optimal tuning. Inputs are deliberately asymmetric. MLP and every standard graph baseline receive all standardized original features, and graph baselines receive the complete original edge set. Candidate-aware methods instead receive the disjoint construction/message views, candidate ordering, reference masses q_{ij}, and evaluation indicators A_{ij}; neither family is a strict information superset. Within the candidate-aware family, EdgeResid is the strict architectural control: it retains the CEDRA backbone and per-layer outcome regressions but replaces inverse residual weighting by normalized propensity-gated residuals. EdgeGate shares the candidate design, evaluated-message path, and construction-derived evaluation score but omits the outcome-regression pathway, so it is a lower-capacity gating ablation rather than an input- or parameter-matched control. Mandatory comparators are evaluated-message mean, full-access candidate mean where available, MLP, GCN (Kipf & Welling, 2016), GraphSAGE (Hamilton et al., 2017), GAT (Veličković et al., 2017), GATv2 (Brody et al., 2021), GIN (Xu et al., 2018), DropEdge (Rong et al., 2019), LINKX (Lim et al., 2021), H2GCN (Zhu et al., 2020), EdgeGate, the architecturally matched EdgeResid control, a Navip-style adaptation, pure HT, and outcome-only aggregation. All 18 arms apply to all six public datasets; reports include realized total, trainable, and fixed parameter counts. GATv2 and LINKX use PyG 2.6.1 implementations (Fey & Lenssen, 2019). No structure-learning or rewiring result is included;

Table 7: Accepted controlled summaries. Factorial rows average all 1,440 condition-by-seed artifacts, and learned rows average all 180 artifacts. ROC-AUC is a generator diagnostic rather than independent predictive validation.

Study and method	Bias	RMSE	ROC-AUC
Factorial, observed mean	0.001	0.312	0.961
Factorial, CEDRA-HT with both nuisances correct	0.001	0.127	0.996
Factorial, CEDRA-HT with outcome nuisance correct	0.001	0.045	1.000
Factorial, CEDRA-HT with propensity nuisance correct	−0.001	0.509	0.966
Factorial, CEDRA-HT with both nuisances wrong	0.199	0.535	0.957
Learned, cross-fitted Navip	0.002	0.306	0.977
Learned, cross-fitted CEDRA-HT	0.002	0.189	0.988
Learned, cross-fitted CEDRA-SN	0.002	0.159	0.990

Table 8: Accepted Optdigits aggregate RMSE means over 20 paired randomized seeds. Bold marks the lowest feasible non-oracle value in each mechanism. Candidate construction is transductive and the graph is imposed semi-synthetically over one fixed public feature table.

Method	Near	Constant	Far
Observed mean	0.0424	0.0506	**0.0586**
Observed-q ratio	0.0403	0.0504	0.0616
Navip ratio	0.0374	0.0504	0.0725
CEDRA-HT with external outcome model	0.0388	0.0532	0.0887
CEDRA-SN with external outcome model	**0.0328**	**0.0443**	0.0635

that missing comparison remains a limitation rather than evidence that such methods recover the same object.

Statistics and reporting. Controlled and predictive analyses use their complete prespecified paired seed sets. Reports for predictive comparisons include seed-level paired effects against the eligible non-CEDRA comparator selected on validation under the shared prespecified configuration, with the Navip-style adaptation always visible. Public summaries are over algorithm/predefined-split run cells on each fixed graph, not population-graph inference. The heterophily split columns reuse nodes from the same graph, and Cora reuses one split, so independence of public paired signs is not established; public bootstrap intervals and sign-test outputs are reported only as descriptive run-cell variability. Optdigits aggregate contrasts are fixed by the aggregate protocol. Its predictive protocol separately defines one ordered family of six contrasts against the validation-selected comparator, candidate mean, and Navip. For the independently randomized synthetic and Optdigits seed designs, contrasts use paired bootstrap intervals, exact two-sided paired sign tests (testing sign balance, not a zero mean), and Holm correction within declared families. Missing, failed, or duplicated seeds fail the aggregation step. Parameter counts, training and total accelerator time, peak memory, and inference latency accompany predictive results.

D PROBABILITY SPACE AND NOTATION

For each fixed receiver i, $\mathcal{F}_i$ contains the finite candidate set, reference weights, and all declared pre-evaluation adjustment information. The draw $J \sim q_i$ is auxiliary: conditionally on $\mathcal{F}_i$, $\Pr(J = j) = q_{ij}$ and J is independent of $\{(A_{ij}, M_{ij}) : j \in \mathcal{C}_i\}$. Expectations and covariances with subscript q_i integrate over J; other expectations integrate over the evaluation/message data-generating law. For vector messages, expectations and covariance with a scalar are componentwise. All displayed conditional expectations are assumed finite.

The target is conditional on the declared design information. If the target instead averages over random candidate-set construction, that additional randomness must be specified and the present conditional identity integrated over it. If q_i or $\mathcal{C}_i$ is selected after seeing evaluated edges or test labels, Assumption 1 fails.

Table 9: Complete Optdigits predictive balanced accuracy over 20 paired randomized graph seeds. The comparator was selected using validation evidence only. Effects are CEDRA minus outcome-only, so negative values favor the comparator.

Method	Mean	Paired effect	Holm p
Outcome-only, selected comparator	0.9688		
CEDRA-HT	0.9626	-0.00611	2.40×10^{-4}
CEDRA-SN	0.9630	-0.00576	0.00201

Table 10 collects this notation for reference.

Table 10: Notation used throughout the paper.

Symbol	Meaning
$\mathcal{C}_i$	Finite candidate support for receiver i
q_{ij}	Reference weight for candidate j; $q_i = (q_{ij})_j$ sums to one and defines the reference law
A_{ij}	Evaluation indicator; $A_{ij} = 1$ iff message ij is evaluated
M_{ij}	Potential first-layer message, observed only if $A_{ij} = 1$
W_{ij}	Cheap pre-evaluation information available for every candidate
$\pi^\star_{ij}$	True evaluation probability, $\pi^\star_{ij} := \Pr(A_{ij} = 1 \mid \mathcal{H}_i)$
m_{ij}	True conditional message mean, $m_{ij} := \mathbb{E}[M_{ij} \mid \mathcal{H}_i]$
$\widehat{\pi}_{ij}$, $\widehat{\mu}_{ij}$	Fitted propensity and message-mean nuisance models, $\widehat{\pi}(W_{ij})$ and $\widehat{\mu}(W_{ij})$
$\tau_i(q)$	Estimand: declared conditional-mean aggregate, $\tau_i(q) := \sum_j q_{ij} m_{ij}$
$T_i(q)$	Realized aggregate under full evaluation, $T_i(q) := \sum_j q_{ij} M_{ij}$, with conditional mean $\tau_i(q)$
$\mathcal{F}_i$	Pre-evaluation filtration: candidate law and declared adjustment information
$\mathcal{H}_i$	Larger filtration including nuisance-training information, $\mathcal{H}_i = \mathcal{F}_i \vee \mathcal{G}$
c_{ij}	Signed covariance remainder, $c_{ij} := \mathrm{Cov}(A_{ij}, M_{ij} \mid \mathcal{H}_i)$
λ	Risk-adaptive blending coefficient; optimal value $\lambda^\star_i = S_{0i}/(S_{0i} + S_{1i})$
S_{0i}	$\sum_j q^2_{ij}$, reference-law dispersion term
S_{1i}	$\sum_j q^2_{ij}(1 - \pi_{ij})/\pi_{ij}$, evaluation-variance term

E PROOF OF THE SELECTION IDENTITY

Proof of Proposition 2. **Step 1 (conditioning setup).** Condition on $\mathcal{F}_i$. The finite candidate set, q_i, and the conditional pair laws are fixed. The auxiliary draw J and the jointly distributed edge/message array remain random, with J conditionally independent of that array.

Step 2 (integrate over the auxiliary candidate). By the conditional law of J and Equation 8,

$$\Pr(A_{iJ} = 1 \mid \mathcal{F}_i) = \sum_{j \in \mathrm{supp}(q_i)} q_{ij} \Pr(A_{ij} = 1 \mid \mathcal{F}_i) = \sum_{j \in \mathrm{supp}(q_i)} q_{ij} \pi^\star_{ij},$$

$$\begin{aligned} \mathbb{E}[M_{iJ} \mathbf{1}\{A_{iJ} = 1\} \mid \mathcal{F}_i] &= \sum_{j \in \mathrm{supp}(q_i)} q_{ij} \mathbb{E}[A_{ij} M_{ij} \mid \mathcal{F}_i] \\ &= \sum_{j \in \mathrm{supp}(q_i)} q_{ij} \pi^\star_{ij} m_{ij}, \end{aligned}$$

Step 3 (condition on selection). Dividing the second identity by the positive first identity proves Equation 9.

Step 4 (identify candidate-law covariance). Subtracting $\tau_i(q) = \mathbb{E}_{q_i}[m_{iJ} \mid \mathcal{F}_i]$ gives

$$\frac{\mathbb{E}_{q_i}[\pi^\star_{iJ} m_{iJ} \mid \mathcal{F}_i]}{\mathbb{E}_{q_i}[\pi^\star_{iJ} \mid \mathcal{F}_i]} - \mathbb{E}_{q_i}[m_{iJ} \mid \mathcal{F}_i] = \frac{\mathbb{E}_{q_i}[\pi^\star_{iJ} m_{iJ} \mid \mathcal{F}_i] - \mathbb{E}_{q_i}[\pi^\star_{iJ} \mid \mathcal{F}_i] \mathbb{E}_{q_i}[m_{iJ} \mid \mathcal{F}_i]}{\mathbb{E}_{q_i}[\pi^\star_{iJ} \mid \mathcal{F}_i]},$$

which is Equation 10 by the definition of candidate-law covariance over J. This covariance is distinct from the fixed-pair evaluation-message covariance c_{ij} in Equation 3. □

F Derivation for the failed evaluated-message-only formula

Restricting the regression baseline to evaluated candidates gives

$$\widetilde{\tau}_i^{\mathrm{obs}} = \sum_{j \in \mathrm{supp}(q_i)} q_{ij} A_{ij} \left\{ \widehat{\mu}_{ij} + \frac{M_{ij} - \widehat{\mu}_{ij}}{\widehat{\pi}_{ij}} \right\}. \tag{14}$$

Even under mean ignorability and a correct message model, its expectation is $\sum_j q_{ij} \pi^\star_{ij} m_{ij}$ rather than $\tau_i(q)$ because evaluation still selects the baseline. CEDRA moves that baseline outside the evaluated-message restriction.

Step 1 (conditioning setup). Condition on $\mathcal{H}_i = \mathcal{F}_i \vee \mathcal{G}$ and additionally impose the mean-ignorability condition of Corollary 1. The nuisance values and conditional first moments are fixed; (A_{ij}, M_{ij}) remains random.

Step 2 (expand one candidate's conditional expectation).

$$\begin{aligned}\mathbb{E}\left[A_{ij} \left\{ \widehat{\mu}_{ij} + \frac{M_{ij} - \widehat{\mu}_{ij}}{\widehat{\pi}_{ij}} \right\} \middle| \mathcal{H}_i \right] &= \pi^\star_{ij} \widehat{\mu}_{ij} + \frac{1}{\widehat{\pi}_{ij}} \left\{ \mathbb{E}[A_{ij} M_{ij} \mid \mathcal{H}_i] - \widehat{\mu}_{ij} \mathbb{E}[A_{ij} \mid \mathcal{H}_i] \right\} \\ &= \pi^\star_{ij} \widehat{\mu}_{ij} + \frac{\pi^\star_{ij}}{\widehat{\pi}_{ij}} (m_{ij} - \widehat{\mu}_{ij}).\end{aligned}$$

Step 3 (evaluate the nuisance branches and sum). If $\widehat{\mu}_{ij} = m_{ij}$, the right-hand side equals $\pi^\star_{ij} m_{ij}$; if $\widehat{\pi}_{ij} = \pi^\star_{ij}$, it equals $m_{ij} + (\pi^\star_{ij} - 1)\widehat{\mu}_{ij}$. Neither branch generally equals m_{ij}. Summing with q_{ij} gives the expectation stated in Section 4 and proves that the evaluated-message-only baseline destroys the usual two-branch identity.

G Complete proof of the conditional bias decomposition

Proof of Theorem 1. **Step 1 (conditioning setup).** Write $\mathcal{H}_i := \mathcal{F}_i \vee \mathcal{G}$ and condition on it. By Assumptions 1, 2, and 4, q_{ij}, $\widehat{\mu}_{ij}$, and $\widehat{\pi}_{ij}$ are fixed under conditioning on $\mathcal{H}_i$, and the conditional first moments remain $\pi^\star_{ij}$ and m_{ij}. The joint edge/message array is the remaining random object; no across-candidate factorization is introduced.

Step 2 (evaluate one candidate). Equation 3 gives $\mathbb{E}[A_{ij} M_{ij} \mid \mathcal{H}_i] = \pi^\star_{ij} m_{ij} + c_{ij}$, hence

$$\begin{aligned}\mathbb{E}\left[\frac{A_{ij}}{\widehat{\pi}_{ij}} (M_{ij} - \widehat{\mu}_{ij}) \middle| \mathcal{H}_i \right] &= \frac{1}{\widehat{\pi}_{ij}} \left\{ \mathbb{E}[A_{ij} M_{ij} \mid \mathcal{H}_i] - \widehat{\mu}_{ij} \mathbb{E}[A_{ij} \mid \mathcal{H}_i] \right\} \\ &= \frac{\pi^\star_{ij}}{\widehat{\pi}_{ij}} (m_{ij} - \widehat{\mu}_{ij}) + \frac{c_{ij}}{\widehat{\pi}_{ij}}.\end{aligned}$$

Step 3 (sum over the finite pool). Linearity of conditional expectation over the fixed finite pool, without any pairwise edge-independence assumption, gives

$$\mathbb{E}[\widehat{\tau}_i^{\mathrm{HT}} - \tau_i(q) \mid \mathcal{H}_i] = \sum_{j \in \mathrm{supp}(q_i)} q_{ij} \left\{ \widehat{\mu}_{ij} + \frac{\pi^\star_{ij}}{\widehat{\pi}_{ij}} (m_{ij} - \widehat{\mu}_{ij}) + \frac{c_{ij}}{\widehat{\pi}_{ij}} - m_{ij} \right\}.$$

Step 4 (regroup the remainder). Candidatewise algebra yields

$$\mathbb{E}[\widehat{\tau}_i^{\mathrm{HT}} - \tau_i(q) \mid \mathcal{H}_i] = \sum_{j \in \mathrm{supp}(q_i)} q_{ij} \left\{ \frac{\pi^\star_{ij} - \widehat{\pi}_{ij}}{\widehat{\pi}_{ij}} (m_{ij} - \widehat{\mu}_{ij}) + \frac{c_{ij}}{\widehat{\pi}_{ij}} \right\},$$

where division is valid by the evaluated-score clause $\widehat{\pi}_{ij} > 0$ in Assumption 3. This proves Equation 4 without imposing mean ignorability. □

Proof of Corollary 1. **Step 1 (remove the covariance term).** Under mean ignorability, every c_{ij} in Equation 4 is zero, leaving Equation 5.

Step 2 (verify the two nuisance branches). If either nuisance is pointwise correct on $\mathrm{supp}(q_i)$, one factor in every candidate's product remainder is zero, so the complete finite sum is zero. □

Proof of Corollary 2. Continue under the theorem's assumptions. **Step 1 (condition on realized candidate messages).** Condition on $\mathcal{K}_i$. Each M_{ij} and $\widehat{\mu}_{ij}$ is fixed, while only the jointly distributed candidate-edge vector remains random and $\mathbb{E}[A_{ij} \mid \mathcal{K}_i] = \pi^\star_{ij}$. With $\widehat{\pi}_{ij} = \pi^\star_{ij}$,

$$\mathbb{E}\left[\widehat{\mu}_{ij} + \frac{A_{ij}}{\pi^\star_{ij}}(M_{ij} - \widehat{\mu}_{ij}) \,\middle|\, \mathcal{K}_i\right] = M_{ij}.$$

Step 2 (sum without factorization). Multiplication by the fixed q_{ij} and finite summation gives Equation 11; no factorization across candidates is used. □

Proof of Corollary 3. **Step 1 (condition and fix the affine map).** Because G_i and b_i are fixed under conditioning on $\mathcal{H}_i$, the only remaining randomness is in $\widehat{\tau}_i^{\mathrm{HT}}$.

Step 2 (apply linearity). The two copies of b_i cancel and linearity of conditional expectation gives

$$\mathbb{E}[L_i(\widehat{\tau}_i^{\mathrm{HT}}) - L_i(\tau_i(q)) \mid \mathcal{H}_i] = G_i\,\mathbb{E}[\widehat{\tau}_i^{\mathrm{HT}} - \tau_i(q) \mid \mathcal{H}_i],$$

which is Equation 12. □

Remark 1 (What the theorem does not condition away)**.** *Treating fitted nuisances as numerically fixed is not enough if their training data reveal the evaluated edge or message residual. The preserved-moment condition in Assumption 2 is substantive. Receiver-level or independent-graph cross-fitting is safer than edge-wise splitting when edges share nodes, but even it must be justified for the data-generating dependence.*

H WHY SELF-NORMALIZATION HAS A DIFFERENT BIAS EQUATION

Step 1 (define the numerator and denominator). Let

$$N_i = \sum_{j \in \mathrm{supp}(q_i)} q_{ij} A_{ij} \widehat{\pi}_{ij}^{-1}(M_{ij} - \widehat{\mu}_{ij}), \qquad D_i = \sum_{j \in \mathrm{supp}(q_i)} q_{ij} A_{ij} \widehat{\pi}_{ij}^{-1}.$$

For a declared denominator threshold $\delta \geq 0$, define $R_i = N_i/D_i$ when $D_i > \delta$ and $R_i = 0$ otherwise. Even when $\widehat{\pi} = \pi^\star$, generally $\mathbb{E}[R_i] \neq \mathbb{E}[N_i]/\mathbb{E}[D_i]$ because N_i and D_i share the evaluation indicators.

