# SPARSEBANK: SPARSE MEASUREMENT FOR FIXED WORKFLOW BANKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

SparseBank studies sparse measurement for selecting a fixed workflow bank and routing queries within it. On the Automated Programming Progress Standard (APPS), support-aware $G$-opt lowers routed regret against Uniform without replacement at 200 and 280 unique edges. The gain persists after matching unique counts and reflects support and weights jointly. Prospective CodeContests supports the preregistered primary $90/10$ $G$-opt–Uniform mixture against Uniform. The unchanged sampler is a secondary replication arm and performs better than the mixture. HumanEval+ marks a boundary: ordinary least squares (OLS) shows no stable acquisition advantage and loses to a development-set predictor. Supporting theory bounds completion error and its transfer to bank and routed regret for realizable linear models with fresh independent noise. A realized-design corollary covers outcome-independent acquisition with full rank. The real-task evidence concerns frozen observed tables; it does not certify expected fresh-execution utility.

## 1 INTRODUCTION

A workflow bank stores complementary agentic procedures and routes each query to one of them. Dense evaluation runs all $K$ workflows on all $N$ queries. Each execution can invoke language models, tools, or external APIs (Yuan et al., 2026; Li et al., 2023). For a query–workflow pair $(i, k)$, let $Y_{ik}$ be fresh-execution utility and $\mu_{ik} = \mathbb{E}[Y_{ik}]$ its mean given frozen pre-target information. One execution per pair produces an observed table $U^{\mathrm{obs}}$. Estimating the mean table $\mu$ may require independent repeats.

SparseBank measures selected pairs, completes the table, and reuses the original curation and routing code (Figure 1). Acquisition and decisions concern the same fixed query–workflow population. Workflow generation and routing unseen queries are outside this transductive setting.

Fresh repeated executions can improve mean estimates; repeated access to a cached score adds no information. We study how acquisition weights translate into discrete measurements when each pair is evaluated once. Real-task replay evaluates decisions against $U^{\mathrm{obs}}$ and its own optimum. Established tools—optimal design, OLS, concentration, and perturbation bounds—provide supporting guarantees for the linear mean model.

Our empirical contribution is a workflow-bank acquisition diagnosis: unique-edge accounting explains part of the APPS routing gap, but a gain remains against matched Uniform-WOR. That gain reflects support and weights together and varies with the representation and budget. Prospective CodeContests tests the allocation on a second population; HumanEval+ shows that estimator and representation quality can matter more than acquisition.

## 2 RELATED WORK

**Workflow banks and algorithm portfolios.** FlowBank curates a compact portfolio and matches queries to reusable workflows (Yuan et al., 2026). The instance–candidate view is also classical in algorithm selection and portfolios (Rice, 1976; Gomes & Selman, 2001; Bischl et al., 2016). SparseBank asks how sparse measurement affects a fixed bank-and-route decision.



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

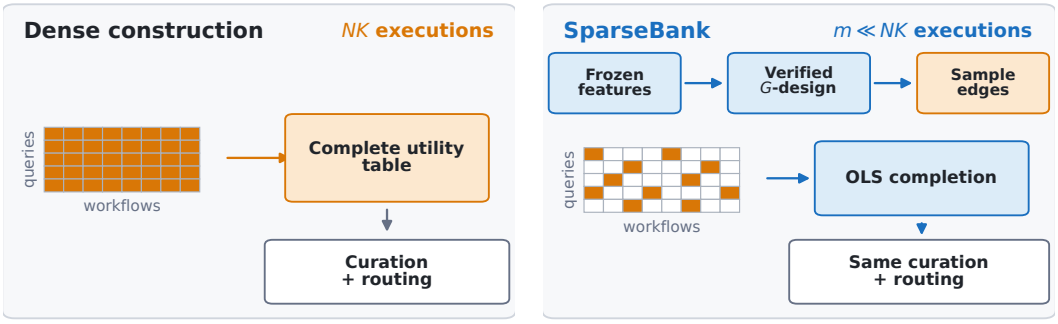



Figure 1: **SparseBank replaces exhaustive measurement with sparse acquisition and completion. Left:** dense construction executes all $NK$ query–workflow edges. **Right:** SparseBank freezes features, verifies a pre-label design, executes $m \ll NK$ edges, completes the table by OLS, and reuses the same curation and routing code. Certification requires the model, noise, and design conditions in Section 5.

**Sample-efficient evaluation and routing.** Active Testing selects labels to estimate model risk (Kossen et al., 2021). tinyBenchmarks estimates aggregate scores with item-response models and anchor examples (Maia Polo et al., 2024); multidimensional IRT models sparse responses through latent factors (Bergner et al., 2022). RouteLLM learns a strong-versus-weak router from preference data (Ong et al., 2025). SparseBank targets bank selection and routing across a fixed population.

**Optimal design and transductive identification.** Classical $G$-optimal design minimizes worst-case prediction variance and is equivalent to $D$-optimality (Kiefer & Wolfowitz, 1960); sequential design constructions are also classical (Wynn, 1970). Design also supports linear best-arm and transductive identification (Soare et al., 2014; Fiez et al., 2019). A complete bank-and-route policy can be encoded as one structured answer in the latter setting. Discrete experimental design already distinguishes repeated measurements from distinct-row selection and studies rounding with and without replacement (Allen-Zhu et al., 2017). SparseBank studies sparse measurement of a fixed query–workflow table with unchanged downstream curation and routing. Its contribution is empirical evidence about the consequences of realizing a continuous design at finite unique-measurement budgets in workflow-bank construction.

**Matrix completion with side information.** SoftImpute uses nuclear-norm regularization (Mazumder et al., 2010). IMC uses row/column side information (Jain & Dhillon, 2013), and goal-directed IMC learns nonlinear feature mappings (Si et al., 2016). Adaptive matrix completion selects entries to recover low-rank structure (Krishnamurthy & Singh, 2013). SparseBank uses frozen pair features and controls worst-edge prediction error. Our fair-feature comparison supplies comparable side information; the factorial holds observations fixed across estimators.

## 3 Problem Formulation

We consider $N$ queries and $K$ candidate workflows. For a positive integer $n$, write $[n] = \{1, \ldots, n\}$. We refer to each query–workflow pair $(i, k)$ as an edge and denote the set of all $M = NK$ edges by $\mathcal{E} = [N] \times [K]$. As in Introduction, $Y_{ik}$ denotes the utility obtained by executing workflow $k$ on query $i$, and $\mu_{ik}$ is its conditional mean.

Each workflow $k$ has a known cost $c_k \geq 0$. Given a fixed tradeoff parameter $\alpha \geq 0$, we define its cost-adjusted expected utility on query $i$ as $u_{ik} = \mu_{ik} - \alpha c_k$. Thus, our objective accounts for the cost of the workflow used to answer each query. It includes no separate charge for retaining workflows in a bank or acquiring observations. We evaluate performance across queries using weights $w_i \geq 0$, with $\sum_{i=1}^{N} w_i = 1$.

A *bank* is a nonempty subset of workflows available for routing. Given an integer budget $1 \leq B \leq K$, a bank may contain at most $B$ workflows. We denote the collection of feasible banks by $\mathcal{S}_B := \{S \subseteq [K] : 1 \leq |S| \leq B\}$. For a chosen bank $S \in \mathcal{S}_B$, a *route* is a mapping $r : [N] \to S$, where $r(i)$ specifies the workflow used to answer query $i$.

For a feasible bank $S \in \mathcal{S}_B$ and a route $r : [N] \to S$, we define

$$F(S) = \sum_{i=1}^{N} w_i \max_{k \in S} u_{ik}, \qquad R(S, r) = \sum_{i=1}^{N} w_i u_{i,r(i)}.$$

Here, $F(S)$ evaluates the bank under optimal routing within $S$, whereas $R(S, r)$ evaluates the assignments specified by $r$.

Several banks or workflows may attain the same maximum utility. To resolve ties, both pipelines use fixed deterministic rules: $T_{\text{bank}}$ selects one maximizing bank, and $T_{\text{route}}$ selects one maximizing workflow index. We define the oracle bank as $S^\star = T_{\text{bank}}(\arg\max_{S \in \mathcal{S}_B} F(S))$. For any bank $S$, its optimal route is $r_S(i) = T_{\text{route}}(\arg\max_{k \in S} u_{ik})$ for each query $i \in [N]$, so that $R(S, r_S) = F(S)$.

A completed table $\widehat{\mu}$ gives estimated cost-adjusted utilities $\widehat{u}_{ik} = \widehat{\mu}_{ik} - \alpha c_k$ and the bank objective $\widehat{F}(S) = \sum_{i=1}^{N} w_i \max_{k \in S} \widehat{u}_{ik}$. Curation selects a feasible bank $\widehat{S}$ within an additive tolerance $\eta \geq 0$ of the global optimum: $\widehat{F}(\widehat{S}) \geq \max_{S \in \mathcal{S}_B} \widehat{F}(S) - \eta$. Exact curation uses $T_{\text{bank}}$ to choose among maximizing banks. Exhaustive search certifies $\eta = 0$, while a heuristic requires a valid upper bound on the global optimum to certify its gap. Each query is then assigned to a workflow maximizing its estimated utility within the selected bank: $\widehat{r}(i) = T_{\text{route}}(\arg\max_{k \in \widehat{S}} \widehat{u}_{ik})$.

The *bank regret*, $\text{Reg}_{\text{bank}} = F(S^\star) - F(\widehat{S})$, measures the loss from selecting $\widehat{S}$, assuming optimal routing within it. The *routed regret*, $\text{Reg}_{\text{route}} = F(S^\star) - R(\widehat{S}, \widehat{r})$, also accounts for the route actually used. Both evaluate the final decisions rather than cumulative losses during acquisition.

We also compare workflow and route identities. Bank overlap, $|\widehat{S} \cap S^\star|/|S^\star|$, is the fraction of oracle-bank workflows retained in $\widehat{S}$. Route agreement, $N^{-1} \sum_{i=1}^{N} \mathbf{1}\{\widehat{r}(i) = r_{S^\star}(i)\}$, is the fraction of queries assigned to the same workflow as the oracle route. Exact bank recovery means $\widehat{S} = S^\star$; exact route recovery means $\widehat{r}(i) = r_{S^\star}(i)$ for every $i \in [N]$. Small regret need not imply exact recovery, since different decisions can have similar or identical utility. Shared tie-breaking specifies a reproducible choice among maximizers but neither makes them unique nor guarantees identical decisions across tables.