Step 2 (exact identity on the positive-denominator event). Let $E_i = \{D_i > \delta\}$ and, for this step, let $\mathbb{E}_+[\cdot] = \mathbb{E}[\cdot \mid E_i, \mathcal{H}_i]$, $\mu_N^+ = \mathbb{E}_+[N_i]$, and $\mu_D^+ = \mathbb{E}_+[D_i]$. The scalar identity $D_i^{-1} = (\mu_D^+)^{-1} - (D_i - \mu_D^+)/(\mu_D^+)^2 + (D_i - \mu_D^+)^2/\{(\mu_D^+)^2 D_i\}$ gives

$$\mathbb{E}[R_i \mid \mathcal{H}_i] = \Pr(E_i \mid \mathcal{H}_i)\left\{\frac{\mu_N^+}{\mu_D^+} - \frac{\mathrm{Cov}_+(N_i, D_i)}{(\mu_D^+)^2} + \mathbb{E}_+\left[\frac{N_i(D_i - \mu_D^+)^2}{(\mu_D^+)^2 D_i}\right]\right\}. \tag{15}$$

This is exact and retains the fallback probability. When the positive-denominator event has probability near one and fluctuations are small, a second-order delta expansion is

$$\mathbb{E}[R_i \mid \mathcal{H}_i] \approx \frac{\mu_N}{\mu_D} - \frac{\mathrm{Cov}(N_i, D_i \mid \mathcal{H}_i)}{\mu_D^2} + \frac{\mu_N \,\mathrm{Var}(D_i \mid \mathcal{H}_i)}{\mu_D^3}, \tag{16}$$

where $\mu_N = \mathbb{E}[N_i \mid \mathcal{H}_i]$ and $\mu_D = \mathbb{E}[D_i \mid \mathcal{H}_i]$. The approximation is descriptive rather than a guarantee unless explicit moment, denominator, and dependence conditions control the omitted higher-order terms.

Step 3 (evaluate a one-candidate case). Take $q_{i1} = 1$, $M_{i1} = m \neq 0$, $\widehat{\mu}_{i1} = 0$, and $0 < p < 1$, with $A_{i1} \sim \mathrm{Bernoulli}(p)$ and $\widehat{\pi}_{i1} = \pi^\star_{i1} = p$. With baseline fallback when $D_i = 0$, only A_{i1} remains random. The SN estimate is m if $A_{i1} = 1$ and 0 otherwise.

Step 4 (compare expectations). The SN expectation is $pm \neq m$, whereas the HT estimate is $A_{i1}m/p$ and has expectation m. This example does not argue that SN is always inferior; it shows only that its finite-sample stability must not be advertised as the exact mean-ignorable, propensity-correct guarantee of Corollary 1 for CEDRA-HT.

I IMPLEMENTATION CONTRACT

A conforming first-layer implementation must satisfy the following checks.

1. Construct $\mathcal{C}_i$ and normalized q_i without evaluated labels or realized candidate-edge indicators.
2. Produce $\widehat{\mu}_{ij}$ for every $j \in \mathcal{C}_i$; produce residual terms only where $A_{ij} = 1$.
3. Keep tensor orientation receiver-major and verify $\sum_{j \in \mathcal{C}_i} q_{ij} = 1$ for every receiver.
4. Separate true design overlap from estimated-score clipping and log the unclipped and clipped weight distributions.
5. Use nuisance folds whose held-out unit matches the conditional independence justification; store fold membership in the artifact.
6. Test the candidate-baseline path, the failed evaluated-message-only formula, each correct-nuisance branch, the both-wrong case, HT versus SN, zero denominators, and component shapes.
7. Label later-layer reuse or recomputation of any score as practical and theorem-ineligible; expose HT, SN, clipping, and depth as separate choices.
8. Record the message-view input projection, the one construction-view propensity head reused across layers, and the separate construction-view regression head for every layer. Verify that each regression output matches its candidate-message width and that no outcome head receives that message or hidden vector as an input.

J EXPANDED LIMITATIONS

Estimand and identification. The reference law q_i is chosen by the application, not discovered by the estimator. Moving toward it can discard useful selected-edge signal, weaken locality, or harm prediction. The theory also requires a fixed candidate law, stable conditional moments, target-support positivity, separated nuisances, and, for the two-branch corollary, mean ignorability. These conditions are generally unverified on observational graphs. A deterministic k-NN graph violates positivity for any target that retains mass outside its selected neighbors; redefining q_i on those neighbors changes the estimand rather than validating the theorem. If all candidate messages can be evaluated directly, correction is unnecessary and the candidate-wide baseline adds cost.

Operator scope. The exact statement concerns the unnormalized first-layer pre-transform HT aggregate. SN is a biased random ratio; active clipping generally removes the propensity-correct branch; candidate subsampling introduces another design; and nuisance reuse requires the stated conditional-moment justification. The guarantee does not pass automatically through hidden layers, learned maps, normalization, nonlinearities, pooling, or prediction. A small residual can be the intended consequence of a good outcome regression, but it does not show that a broad candidate baseline preserves local signal or prevents collapse. Component-zeroing and representation diagnostics probe these concerns without providing causal attribution or a non-collapse guarantee.

Empirical coverage. The accepted results are endpoint and mechanism dependent. In the learned controlled suite, self-normalization improved RMSE without a statistically distinguishable improvement in absolute bias or ROC-AUC after Holm correction, and clipping at 0.02 produced only small RMSE changes. In Optdigits, the near-mechanism primary RMSE gain for CEDRA-HT coexists with higher MAE than the evaluated-q ratio and higher RMSE than Navip; under the far mechanism, the unweighted evaluated-message mean outperformed CEDRA-SN. In the predictive Optdigits study, outcome-only outperformed both CEDRA variants. In the historical public study, the best CEDRA balanced accuracy trailed the selected comparator on every graph, although CEDRA-SN improved over direct candidate aggregation on all six and improved Amazon-ratings macro-F1 over the selected arm. The matched-Q3 study resolves the information mismatch for its own aggregation comparisons, not retrospectively for those earlier studies. Its final balanced-head results lift Tolokers across all aggregation rules, but Minesweeper remains near chance and outcome-only is competitive. The head protocol was developed on the same datasets and seeds, so these scores are not independent confirmation or evidence of CEDRA-specific superiority. The evidence supports recovery gains in specified

regimes rather than universal predictive or recovery dominance. The Optdigits intervals quantify evaluation-draw and optimization variability on one fixed feature table, while matched-Q3 intervals quantify three paired seed units per fixed dataset. Neither establishes causal, writer-population, or population-graph inference.

Tested sensitivities and remaining comparisons. The matched extension evaluates candidate sizes $10, 20, 40, 80, 160$, uniform and scaled-rank exponential laws of strengths $0.5, 1, 2$, evaluation fractions $0.1, 0.2, 0.3, 0.5, 0.8, 1$, two SAGE/GIN-style candidate-graph backbones, and depths $1, 2, 3$. A separate synthetic experiment varies a known signed nonzero c_{ij} through residual-dependent evaluation. These are additive sweeps on previously examined datasets, not a full factorial, untouched-dataset confirmation, or natural selective acquisition. The $K = 40$ nuisance pilot is reused beyond rank 40, changing q_i changes the target, and increasing K at a fixed fraction increases query count. There is still no verified naturally selective log, kernel-bandwidth sweep, structure-learning comparator, or broad search over optimizers, losses, widths, outcome/nuisance capacity, and cross-fitting folds. The historical router was removed rather than validated. Consequently, the tested ranges do not establish broad architectural, task-optimality, scalability, or hyperparameter robustness.

Controls and diagnostics that remain separate. The controlled synthetic grid contains neither the promised matched-input EdgeGate nor a total-parameter-matched neural control; analogous controls on public graphs do not close that gap. Controlled degree and mixing interventions are separate from the public neural-collapse diagnostics, so the study does not measure neural collapse under those interventions. It also does not provide the promised decompositions of candidate-baseline variance and inverse-weight variance. We retain these points as limitations rather than treating adjacent diagnostics as substitutes.

K Detailed experiment record and reporting plan

K.1 Matched-Q3 selective-evaluation protocol

Data, pairing, and declared population. The study retains Cora, Roman-empire, Amazon-ratings, Minesweeper, Tolokers, and Questions with seeds $1, 2, 3$. Cora uses the loader's fixed split column 0, including its completion of unassigned nodes into training; the other graphs use supplied split columns $1, 2, 3$ with the corresponding seeds. The replicate is the complete split/initialization/mask-seed unit, not a node or candidate pair. The datasets were studied previously, so this is a newly frozen matched comparison rather than confirmation on untouched graphs. Training-node statistics standardize source features, with even-indexed columns forming the cheap view and odd-indexed columns the expensive sender view. Cosine similarity in the cheap view defines one stable top-160 non-self candidate ordering, with sender index breaking ties; smaller pools are its prefixes. Realized adjacency and labels do not determine candidate membership. Candidate construction and later propagation are transductive through unlabeled cheap features.

Shared exact messages and nuisance predictions. Training receivers are divided into a stratified calibration half and a disjoint nuisance-training half before fitting. A nonlinear pair encoder uses cheap receiver features and expensive sender features; it is task-trained for 40 epochs on the calibration half with full $K = 40$ uniform-reference messages and then frozen. These full calibration queries are shared training costs, not free test-time information. A separate $K = 40$, expected-budget-0.5 pilot supplies evaluated messages for the outcome model on nuisance-training receivers. Three stratified receiver folds produce out-of-fold predictions there; a fit on all nuisance-training receivers predicts calibration, validation, and test receivers. Encoder and nuisance training use fixed epoch counts, not test or validation checkpoint selection. All applicable arms and all sweeps reuse the same frozen encoder and nuisance predictions. A counted selected-pair interface is the only source of exact messages for feasible arms; complete reference arrays are separate evaluator inputs. Test messages are never nuisance-fitting targets, and the expensive feature view never enters candidate construction, nuisance inputs, or the predictor's self-feature input. Calibration receivers are not treated as held-out evidence for their own encoder.

Table 11: Matched-Q3 aggregation registry. All budget-feasible methods have the same available information and applicable frozen nuisances; the full reference uses all exact messages. A ratio with no observed denominator contributes zero, leaving a candidate baseline intact when present.

Arm	Aggregate
Evaluated-q	$\left(\sum_j q_{ij}A_{ij}M_{ij}\right)/\left(\sum_j q_{ij}A_{ij}\right)$
Navip-style	$D_i^{-1}\sum_j q_{ij}A_{ij}M_{ij}/\pi_{ij}$
HT/IPW	$\sum_j (q_{ij}A_{ij}/\pi_{ij})M_{ij}$
Model-only	B_i
EdgeResid	$B_i + \dfrac{\sum_j q_{ij}A_{ij}\pi_{ij}(M_{ij}-\widehat{\mu}_{ij})}{\sum_j q_{ij}A_{ij}\pi_{ij}}$
CEDRA-HT	$B_i + R_i$
CEDRA-SN	$B_i + R_i/D_i$
CEDRA-RA	$B_i + \lambda_i^\star R_i$, with Equation 6's design-known weight
Full-message reference	$\sum_j q_{ij}M_{ij}$, using all exact messages

Reference law and acquisition. For candidate rank $r = 0, \ldots, K-1$, define $t_r = r/\max(K-1, 1)$. The reference laws are uniform or normalized scaled-rank exponentials,

$$q_{ir}^{\text{unif}} = \frac{1}{K}, \qquad q_{ir}^{(a)} = \frac{\exp(-at_r)}{\sum_{s=0}^{K-1}\exp(-at_s)}, \qquad a \in \{0.5, 1, 2\}. \tag{17}$$

The primary law has $a = 1$. This is not the older unscaled $\exp(-r/10)$ law. For a requested evaluation fraction $b < 1$, deterministic bisection chooses $c_{K,b}$ so that

$$\pi_{ir}(b) = \text{sigmoid}(c_{K,b} + 2(1-2t_r)), \qquad \frac{1}{K}\sum_{r=0}^{K-1}\pi_{ir}(b) = b. \tag{18}$$

At $b = 1$, every probability is exactly one. No propensity clipping, forced observation, empty-receiver exclusion, or mask resampling is applied. One independently seeded uniform array of size number-of-nodes by 160 is shared across methods and settings, with $A_{ir}(b) = \mathbf{1}\{U_{ir} < \pi_{ir}(b)\}$. Masks are nested across budgets at fixed K; changing K reuses prefix uniforms but recomputes normalized ranks and probabilities. The budget is an expected Bernoulli fraction, not an exact fixed query count. All propensity-based methods receive the known marginal design probability, so their comparison does not confound aggregation with separately fitted evaluation models.

Exact controls. Let $B_i = \sum_j q_{ij}\widehat{\mu}_{ij}$, $R_i = \sum_j q_{ij}A_{ij}(M_{ij}-\widehat{\mu}_{ij})/\pi_{ij}$, and $D_i = \sum_j q_{ij}A_{ij}/\pi_{ij}$, with all sums over the declared pool. The eight feasible arms and the full-cost reference are fixed before execution: Navip-style is a q-weighted ratio adaptation, not a reproduction of the full recommender architecture. EdgeResid uses the same propensity scores as non-inverse residual gates and the same outcome predictions as CEDRA, without an additional learned gate. Evaluated-q, Navip-style, and pure HT return zero for empty observations; the outcome-baseline methods retain B_i. At full evaluation, every non-model-only aggregate equals the full candidate aggregate to numerical tolerance, an implementation identity rather than a predictive superiority test. The reference has a different acquisition cost and is excluded from the feasible-arm inferential family.

Matched predictors. Within each backbone, all arms share architecture, dimensions, normalization, dropout, initialization, optimizer, minibatch order, epoch ceiling, and stopping rule; random streams are reset so training one arm does not alter the next arm's initialization or dropout. Each arm trains its own predictor, so its final fitted weights can differ. The exact-message width is 32, the cheap projection width 16, and predictor and nuisance hidden widths 64. The pair encoder has tanh receiver/sender maps followed by a GELU interaction network; the outcome model has two GELU hidden layers and trains for 25 epochs. Pair batches contain 4,096 candidates and calibration receiver batches 128 receivers. AdamW uses learning rate 0.001, weight decay 0.0001, and gradient-norm clipping at 2. Prediction runs for at most 100 epochs with patience 20 and dropout 0.2; validation balanced accuracy chooses the checkpoint, with the earliest epoch breaking exact ties. SAGE-style updates concatenate self and neighbor states; GIN-style updates use a learned epsilon and a sum

before a width-64 two-linear-layer ReLU update. Both use LayerNorm and the same linear classifier. Later layers consume only cheap features and already accessible representations. They are matched candidate-graph backbones, not canonical native-adjacency reproductions or jointly end-to-end learned encoder/nuisance systems.

Completion and provenance. The frozen additive design contains 18 dataset-seed tasks, 540 setting-backbone configurations, and 4,860 arm-output slots. Exact-input fits may be reused only with recorded identity, so output slots are not necessarily distinct fits. The complete analysis combines five original complete tasks with 13 predetermined recovery tasks under unchanged scientific source and seed definitions after an operational failure; failed and partial attempts are preserved but contribute no scores. All source identities, complete task receipts, output-tree hashes, and the final full-stress receipt were checked before manuscript integration. There was no effect-dependent filtering, arm removal, or dataset/seed replacement.

K.1.1 Balanced-head protocol and complete results

Protocol. For Minesweeper and Tolokers, every aggregation arm uses the same train-only balanced cross-entropy with $w_c = n_{\text{train}}/(2n_{\text{train},c})$, balanced-validation-NLL checkpoint selection and stopping, and a threshold maximizing validation balanced accuracy. The epoch ceiling is 100 with patience 20; threshold ties are resolved by proximity to zero and then the lower threshold. The pair encoder, nuisance models, candidates, features, reference law, acquisition masks and probabilities, public splits, initialization, and aggregation arms are held fixed. Every checkpoint and threshold choice is recorded before held-out scoring. The protocol was developed on these same benchmarks, so the results are exploratory rather than independent confirmation. No best-test policy, last-epoch test score, test-set threshold search, checkpoint reselection, arm filtering, or new hypothesis test is used.

Table 12: Complete balanced-head results. Test balanced accuracy is mean ± sample SD in percent over three paired seeds. Every method uses the same head protocol and frozen upstream inputs. The full-message row remains an information-rich reference, not a budget-feasible arm.

Method	Minesweeper SAGE	Minesweeper GIN	Tolokers SAGE	Tolokers GIN
Evaluated-q	50.43 ± 1.86	51.49 ± 0.05	68.36 ± 0.74	67.84 ± 0.39
Navip-style	49.71 ± 1.96	51.35 ± 0.48	68.37 ± 0.69	68.21 ± 0.58
HT/IPW	50.94 ± 0.70	51.20 ± 0.21	68.40 ± 0.68	68.56 ± 0.80
Model-only	51.07 ± 0.78	51.49 ± 0.47	68.67 ± 0.75	68.53 ± 0.79
EdgeResid	50.04 ± 0.29	51.38 ± 0.61	68.16 ± 0.59	68.24 ± 0.45
CEDRA-HT	50.74 ± 0.36	50.58 ± 1.83	68.06 ± 0.54	68.28 ± 0.65
CEDRA-SN	50.86 ± 0.81	50.87 ± 0.95	68.42 ± 0.84	68.20 ± 0.67
CEDRA-RA	50.10 ± 1.03	50.22 ± 1.30	68.09 ± 0.56	68.15 ± 0.61
Full-message reference	51.05 ± 0.70	51.34 ± 0.97	68.56 ± 0.72	68.24 ± 0.67

Result. Every final dataset-backbone-arm group has zero constant-class seeds. Tolokers scores range from 67.84% to 68.67%, with CEDRA-HT at 68.06% for SAGE and 68.28% for GIN. Minesweeper remains near chance, ranging from 49.71% to 51.49%. Outcome-only and HT/IPW are competitive with the CEDRA variants, so the results do not establish CEDRA-specific predictive dominance.

K.2 Matched candidate-size, reference-law, budget, and depth sweeps

The reference setting is $K = 40$, expected budget $b = 0.3$, exponential-rank strength $a = 1$, and depth 2. The additive matrix varies $K \in \{10, 20, 40, 80, 160\}$, reference law in uniform and $a \in \{0.5, 1, 2\}$, $b \in \{0.1, 0.2, 0.3, 0.5, 0.8, 1\}$, or depth in $\{1, 2, 3\}$ while keeping the other reference values fixed. Deduplicating the reference gives 15 settings per dataset-seed-backbone combination, not a full Cartesian interaction study. Both backbones, all seeds, every feasible arm, and each setting's full-message reference are retained.