For replay evaluation, the frozen observed utility table $U^{\text{obs}}$ replaces $\mu$ throughout these definitions, including in the dense oracle. Replay therefore evaluates decisions against the oracle constructed from that same observed table.

# 4 SPARSEBANK

## 4.1 FROZEN FEATURES AND CERTIFIED DESIGN

Before target utilities are observed, each edge $e = (i, k)$ receives features $z_e \in \mathbb{R}^d$. Features, design choices, and thresholds are then fixed. Let $Z$ contain the rows $z_e^\top$ and let $d = \text{rank}(Z) \geq 1$ be the effective dimension (compact-SVD rank $r = d$). Orthogonal span compression preserves row norms. A common rescaling ensures $\|z_e\|_2 \leq 1$ and is absorbed into the parameter (Appendix A.2).

APPS and CodeContests use features derived from public text and workflow descriptors. HumanEval+ uses workflow indicators and a utility prediction trained on DEV. Table 4 specifies their construction, normalization, ranks, and freeze timing.

A design distribution $\pi \in \Delta(\mathcal{E}) = \{\pi : \pi_e \geq 0, \sum_e \pi_e = 1\}$ assigns sampling probabilities to edges, with support $\{e : \pi_e > 0\}$. Its information matrix is $\Sigma(\pi) = \sum_e \pi_e z_e z_e^\top$, an uncentered second moment that we call the design covariance. Leverage $z_e^\top \Sigma(\pi)^{-1} z_e$ measures prediction uncertainty along edge $e$'s direction. A $G$-optimal design minimizes the maximum leverage over all population edges, with singular designs assigned infinite objective value.

We compute the design using an active-set solver for the equivalent $D$-optimal problem. To certify the result, a separate scan of all edges must verify that $\Sigma(\pi)$ is nonsingular and that $\max_{e\in\mathcal{E}} z_e^\top \Sigma(\pi)^{-1} z_e \le \kappa d$. The coverage factor $\kappa \ge 1$ bounds maximum leverage divided by $d$; it is not a condition number. The computed ratio $\widehat{\kappa}$ alone is a numerical estimate, and does not constitute a certificate. If either check cannot be verified, no certificate is issued.

### 4.2 Sparse execution and OLS completion

Canonical SparseBank samples $E_t \overset{\text{i.i.d.}}{\sim} \pi$ with replacement for $t = 1, \dots, m$ and observes $Y_t = \mu_{E_t} + \xi_t$. Under the theoretical sampling model, each draw triggers a fresh execution. Let $X \in \mathbb{R}^{m\times d}$ have row $z_{E_t}^\top$ and let $Y = (Y_1, \dots, Y_m)^\top$. Unregularized minimum-norm OLS gives $\widehat{\theta} = X^\dagger Y$ and $\widehat{\mu}_e = z_e^\top \widehat{\theta}$, where $X^\dagger$ is the Moore–Penrose inverse. We use the unnormalized Gram matrix $V = X^\top X$, so $V/m$ is the empirical second moment. Costs remain $\widehat{u}_{ik} = \widehat{\mu}_{ik} - \alpha c_k$.

The fresh-execution budget $m$ differs from the distinct-edge count $|\{E_1, \dots, E_m\}| \le m$. Our without-replacement experiments use a unique-edge budget, with one OLS row per selected pair. To preserve the original APPS experimental protocol, its replay retains repeated draws as OLS rows with the same stored score, maintaining the original sampling stream and its observation weights.

The original curation routine returns a bank satisfying

$$\widehat{F}(\widehat{S}) \ge \max_{S\in\mathcal{S}_B} \widehat{F}(S) - \eta, \tag{1}$$

with $\eta$, $\widehat{F}$, and the maximizing route defined in Section 3. A multiplicative approximation guarantee alone does not certify an arbitrary additive $\eta$.

### 4.3 Certification conditions

Canonical certification requires the stated realizability and noise assumptions, a verified leverage bound, the sample-size condition (2), and sampled full rank. Corollary 4 instead computes an error bound directly from the realized Gram matrix, without requiring the population leverage bound or sample-size condition. It applies when acquisition and stopping are independent of target outcomes and $X$ has full column rank. Both guarantees use the prescribed curation, costs, routing, and tie-breaking rules. Design checks alone do not establish the model and noise assumptions; Appendix C lists the full requirements. These assumptions are not established for our real-task studies, whose results we evaluate empirically.

### 4.4 Empirical support-aware deduplicated acquisition

For experiments with a unique-edge budget, we sample distinct edges using the frozen $G$-optimal weights $\pi$. Let $\mathcal{O}_{t-1} = \{E_1, \dots, E_{t-1}\}$ contain the previously selected edges, and fix a threshold $\tau \ge 0$. At each step, form the eligible set $\mathcal{A}_t = \{e \notin \mathcal{O}_{t-1} : \pi_e > \tau\}$ and, when it is nonempty, select $E_t = e$ with probability $\pi_e / \sum_{a\in\mathcal{A}_t} \pi_a$. This preserves the relative weights among eligible edges without refitting the design. Sampling stops after $m$ distinct edges have been selected or the eligible set is exhausted.

Plain deduplication corresponds to $\tau = 0$. For support-aware deduplication on APPS, we chose $\tau = 10^{-10}$ based on the scale of the original weights before evaluating the thresholded variant. Thresholds of $10^{-8}$ and $10^{-6}$ retain the same support. For CodeContests, both the sampler and threshold were fixed before target execution.

OLS and downstream decisions remain unchanged. Since selection and stopping depend only on fixed weights and previously selected edge identities, the relevant guarantee is Corollary 4, provided its model, noise, and rank assumptions hold. If the eligible support is exhausted, $X$ contains the actual number $n \le m$ of selected rows. Because thresholding can remove feature directions, full column rank must be checked on this realized matrix.

# 5 THEORY

We first bound the uniform prediction error of the completed utility table, then translate this bound into bank and routing regret. Proofs are in Appendix A. For Theorem 1 and Corollary 3, assume the canonical sampling and OLS procedure, frozen features spanning $\mathbb{R}^d$, the verified leverage bound, and exact realizability $\mu_e = z_e^\top \theta^\star$. Fix $0 < \delta < 1$ before observing target outcomes. We condition on the frozen pre-target information and assume that, given $E_{1:m}$, the execution noises are independent and satisfy $\mathbb{E}[e^{\lambda \xi_t} \mid E_{1:m}] \le e^{\lambda^2 \sigma^2/2}$ for every $\lambda \in \mathbb{R}$, with a common sub-Gaussian parameter $\sigma \ge 0$.

**Theorem 1** (Uniform completion error)**.** *Under the assumptions above, if*

$$m \ge 8\kappa d \log\left(\frac{2d}{\delta}\right), \tag{2}$$

*then with probability at least* $1 - \delta$*,*

$$\varepsilon_\infty := \max_{e \in \mathcal{E}} |\widehat{\mu}_e - \mu_e| \le r_m(\delta) := 2\sigma \sqrt{\frac{\kappa d \log(4|\mathcal{E}|/\delta)}{m}}. \tag{3}$$

The proof first controls the sampled Gram matrix, then applies conditional OLS concentration and a union bound over all edges.

Uniform prediction accuracy also controls the quality of the selected bank and routes. The following result applies to any completed table, regardless of how it is estimated.

**Theorem 2** (Deterministic decision transfer)**.** *Suppose* $\max_{i,k} |\widehat{\mu}_{ik} - \mu_{ik}| \le \varepsilon$*, curation satisfies Eq.* (1)*, and* $\widehat{r}(i) \in \arg\max_{k \in \widehat{S}} \widehat{u}_{ik}$*. Then*

$$\mathrm{Reg}_{\mathrm{bank}}(\widehat{S}) \le \eta + 2\varepsilon, \tag{4}$$

$$\mathrm{Reg}_{\mathrm{route}}(\widehat{S}, \widehat{r}) \le \eta + 2\varepsilon. \tag{5}$$

Since $\widehat{r}$ maximizes the completed utilities within $\widehat{S}$, its value on the completed table equals $\widehat{F}(\widehat{S})$. Comparing this value directly with the dense optimum yields the same $2\varepsilon$ term for both regrets, without any statistical assumptions. Combining Theorems 1 and 2 gives simultaneous bank and routing regret bounds under canonical sampling.

**Corollary 3** (Certified bank and route guarantees)**.** *Under the assumptions of Theorems 1 and 2, with probability at least* $1 - \delta$*,*

$$\mathrm{Reg}_{\mathrm{bank}}(\widehat{S}) \le \eta + 2r_m(\delta), \tag{6}$$

$$\mathrm{Reg}_{\mathrm{route}}(\widehat{S}, \widehat{r}) \le \eta + 2r_m(\delta). \tag{7}$$

These bounds separate the curation error $\eta$ from the statistical error $2r_m(\delta)$, which decreases at rate $m^{-1/2}$ once the sample-size condition holds. A complementary guarantee uses the realized sample matrix directly and also accommodates sampling without replacement.

**Corollary 4** (Completion conditional on the realized design)**.** *Fix* $0 < \delta < 1$ *before observing target outcomes. Consider acquisition and stopping rules that do not use observed target utilities, and let* $\mathcal{D}$ *collect the frozen information, selected edges, and actual sample count* $n$*. Conditional on* $\mathcal{D}$*, assume exact realizability and independent centered* $\sigma$*-sub-Gaussian noises. For every full-rank* $X \in \mathbb{R}^{n \times d}$*, let* $V = X^\top X$ *and define*

$$L(V) = \max_{e \in \mathcal{E}} z_e^\top V^{-1} z_e, \qquad r_{\mathrm{obs}}(\delta) = \sigma \sqrt{2L(V) \log(2M/\delta)}. \tag{8}$$

*Then*

$$\Pr(\|\widehat{\mu} - \mu\|_\infty \le r_{\mathrm{obs}}(\delta) \mid \mathcal{D}) \ge 1 - \delta.$$

*If curation satisfies* (1) *and routing maximizes the completed utilities within* $\widehat{S}$*, both decision regrets are at most* $\eta + 2r_{\mathrm{obs}}(\delta)$ *on the same event.*

This bound requires neither a population leverage bound nor the canonical sample-size condition. It applies to the actual retained observations, provided the full-column-rank and conditional model assumptions hold.