For candidate size, the measured diagnostics include balanced accuracy, aggregate RMSE/MAE, actual pair-query count, effective candidate count $1/\sum_j q_{ij}^2$, prediction-head inference time and memory, and nuisance error by rank. A fixed evaluation fraction does not fix absolute query count, and predictions at ranks above 40 extrapolate beyond the shared nuisance pilot. For the reference-law sweep, target errors and full-message prediction refer to each law's own $T_i(q)$; a change in accuracy is not an estimator comparison holding the estimand constant. These rank-law tests do not constitute a kernel-bandwidth sweep. Depth includes the first selective update and subsequent propagation layers; correction is not newly injected at every hidden layer, and no deeper-layer double-robust theorem is claimed.

Budget curves use realized test query fractions on the horizontal axis and retain empty receivers. The prespecified area summary instead uses the requested budgets as coordinates for an unnormalized trapezoidal integral over $[0.1, 1.0]$; these definitions are not silently interchanged. The archived analysis reports every fixed budget-area contrast. For HT minus model-only, the SAGE area difference is 0.0102013 with paired-seed interval $[0.0045253, 0.0154964]$; GIN is 0.00804405 with interval $[-0.00207132, 0.0180792]$. These are balanced-accuracy-fraction times budget-fraction units, not accuracy at one budget, and are descriptive rather than a new confirmatory family. Raw records also retain realized query counts, zero-observation frequencies, unique queried senders, denominator/ESS diagnostics, and the separate full-access identity check. Shared calibration and pilot acquisition are accounted separately from test queries; prediction-head timing omits parts of the pipeline and cannot establish overall acceleration.

Figure 2 gives the pooled budget sweep in the main text, Figure 3 summarizes the candidate-size, reference-law, and depth sweeps across the six fixed graphs, and Figures 4–11 retain every graph separately, including unfavorable outcomes. The shaded regions in pooled summaries are pointwise paired-seed intervals, whereas the per-graph regions show sample standard deviations; neither is a simultaneous confidence band or a new confirmatory test family. Model-only requires no test-message queries, and the full-message reference always requires all test messages, even when drawn at a selective-budget position for comparison.

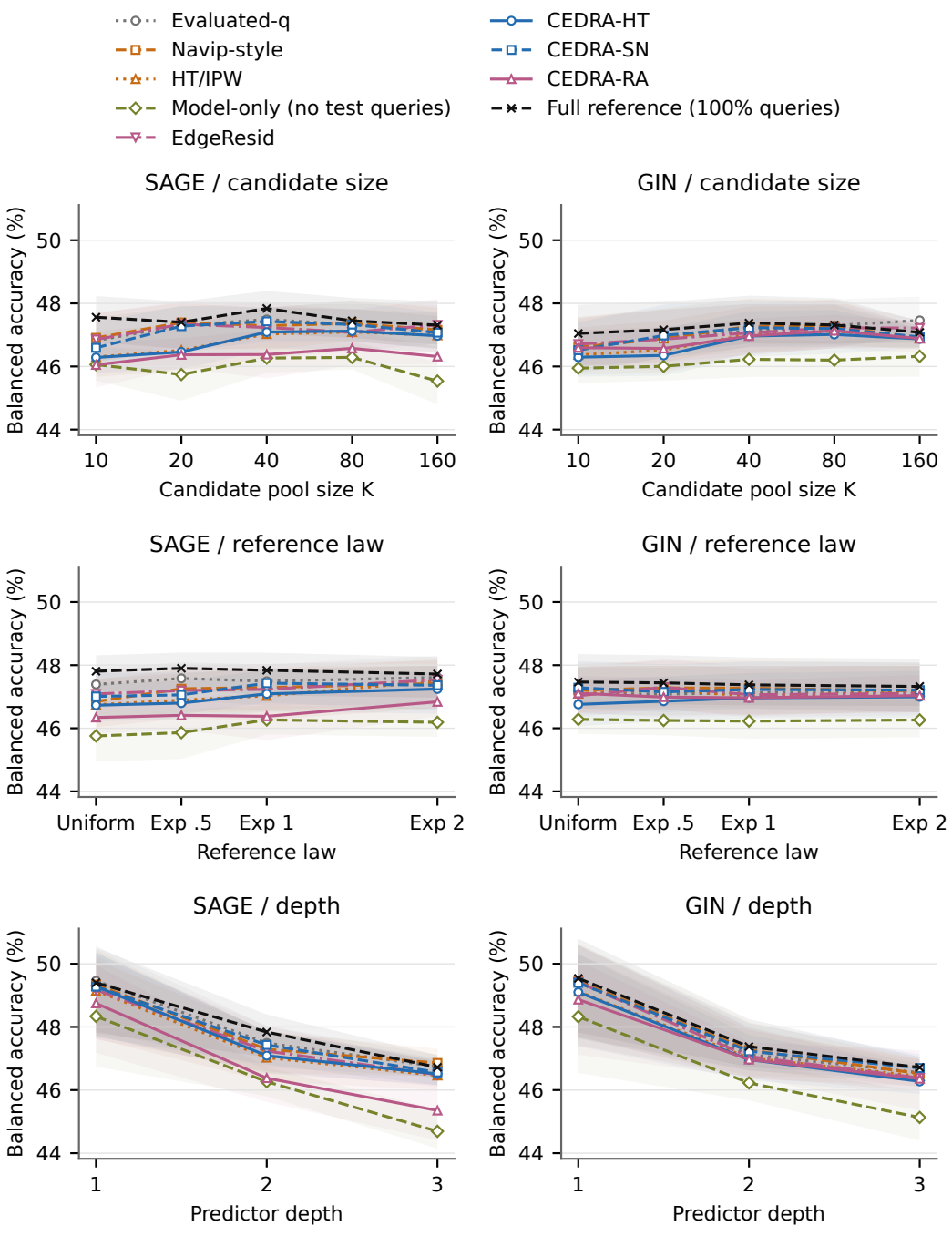


Figure 3: Matched sensitivity overview for the SAGE- and GIN-style predictors. Lines give equal-graph mean test balanced accuracy; bands are pointwise 95% within-dataset paired-seed percentile intervals on six fixed graphs. Each row changes one factor from the reference setting. Changing q_i changes the target as well as the aggregation weights. The depth sweep applies selective correction only in the first layer. The full-message reference incurs full query cost, while model-only uses no evaluated test messages.

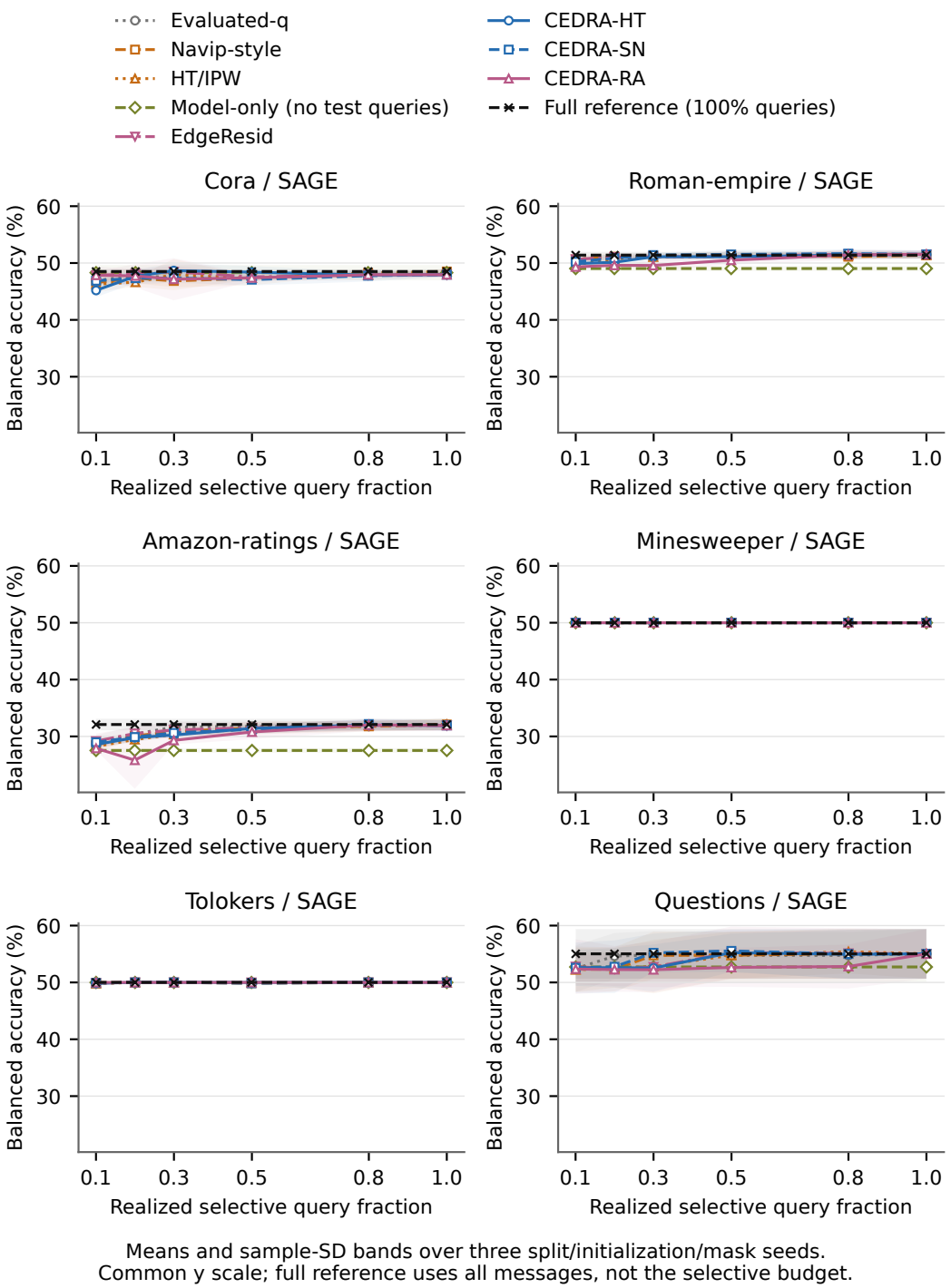


Figure 4: SAGE-style prediction versus evaluation budget on each fixed graph. Points are means and bands are sample standard deviations over three paired split/initialization/mask seeds, with Cora retaining split 0. The horizontal coordinate is the realized selective test-query fraction, not the requested budget used in the area summary. The full-message reference uses 100% of queries at every plotted position; model-only consumes no evaluated test messages. All nine arms are retained.

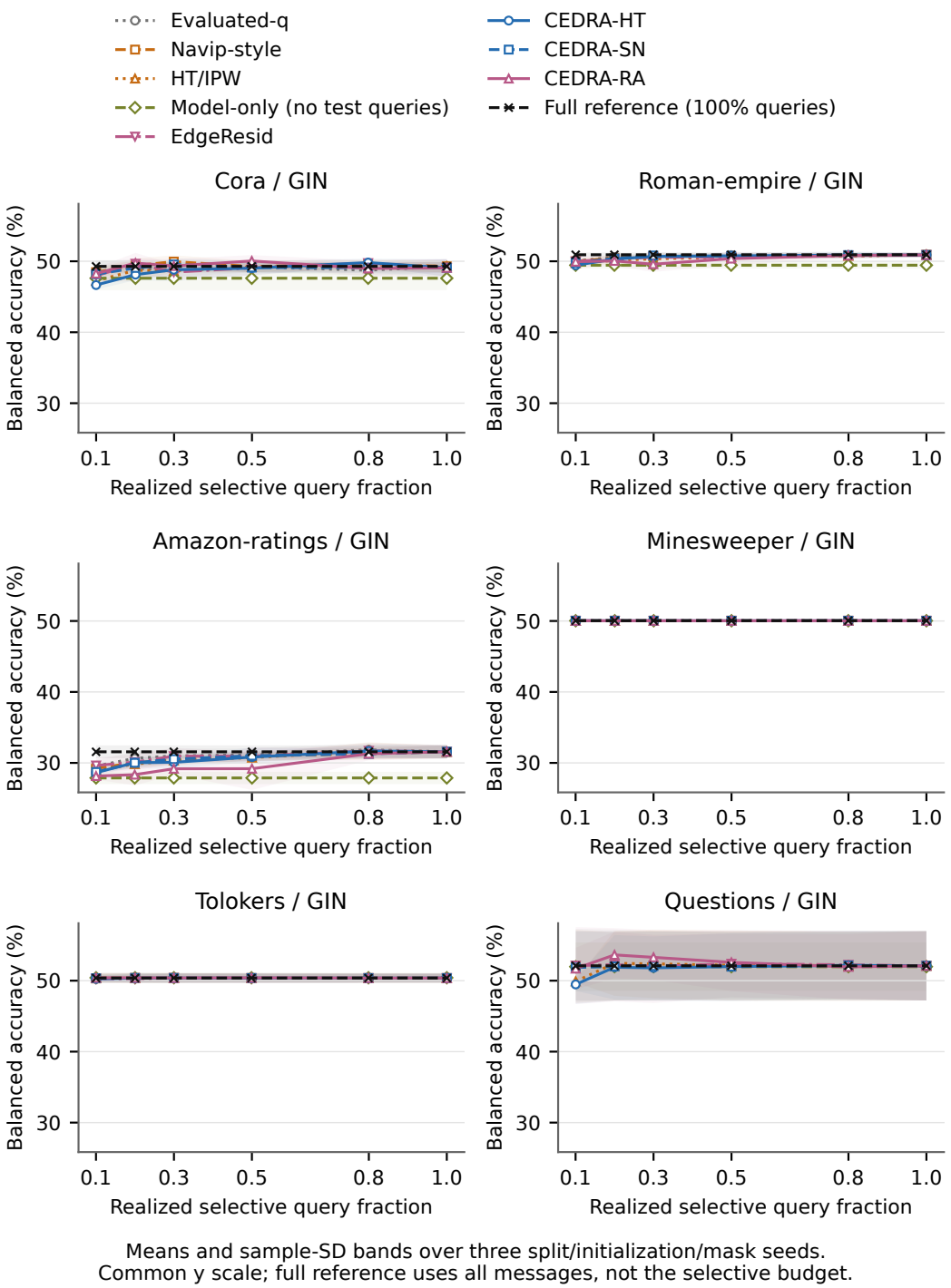


Figure 5: GIN-style prediction versus evaluation budget on each fixed graph, under the same protocol as Figure 4. Means and sample-standard-deviation bands use the three paired seeds. Realized selective test-query fraction is the horizontal coordinate; the full-message reference always uses 100% of queries and model-only uses none. Every arm and graph is shown on a common vertical scale, including near-chance performance.

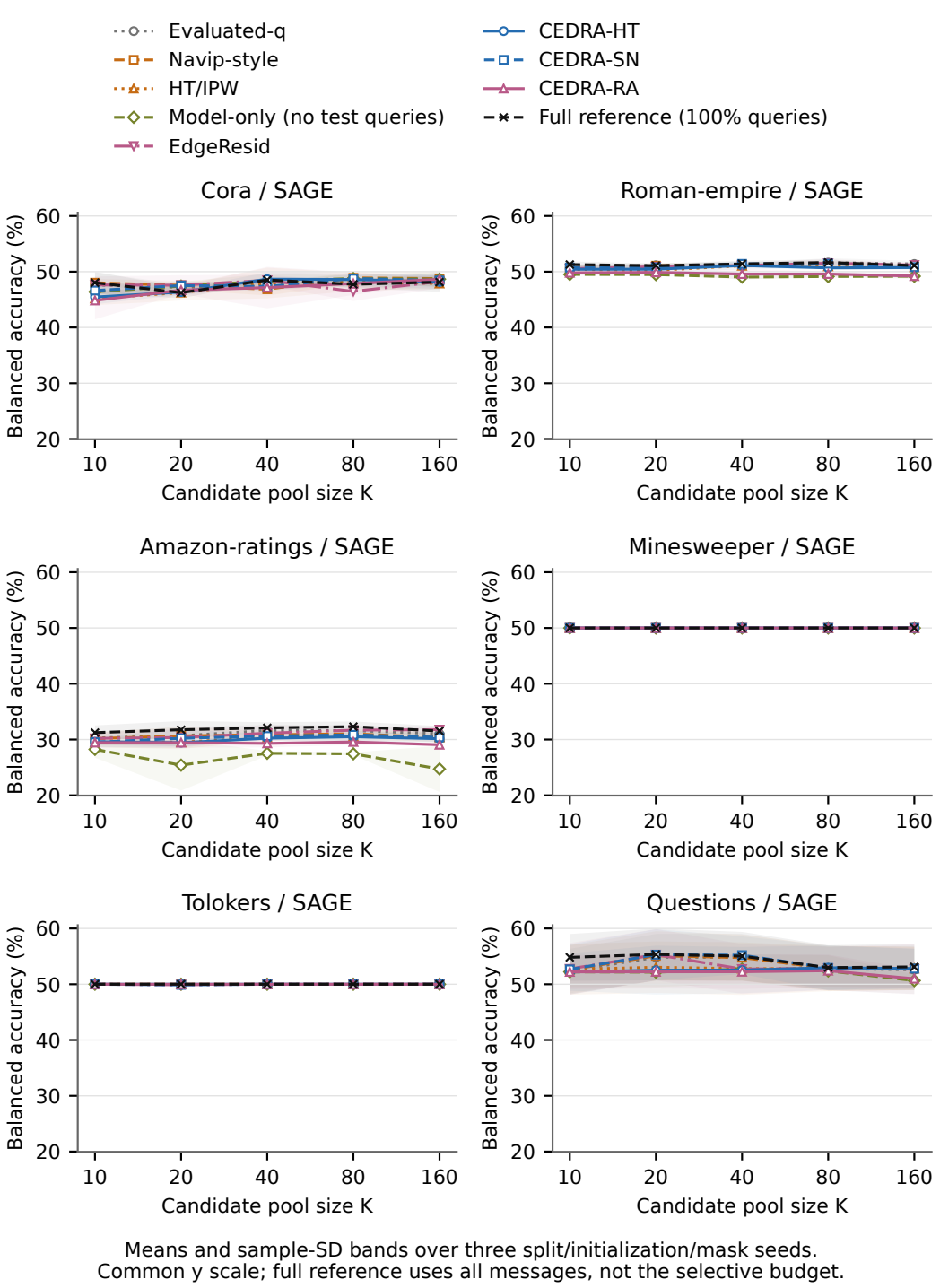


Figure 6: SAGE-style candidate-pool-size sweep on each fixed graph. Means and sample-standard-deviation bands cover three paired seeds at $K \in \{10, 20, 40, 80, 160\}$, with expected evaluation fraction 0.3, exponential-rank strength 1, and depth 2. A fixed fraction does not hold the absolute query count or candidate target constant as K changes. Candidates beyond rank 40 extrapolate beyond the shared nuisance pilot. The full-message reference uses every candidate message; model-only uses no evaluated test messages.