Small regret need not imply identical decisions. Strict decision margins provide sufficient conditions for exact bank and route recovery. Define $\Gamma_P = F(S^\star) - \max_{S \neq S^\star} F(S)$ as the gap to the best alternative bank, and $\Gamma_R = \min_i [u_{i,r_{S^\star}(i)} - \max_{k \in S^\star \setminus \{r_{S^\star}(i)\}} u_{ik}]$ as the smallest gap between the optimal route and an alternative workflow within $S^\star$. Empty maxima are $-\infty$, so $\Gamma_R = +\infty$ when $S^\star$ is a singleton.

Under the curation and routing conditions of Theorem 2, a uniform completion error of at most $\varepsilon$ guarantees $\widehat{S} = S^\star$ and $\widehat{r}(i) = r_{S^\star}(i)$ for every query whenever

$$\Gamma_P > \eta + 2\varepsilon, \qquad \Gamma_R > 2\varepsilon. \tag{9}$$

Appendix A.5 proves this result. Redundant additions to an optimal bank can make $\Gamma_P = 0$, and deterministic tie-breaking alone does not create a positive margin. Thus, low regret can coexist with a different bank or route.

**Theorem 5** (Dimension-dependent lower bound)**.** *For any integer $d \geq 1$ and $\sigma > 0$, consider $N = d$ queries and $K = B = 2$ workflows, with a known bank $S_0 = \{1, 2\}$, uniform query weights, and zero costs. For $a > 0$, define the family $\mu_{i1}^{(s)} = as_i$ and $\mu_{i2}^{(s)} = -as_i$, indexed by $s \in \{-1, +1\}^d$. This family has a fixed $d$-dimensional linear feature representation for which the uniform design is exactly $G$-optimal with $\kappa = 1$. Assume fresh independent Gaussian execution noise with variance $\sigma^2$. For any $\varepsilon > 0$, set $a = 4\varepsilon$. Every procedure that selects edges using past observations and internal randomness with a fixed budget $m \leq \sigma^2 d/(64\varepsilon^2)$ satisfies $\sup_s \mathbb{E}_s[\mathrm{Reg}_{\mathrm{route}}] \geq 2\varepsilon$. Thus worst-case expected regret at most $\varepsilon$ requires*

$$m = \Omega\left(\frac{\sigma^2 d}{\varepsilon^2}\right) \tag{10}$$

*For a fixed amplitude $a > 0$, the within-query utility gap is $\Delta = 2a$. Recovering all routes within $S_0$ with probability at least $3/4$ uniformly over $s$ requires $m = \Omega(\sigma^2 d/\Delta^2)$.*

Each query in this construction requires identifying a hidden sign. An Assouad argument relates the information needed to recover these signs to routed regret (Yu, 1997). The result shows that linear dependence on $d$ and inverse-quadratic dependence on the target regret are unavoidable in general, even when the bank is already known. It does not establish bank-selection difficulty or the necessity of the coverage and logarithmic factors in the uniform completion bound. Appendix A.6 gives the construction and proof, together with a confidence-dependent refinement for exact route recovery.

**Misspecification and sparse budgets.** Under the outcome-independent acquisition and conditional noise assumptions of Corollary 4, suppose $X$ is full rank and $\mu = Z\theta + b$ with $|b|_\infty \leq b_\infty$. Appendix A.7 gives the error bound $r_{\mathrm{obs}}(\delta) + (1 + \Lambda_X)b_\infty$, where $\Lambda_X = \max_e |z_e^\top X^\dagger|_1$. The additional term accounts for approximation error and its amplification by OLS. A fitted residual alone does not establish the required uniform bound $b_\infty$. Appendix A.8 gives an explicit sufficient budget for a prescribed decision regret. When this budget is below $M = NK$, the canonical guarantees apply using fewer executions than exhaustive evaluation.

## 6 Experiments

Two controlled studies test certification and separate acquisition from estimator effects. The real-task evidence follows APPS from equal draws to equal unique counts and matched Uniform-WOR, then tests CodeContests replication and the HumanEval+ boundary. Metrics are defined in Appendix B.1. We separate prospective decisions from replay on the completed observed table.

### 6.1 Theorem-valid certification

A controlled population has $N = 100$, $K = 20$, $d = 8$, exactly realizable utilities, and independent Gaussian noise with $\sigma = 0.002$. An all-edge scan verifies $\kappa = 1.00000006$, giving the gate $m \geq 370$ at $\delta = 0.05$. Canonical acquisition and OLS pass certification for 100 seeds at each budget $m =$

Table 1: **APPS allocation matters after matching unique counts.** Means over 100 matched seeds with identical OLS and decisions. Support-aware $G$-opt and Uniform-WOR each observe exactly $m$ unique edges; the former retains original relative weights on edges with $\pi_e > 10^{-10}$. The remaining arms show the equal-draw comparison.

| | $m = 200$ | | $m = 280$ | |
|---|---|---|---|---|
| Method | Routed regret | Bank regret | Routed regret | Bank regret |
| Support-aware $G$-opt | 0.12465 | 0.00482 | 0.09205 | 0.00352 |
| Uniform-WOR | 0.17010 | 0.01931 | 0.12446 | 0.01071 |
| Canonical $G$-opt | 0.18684 | 0.01699 | 0.16006 | 0.00663 |
| Uniform-WR | 0.20039 | 0.03206 | 0.14119 | 0.01276 |
| Task-balanced | 0.16905 | 0.02216 | 0.12103 | 0.00983 |

$400, 600, 800, 1200$ (20–60% of $NK$). Exhaustive bank search gives $\eta = 0$. The largest maximum error, $0.001293$, is below its $m = 400$ radius $0.001958$. This checks the guarantee in its model setting; practical tightness relative to decision margins remains unestablished (Appendix B.5).

### 6.2 Acquisition-by-estimator factorial

A $3 \times 2$ factorial crosses $G$-opt, Uniform, and Leverage with OLS and ridge ($\lambda = 0.01$) on one $N = 60$, $K = 12$, $d = 12$ population. Each acquisition supplies identical observations to both estimators at $m = 72, 144, 216, 288, 360$ over ten matched seeds. Noise streams differ across some acquisitions. These empirical-only runs fall below the 593-sample canonical gate and have no established realized-design certificate.

For realizable utilities, $G$-opt has lower mean RMSE with either estimator (Table 7). At misspecification strengths $0.05$ and $0.10$, Uniform and Leverage have lower RMSE with both estimators. Yet $G$-opt retains lower mean routed regret: prediction accuracy and decision quality can rank acquisitions differently. Appendix B.4 gives the full factorial; Appendix B retains historical controls.

### 6.3 APPS: geometry and finite-budget realization

APPS (Hendrycks et al., 2021) uses $N = 100$, $K = 8$, $B = 3$, and $\alpha = 0$. Tasks, rank-78 features, and masks were frozen before target utilities. The original sparse union covered 555 edges; 245 further executions completed the table. All follow-ups reuse this observed $100 \times 8$ table with unchanged OLS, exhaustive curation, routing, and 100 matched seeds. The score is the number of official tests passed before the first failure divided by the full test count. We distinguish Uniform with replacement (Uniform-WR) from Uniform without replacement (Uniform-WOR).

**Equal draws favor Task-balanced acquisition.** Canonical $G$-opt has lower bank regret and higher overlap than Uniform-WR and Leverage at $m = 200, 280, 400$. Its routed regret is higher at $m = 280, 400$. At 200 and 280 draws, it observes only 131.40 and 160.67 unique edges on average. Task-balanced acquisition cycles through randomized tasks and distinct workflows. It observes exactly $m$ unique edges and lowers routed regret (Table 1). Equal draws therefore compare different amounts of unique information.

**Equal unique observations reverse that ranking.** We continue each unchanged stream until its $U$-th distinct edge and retain repeats as OLS rows. At $U = 120, 160, 190$, $G$-opt has lower routed regret than Task-balanced; all paired intervals exclude zero (Table 29). Task-balanced's three contrasts with Uniform include zero. Reaching $U = 190$ takes $G$-opt 394.58 draws on average versus Task-balanced's 190. Matching unique counts reverses the ranking, but edge identities and duplicate OLS weights still differ.

**Deduplication improves the finite-budget realization.** Support-aware $G$-opt selects each eligible edge at most once (Section 4.4). At $m = 200/280$, routed regret falls from $0.18684/0.16006$ to $0.12465/0.09205$ and bank regret from $0.01699/0.00663$ to $0.00482/0.00352$. Paired improve-

Table 2: **CodeContests replicates the routed-regret advantage at both unique-edge budgets.** Prospective seed 0 uses decisions frozen before dense execution; replay means use 100 frozen acquisition seeds. $M$ versus $U$ is primary; $G$ is the unchanged secondary arm. Paired differences and pointwise 95% intervals condition on the completed table.

| $m$ | Method | Prospective | Replay mean | Replay $\Delta$ vs. $U$ [95% interval] |
|---|---|---|---|---|
| 200 | $M$ (primary) | 0.197956 | 0.203203 | $-0.017245$ $[-0.028662, -0.006113]$ |
| 200 | $U$ | 0.276309 | 0.220449 | — |
| 200 | $G$ (secondary) | 0.160358 | 0.188634 | $-0.031814$ $[-0.041989, -0.021652]$ |
| 280 | $M$ (primary) | 0.147861 | 0.172516 | $-0.016311$ $[-0.025003, -0.008125]$ |
| 280 | $U$ | 0.230364 | 0.188827 | — |
| 280 | $G$ (secondary) | 0.142361 | 0.141710 | $-0.047118$ $[-0.054429, -0.040587]$ |

ments over canonical $G$-opt, Task-balanced, and Uniform-WR exclude zero at both budgets. This intervention changes both the acquired edges and OLS row multiplicities.

**Matched Uniform-WOR leaves an allocation gain.** Uniform-WOR uses the same successive sampler with uniform weights. Both arms observe exactly $m$ unique edges and fit one OLS row per edge. Support-aware $G$-opt still has lower routed and bank regret at $m = 200, 280$, with paired intervals excluding zero (Tables 1 and 31). Replacing Uniform-WR with Uniform-WOR reduces the mean routed-regret gap by 40.0% and 34.0%. These descriptive reductions leave a gain from support and weights jointly; they do not isolate the weights or identify a causal mediator (Appendix B.16).

Lower regret need not preserve bank identity. At $m = 280$, support-aware $G$-opt's overlap is $0.66667$, below Uniform-WOR's $0.72333$ and canonical $G$-opt's $0.84333$. Exact recovery falls from 57 to zero of 100 seeds relative to canonical $G$-opt. The sampler selects a different, near-optimal bank and still visits few tasks (Appendix B.14).

**Support exhaustion limits the usable budget.** Thresholds $10^{-10}$, $10^{-8}$, and $10^{-6}$ retain the same 288 edges and leave all $m = 200/280$ results unchanged. A 400-edge request exceeds this support. Plain deduplication then uses 112 numerical-tail edges and reaches routed regret $0.05715$. The support-aware sampler stops at 288 edges with regret $0.09425$ and no clear advantage over 400-draw Task-balanced. These unequal-budget endpoints are sensitivity results, not the matched evidence above (Appendix B.15).

### 6.4 Prospective CodeContests replication

CodeContests (Li et al., 2022) uses 100 deterministically selected TEST tasks with no prior workflow artifacts in the recorded project history. We freeze $K = 8$, $B = 3$, $\alpha = 0$, and rank-68 features. Utility is the fraction of all public, private, and generated test records passed, including tests after failures. All 492 sparse-union edges completed before the six fitted tables, banks, and routes were frozen. The 308 dense-remainder executions then completed the 800-edge table. Appendix B.17 specifies eligibility, selection, scoring, and timestamps.

The preregistered primary comparison is mixture-WOR ($M$) versus Uniform-WOR ($U$). The mixture puts 90% of its mass on $G$-opt and 10% on Uniform. A predeclared coverage rule chose this floor because $G$-opt support spans 288 edges but only 36 tasks. Unchanged support-aware $G$-opt ($G$) is the preregistered secondary replication arm. All arms acquire exactly $m$ unique edges with identical OLS and downstream decisions.

The primary $M - U$ contrast favors $M$ at both budgets in the prospective run and the 100-seed replay (Table 2). Both replay intervals are below zero. At $\alpha = 0$, the secondary $G$ arm replicates the APPS direction and has the lowest regret. The exploration floor hurts relative to $G$: replay $M - G$ differences are $+0.014569$ and $+0.030807$, with both intervals above zero.

Post-hoc CodeContests replay favors both $G$ and $M$ over Task-balanced-WOR on routed regret at both budgets ($\alpha = 0$). A downstream penalty on recorded model calls preserves $M$'s advantage over Uniform and Task-balanced across the tested grid; $G$'s advantage is unresolved at the highest penalty and $m = 280$ (Appendix B.18).

### 6.5 PROSPECTIVE HUMANEVAL+ BOUNDARY EVIDENCE

HumanEval+ (Liu et al., 2023) uses 60 previously unexecuted TEST2 tasks, four workflows, $B = 2$, and $\alpha = 0$. Features and acquisition masks were frozen before target utilities; sparse decisions were frozen before dense revelation. Weighted sampling without replacement uses $m = 60, 84, 120$ unique edges (25/35/50% of $NK = 240$). The study collected all 240 executions for evaluation.

The preregistered realization favored SparseBank. The 100-seed replay shows no stable acquisition advantage: budget-averaged routed regret is $0.068446/0.067444/0.067263$ for OLS with $G$-opt/Uniform/Leverage, and both paired acquisition intervals cross zero. DEV-only is stronger at $0.059476$. OLS is worse by $0.008970$ (95% interval $[0.006374, 0.011792]$).

Post-hoc Anchor adds a shrunk residual estimate to the DEV prediction, with shrinkage selected on DEV. Its $G$-opt routed regret is $0.057135/0.056016/0.056430$. Bank regret and route agreement remain worse than DEV-only, and acquisition superiority remains unestablished. Anchor is outside the canonical theorem. Appendix B.10 reports the estimator ablations and SoftImpute's advantage at the two larger budgets.

## 7 LIMITATIONS

Unique-edge accounting is necessary but does not explain the whole APPS routing gap. The residual gain reflects support and weights jointly, with its value depending on the representation and budget. HumanEval+ shows that estimator and representation quality can dominate acquisition.

The real-task studies do not establish the certification model. APPS's sufficient canonical gate is $5021 > NK = 800$ even at $\kappa = 1, \delta = 0.05$; it is not a necessary threshold. CodeContests' observed-table projection RMSE is $0.387952$ with $R^2 = 0.149229$. These residuals cannot separate execution noise from mean-model misspecification.

APPS follow-ups and the new CodeContests checks are exploratory. APPS and CodeContests use pointwise seed-bootstrap intervals on one observed table, without multiplicity adjustment. They do not quantify fresh-execution or population uncertainty. The original real-task protocols use $\alpha = 0$; only post-hoc CodeContests replay varies a downstream call-count penalty. Monetary and latency tradeoffs remain untested; construction budgets exclude DEV and dense-evaluation costs.

## 8 CONCLUSION

Matching unique measurements removes part of the APPS routing gap; the gain that remains against Uniform-WOR is replicated by CodeContests' secondary arm. This evidence concerns support and weights together at the tested representations and budgets. HumanEval+'s OLS loss to DEV-only shows why acquisition must be judged alongside the estimator and representation.

## AI USE STATEMENT

LLM-based assistants were used for code and analysis support, experiment planning, literature triage, and language editing. Checks covered source provenance, proofs, numerical outputs, and the LaTeX build. Assistants did not supply experimental measurements.

## REPRODUCIBILITY STATEMENT

The appendix gives proofs, protocols, budgets, metrics, and the feature specification in Table 4. The anonymous artifact `README.md` maps central tables to frozen inputs and replay commands. Appendix B.1 identifies legacy provenance gaps.

## REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal design of experiments via regret minimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 126–135. PMLR, 2017. URL https://proceedings.mlr.press/v70/allen-zhu17e.html.

Yoav Bergner, Peter Halpin, and Jill-Jênn Vie. Multidimensional item response theory in the style of collaborative filtering. *Psychometrika*, 87(1):266–288, 2022. doi: 10.1007/s11336-021-09788-9.

Bernd Bischl, Pascal Kerschke, Lars Kotthoff, Marius Lindauer, Yuri Malitsky, Alexandre Fréchette, Holger Hoos, Frank Hutter, Kevin Leyton-Brown, Kevin Tierney, and Joaquin Vanschoren. ASlib: A benchmark library for algorithm selection. *Artificial Intelligence*, 237:41–58, 2016. doi: 10.1016/j.artint.2016.04.003.

Tanner Fiez, Lalit Jain, Kevin G. Jamieson, and Lillian Ratliff. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/8ba6c657b03fc7c8dd4dff8e45defcd2-Abstract.html.

Carla P. Gomes and Bart Selman. Algorithm portfolios. *Artificial Intelligence*, 126(1–2):43–62, 2001. doi: 10.1016/S0004-3702(00)00081-3.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with APPS. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c24cd76e1ce41366a4bbe8a49b02a028-Abstract-round2.html.

Prateek Jain and Inderjit S. Dhillon. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*, 2013. doi: 10.48550/arXiv.1306.0626.

J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960. doi: 10.4153/CJM-1960-030-4.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Akshay Krishnamurthy and Aarti Singh. Low-rank matrix and tensor completion via adaptive sampling. In *Advances in Neural Information Processing Systems 26*, pp. 836–844, 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/hash/2050e03ca119580f74cca14cc6e97462-Abstract.html.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-Bank: A comprehensive benchmark for tool-augmented LLMs.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3102–3116. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.187.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097, 2022. doi: 10.1126/science.abq1158.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation. In *Advances in Neural Information Processing Systems 36*, pp. 21558–21572, 2023. doi: 10.52202/075280-0943.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80):2287–2322, 2010. URL https://www.jmlr.org/papers/v11/mazumder10a.html.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. RouteLLM: Learning to route LLMs from preference data. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8sSqNntaMr.

John R. Rice. The algorithm selection problem. In *Advances in Computers*, volume 15, pp. 65–118. Academic Press, 1976. doi: 10.1016/S0065-2458(08)60520-3.

Si Si, Kai-Yang Chiang, Cho-Jui Hsieh, Nikhil Rao, and Inderjit S. Dhillon. Goal-directed inductive matrix completion. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1165–1174, 2016. doi: 10.1145/2939672.2939809.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems 27*, 2014.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Henry P. Wynn. The Sequential Generation of D-Optimum Experimental Designs. *The Annals of Mathematical Statistics*, 41(5):1655–1664, 1970. doi: 10.1214/aoms/1177696809.

Bin Yu. Assouad, Fano, and Le Cam. In David Pollard, Erik Torgersen, and Grace L. Yang (eds.), *Festschrift for Lucien Le Cam*, chapter 29, pp. 423–435. Springer, 1997. doi: 10.1007/978-1-4612-1880-7_29.

Lingzhi Yuan, Chenghao Deng, Fangxu Yu, Souradip Chakraborty, Mohammad Rostami, and Furong Huang. FlowBank: Query-adaptive agentic workflows optimization through precompute-and-reuse. *arXiv preprint arXiv:2606.11290*, 2026. doi: 10.48550/arXiv.2606.11290.

## A Complete Proofs and Technical Details

This appendix supplies the proofs of the results in Section 5. We first make the theorem-mode assumptions precise, then prove the uniform completion theorem, the deterministic decision-transfer theorem and exact-recovery statement, and finally the dimension-dependent lower bound. The last sections record the misspecification boundary and several load-bearing implementation facts used by the fail-closed protocol.