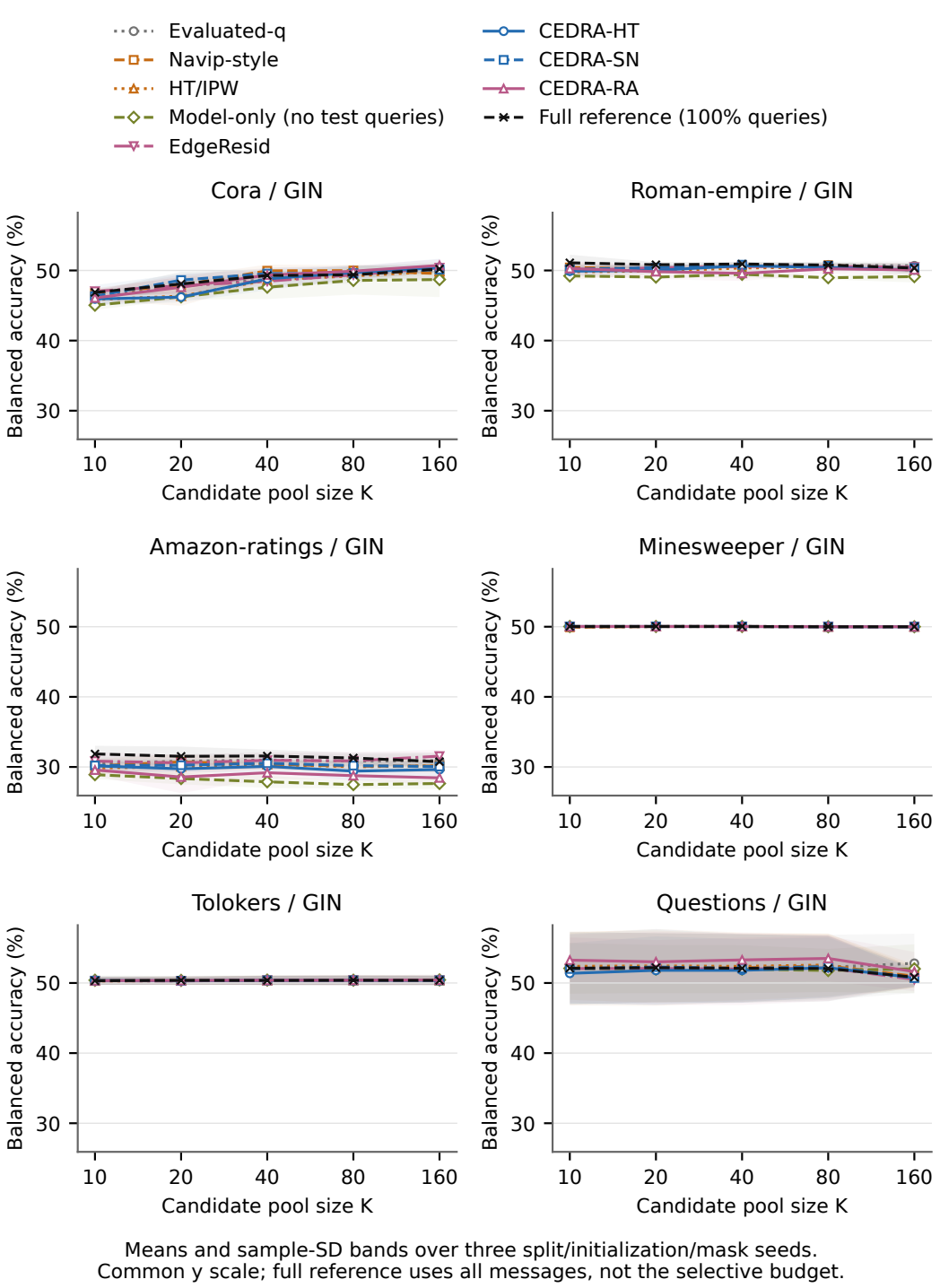


Figure 7: GIN-style candidate-pool-size sweep under the settings of Figure 6. Every graph and arm is retained, with three-seed means and sample-standard-deviation bands on a common vertical scale. Larger pools change the candidate target and absolute evaluation cost; the tested range alone does not establish computational scaling beyond $K = 160$. The full-message and model-only arms retain their distinct query costs.

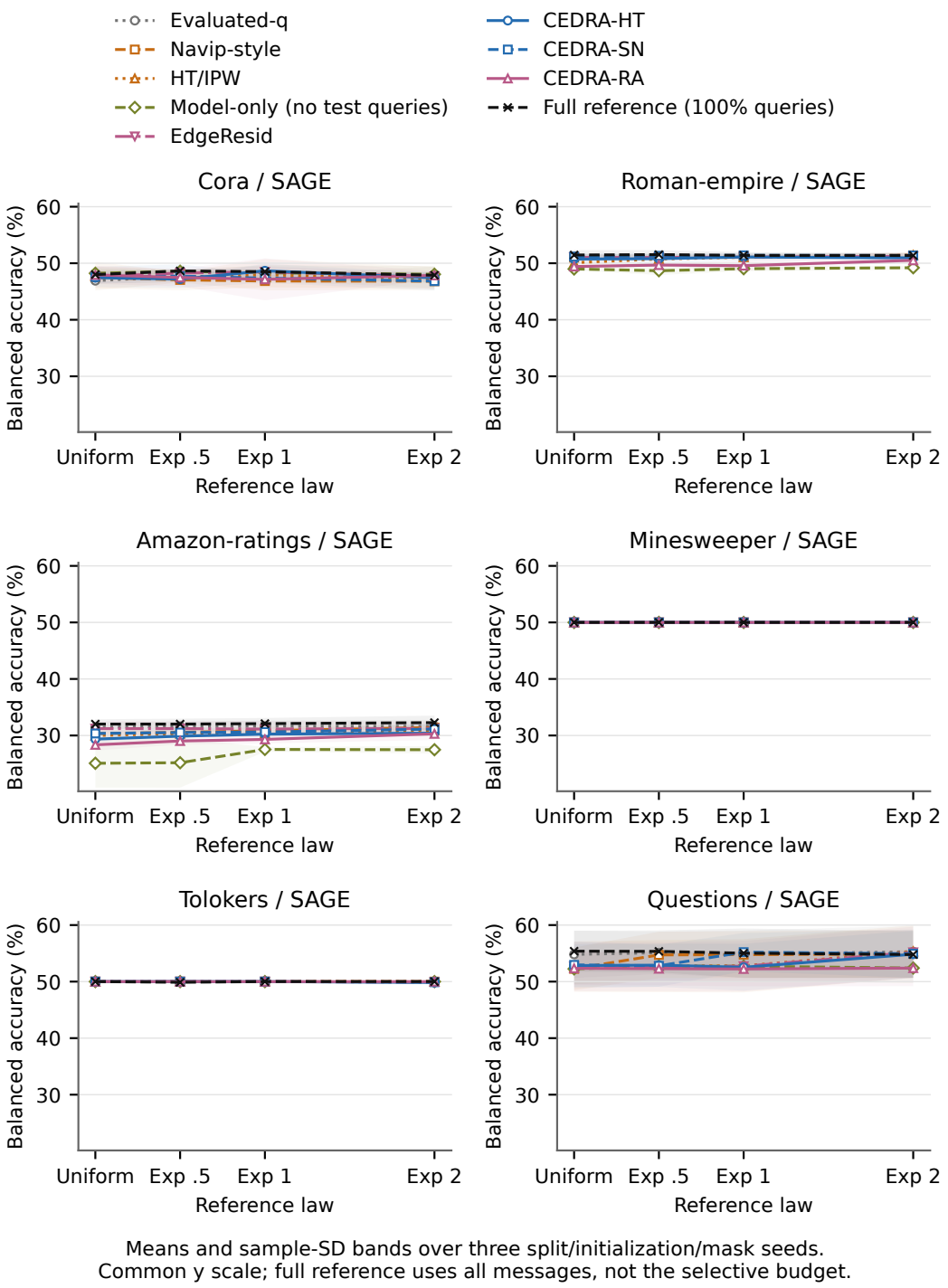


Figure 8: SAGE-style reference-law sensitivity on each fixed graph. Uniform weights and normalized exponential-rank laws of strengths 0.5, 1, and 2 are compared at $K = 40$, expected budget 0.3, and depth 2. Points and bands are three-seed means and sample standard deviations. Each law defines its own target and full-message reference, so cross-law differences do not hold the estimand fixed. This is a rank-weighting comparison, not a kernel-bandwidth sweep.

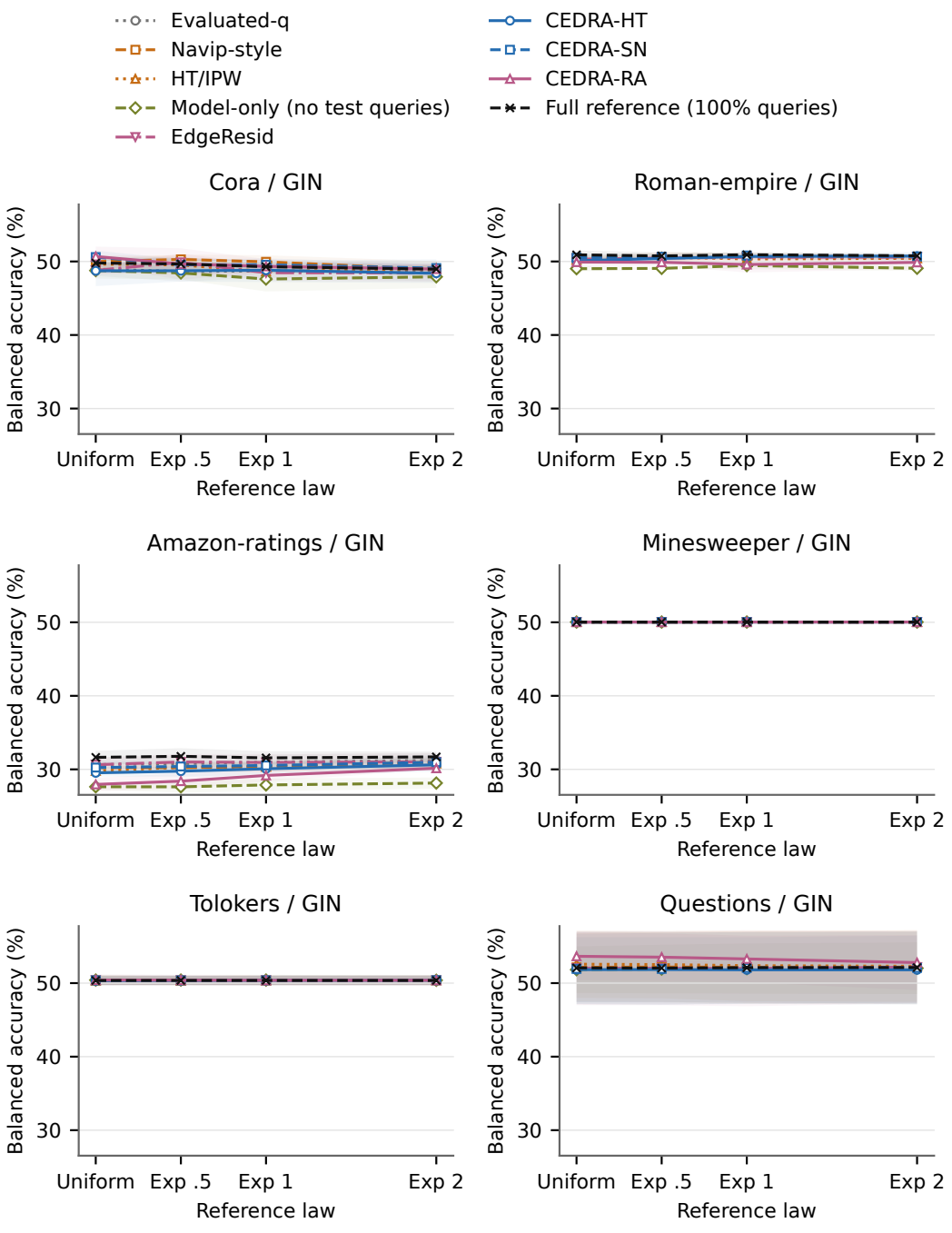


Figure 9: GIN-style reference-law sensitivity under the settings of Figure 8. Three-seed means and sample-standard-deviation bands retain all methods and fixed graphs. Changing the reference law changes the target for every arm, including the full-message reference. Model-only consumes no evaluated test messages, whereas the full-message reference requires all of them.

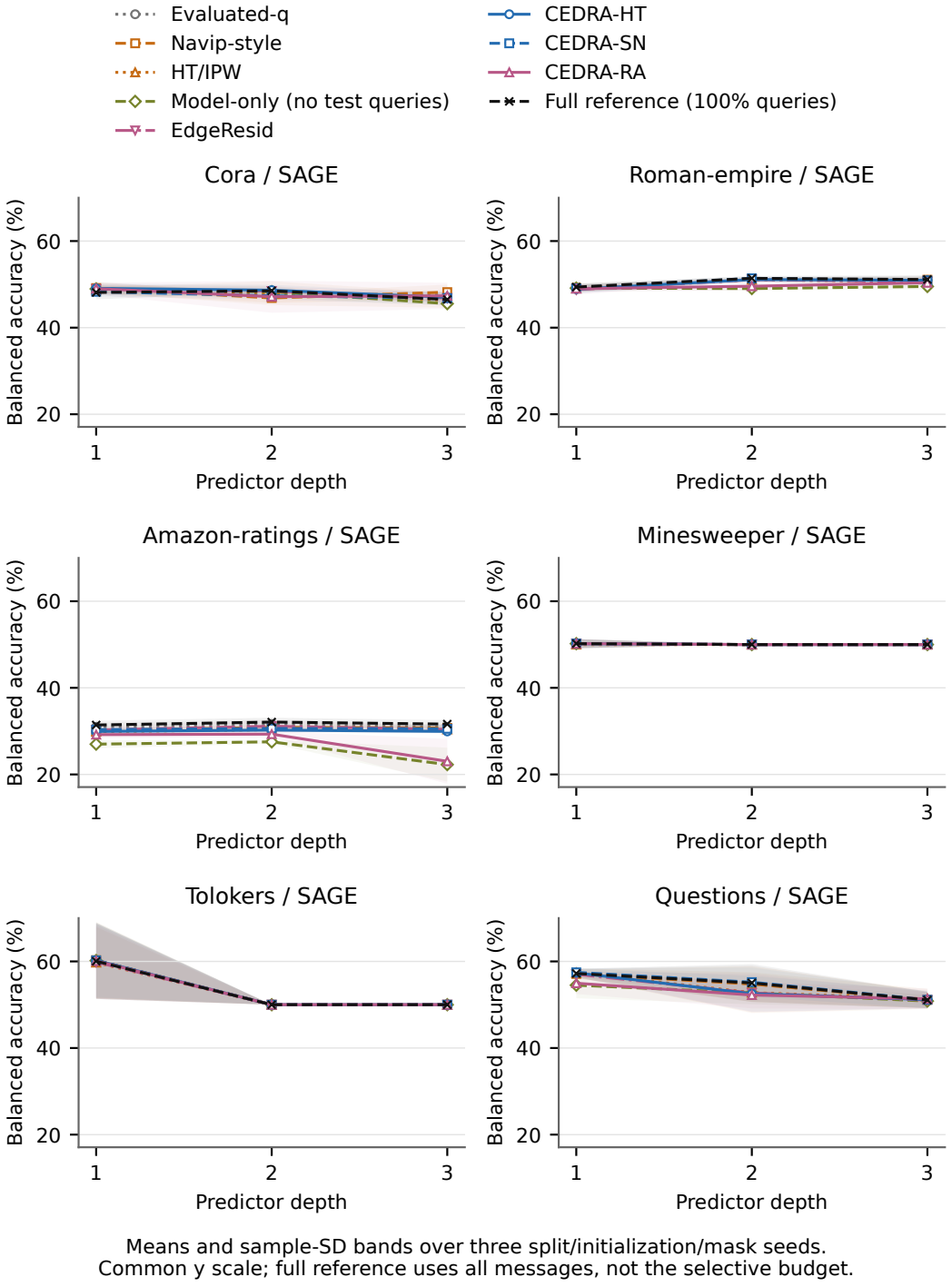


Figure 10: SAGE-style predictor-depth sweep on each fixed graph at $K = 40$, expected budget 0.3, and exponential-rank strength 1. Depths 1, 2, and 3 count the initial selective update and subsequent propagation layers. CEDRA correction occurs only in the first layer; later layers use cheap features and accessible representations. Lines and bands are three-seed means and sample standard deviations. These predictive comparisons do not extend the first-layer pre-transform double-robust guarantee to deeper representations.