### A.1 Notation and theorem-mode assumptions

Let $\mathcal{E} = [N] \times [K]$ and $M = |\mathcal{E}| = NK$. Every edge $e = (i, k)$ has a feature $z_e \in \mathbb{R}^d$ fixed before target utilities are observed. For the canonical completion result, $0 < \delta < 1$ and $\sigma \geq 0$. We condition on frozen features and pre-target information, including any separate DEV split. The canonical assumptions below apply to Theorem 1 and Corollary 3. Lemma 8 states the separate conditional assumptions used by Corollary 4; the adaptive lower bound uses its own sequential model. The exact-rank coordinates used by the design certificate and by OLS satisfy

$$\operatorname{span}\{z_e : e \in \mathcal{E}\} = \mathbb{R}^d, \qquad \max_{e \in \mathcal{E}} \|z_e\|_2 \leq 1. \tag{11}$$

A pre-label design $\pi \in \Delta(\mathcal{E})$ has covariance and certified maximum leverage

$$\Sigma := \sum_{e \in \mathcal{E}} \pi_e z_e z_e^\top \succ 0, \qquad \max_{e \in \mathcal{E}} z_e^\top \Sigma^{-1} z_e \leq \kappa d. \tag{12}$$

The realizable observation model is

$$\mu_e = z_e^\top \theta^\star, \qquad E_t \overset{\text{i.i.d.}}{\sim} \pi, \qquad Y_t = \mu_{E_t} + \xi_t, \qquad t \in [m]. \tag{13}$$

Conditionally on $E_{1:m} := (E_1, \ldots, E_m)$, the noises are independent, mean zero, and $\sigma$-sub-Gaussian:

$$\mathbb{E}[\xi_t \mid E_{1:m}] = 0, \qquad \mathbb{E}[\exp(\lambda \xi_t) \mid E_{1:m}] \leq \exp(\lambda^2 \sigma^2 / 2) \quad \forall \lambda \in \mathbb{R}. \tag{14}$$

Let $X \in \mathbb{R}^{m \times d}$ have rows $z_{E_t}^\top$, let $Y = (Y_1, \ldots, Y_m)^\top$, and retain repeated executions as repeated rows. SparseBank uses

$$\widehat{\theta} = X^\dagger Y, \qquad \widehat{\mu}_e = z_e^\top \widehat{\theta}, \qquad V := X^\top X = \sum_{t=1}^{m} z_{E_t} z_{E_t}^\top. \tag{15}$$

All costs, query weights, feasible banks, curation rule, optimizer certificate $\eta$, routing rule, and tie-breaking convention are shared by the dense and completed-table pipelines.

### A.2 Exact rank-span coordinates

The theorem is stated in the exact feature span rather than an arbitrary raw coordinate system. The following lemma shows that this change of coordinates does not alter any finite-population prediction, leverage score, or minimum-norm OLS prediction.

**Lemma 6** (Rank-span invariance)**.** *Fix an ordering $\mathcal{E} = \{e_1, \ldots, e_M\}$ and let $A \in \mathbb{R}^{M \times p}$ contain raw frozen feature rows $a_{e_j}^\top$. Write its compact singular-value decomposition as $A = U_r D_r V_r^\top$ and define $Z := U_r D_r$, whose $j$th row is $z_{e_j}^\top$. Then:*

$$a_e^\top \beta = z_e^\top \theta \qquad \text{for } \theta = V_r^\top \beta, \tag{16}$$

$$AA^\top = ZZ^\top, \tag{17}$$

$$a_e^\top \Sigma_A(\pi)^\dagger a_e = z_e^\top \Sigma_Z(\pi)^{-1} z_e, \tag{18}$$

$$a_e^\top \widehat{\beta}_{\min} = z_e^\top \widehat{\theta}_{\min}. \tag{19}$$

*Here $r = d \geq 1$, $\Sigma_A(\pi) = \sum_e \pi_e a_e a_e^\top$, and $\Sigma_Z(\pi) = \sum_e \pi_e z_e z_e^\top$. For the inverse identity, require the positive-probability support to span $\mathbb{R}^r$; this ensures nonsingularity. The remaining identities hold for any sampled rank. Compression preserves row norms, rather than normalizing them. A common scalar normalization may be applied to both $A$ and $Z$, with reciprocal parameter scaling, preserving the Gram identity for those rescaled matrices.*

*Proof.* For the $j$th edge, $a_{e_j} = A^\top \mathbf{e}_j = V_r D_r U_r^\top \mathbf{e}_j = V_r z_{e_j}$. This immediately gives (16). Since $V_r^\top V_r = I_r$,

$$AA^\top = U_r D_r V_r^\top V_r D_r U_r^\top = U_r D_r^2 U_r^\top = ZZ^\top,$$

which proves (17). Moreover,

$$\Sigma_A(\pi) = V_r \Sigma_Z(\pi) V_r^\top, \qquad \Sigma_A(\pi)^\dagger = V_r \Sigma_Z(\pi)^{-1} V_r^\top,$$

so substituting $a_e = V_r z_e$ proves (18). For a sampled design, $X_A = X_Z V_r^\top$ and $(X_Z V_r^\top)^\dagger = V_r X_Z^\dagger$ because $V_r$ has orthonormal columns. Thus $a_e^\top \widehat{\beta}_{\min} = z_e^\top V_r^\top V_r X_Z^\dagger Y = z_e^\top \widehat{\theta}_{\min}$. □

### A.3 UNIFORM ACTIVE COMPLETION

We split the statistical proof into a covariance-concentration lemma and a conditional finite-edge prediction lemma.

**Lemma 7** (Whitened covariance concentration)**.** *Under* (12)*, define* $q_e := \Sigma^{-1/2} z_e$*. If*

$$m \geq 8\kappa d \log\left(\frac{2d}{\delta}\right), \tag{20}$$

*then, with probability at least* $1 - \delta/2$*,*

$$\sum_{t=1}^{m} q_{E_t} q_{E_t}^\top \succeq \frac{m}{2} I_d, \qquad \textit{equivalently} \qquad V \succeq \frac{m}{2}\Sigma. \tag{21}$$

*In particular,* $V$ *is nonsingular on this event.*

*Proof.* The whitened vectors obey

$$\mathbb{E}_{E\sim\pi}[q_E q_E^\top] = I_d, \qquad \|q_e\|_2^2 = z_e^\top \Sigma^{-1} z_e \leq \kappa d. \tag{22}$$

Set $A_t = q_{E_t} q_{E_t}^\top$. These matrices are independent and positive semidefinite, satisfy $\lambda_{\max}(A_t) \leq \kappa d$, and have $\mathbb{E}\sum_t A_t = mI_d$. The lower-tail matrix Chernoff inequality (Tropp, 2012), evaluated at relative deviation $1/2$, gives

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_{t=1}^{m} A_t\right) \leq \frac{m}{2}\right\} \leq d\exp\Big(-\frac{m}{8\kappa d}\Big).$$

The right-hand side is at most $\delta/2$ under (20). Conjugating the resulting matrix inequality by $\Sigma^{1/2}$ yields $V = \Sigma^{1/2}(\sum_t A_t)\Sigma^{1/2} \succeq (m/2)\Sigma$. □

**Lemma 8** (Conditional finite-edge OLS concentration)**.** *Fix* $0 < \delta < 1$ *before target outcomes and* $\sigma \geq 0$*. Let* $\mathcal{D}$ *include frozen information, selected edges, and actual count* $n$*, and let* $X \in \mathbb{R}^{n\times d}$ *contain the actual retained feature rows. Acquisition and stopping are independent of target outcomes. Conditional on* $\mathcal{D}$*, assume* $\mu = Z\theta^\star$ *and independent centered noises with conditional MGF bounded by* $\exp(\lambda^2\sigma^2/2)$*. For every full-column-rank* $X$*, let* $V = X^\top X$ *and* $L(V) = \max_{e\in\mathcal{E}} z_e^\top V^{-1} z_e$*. Then*

$$\Pr\Big(\|\widehat{\mu} - \mu\|_\infty \leq \sigma\sqrt{2L(V)\log(2M/\delta)} \;\Big|\; \mathcal{D}\Big) \geq 1 - \delta. \tag{23}$$

*Proof.* Full rank gives $X^\dagger = V^{-1}X^\top$ and

$$\widehat{\theta} - \theta^\star = V^{-1}X^\top\xi, \qquad \widehat{\mu}_e - \mu_e = \sum_{t=1}^{n} c_{e,t}\xi_t, \quad c_{e,t} = z_e^\top V^{-1} z_{E_t}. \tag{24}$$

Conditional on $\mathcal{D}$, these coefficients are fixed and

$$\sum_t c_{e,t}^2 = z_e^\top V^{-1}X^\top X V^{-1} z_e = z_e^\top V^{-1} z_e \leq L(V). \tag{25}$$

The product MGF bound gives variance proxy $\sigma^2 z_e^\top V^{-1} z_e$. For $\sigma > 0$, the two-sided tail at $r_{\mathrm{obs}}(\delta)$ is at most $\delta/M$ for each nonzero feature (a zero feature has zero error). A union bound over all $M$ population edges proves the claim; prediction errors need not be mutually independent. If $\sigma = 0$, every noise and prediction error vanishes almost surely. □

On the canonical covariance event (21), $L(V) \leq 2\kappa d/m$. Applying Lemma 8 with failure budget $\delta/2$ recovers exactly $2\sigma\sqrt{\kappa d\log(4M/\delta)/m}$, leaving $\delta/2$ for rank/covariance failure.

*Proof and interpretation of Corollary 4.* Apply Lemma 8, then Theorem 2 on the same completion event; both regrets are bounded simultaneously with no additional union bound. If $C = \{\mathrm{rank}(X) = d\}$, pointwise conditioning implies

$$\Pr(C \cap \{\text{error} > r_{\mathrm{obs}}(\delta)\}) \le \delta \Pr(C).$$

Thus coverage conditional on $C$ is at least $1-\delta$ when $\Pr(C) > 0$. An unconditional claim without abstention must also account for rank failure. Restricting to a geometry-only acceptance event preserves this conclusion. Acceptance based on small observed residuals need not do so.

All population edges enter $L(V)$, including unselected and ineligible edges. Use actual retained observations at support exhaustion; original full-rank $\Sigma(\pi)$ does not imply thresholded or sampled full rank. Neither a pseudoinverse nor shrinking $d$ to the sampled rank repairs missing population directions. The confidence level is fixed before target outcomes. This selected-design guarantee is not an anytime confidence sequence for label-dependent stopping.

The canonical and realized radii are separate guarantees. Taking their minimum at the same nominal $\delta$ is not automatically a $1-\delta$ guarantee. Such a comparison requires applicable assumptions for both and failure budgets summing to at most $\delta$; confidence splitting cannot make the canonical sampling premise hold for deduplication. □

*Proof of Theorem 1.* Let $\mathcal{A}$ denote the covariance event in (21). Lemma 7 shows $\mathbb{P}(\mathcal{A}^c) \le \delta/2$ whenever $m \ge 8\kappa d \log(2d/\delta)$. Conditional on $\mathcal{A}$, Lemma 8, applied with failure budget $\delta/2$ and the preceding leverage bound, fails with probability at most $\delta/2$. Therefore

$$\begin{aligned}\mathbb{P}\left[\max_{e\in\mathcal{E}} |z_e^\top(\widehat{\theta}-\theta^\star)| > 2\sigma\sqrt{\frac{\kappa d\log(4M/\delta)}{m}}\right] &\le \mathbb{P}(\mathcal{A}^c) + \mathbb{E}[\mathbf{1}_{\mathcal{A}}\mathbb{P}(\text{noise failure} \mid E_{1:m})] \\ &\le \delta. \end{aligned} \tag{26}$$

Finally, realizability implies $\widehat{\mu}_e - \mu_e = z_e^\top(\widehat{\theta}-\theta^\star)$, so this is exactly the entrywise completion bound in (3). □

### A.4 Deterministic decision transfer

The transfer result requires no realizability or noise assumption after a uniform table-error bound has been established.

**Lemma 9** (Stability of a fixed bank and its completed-table route)**.** *Suppose* $\max_{i,k} |\widehat{\mu}_{ik} - \mu_{ik}| \le \varepsilon$ *and the dense and completed pipelines subtract the same costs. Then, for every feasible bank* $S$*,*

$$|\widehat{F}(S) - F(S)| \le \varepsilon. \tag{27}$$

*Moreover, if* $\widehat{r}_S(i)$ *maximizes* $\widehat{u}_{ik}$ *over* $k \in S$ *using the declared tie-breaking rule, then*

$$F(S) - R(S, \widehat{r}_S) \le 2\varepsilon. \tag{28}$$

*Proof.* Because the same deterministic cost $\alpha c_k$ is subtracted in both pipelines, $\max_{i,k} |\widehat{u}_{ik} - u_{ik}| \le \varepsilon$. For each query, the maximum is 1-Lipschitz in the sup norm:

$$\left|\max_{k\in S} \widehat{u}_{ik} - \max_{k\in S} u_{ik}\right| \le \varepsilon.$$

Multiplying by $w_i$, summing, and using $\sum_i w_i = 1$ proves (27).

For the route claim, let $k_i^\star \in \arg\max_{k\in S} u_{ik}$. Since $\widehat{r}_S(i)$ maximizes the completed utility,

$$u_{i,k_i^\star} \le \widehat{u}_{i,k_i^\star} + \varepsilon \le \widehat{u}_{i,\widehat{r}_S(i)} + \varepsilon \le u_{i,\widehat{r}_S(i)} + 2\varepsilon. \tag{29}$$

Weighted summation gives (28). □

*Proof of Theorem 2.* By Lemma 9 and the $\eta$-optimality condition (1),

$$\begin{aligned} F(S^\star) - F(\widehat{S}) &\le \widehat{F}(S^\star) + \varepsilon - \widehat{F}(\widehat{S}) + \varepsilon \\ &\le \eta + 2\varepsilon, \end{aligned} \tag{30}$$

which proves (4). For routing, the completed-table route maximizes the completed utility inside $\widehat{S}$, so its completed routed utility equals $\widehat{F}(\widehat{S})$. Applying the curation condition directly therefore gives

$$\begin{aligned} R(\widehat{S}, \widehat{r}) &\geq \widehat{F}(\widehat{S}) - \varepsilon \\ &\geq \widehat{F}(S^\star) - \eta - \varepsilon \\ &\geq F(S^\star) - \eta - 2\varepsilon, \end{aligned} \tag{31}$$

where the first and last inequalities are the two uniform table-error perturbations. Rearranging proves (5). □

*Proof of Corollary 3.* On the probability-$1-\delta$ event from Theorem 1, take $\varepsilon = r_m(\delta)$ in Theorem 2. □

### A.5 Exact recovery under strict margins

Assume a unique dense-optimal bank and unique within-bank routes. The margins $\Gamma_P, \Gamma_R$ are defined in Section 5; their labels below identify the same objects:

$$\Gamma_P = F(S^\star) - \max_{S \in \mathcal{S}_B,\, S \neq S^\star} F(S), \tag{32}$$

$$\Gamma_R = \min_i \left[ u_{i, r_{S^\star}(i)} - \max_{k \in S^\star \setminus \{r_{S^\star}(i)\}} u_{ik} \right]. \tag{33}$$

An empty maximum is $-\infty$. With $|S| \leq B$, adding redundant workflows may give equal-value optimal banks. Deterministic selection of one does not make $\Gamma_P$ positive.

**Proposition 10** (Exact bank and route recovery)**.** *If $\max_{i,k} |\widehat{\mu}_{ik} - \mu_{ik}| \leq \varepsilon$ and curation satisfies* (1)*, $\Gamma_P > \eta + 2\varepsilon$, and $\Gamma_R > 2\varepsilon$, then $\widehat{S} = S^\star$ and $\widehat{r}(i) = r_{S^\star}(i)$ for every query.*

*Proof.* For any $S \neq S^\star$, bank-objective stability gives

$$\widehat{F}(S^\star) - \widehat{F}(S) \geq F(S^\star) - \varepsilon - [F(S) + \varepsilon] \geq \Gamma_P - 2\varepsilon > \eta. \tag{34}$$

Every wrong bank therefore has completed-objective gap greater than $\eta$ and cannot be returned by an $\eta$-optimal routine. Hence $\widehat{S} = S^\star$.

For a query $i$, let $k_i^\star = r_{S^\star}(i)$. Every competitor $k \in S^\star \setminus \{k_i^\star\}$ satisfies

$$\widehat{u}_{i,k_i^\star} - \widehat{u}_{ik} \geq u_{i,k_i^\star} - \varepsilon - [u_{ik} + \varepsilon] \geq \Gamma_R - 2\varepsilon > 0. \tag{35}$$

Hence the completed-table route is uniquely $k_i^\star$ for every query. This proves the margin condition (9). □

### A.6 Dimension-dependent lower bound

We now give a complete adaptive-information proof of Theorem 5. The construction uses a fixed bank containing two workflows, so the only remaining decision is the per-query route.

**Hard family.** Set $N = d$, $K = 2$, $B = 2$, $S_0 = \{1, 2\}$, $w_i = 1/d$, and zero costs. For the $i$th standard basis vector $e_i \in \mathbb{R}^d$, define

$$z_{i,1} = e_i, \qquad z_{i,2} = -e_i. \tag{36}$$

For each hidden sign vector $s \in \{-1, +1\}^d$ and amplitude $a > 0$, let

$$\theta^{(s)} = as, \qquad \mu_{i,1}^{(s)} = as_i, \qquad \mu_{i,2}^{(s)} = -as_i. \tag{37}$$

The table is exactly realizable, and the optimal route is workflow 1 when $s_i = +1$ and workflow 2 when $s_i = -1$. The uniform design over the $2d$ edges has

$$\pi_{i,k} = \frac{1}{2d}, \qquad \Sigma(\pi) = \frac{1}{d} I_d, \qquad \max_{i,k} z_{i,k}^\top \Sigma(\pi)^{-1} z_{i,k} = d. \tag{38}$$

It therefore has $\kappa = 1$. It is exactly $G$-optimal because every design with nonsingular covariance satisfies

$$\max_e z_e^\top \Sigma(\pi)^{-1} z_e \geq \sum_e \pi_e z_e^\top \Sigma(\pi)^{-1} z_e = \mathrm{tr}(\Sigma(\pi)^{-1}\Sigma(\pi)) = d.$$

An arbitrary adaptive procedure chooses an edge $A_t = (I_t, K_t)$ as a function of history $\mathcal{H}_{t-1}$ (previous actions and observations) and internal randomization, with policy kernels independent of the unknown $s$ except through history. It observes $Y_t \mid \mathcal{H}_{t-1}, A_t \sim \mathcal{N}(\mu_{A_t}^{(s)}, \sigma^2)$. We do not condition on future adaptive actions. The budget $m$ is fixed. Let $P_s$ denote the law of the full transcript and let $T_i := \sum_{t=1}^m \mathbf{1}\{I_t = i\}$ be the number of samples allocated to query $i$.