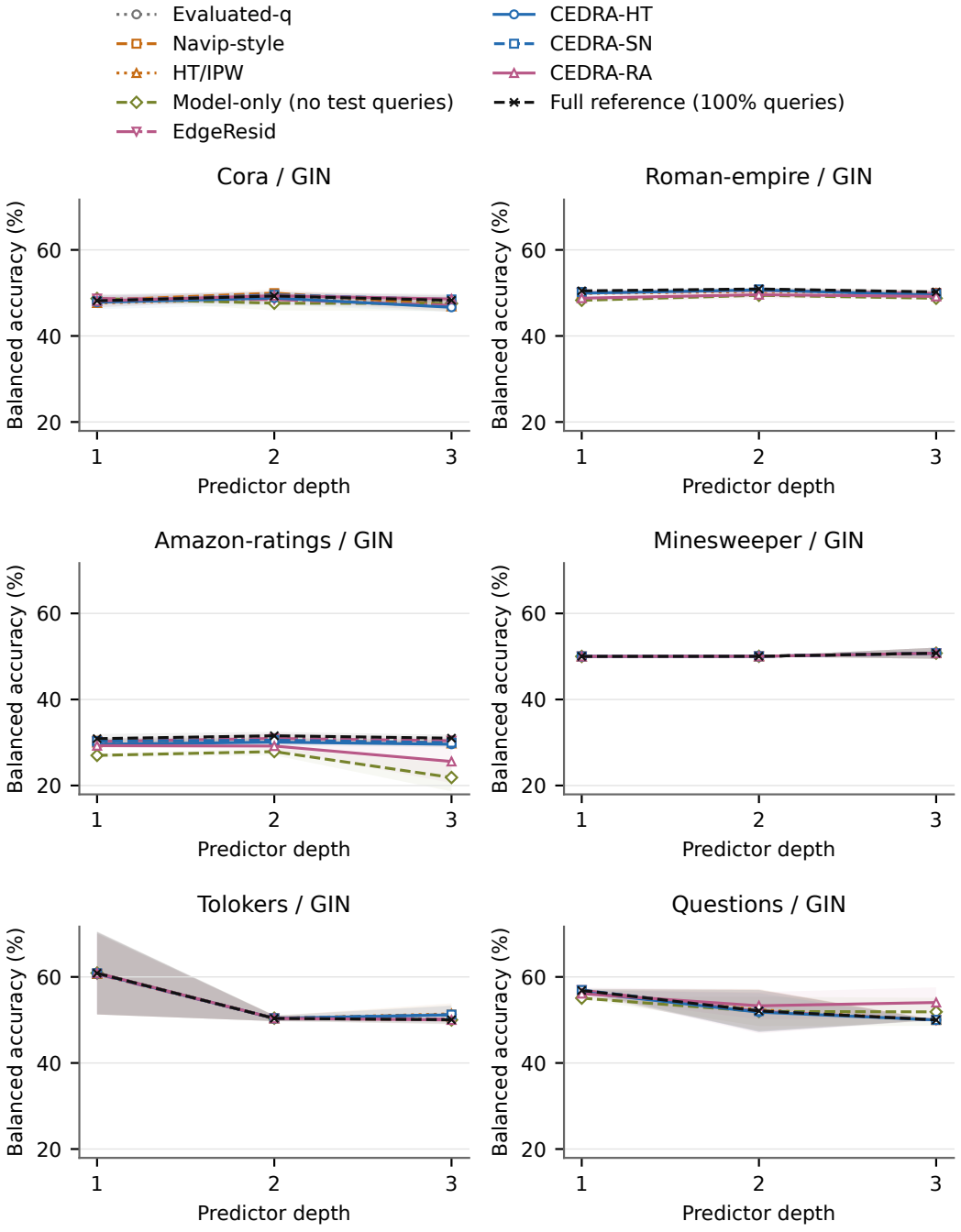


Figure 11: GIN-style predictor-depth sweep under the settings of Figure 10. Every graph and arm is retained with three-seed means and sample-standard-deviation bands. Only the first update uses selective correction, not each hidden layer. The full-message reference retains full query cost and model-only uses no evaluated test messages. Depth-dependent prediction is empirical and is not covered by the first-layer HT unbiasedness theorem.

K.3 CONTROLLED VIOLATIONS OF MEAN IGNORABILITY

The separate prospective synthetic stress uses 64 receivers, $K = 40$, the normalized exponential-rank law with $a = 1$, expected evaluation fraction 0.3, and design seeds $101, 202, 303$. For generated scalar cheap covariates W_{ij}, the conditional message mean is $m_{ij} = \tanh(W_{ij}) + 0.2\sin(2W_{ij})$, and the message is $M_{ij} = m_{ij} + \sigma E_{ij}$ with $\sigma = 0.5$ and independent equiprobable residual signs $E_{ij} \in \{-1, +1\}$. The cheap acquisition score is $h_{ij} = 2(1 - 2t_j) + 0.25\tanh(W_{ij})$. For $\gamma \in \{-2, -1, 0, 1, 2\}$, choose a receiver-specific intercept c_i so that the mean marginal evaluation probability is 0.3, and set

$$\begin{aligned} p_{ij}^{+} &= \mathrm{sigmoid}(c_i + h_{ij} + \gamma), & p_{ij}^{-} &= \mathrm{sigmoid}(c_i + h_{ij} - \gamma), \\ \pi_{ij}^{\star} &= \tfrac{1}{2}(p_{ij}^{+} + p_{ij}^{-}), & c_{ij} &= \tfrac{\sigma}{2}(p_{ij}^{+} - p_{ij}^{-}). \end{aligned} \tag{19}$$

Evaluation uses p_{ij}^{+} when $E_{ij} = +1$ and p_{ij}^{-} otherwise. The feasible propensity input is the marginal $\pi_{ij}^{\star}$, never the unavailable residual-conditioned probability. Correct versus shifted outcome predictions, m_{ij} versus $m_{ij} + 0.25 + 0.1\sin(W_{ij})$, are crossed with correct marginal versus constant-0.3 propensity predictions, yielding four nuisance regimes. Independent residual/evaluation draws are repeated 4,096 times per design-seed/gamma cell, reusing random streams across gamma and nuisance regimes. The full record contains 540 method rows; the separate smoke is excluded from scientific claims.

The stress compares empirical conditional-mean bias and both conditional-target and realized-target RMSE with exact moments for IPW, model-only, HT, and RA. Exact moments enumerate the residual sign and evaluation indicator for each candidate and sum independent-candidate variances. Under both-correct nuisances, HT's predicted conditional bias is $\sum_j q_{ij} c_{ij} / \pi_{ij}^{\star}$, which is generally nonzero for $\gamma \neq 0$. For example, the reported analytical biases for design seed 101 under both-correct nuisances are approximately $-0.434605, -0.280383, 0, +0.280383, +0.434605$ in increasing gamma order. The Monte Carlo fields follow this signed progression. At gamma zero, the both-correct and two singly-correct HT branches have zero analytical bias to numerical precision; the both-wrong branch does not. These results test the covariance remainder directly rather than asserting robustness to nonignorable selection. Analytical fields are unavailable for the ratio controls, SN, EdgeResid, and full-reference rows; null fields are not interpreted as zero, and deterministic model-only bias-standardization fields remain undefined where appropriate.

Figure 12 compares HT's empirical signed bias with its exact analytical counterpart. Figures 13–15 retain all nine methods for signed conditional-target bias and both error targets. Error bars describe variation across the three fixed synthetic designs, not Monte Carlo standard errors or confidence intervals. The conditional target averages over message residuals, whereas the realized target retains each draw's messages; their RMSE values answer different questions.

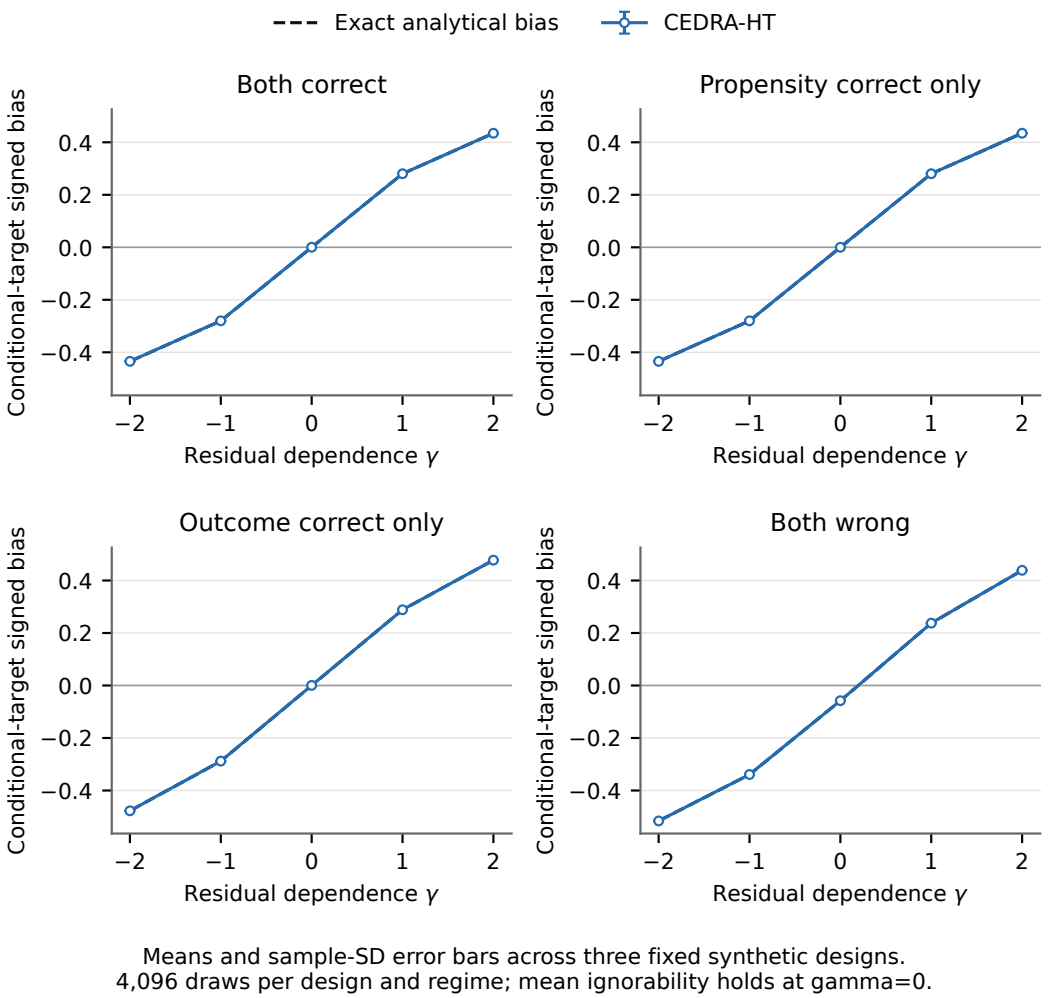


Figure 12: CEDRA-HT signed conditional-target bias under controlled violations of mean ignorability. The four panels cross correct or misspecified message and marginal evaluation models. Colored points are empirical means across three fixed design seeds, with sample-standard-deviation error bars; dashed curves are means of the exact analytical biases. Each design/regime uses 4,096 residual/evaluation draws. Mean ignorability holds at $\gamma = 0$; nonzero γ introduces residual-dependent evaluation, whose covariance contribution remains even with both nuisances correct.

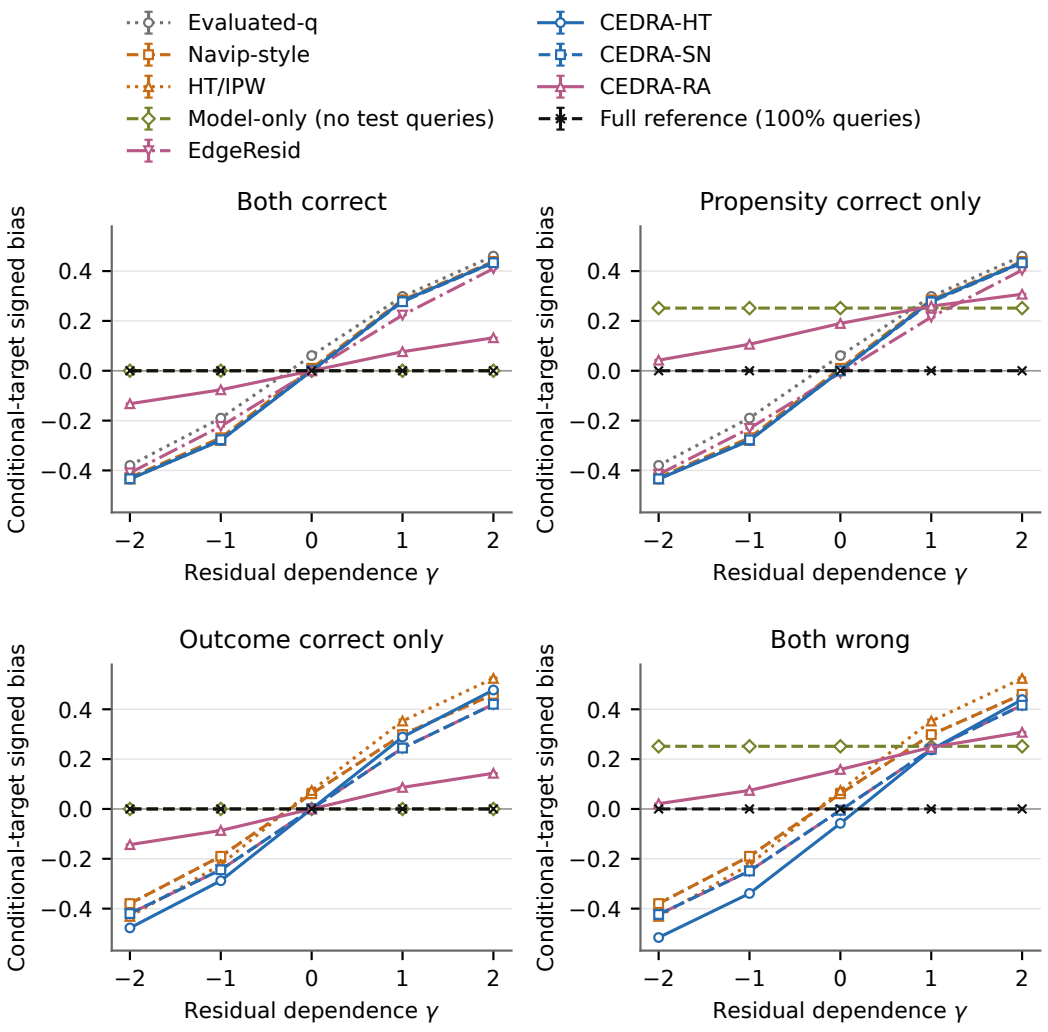
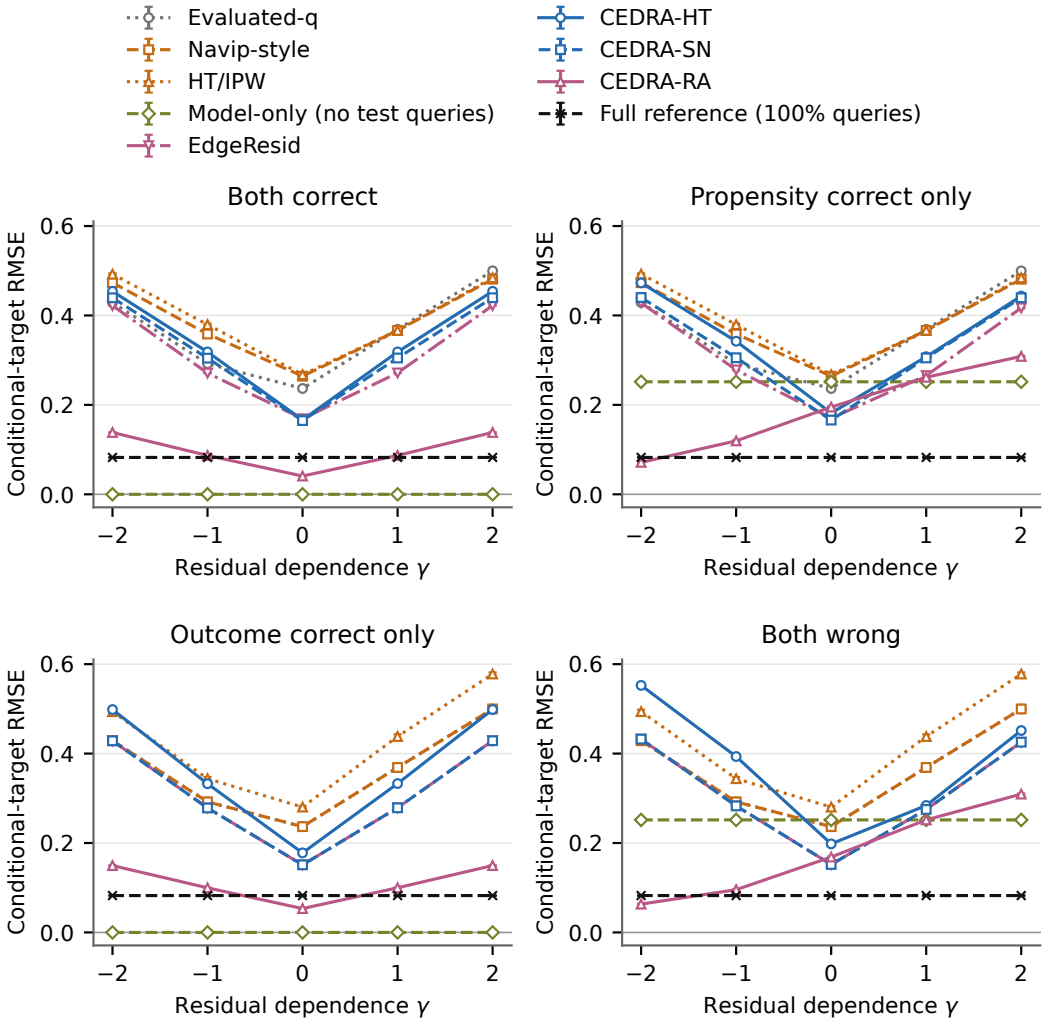
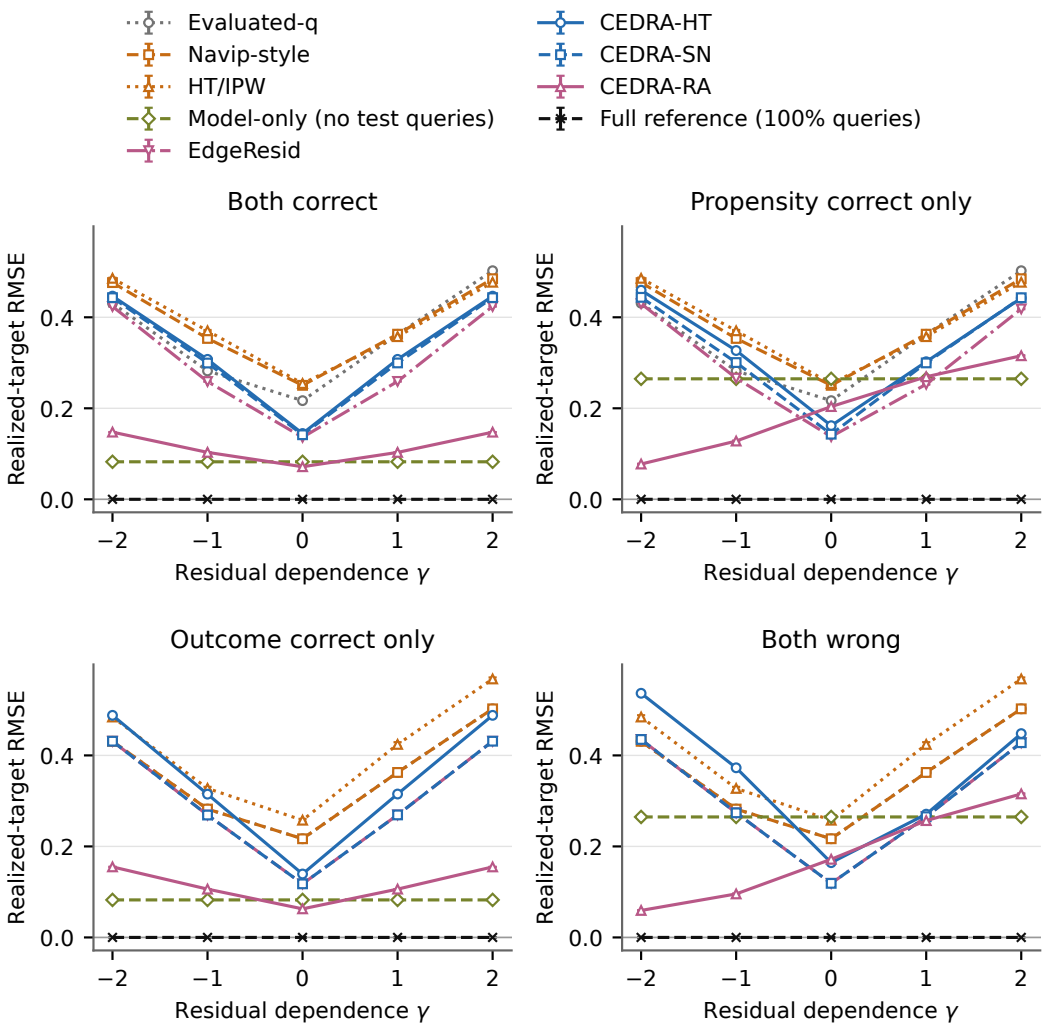


Figure 13: Signed conditional-target bias for all nine arms across all four nuisance regimes and five residual-dependence values. Points and error bars are means and sample standard deviations across three fixed synthetic designs, with 4,096 draws per design/regime. Signs are retained rather than replaced by absolute bias. The generic RMSE sentence in the plot footer refers only to the companion RMSE figures, Figures 14 and 15; this figure displays signed bias.