**Lemma 11** (Neighboring-hypothesis information)**.** *Let $s^{(i)}$ be $s$ with coordinate $i$ flipped. For any adaptive sampling procedure,*

$$D_{\mathrm{KL}}(P_s \| P_{s^{(i)}}) = \frac{2a^2}{\sigma^2}\,\mathbb{E}_s[T_i]. \tag{39}$$

*Consequently,*

$$\frac{1}{d2^d}\sum_{i=1}^d \sum_{s\in\{-1,+1\}^d} D_{\mathrm{KL}}(P_s \| P_{s^{(i)}}) = \frac{2a^2 m}{\sigma^2 d}. \tag{40}$$

*Proof.* Apply the chain rule for KL divergence to the adaptive transcript. The policy kernel selecting $A_t$ conditional on the past is the same measurable rule under both parameters and therefore contributes no KL term. Conditional on $A_t$, the two Gaussian observation laws differ only when $I_t = i$. In that case their means differ in magnitude $2a$, regardless of which workflow is chosen, and

$$D_{\mathrm{KL}}\big(\mathcal{N}(m, \sigma^2) \| \mathcal{N}(m \pm 2a, \sigma^2)\big) = \frac{(2a)^2}{2\sigma^2} = \frac{2a^2}{\sigma^2}.$$

Summing the conditional contributions and taking expectation under $P_s$ gives (39). Averaging and using $\sum_i T_i = m$ almost surely proves (40). □

**Lemma 12** (Adaptive sign-estimation lower bound)**.** *For any estimator $\widehat{s}$ based on the transcript,*

$$\sup_s \mathbb{E}_s[d_H(\widehat{s}, s)] \geq \frac{d}{2}\left(1 - \sqrt{\frac{a^2 m}{\sigma^2 d}}\right), \tag{41}$$

*where $d_H$ is Hamming distance.*

*Proof.* It suffices to lower-bound the Bayes risk under the uniform prior on $s$. For coordinate $i$, let

$$Q_{i,+} = 2^{-(d-1)} \sum_{s: s_i = +1} P_s, \qquad Q_{i,-} = 2^{-(d-1)} \sum_{s: s_i = -1} P_s.$$

The minimum probability of estimating bit $i$ incorrectly is $\frac{1}{2}(1 - \mathrm{TV}(Q_{i,+}, Q_{i,-}))$. Pairing every $s$ with $s^{(i)}$, convexity of total variation, Pinsker's inequality, and Jensen's inequality yield

$$\begin{aligned}
\mathrm{TV}(Q_{i,+}, Q_{i,-}) &\leq 2^{-(d-1)} \sum_{s: s_i = +1} \mathrm{TV}(P_s, P_{s^{(i)}}) \\
&= 2^{-d} \sum_s \mathrm{TV}(P_s, P_{s^{(i)}}) \\
&\leq 2^{-d} \sum_s \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(P_s \| P_{s^{(i)}})}.
\end{aligned} \tag{42}$$

The equality is the paired-mixture identity: the all-$s$ sum counts each unordered neighboring pair twice, while total variation is symmetric. Averaging (42) over $i$ and using Jensen together with (40) gives

$$\frac{1}{d}\sum_{i=1}^d \mathrm{TV}(Q_{i,+}, Q_{i,-}) \leq \sqrt{\frac{a^2 m}{\sigma^2 d}}. \tag{43}$$

Summing the coordinate-wise Bayes errors therefore gives

$$\inf_{\widehat{s}} \mathbb{E}[d_H(\widehat{s}, S)] \geq \frac{1}{2} \sum_{i=1}^{d} \big(1 - \mathrm{TV}(Q_{i,+}, Q_{i,-})\big) \geq \frac{d}{2} \left(1 - \sqrt{\frac{a^2 m}{\sigma^2 d}}\right).$$

Minimax risk is at least Bayes risk, proving the lemma. □

*Proof of Theorem 5.* Any route on the hard family defines a sign estimate: output $+1$ when the route chooses workflow 1 and $-1$ when it chooses workflow 2. A sign error on query $i$ selects utility $-a$ instead of $+a$, so, under weights $1/d$,

$$\mathrm{Reg}_{\mathrm{route}} = \frac{2a}{d}\, d_H(\widehat{s}, s). \tag{44}$$

If $m \leq \sigma^2 d/(4a^2)$, Lemma 12 gives $\sup_s \mathbb{E}_s[d_H(\widehat{s}, s)] \geq d/4$ and hence worst-case expected routed regret at least $a/2$. Choose $a = 4\varepsilon$. Then every method with

$$m \leq \frac{\sigma^2 d}{64\varepsilon^2} \tag{45}$$

has worst-case expected routed regret at least $2\varepsilon$, and therefore cannot guarantee regret at most $\varepsilon$. This proves $m = \Omega(\sigma^2 d/\varepsilon^2)$, including adaptive procedures.

For exact recovery, suppose a method recovers all routes with probability at least $3/4$ uniformly over $s$. Its Bayes expected Hamming loss is then at most $d/4$. If $m \leq \sigma^2 d/(16a^2)$, however, Lemma 12 gives Bayes Hamming loss at least $3d/8$, a contradiction. Thus uniformly high-probability exact recovery also needs $m = \Omega(\sigma^2 d/a^2) = \Omega(\sigma^2 d/\Delta^2)$, where $\Delta = \min_i |u_{i1} - u_{i2}| = 2a$. This concerns the supplied bank $S_0$, not bank selection. The full hypercube, including constant signs and $d = 1$, is retained; optimal singleton banks need not have the same margin. □

**Exact-route confidence refinement.** For uniform all-route success at least $1 - \delta$, $0 < \delta < 1/2$, data processing applied to the event $\{\widehat{s}_i = s_i\}$ gives $D_{\mathrm{KL}}(P_s \| P_{s^{(i)}}) \geq \mathrm{kl}(1-\delta, \delta)$. Here $\mathrm{kl}(p, q) = p \log(p/q) + (1-p)\log((1-p)/(1-q))$ is binary KL. Under $s$ this event has probability at least $1 - \delta$ and under its neighbor at most $\delta$. Summing (39) over coordinates at any fixed $s$ gives

$$m \geq \frac{2\sigma^2 d}{\Delta^2}\, \mathrm{kl}(1-\delta, \delta) = \frac{2\sigma^2 d}{\Delta^2}(1-2\delta) \log \frac{1-\delta}{\delta}. \tag{46}$$

This is a fixed-budget confidence refinement, not a logarithmic-in-$d$ lower bound or a claim about random stopping times.

### A.7 Projected targets and misspecification

Let the unweighted finite-population projection and residual be

$$\theta_{\mathrm{proj}} \in \arg\min_{\theta} \sum_{e \in \mathcal{E}} (\mu_e - z_e^\top \theta)^2, \qquad \mu_e^{\mathrm{proj}} = z_e^\top \theta_{\mathrm{proj}}, \qquad b_\infty = \max_e |\mu_e - \mu_e^{\mathrm{proj}}|. \tag{47}$$

**Proposition 13** (Valid projected-target accounting)**.** *If the observed labels are generated around the declared projected target, $Y_t = \mu_{E_t}^{\mathrm{proj}} + \xi_t$, then Theorem 1 applies to $\mu^{\mathrm{proj}}$ and, with probability at least $1 - \delta$,*

$$\max_e |\widehat{\mu}_e - \mu_e| \leq r_m(\delta) + b_\infty. \tag{48}$$

*Theorem 2 may then be applied with $\varepsilon = r_m(\delta) + b_\infty$.*

*Proof.* Theorem 1 gives $\max_e |\widehat{\mu}_e - \mu_e^{\mathrm{proj}}| \leq r_m(\delta)$. The triangle inequality then yields

$$\max_e |\widehat{\mu}_e - \mu_e| \leq \max_e |\widehat{\mu}_e - \mu_e^{\mathrm{proj}}| + \max_e |\mu_e^{\mathrm{proj}} - \mu_e| \leq r_m(\delta) + b_\infty.$$

□

**Proposition 14** (Naive additive projection bias can fail)**.** *There is a noiseless two-edge example with $b_\infty = \rho$ for which OLS trained on the true misspecified labels has maximum prediction error $2\rho$. Thus a theorem of the form "estimation radius plus $b_\infty$" does not hold for arbitrary misspecified OLS labels without additional residual-control or orthogonality assumptions.*

*Proof.* Take two edges with identical one-dimensional features $z_1 = z_2 = 1$ and true labels $\mu_1 = \rho$, $\mu_2 = -\rho$. The unweighted least-squares projected constant table is $(0, 0)$, hence $b_\infty = \rho$. Take $\pi = (1, 0)$ and $\sigma = 0$. Then $\Sigma = 1$, $d = \kappa = 1$, and any sample size above (2) satisfies all canonical conditions except realizability. Every sample contains only edge 1, so noiseless OLS fits $\widehat{\theta} = \rho$ and predicts $\widehat{\mu}_1 = \widehat{\mu}_2 = \rho$. Its error at edge 2 is $|\rho - (-\rho)| = 2\rho$. The unweighted residual already obeys $Z^\top b = 0$; that orthogonality does not cancel the design-weighted sampled residual. □

**Proposition 15** (Bounded misspecification with true-label observations)**.** *Fix a decomposition $\mu = Z\theta + b$, $\|b\|_\infty \le b_\infty$, and suppose $Y_t = \mu_{E_t} + \xi_t$ under the outcome-independent conditional noise model of Lemma 8. For full-rank $X$, define $\Lambda_X = \max_e \|z_e^\top X^\dagger\|_1$. Conditional on that design, with probability at least $1 - \delta$,*

$$\|\widehat{\mu} - \mu\|_\infty \le r_{\mathrm{obs}}(\delta) + (1 + \Lambda_X) b_\infty.$$

*Under the canonical sampling, noise, and sample-threshold assumptions, the bound becomes, with probability at least $1 - \delta$,*

$$\|\widehat{\mu} - \mu\|_\infty \le r_m(\delta) + (1 + \sqrt{2\kappa d}) b_\infty.$$

*Proof.* Let $b_X$ collect sampled residuals. Prediction error equals $z_e^\top X^\dagger \xi + z_e^\top X^\dagger b_X - b_e$. The first term has the conditional concentration bound; the remaining terms are at most $(\Lambda_X + 1) b_\infty$. On the canonical covariance event, Cauchy–Schwarz gives $\Lambda_X \le \max_e \sqrt{m z_e^\top V^{-1} z_e} \le \sqrt{2\kappa d}$. Splitting the canonical covariance and noise failure budgets as before proves the second claim. □

Both decision regrets are at most $\eta$ plus twice the applicable radius by Theorem 2. An empirical fitted residual is not automatically a valid uniform bound $b_\infty$; this proposition does not certify APPS or other retained real-task results.

### A.8 Explicit sparse-budget regime

For desired decision regret $\varepsilon_{\mathrm{dec}} > \eta$, the canonical result suffices at the integer-rounded budget

$$m \ge \left\lceil \max\left\{ 8\kappa d \log \frac{2d}{\delta}, \quad \frac{16\sigma^2 \kappa d \log(4M/\delta)}{(\varepsilon_{\mathrm{dec}} - \eta)^2} \right\} \right\rceil. \tag{49}$$

The first term controls sampled rank/covariance; the second ensures $\eta + 2r_m(\delta) \le \varepsilon_{\mathrm{dec}}$. Certified savings relative to one execution of all $M$ edges require an applicable sufficient budget below $M$. Thus $m \ll M$ is a parameter regime, not a universal guarantee, and the dense single-execution table is not an exact mean oracle.