All nine arms and all four nuisance regimes; mean and sample SD over three designs.
RMSE summaries average design-level RMSE; no outcome-selected cells.

Figure 14: Conditional-target RMSE for every arm and nuisance regime under residual-dependent evaluation. The target is $\tau_i(q) = \sum_j q_{ij}\mathbb{E}[M_{ij} \mid \mathcal{F}_i]$. Points average design-level RMSE over three fixed designs, and error bars are sample standard deviations; these are not RMSE values recomputed after pooling designs. Each design/regime contains 4,096 draws. The full-message reference uses all exact messages, which still fluctuate around the conditional target.



All nine arms and all four nuisance regimes; mean and sample SD over three designs.
RMSE summaries average design-level RMSE; no outcome-selected cells.

Figure 15: Realized-target RMSE for every arm and nuisance regime under residual-dependent evaluation. The target is $T_i(q) = \sum_j q_{ij} M_{ij}$ for each message draw, rather than its conditional expectation. Points and error bars are means and sample standard deviations of design-level RMSE over three fixed designs, with 4,096 draws per design/regime. The full-message reference accesses the complete target; model-only consumes no evaluated messages. All signed dependence values and nuisance regimes are retained.

K.4 CONTROLLED FACTORIAL

The controlled oracle factorial specifies 600 nodes per run, 40 candidates per receiver, and seeds $0, \ldots, 19$. It crosses edge-feature coupling $\{0, 1.5, 3\}$, label-message alignment $\{-1, 0, 1\}$, expected degree $\{8, 16\}$, community mixing $\{0.15, 0.50\}$, and overlap bound $\{0.02, 0.10\}$, giving 72 conditions and 1,440 condition-by-seed artifacts. The generator uses six construction features, four message covariates, three communities, community separation 1.5, message-residual scale 0.75, message-realization noise scale 0.35, and a 60/20/20 split. Every retained run records a fixed candidate population, normalized q_i, known bounded Bernoulli evaluation probabilities, the exact nuisance protocol, and `assumption_status=satisfied`. Misspecified oracle nuisances are the candidate-constant laws $\widehat{\pi} = 0.5$ and $\widehat{\mu} = 0.5$, fixed before generation and independent of candidates, evaluation indicators, messages, targets, and labels. The learned grid fixes expected degree 12, mixing 0.30, and overlap 0.05 while crossing the same coupling and alignment levels over 20 seeds; its 180 artifacts use two receiver-level nuisance folds and 150 fitting epochs. Each learned artifact contains 10 common methods and 12 learned methods: six explicitly unclipped primary arms and six otherwise matched clipped-at-0.02 sensitivity arms, all from the same two nuisance fits. The accepted analysis checked all 1,620 artifacts, their hashes, and their common source identity before producing any table, figure, interval, or test. The accepted analysis included all 1,620 artifacts and 18,360 unique run-method rows, with complete coverage, finite primary metrics, and 132 valid paired comparisons. Informative rather than misspecified learned nuisances improved CEDRA-HT RMSE by 0.304 and ROC-AUC by 0.0379, both with Holm-adjusted $p = 4.58 \times 10^{-5}$; outcome-nuisance R^2 was 0.876 versus -0.441, while the ECE contrast was null after correction. The final analyzer included a one-line validator correction recognizing that the recorded validation threshold is on the finite raw-score scale rather than a probability; no experimental output was changed.

K.5 OPTDIGITS RANDOMIZED AGGREGATE PROTOCOL

The complete UCI release is read in official file order: 3,823 training rows followed by 1,797 writer-separated test rows, with each of the first 64 integer fields scaled by $1/16$ (Alpaydin & Kaynak, 1998). The construction view contains the 32 coordinates with even row-plus-column parity and the message view the complementary 32 coordinates. For each receiver, squared Euclidean distance in the construction view fixes the 40 closest non-self candidates, with sender index breaking ties, and $q_{ij} \propto \exp(-(r_{ij} - 1)/10)$ for rank $r_{ij} \in \{1, \ldots, 40\}$. Candidate construction uses all unlabeled feature rows and is therefore transductive. For the aggregate study, candidate messages with $A_{ij} = 0$ are deliberately masked to create a controlled selective-evaluation stress test, even though the source table contains them; this is not natural missingness. The external outcome nuisance is fitted without evaluated test messages, labels, or evaluation indicators and is not an oracle conditional mean. The ten controls include a Navip-style inverse-ratio adaptation and the direct full-candidate aggregate; the latter computes $T_i(q)$ exactly and is simpler whenever full message access is operationally allowed.

The three known directed evaluation-probability laws are $\pi_{ij} = 0.50 - 0.40(r_{ij} - 1)/39$, $\pi_{ij} = 0.30$, and $\pi_{ij} = 0.10 + 0.40(r_{ij} - 1)/39$. Each is bounded in $[0.10, 0.50]$ and has expected receiver degree 12. The same PCG64 uniforms are shared across the three mechanisms within each of 20 fixed seeds. Candidate/message views, candidate ranks and q, propensity tables, seeds, and official source hashes were specified before aggregation. This protocol covers ten prespecified aggregate estimators and diagnostics; it generates no predictive neural result. The accepted seed-level outputs and aggregate analysis passed manifest, completeness, duplicate, finiteness, configuration, and hash checks. The accepted aggregate analysis contains exactly 600 unique seed-by-mechanism-by-method rows and 90 complete method summaries. All values are finite, every interval contains its reported mean, and exactly one contrast is designated primary. The uncertainty summaries cover the declared evaluation-draw and optimization seeds on this fixed finite dataset.

K.6 ADDITIONAL REAL-FEATURE RANDOMIZED AGGREGATE PROTOCOLS

The additional breadth extension was frozen before outcome inspection and applied without dataset replacement to UCI Dry Bean and Letter Recognition (Köklü & Özkan, 2020; Slate, 1991). Dry Bean contains 13,611 rows, 16 numeric image-shape features, and seven classes; Letter Recognition contains 20,000 rows, 16 integer image features, and 26 classes. Labels were used only to validate the source tables. Alternating feature columns define disjoint eight-dimensional construction and message

Table 13: Additional real-feature target-recovery RMSE over 20 paired seeds. Values are means; lower is better. The table reports the two direct weighting controls and both CEDRA variants.

Dataset	Mechanism	Observed-q	Navip	CEDRA-HT	CEDRA-SN
Dry Bean	Near	0.10622	0.09754	**0.09170**	0.09175
Dry Bean	Constant	0.13163	0.13163	0.12642	**0.12434**
Dry Bean	Far	**0.15362**	0.18527	0.20814	0.17440
Letter Recognition	Near	0.15264	0.14195	0.18764	**0.13960**
Letter Recognition	Constant	0.19127	0.19127	0.25627	**0.18799**
Letter Recognition	Far	**0.24105**	0.27941	0.42720	0.27430

Table 14: Best CEDRA and strongest non-CEDRA near-mechanism target-recovery RMSE in the complete breadth panel. Lower is better. Family selection and seed-level Wilcoxon values are descriptive.

Dataset	Best CEDRA	Best non-CEDRA	Difference
Abalone	**0.06627**	0.06984	−0.00357
Airfoil	**0.18482**	0.19028	−0.00545
Wine Quality	**0.19596**	0.20147	−0.00551
MAGIC	**0.18088**	0.18695	−0.00607
Rice	**0.01782**	0.02937	−0.01155
Optdigits	**0.15549**	0.16920	−0.01371
Dry Bean	**0.08811**	0.09374	−0.00564
Letter Recognition	**0.14164**	0.14439	−0.00276
PenDigits	0.12110	**0.12050**	+0.00060
Landsat	**0.05796**	0.06627	−0.00830

views, standardized from a deterministic 70% training split. The candidate pool, rank-decay law, three propensity mechanisms, shared PCG64 edge draws, ten estimators, 20 seeds, endpoints, paired bootstrap, exact sign tests, and Holm families match the declared Optdigits aggregate design. Every one of the 40 seed artifacts passed hash, completeness, and finiteness checks, and an independent recomputation matched all 18 central RMSE contrasts.

Across all six dataset-mechanism cells, CEDRA-SN improved on Navip in all 20 paired seeds. It also improved on observed-q in every near and constant seed, but not in any far-mechanism seed. This pattern was not inferred from a favorable subset: all datasets, mechanisms, methods, endpoints, and seeds fixed by the extension protocol are retained. The tables are semi-synthetic target-recovery results on fixed public feature tables and do not establish population, causal, or predictive improvement.

K.7 TEN-DATASET BREADTH PANEL AND AUDIT

The later breadth panel fixes ten UCI datasets covering 90,066 rows and three source-task types. It standardizes features from a deterministic training split, uses alternating columns for disjoint construction and message views, fixes 40 construction-view candidates per receiver, and applies one shared rank-decay reference law. All 200 seed artifacts and 660 aggregate summary rows passed completeness, uniqueness, finiteness, design-hash, and exact recomputation checks; the maximum absolute difference between recomputed and published means was zero. The corrected analyzer excludes both CEDRA variants from the non-CEDRA comparator set. Its immutable analysis digest begins `1479bcd919ab106e`, and its near-recovery table digest begins `cb8a2c19f916181a`.

The execution used PCG64 seed $51000 + s$ rather than the $10000 + s$ stated in the frozen prose protocol. This fixed offset was shared across all datasets and mechanisms, no dataset was replaced or omitted, and all outcomes were retained, but the deviation prevents describing the panel as perfectly preregistered. The executed downstream heads were ridge and logistic regressions rather than the declared neural heads and GraphSAGE baseline. Their results are retained as exploratory linear probes and are not used for confirmatory predictive claims. The accepted breadth claim is restricted to target recovery under the executed imposed selection mechanisms.

Table 15: Risk-adaptive aggregate RMSE averaged over two evaluation mechanisms, three budgets, and five masks per group. Lower is better. The final three datasets are the untouched confirmation suite. All 180 cells are retained.

Dataset	Observed mean	Model-only	IPW	Full CEDRA-HT	CEDRA-RA
Cora	0.36351	0.20357	0.74007	0.33611	**0.16670**
Roman-empire	0.34723	0.16314	0.63928	0.27204	**0.13334**
Questions	0.25292	0.21886	0.84878	0.37883	**0.17863**
Amazon-ratings	0.30932	0.15556	0.42409	0.19522	**0.11964**
Minesweeper	0.27402	0.03844	0.31850	0.06118	**0.03140**
Tolokers	0.19293	0.07307	0.32693	0.10435	**0.05726**

K.8 RISK-ADAPTIVE SELECTIVE-MESSAGE CONFIRMATION

The risk-adaptive estimator was developed on Cora, Roman-empire, and Questions and then frozen with seven advancement gates before any held-out result was generated. Untouched confirmation used Amazon-ratings, Minesweeper, and Tolokers. Both suites used the original directed non-self graph as candidate support, a fixed label-independent random projection of source features as cheap pre-evaluation state, and one fixed nonlinear edge-conditioned teacher as the selectively evaluated 32-dimensional message. Two fixed selection mechanisms, three probability budgets, and five mask seeds gave 90 cells per suite and 180 total. The known Bernoulli evaluation probabilities ranged from $[0.05, 0.25]$ through $[0.20, 0.80]$. Receiver-separated ridge nuisance fits used only training receivers and inverse-probability weights. Every result and manifest was present, all errors were finite, and no dataset, mechanism, budget, or seed was omitted. The frozen confirmation gates required at least 20% lower RMSE than full HT correction, 50% lower than IPW, 5% lower than model-only, at least 12 of 18 wins against each simple comparator, and no group more than 10% worse than the better simple comparator. All gates passed. This is confirmation of aggregate recovery under a controlled message teacher and algorithmic evaluation design; it is not natural-edge or downstream prediction evidence.

External-validity robustness extension. A separately frozen matrix crossed the same six public supports with two candidate laws, two structurally different fixed nonlinear teachers, two acquisition mechanisms, three probability budgets, and five fresh masks, giving 720 complete cells in 144 groups. The original-edge law uses the directed non-self graph; the projected-neighbor law uses the eight preceding and eight following nodes under a fixed label-independent feature projection. The additive-tanh teacher and an independently seeded gated-ReLU-plus-interaction teacher both output 32-dimensional messages. All eight frozen completeness and performance gates passed. Independent recomputation from every raw cell reproduced RMSE 0.091071 for CEDRA-RA, 0.147570 for full CEDRA-HT, 0.342540 for IPW, 0.116900 for model-only regression, and 0.196169 for observed-message averaging. CEDRA-RA wins 144 of 144 groups against model-only and 140 of 144 against observed-message averaging. The four observed-message losses, all on Roman-empire with projected-neighbor candidates under diversity acquisition at medium or high budget, remain included. The raw analyzer records `decision=ADVANCE` while retaining `eligible_for_paper_claim=false` and `paper_claim=null`; we therefore do not treat the analyzer as self-authorizing. A separate independent audit verified all 720 artifacts and judged the matrix ready within the reviewed selective-aggregate scope. Manuscript claims follow that narrower audit verdict. This extension broadens controlled external validity within two fixed message teachers and randomized evaluation mechanisms on public graph supports.

K.9 FROZEN TEN-DATASET NONLINEAR-HEAD BENCHMARK

The follow-up nonlinear-head benchmark was frozen before any outcome was generated or inspected. It reconstructs the accepted near-selection aggregates for all ten breadth datasets and all 20 seeds, first requiring exact agreement with every accepted recovery RMSE and MAE. The source tasks comprise three regressions, two binary classifications, and five multiclass classifications. The construction-only arm receives the construction view, while each of the other seven arms concatenates that same view with observed mean, observed-q, Navip, pure HT, outcome-only, CEDRA-HT, or CEDRA-SN aggregation. Every arm then uses the same two-hidden-layer MLP with widths 128 and 64, GELU

Table 16: Frozen CEDRA-RA robustness results by message teacher and candidate law. Entries are group-mean aggregate RMSE over six graphs, two acquisition mechanisms, three budgets, and five masks. Lower is better; every one of the 720 cells is retained.

Teacher	Candidate law	Observed	Model	IPW	Full HT	CEDRA-RA
Additive tanh	Original edges	0.29028	0.14106	0.54816	0.22099	**0.11411**
Additive tanh	Projected neighbors	0.17605	0.11362	0.24997	0.09333	**0.08360**
Gated ReLU	Original edges	0.19680	0.12607	0.39874	0.19797	**0.10204**
Gated ReLU	Projected neighbors	0.12155	0.08684	0.17329	0.07799	**0.06453**
Overall	All four families	0.19617	0.11690	0.34254	0.14757	**0.09107**

Table 17: Complete frozen nonlinear-head benchmark over 20 seeds per dataset. Each entry compares the better prespecified CEDRA input with the strongest prespecified non-CEDRA input under the same MLP protocol. Positive improvements favor CEDRA. Family selection is descriptive; intervals resample the 20 paired seeds.

Dataset	Task and metric	Best CEDRA	Best non-CEDRA	Improvement	95% interval	Seed W/L
Abalone	Regression RMSE	HT 2.39800	Construction 2.39574	-0.00225	$[-0.00430, -0.00020]$	8/12
Airfoil	Regression RMSE	HT 3.45264	Outcome 3.40989	-0.04275	$[-0.06856, -0.01378]$	4/16
Wine Quality	Regression RMSE	**SN 0.72516**	Navip 0.72530	$+0.00013$	$[-0.00025, 0.00051]$	12/8
MAGIC	Binary BA	HT 0.81787	Outcome 0.81952	-0.00165	$[-0.00292, -0.00038]$	5/15
Rice	Binary BA	**HT 0.92004**	Observed mean 0.92000	$+0.00005$	$[-0.00174, 0.00185]$	10/9
Optdigits	Multiclass BA	**HT 0.94639**	Outcome 0.94400	$+0.00239$	$[-0.00030, 0.00501]$	12/8
Dry Bean	Multiclass BA	SN 0.92947	Construction 0.92992	-0.00046	$[-0.00124, 0.00027]$	10/10
Letter	Multiclass BA	SN 0.70869	Observed-q 0.71037	-0.00168	$[-0.00274, -0.00064]$	6/14
PenDigits	Multiclass BA	HT 0.87865	Observed-q 0.88237	-0.00372	$[-0.00511, -0.00235]$	3/17
Landsat	Multiclass BA	**HT 0.82134**	Navip 0.82041	$+0.00093$	$[-0.00304, 0.00497]$	10/10

activations, dropout 0.2, AdamW optimization, and train-only input scaling. Checkpoints are selected exclusively by validation RMSE for regression or validation balanced accuracy for classification, and test labels are evaluated once after selection. The complete design contains 200 dataset-seed cells and 1,600 method results. Its family-level best-CEDRA comparison is descriptive because the better of two CEDRA inputs and the strongest of six non-CEDRA inputs are identified from the complete dataset-level test means. All method and seed outcomes remain visible, and no predictive result is covered by the first-layer aggregate theorem. The result bundle records full SHA-256 identities for the protocol, runner, analyzer, wrappers, 200 cell artifacts, and four analysis outputs.

The frozen nonlinear-head benchmark retained all 200 dataset-seed cells and 1,600 method outcomes. The better CEDRA input had the strongest dataset mean on four of ten datasets, with improvements of 0.00013 RMSE on Wine Quality and 0.00005, 0.00239, and 0.00093 balanced-accuracy points on Rice, Optdigits, and Landsat. It trailed the strongest non-CEDRA input on six datasets; intervals excluded zero for losses on Abalone, Airfoil, MAGIC, Letter, and PenDigits, while none of the four positive intervals excluded zero.

K.10 OPTDIGITS TRANSDUCTIVE PREDICTIVE PROTOCOL

The predictive study is a separate 18-arm, 20-seed experiment on the single prespecified "near" randomized mechanism. Labels are split into 3,058 model-training, 765 validation, and 1,797 final-test nodes, while all feature rows and the randomized graph are available transductively. The 18 methods are MLP; GCN, GraphSAGE, GAT, GATv2, GIN, DropEdge, LINKX, and H2GCN; observed and candidate means; EdgeGate and an architecturally matched residual control; Navip; pure HT; outcome-only; CEDRA-HT; and CEDRA-SN. Conventional arms use all 64 training-standardized features, with no graph for MLP and the realized graph for graph baselines. Candidate-aware arms instead use disjoint 32-coordinate checkerboard construction and message views scaled by $1/16$, together with the candidate population, A, and q. These are deliberately different information sets and are reported as such; estimator attribution relies on matched within-family contrasts, particularly EdgeResid.

Training receives sanitized features, training labels, and validation labels, but no materialized final-test labels. For each seed, training writes all 18 validation-selected checkpoints and their hashes before

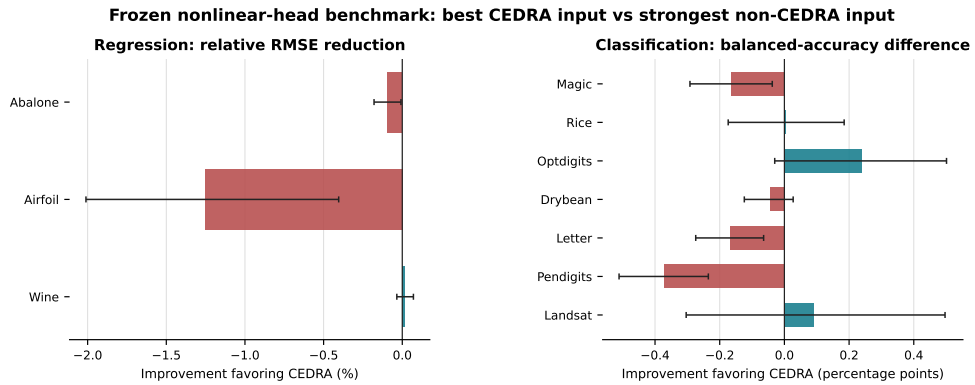


Figure 16: Complete frozen nonlinear-head effects. Regression values are relative RMSE reductions and classification values are balanced-accuracy percentage-point differences; positive values favor the better CEDRA input. Bars show dataset means and whiskers show paired seed-bootstrap 95% intervals. The axes use different units and should not be pooled.

final evaluation. A separate evaluator first validates all 20 seed directories and 360 checkpoints, independently recomputes every validation score, and freezes the validation-selected strongest non-CEDRA comparator. Only then may it load final-test labels once and emit aggregate balanced accuracy; it writes no row-level labels, predictions, probabilities, embeddings, or logits. The primary contrast is CEDRA-HT against that validation-selected comparator. Together with the corresponding CEDRA-SN contrast and fixed candidate-mean and Navip contrasts, the six ordered tests form one Holm-adjusted family. Paired 20,000-resample bootstrap intervals and exact two-sided paired sign tests use graph seed as the unit. Reports also include parameter counts, training time, peak GPU memory, and pre-label-load inference latency under their prespecified measurement scopes. All findings must be labeled semi-synthetic, transductive, and theorem-ineligible. Across all 18 arms, mean held-out balanced accuracy ranged from 0.9525 for GAT to 0.9693 for GraphSAGE. The validation-selected non-CEDRA comparator was outcome-only at 0.9688. Against direct candidate aggregation at 0.9653, CEDRA-HT was lower by 0.00262 with Holm-adjusted $p = 0.0103$, while the CEDRA-SN difference of -0.00226 did not survive correction. Against Navip at 0.9633, neither CEDRA contrast was distinguishable after correction. The most expensive CEDRA arm used about 2.28 GiB peak GPU memory and about 12.6 standalone accelerator-seconds per seed under the declared accounting scope.

K.11 Frozen end-to-end residual-adapter study

The CEDRA-Fuse study was frozen before its outcomes were inspected. It uses Cora, Roman-empire, Amazon-ratings, Minesweeper, Tolokers, and Questions; seeds 1 through 9; one matched GraphSAGE baseline; and one CEDRA-Fuse adapter that adds candidate-residual corrections through learned sigmoid gates. Both arms share the dataset splits, optimization protocol, stopping rule, and GraphSAGE backbone. All 54 paired cells are retained. The primary endpoint is the mean paired test balanced-accuracy difference, CEDRA-Fuse minus matched GraphSAGE, with a 200,000-resample seed-stratified bootstrap interval and an exact sign-flip test over the six dataset means. The public candidate supports depend on realized adjacency, and the adapter acts in hidden layers, so this experiment is explicitly theorem-ineligible. A frozen 54-cell end-to-end study compared a CEDRA residual adapter with matched GraphSAGE on six public graphs and nine retained seeds per graph. The mean paired balanced-accuracy difference was -0.00311, with a stratified seed-bootstrap 95% interval $[-0.00514, -0.00111]$: CEDRA-Fuse won 22 cells and lost 32, with positive dataset means on Roman-empire and Questions and negative means on the other four datasets.

Design diagnosis. The evaluated parameterization uses strictly positive sigmoid gates initialized near 0.05 and injects residual corrections in hidden layers. It therefore cannot represent the matched GraphSAGE predictor exactly during optimization, while the estimator guarantee covers neither those hidden-layer corrections nor the public candidate supports built partly from realized adjacency. A

Table 18: Frozen end-to-end CEDRA-Fuse comparison over nine seeds per public graph. Differences are CEDRA-Fuse minus matched GraphSAGE balanced accuracy; positive values favor CEDRA-Fuse. All datasets and seeds are shown in the analysis bundle.

Dataset	CEDRA-Fuse	GraphSAGE	Difference
Cora	0.83980	0.84574	-0.00594
Roman-empire	**0.72677**	0.72388	$+0.00289$
Amazon-ratings	0.45164	**0.45200**	-0.00036
Minesweeper	0.82008	**0.82158**	-0.00150
Tolokers	0.73193	**0.74775**	-0.01583
Questions	**0.67692**	0.67484	$+0.00208$
Mean paired difference			-0.00311

Table 19: Complete CEDRA-Guard successor study over nine seeds per public graph. Differences are CEDRA-Guard minus matched GraphSAGE test balanced accuracy. Every dataset and seed is retained.

Dataset	Guard	SAGE	Difference	95% interval	Zero / 9
Cora	0.8416	0.8457	-0.0041	[-0.0098, +0.0012]	8
Roman-empire	0.7261	0.7239	+0.0023	[-0.0012, +0.0064]	2
Amazon-ratings	0.4523	0.4520	+0.0003	[-0.0050, +0.0052]	1
Minesweeper	0.8187	0.8216	-0.0029	[-0.0064, +0.0008]	3
Tolokers	0.7326	0.7478	-0.0152	[-0.0225, -0.0080]	5
Questions	0.6813	0.6748	+0.0065	[-0.0024, +0.0156]	3

better predictive adapter should make the uncorrected predictor an exact zero-gate member of the model class, restrict the correction to the first pre-transform aggregate, normalize its scale, and select or reject the adapter using validation data only. These are testable design requirements rather than a claim of an unobserved improvement.

K.12 CEDRA-GUARD: A TASK-ALIGNED FIRST-LAYER ADAPTER

The CEDRA-Guard protocol was designed after the complete CEDRA-Fuse study and frozen before any successor test result was inspected. It is therefore a documented successor study rather than an untouched confirmation. It retains the same six datasets, seeds 1 through 9, matched GraphSAGE cells, primary balanced-accuracy endpoint, 200,000-resample stratified bootstrap, and nonselective reporting rule. Candidate supports are cosine top-32 sets constructed without labels or realized adjacency. Even feature columns define candidates and feed the nuisance inputs, while disjoint odd columns define candidate messages. Propensity and message nuisances are fitted on training receivers, selected on validation receivers, and frozen before supervised training. A clipped CEDRA-SN residual enters only the first layer after train-receiver RMS scale matching. Its coefficient is $0.5\tanh(\gamma)$ with $\gamma = 0$ initially, and each task checkpoint is compared with the exact $\alpha = 0$ path by validation balanced accuracy. The realized graph still supplies only one deterministic evaluation pattern, and the practical operator remains clipped and self-normalized, so this design reduces candidate and optimization leakage while being evaluated as a separate task-trained adapter.

A complete 54-cell successor study tested the task-aligned CEDRA-Guard adapter with feature-only candidates, separately fitted and frozen nuisances, first-layer-only correction, scale normalization, a signed zero-centred gate, and validation selection of an exact zero path. Relative to matched GraphSAGE, its mean paired balanced-accuracy difference was -0.00219 with a stratified seed-bootstrap 95% interval $[-0.00467, 0.00027]$, 25 cell wins and 29 losses, and three positive and three negative dataset means. It improved on CEDRA-Fuse by 0.00091 on average; the pooled interval includes zero, the exact zero path was selected in 22 of 54 cells, and Tolokers remained negative at -0.01520 with interval $[-0.02249, -0.00804]$. These outcomes show that the zero-path design improves integration relative to the earlier adapter while leaving task alignment as the principal opportunity for predictive gains.

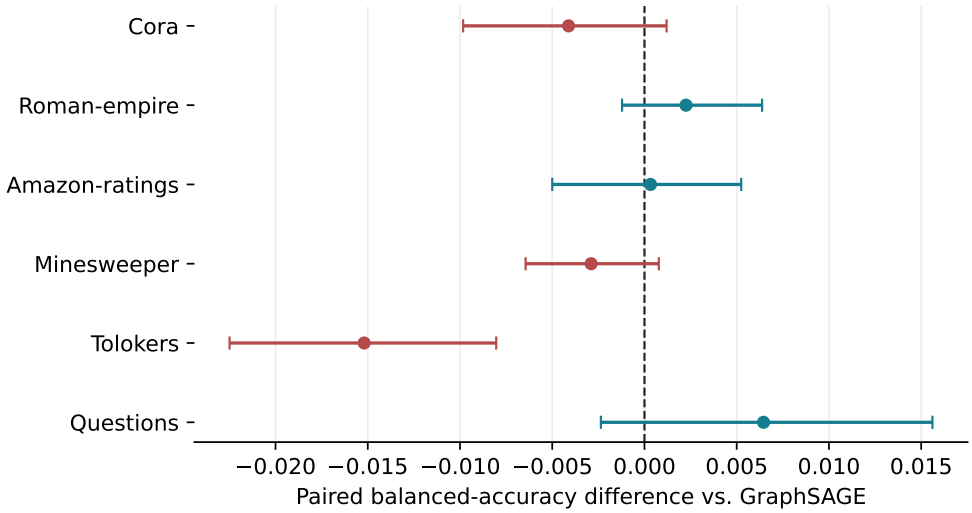


Figure 17: Complete CEDRA-Guard paired effects relative to matched GraphSAGE. Points are dataset means over nine retained seeds and bars are 95% seed-bootstrap intervals. Positive values favor CEDRA-Guard.

The successor improved over CEDRA-Fuse by 0.00091 in balanced accuracy (0.091 percentage points) on average and changed the dataset-mean count from two wins and four losses to three wins and three losses. Its GraphSAGE contrast nevertheless remained -0.00219, the interval crossed zero, and the exact two-sided sign-flip test over dataset means gave $p = 0.5625$. Roman-empire, Amazon-ratings, and Questions had positive means, while Cora, Minesweeper, and Tolokers had negative means. The validation-selected zero path in 22 cells confirms that the correction can now be rejected without changing the backbone parameters at inference, but it does not make the resulting checkpoint identical to a separately optimized GraphSAGE run. The remaining result supports a narrower conclusion: the original adapter defects explained part, but not all, of the predictive deficit, and aligning a first-layer candidate mean is not itself sufficient to optimize downstream classification.

L Expanded empirical results

This section displays the complete generated views used to interpret the summary tables in the main paper. The controlled figures expose the help, neutral, and harm regimes rather than reporting only an overall average. The Optdigits panels show every prespecified aggregate method and paired RMSE comparison. The public panels retain all six graphs, the selected comparators, paired CEDRA effects, and the measured performance-cost relationship. No panel is filtered by outcome.

The plots retain their archived labels: AIPW-HT and DRAPE-HT denote the HT implementations described here as CEDRA-HT; AIPW-SN and DRAPE-SN denote the corresponding CEDRA-SN implementations. Each plot retains its original nuisance, clipping, and data protocol, so this naming correspondence does not make distinct experimental settings interchangeable. The archived result bundle binds every study to its manifests, assumptions, comparators, metrics, uncertainty procedures, and evidence boundary; no numerical output has been changed by the renaming.

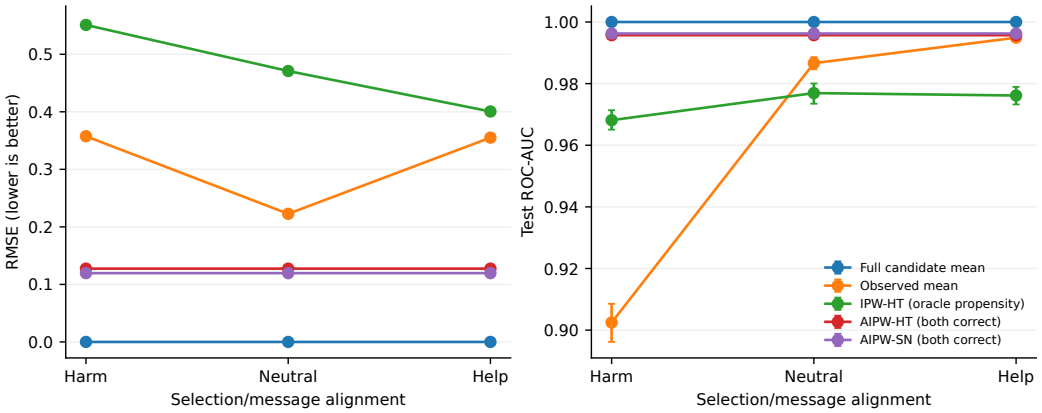


Figure 18: Controlled synthetic results, component A: target-recovery regimes across the complete factorial and learned-nuisance suites. The panels retain settings in which correction helps, is nearly neutral, or harms target recovery. Component B in Figure 19 shows the corresponding HT-versus-SN bias-variance comparison.

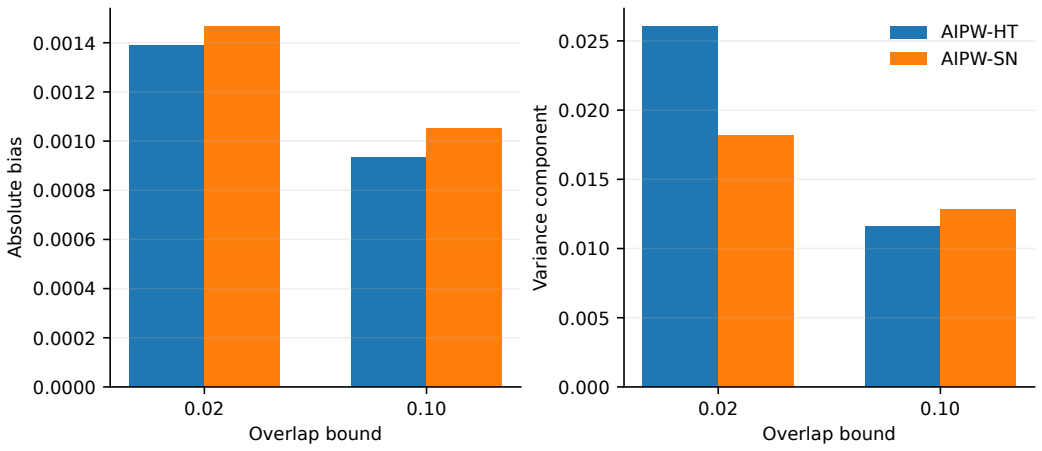


Figure 19: Controlled synthetic results, component B: the empirical bias-variance relationship between the exact HT form and the practical self-normalized form across the complete factorial and learned-nuisance suites. Target-recovery regimes are shown separately in Figure 18.

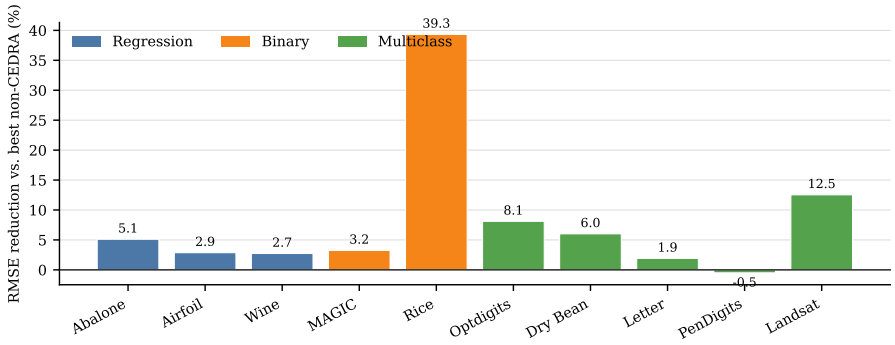


Figure 20: Relative target-recovery RMSE reduction of the best CEDRA variant against the strongest non-CEDRA estimator under near selection. Positive values favor CEDRA. All ten fixed datasets are shown; best-of-family selection is descriptive.

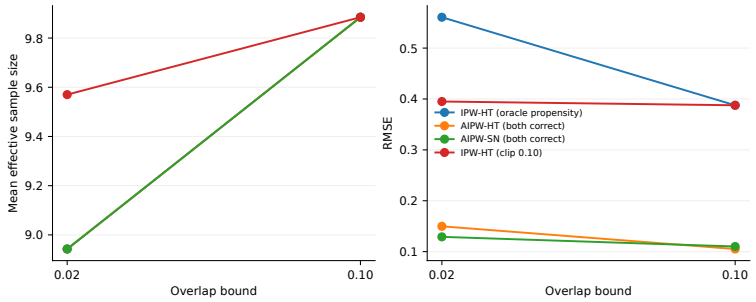


Figure 21: Controlled overlap and effective-sample-size diagnostics. The plot makes the positivity and weight-instability regimes visible rather than treating the overlap bound as a hidden implementation detail.

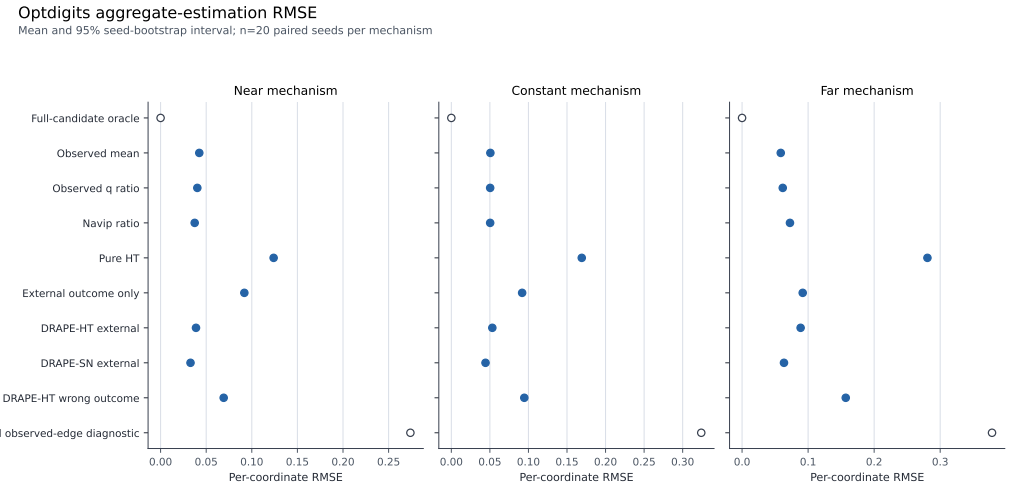


Figure 22: Complete Optdigits aggregate study, component A: RMSE for all declared estimators under near, constant, and far edge-selection mechanisms over 20 paired seeds. All preregistered paired effects and their uncertainty appear in component B, Figure 23.

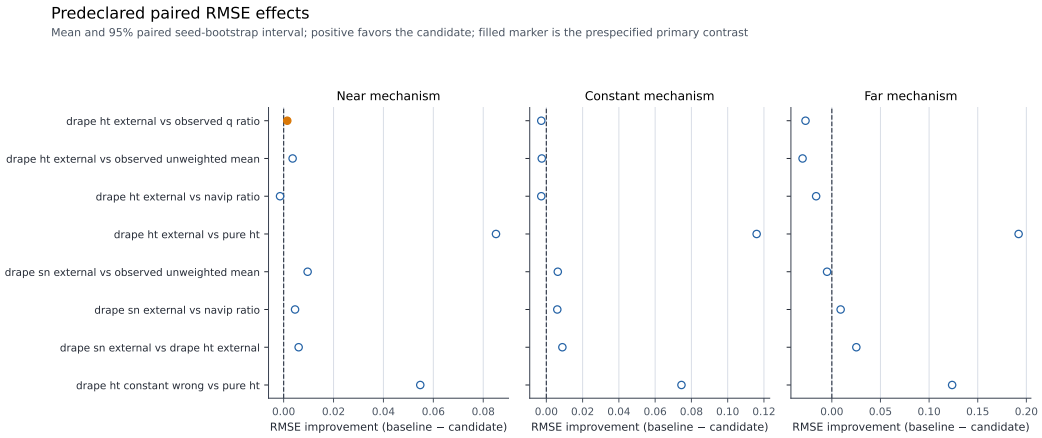


Figure 23: Complete Optdigits aggregate study, component B: all preregistered paired RMSE effects with uncertainty over 20 paired seeds. The complete comparison includes the far-mechanism result that favors the unweighted mean over CEDRA-SN. Component A in Figure 22 reports the estimator-level RMSE values.

Table 20: Aggregate estimation on official UCI Optdigits test receivers. Entries are means over 20 paired randomized seeds; intervals are 95% percentile bootstrap intervals. The graph is imposed semi-synthetically and candidate construction is transductive.

Mechanism	Method	RMSE	Grand-mean error norm	MAE
near	Full-candidate oracle	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
near	Observed mean	0.0424 [0.0423, 0.0426]	0.0015 [0.0015, 0.0016]	0.0243 [0.0243, 0.0244]
near	Observed q ratio	0.0403 [0.0401, 0.0405]	0.0009 [0.0009, 0.0010]	0.0226 [0.0225, 0.0227]
near	Navip ratio	0.0374 [0.0372, 0.0376]	0.0008 [0.0008, 0.0009]	0.0211 [0.0210, 0.0211]
near	Pure HT	0.1239 [0.1232, 0.1247]	0.0022 [0.0019, 0.0024]	0.0681 [0.0677, 0.0684]
near	External outcome only	0.0918 [0.0911, 0.0926]	0.0084 [0.0081, 0.0086]	0.0606 [0.0600, 0.0612]
near	CEDRA-HT external	0.0388 [0.0387, 0.0390]	0.0009 [0.0008, 0.0010]	0.0246 [0.0245, 0.0247]
near	CEDRA-SN external	0.0328 [0.0327, 0.0329]	0.0008 [0.0007, 0.0008]	0.0202 [0.0202, 0.0203]
near	CEDRA-HT wrong outcome	0.0692 [0.0689, 0.0695]	0.0015 [0.0013, 0.0016]	0.0431 [0.0430, 0.0433]
near	Failed observed-edge diagnostic	0.2739 [0.2734, 0.2743]	0.2384 [0.2381, 0.2388]	0.1852 [0.1849, 0.1854]
constant	Full-candidate oracle	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
constant	Observed mean	0.0506 [0.0505, 0.0508]	0.0022 [0.0021, 0.0024]	0.0291 [0.0290, 0.0292]
constant	Observed q ratio	0.0504 [0.0501, 0.0506]	0.0011 [0.0010, 0.0012]	0.0283 [0.0282, 0.0284]
constant	Navip ratio	0.0504 [0.0501, 0.0506]	0.0011 [0.0010, 0.0012]	0.0283 [0.0282, 0.0284]
constant	Pure HT	0.1691 [0.1677, 0.1705]	0.0028 [0.0023, 0.0034]	0.0927 [0.0920, 0.0934]
constant	External outcome only	0.0918 [0.0911, 0.0926]	0.0084 [0.0081, 0.0086]	0.0606 [0.0600, 0.0612]
constant	CEDRA-HT external	0.0532 [0.0529, 0.0535]	0.0012 [0.0011, 0.0013]	0.0336 [0.0334, 0.0337]
constant	CEDRA-SN external	0.0443 [0.0442, 0.0445]	0.0010 [0.0009, 0.0011]	0.0273 [0.0273, 0.0274]
constant	CEDRA-HT wrong outcome	0.0946 [0.0940, 0.0953]	0.0021 [0.0019, 0.0023]	0.0590 [0.0586, 0.0593]
constant	Failed observed-edge diagnostic	0.3241 [0.3237, 0.3245]	0.2829 [0.2826, 0.2833]	0.2203 [0.2200, 0.2205]
far	Full-candidate oracle	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]	0.0000 [0.0000, 0.0000]
far	Observed mean	0.0586 [0.0584, 0.0587]	0.0029 [0.0028, 0.0030]	0.0338 [0.0337, 0.0339]
far	Observed q ratio	0.0616 [0.0614, 0.0618]	0.0018 [0.0016, 0.0019]	0.0349 [0.0348, 0.0350]
far	Navip ratio	0.0725 [0.0723, 0.0727]	0.0018 [0.0016, 0.0019]	0.0409 [0.0408, 0.0410]
far	Pure HT	0.2807 [0.2787, 0.2826]	0.0065 [0.0055, 0.0077]	0.1515 [0.1506, 0.1524]
far	External outcome only	0.0918 [0.0911, 0.0926]	0.0084 [0.0081, 0.0086]	0.0606 [0.0600, 0.0611]
far	CEDRA-HT external	0.0887 [0.0879, 0.0894]	0.0019 [0.0017, 0.0020]	0.0531 [0.0526, 0.0535]
far	CEDRA-SN external	0.0635 [0.0633, 0.0636]	0.0014 [0.0013, 0.0015]	0.0395 [0.0394, 0.0396]
far	CEDRA-HT wrong outcome	0.1570 [0.1559, 0.1581]	0.0035 [0.0032, 0.0038]	0.0959 [0.0953, 0.0964]
far	Failed observed-edge diagnostic	0.3785 [0.3781, 0.3789]	0.3275 [0.3271, 0.3279]	0.2573 [0.2570, 0.2575]

Table 21: Optdigits predictive held-out balanced accuracy over 20 paired randomized graph seeds. Intervals are two-sided 95% paired-seed percentile bootstrap intervals. This is a transductive semi-synthetic prediction benchmark and is theorem-ineligible.

Model	Propensity	Mean	95% CI
MLP	joint	0.9604	[0.9594, 0.9614]
GCN	joint	0.9550	[0.9535, 0.9564]
GraphSAGE	joint	0.9693	[0.9682, 0.9704]
GAT	joint	0.9525	[0.9505, 0.9546]
GATv2	joint	0.9575	[0.9555, 0.9594]
GIN	joint	0.9530	[0.9515, 0.9544]
DropEdge	joint	0.9540	[0.9523, 0.9557]
LINKX	joint	0.9650	[0.9631, 0.9668]
H2GCN	joint	0.9653	[0.9633, 0.9672]
Observed mean	joint	0.9647	[0.9630, 0.9663]
Candidate mean	joint	0.9653	[0.9641, 0.9666]
EdgeGate	external	0.9636	[0.9625, 0.9647]
EdgeResid	external	0.9635	[0.9622, 0.9648]
Navip	external	0.9633	[0.9615, 0.9652]
Pure HT	external	0.9612	[0.9598, 0.9625]
Outcome only	joint	0.9688	[0.9679, 0.9696]
CEDRA-HT	external	0.9626	[0.9612, 0.9642]
CEDRA-SN	external	0.9630	[0.9616, 0.9643]

Table 22: All six predeclared Optdigits predictive contrasts form one Holm family at familywise $\alpha = 0.05$. Effects are left minus right, so positive values favor the CEDRA arm. Exact two-sided sign tests exclude exact-zero ties. Raw and Holm-adjusted p-values, multiplicity decisions, and win/tie/loss counts are generated from the validated aggregate artifact.

CEDRA arm	Comparator	Role	Effect	95% CI	Raw p	Holm p	Reject	W/T/L
CEDRA-HT	Outcome only	primary	-0.0061	[-0.0077, -0.0044]	4.005×10^{-5}	0.0002	yes	1/0/19
CEDRA-SN	Outcome only	secondary	-0.0058	[-0.0077, -0.0039]	0.0004	0.0020	yes	2/0/18
CEDRA-HT	Candidate mean	fixed	-0.0026	[-0.0045, -0.0005]	0.0026	0.0103	yes	3/0/17
CEDRA-SN	Candidate mean	fixed	-0.0023	[-0.0040, -0.0005]	0.1153	0.3460	no	6/0/14
CEDRA-HT	Navip	fixed	-0.0006	[-0.0020, 0.0009]	0.5034	1.0000	no	8/0/12
CEDRA-SN	Navip	fixed	-0.0003	[-0.0022, 0.0015]	0.5034	1.0000	no	12/0/8

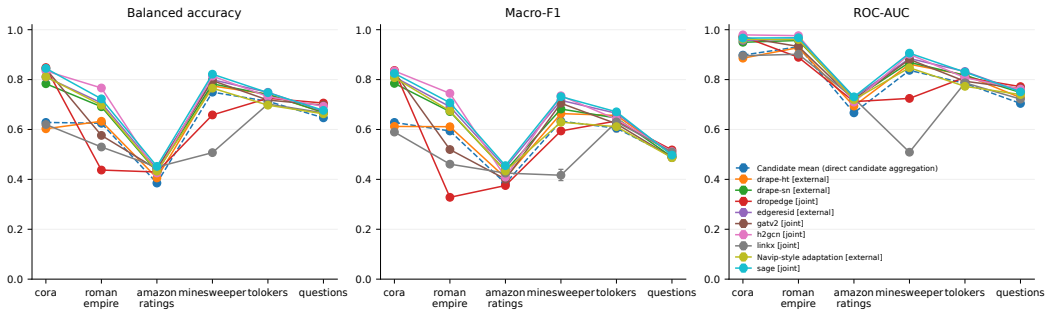


Figure 24: Complete public-graph performance record, component A: all declared arms on Cora, Roman-empire, Amazon-ratings, Minesweeper, Tolokers, and Questions. Component B in Figure 25 gives paired CEDRA effects. These are descriptive fixed-graph stress tests because the candidate sets use realized adjacency.

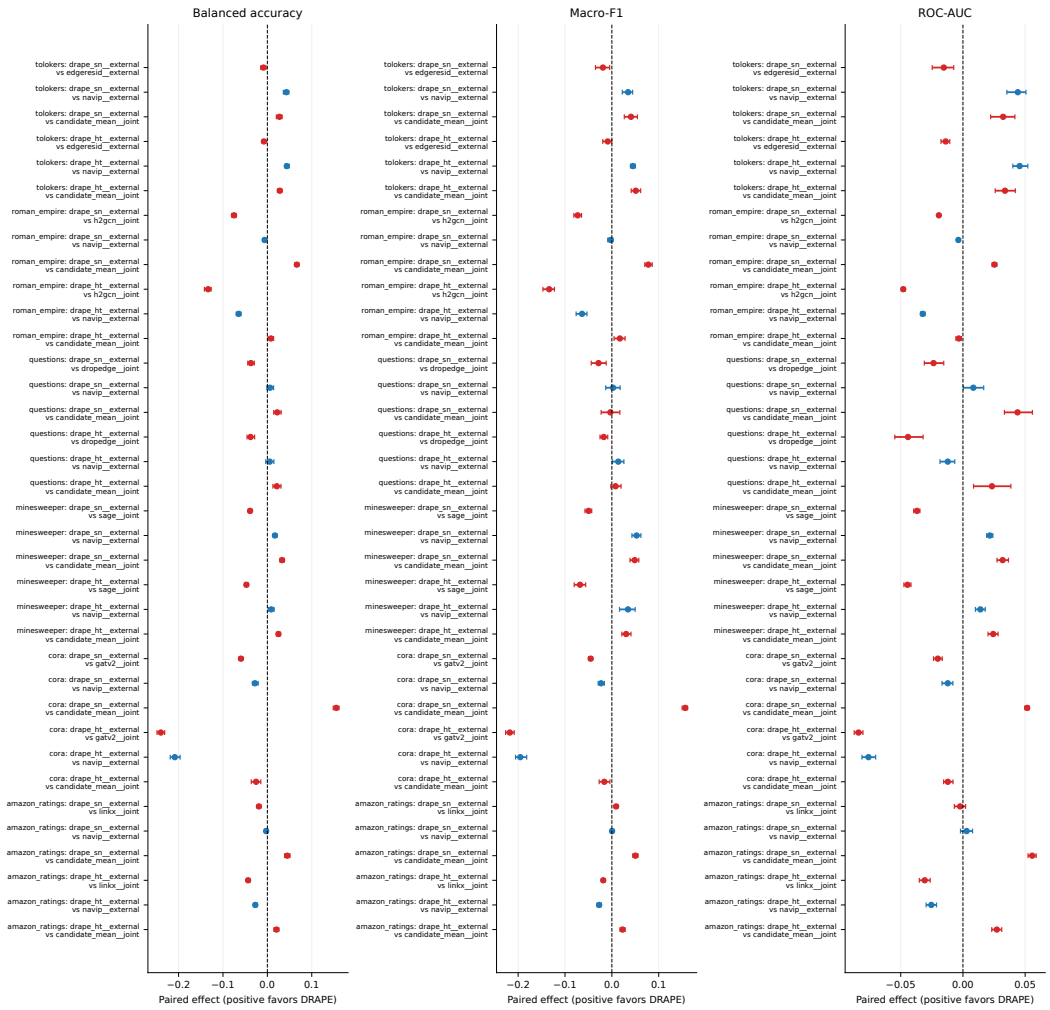


Figure 25: Complete public-graph performance record, component B: paired CEDRA effects against the validation-selected comparator and matched candidate-aware controls for all six fixed graphs. All declared-arm scores are shown in component A, Figure 24. The candidate sets use realized adjacency, so these comparisons remain descriptive fixed-graph stress tests.

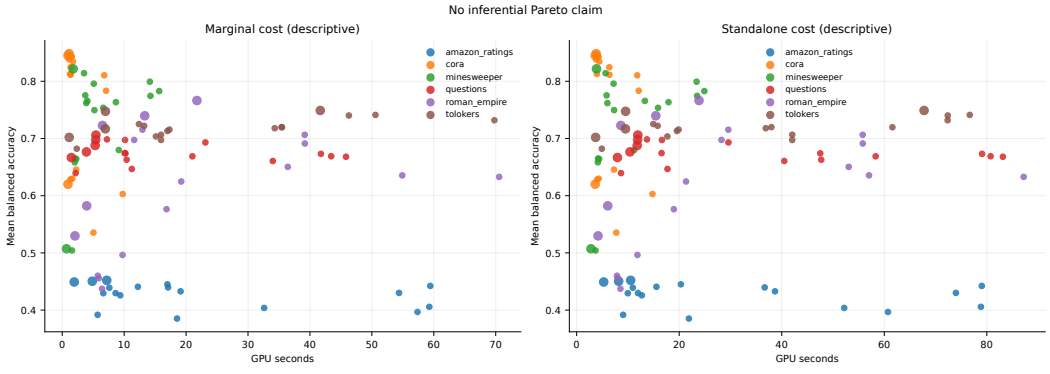


Figure 26: Public-graph performance and measured computational cost for the complete benchmark matrix. The figure reports the cost-quality tradeoff without claiming that the shared fixed configuration is literature-optimal for every baseline.