### A.9 Additional fail-closed facts

**Proposition 16** (A singular certified design cannot support uniform recovery)**.** *Suppose the full finite feature population spans $\mathbb{R}^d$ but $\Sigma(\pi)$ is singular. Then there exist two realizable parameter vectors that induce identical observation laws under $\pi$ but disagree on at least one finite edge. Consequently, over the unrestricted parameter class, no method sampling only from $\pi$ can guarantee a fixed finite uniform-error radius with success probability greater than $1/2$.*

*Proof.* Choose nonzero $v \in \ker \Sigma(\pi)$. Since $v^\top \Sigma(\pi) v = \sum_e \pi_e (v^\top z_e)^2 = 0$, every edge with $\pi_e > 0$ satisfies $v^\top z_e = 0$. Because the full population spans $\mathbb{R}^d$, some edge $e_0$ has $v^\top z_{e_0} \ne 0$. The parameters $\theta_+ = tv$ and $\theta_- = -tv$ therefore agree on every edge in the sampling support but give different utilities at $e_0$. Their sampled-data laws are identical. For any proposed finite radius $q$, choose $t > q/|v^\top z_{e_0}|$. The two predictions differ by more than $2q$, so their success events are disjoint under the same observation law. One parameter therefore has failure probability at least $1/2$. This uses the unrestricted parameter class. □

**Lemma 17** (Retained repeats equal count-weighted aggregation)**.** *Let $n_e$ be the number of times edge $e$ is sampled and let $\bar{Y}_e = n_e^{-1} \sum_{t:E_t=e} Y_t$ for $n_e > 0$. The retained-row OLS objective equals the count-weighted aggregated objective up to a constant that does not depend on $\theta$:*

$$\sum_{t=1}^{m} (Y_t - z_{E_t}^\top \theta)^2 = \sum_{e:n_e>0} n_e (\bar{Y}_e - z_e^\top \theta)^2 + \sum_{e:n_e>0} \sum_{t:E_t=e} (Y_t - \bar{Y}_e)^2. \tag{50}$$

*Hence the objectives have identical minimizer sets. Both implementations must select the minimum-norm minimizer to guarantee identical full-population predictions when sampled rank is deficient; at full rank the minimizer is unique. Unweighted aggregation generally changes the objective.*

*Proof.* For each sampled edge, write $Y_t - z_e^\top \theta = (Y_t - \bar{Y}_e) + (\bar{Y}_e - z_e^\top \theta)$, square, and sum over its repeats. The cross term vanishes because $\sum_{t:E_t=e} (Y_t - \bar{Y}_e) = 0$. Summing over edges gives the identity. Dropping the factors $n_e$ generally changes the objective when counts differ. □

**Proposition 18** (The curation gap is necessary)**.** *The additive $\eta$ term in Theorem 2 cannot in general be removed. Even with an exact completed table ($\varepsilon = 0$), an optimizer whose only guarantee is $\eta$-optimality may return a bank whose dense regret is arbitrarily close to $\eta$.*

*Proof.* Take $N = 1$, $K = 2$, $B = 1$, zero costs, weight one, and utilities $(\eta, 0)$. The singleton $\{1\}$ is globally optimal. With $\widehat{F} = F$, returning $\{2\}$ is permitted by $\eta$-optimality and incurs regret exactly $\eta$. □

**Proposition 19** (Tie-breaking is necessary for exact identity)**.** *Without a shared deterministic tie-breaking convention or a strict route margin, zero completion error and zero routed regret do not imply identical workflow identities.*

*Proof.* Consider one query and a bank containing two workflows with equal utility. The dense pipeline may choose workflow 1 while the completed pipeline chooses workflow 2. The tables and realized utilities are identical, but the route identities differ. A common deterministic rule or a strict margin excludes this ambiguity. □

# B Additional Experiments

This section records the supporting diagnostics behind Section 6. The core studies separate fail-closed theorem-status checks, mechanism, robustness, and lower-bound diagnostics. The fair feature-aware comparison uses a separable side-information population so that SparseBank, IMC, and same-feature regression receive comparable pre-label information. We also report broader feature-aware and published completion baselines, selected archived sensitivity sweeps, solver scalability, and seed-level decision audits.

## B.1 Empirical targets, metrics, and measurement accounting

Theory targets $\mu_{ik} = \mathbb{E}[Y_{ik}]$. Real-task evaluation instead uses the stored table $U^{\mathrm{obs}}$ and its own dense optimum, applying the same costs, uniform task weights $w_i = 1/N$, bank search, and routing. Each real edge records one workflow execution, which may contain several model calls. Replays reuse its score and make no fresh executions. Acquisition-seed intervals condition on this table; they do not quantify fresh-execution uncertainty.

HumanEval+ stores the fraction of base-plus inputs passed by the generated program. APPS stores a different score: the number of official tests passed before the first failure, divided by the total number of official tests. The grader stops on that failure; unexecuted tests contribute nothing to the numerator. Policy-rejected programs receive zero. Thus this APPS score depends on test order and is not the fraction that would pass if every test were executed. Parsed public examples guide workflows but are excluded from the APPS objective tests. All APPS comparisons below condition on this unchanged scoring rule.

CodeContests uses the fraction of all provided public, private, and generated test records passed, retaining their original order and multiplicity. Its grader attempts every record after scientific failures

or per-input timeouts. The prospective realization freezes decisions after sparse acquisition and before dense execution; its subsequent seed replay makes no new executions.

Bank overlap and route agreement use Section 3's definitions with the observed-table oracle. Completion RMSE is $\sqrt{M^{-1}\sum_{i,k}(\widehat{\mu}_{ik} - U^{\mathrm{obs}}_{ik})^2}$ in real-task replay (replace $U^{\mathrm{obs}}$ by $\mu$ for controlled mean-table studies). Experimental workflow IDs are zero-based, corresponding to theoretical index $k$ via ID $k-1$. The retained experimental optima have size $B$; the implementation divides overlap by $B$. Exhaustive bank search visits sizes $1, \ldots, B$ in lexicographic order and replaces the incumbent only on strict improvement. Routing selects the smallest workflow index among exact ties. The visit rate is the fraction of tasks with at least one acquired edge, averaged over seeds. Unique workflows per task counts distinct acquired edges for that task, including zeros for unvisited tasks. The APPS "catastrophic-task union" is the union of the top 20 tasks ranked by mean excess regret against each comparator, with stable index ordering for ties; it has no absolute severity threshold. A separate route-switch diagnostic uses the 95th percentile of positive task regret within each method/budget.

Table 3: **Draws, unique edges, and executions have different meanings.** Replay reveals stored outcomes; the original dense data collection is a separate cost. HumanEval+, CodeContests, and deduplicated APPS masks have no repeats; canonical APPS replay draws are with replacement.

| Study | Draw budgets $m$ | Unique edges per mask | Physical execution accounting |
|---|---|---|---|
| HumanEval+ | $60, 84, 120$ | $m$; no within-mask repeats | 240 in the original dense ledger; zero in replay |
| Canonical APPS | $200, 280, 400$ | At most $m$; repeated indices reuse scores | 555 across the three maximum-budget masks, then 245 remaining edges; zero in replay |
| APPS follow-ups | Fixed draws or unique targets | Explicit unique counts; support-aware cap 288 | Zero fresh executions; all scores cached |
| CodeContests | $200, 280$ | $m$ in all three arms; no repeats | 492 sparse-union executions, decisions frozen, then 308 dense-remainder executions; zero in replay |
| Certification | $400, 600, 800, 1200$ | At most $m$ | $m$ independent simulated noisy observations per run |

The fair-feature budgets are $72, 144, 216, 288, 360$ and the geometry budgets are $24, 48, 72, 96, 120$; these fractions are integral. Margin-study percentages refer to $NK = 144$: nearest-integer rounding gives $14, 29, 43, 58, 72$. Real-study and certification budgets above are exact integers.

The accompanying artifact provides frozen inputs, replay code, and file hashes. The legacy heavy-tail summary and two raster figures lack complete generation provenance and have not been independently reproduced.

### B.2 Real-task feature specification

Table 4 records the representations used in the reported runs. APPS and CodeContests fit their text transforms on the selected public task texts before utility labels. HumanEval+ fits its utility predictor on DEV only. The accompanying `FEATURE_SPECIFICATION.json` records exact feature names, extraction rules, seeds, and source hashes recovered from the existing code and manifests.

For HumanEval+, the DEV gates require positive held-out $R^2$, route agreement at least $0.35$, improved agreement or regret over workflow means, and task-centered prediction RMSE at most twice the median positive route margin. Ties in selection use agreement, pairwise preference accuracy, and $R^2$, then fixed feature/model identifiers. Exact descriptor lists and candidate grids are retained in the specification rather than expanded into the table.

### B.3 Fair feature-aware baselines

The canonical fair feature-aware population uses four-dimensional query descriptors and three-dimensional workflow descriptors. Their tensor product gives the 12-dimensional pair feature used by SparseBank and same-feature ridge; the separable descriptors give IMC its legitimate row/column interface. All features are frozen before utilities. The current canonical SparseBank implementation

Table 4: **Frozen real-task representations.** APPS and CodeContests share a feature builder; HumanEval+ uses a DEV-trained predictor. Dimensions and normalization describe the stored matrices.

| Component | Recovered specification |
|---|---|
| APPS / CodeContests query blocks | 15 coordinates: $\log(1+x)$ of character, whitespace-word, line, and digit counts; three APPS difficulty indicators; eight text components. Text is the public statement plus the fixed Python stdin/stdout task contract; CodeContests also includes the title. Its unavailable APPS difficulty indicators are zero. |
| Text transforms | Lowercase word unigram/bigram TF–IDF, at most 512 terms, minimum document frequency 1, followed by eight-component truncated SVD with a fixed hash-derived seed. The seven static coordinates and eight text components are standardized separately on the selected population. The archived builder fixes tokenization and transform settings. |
| Workflow blocks | Six coordinates: 27B-model indicator; maximum model calls divided by 4; maximum internal tool calls divided by 3; sum of stage repetitions divided by 3; reasoning indicator; verification/selection indicator. The last is true for tool-using, critique/revise, public-test repair, or candidate-selection procedures. |
| Pair construction | Concatenate 15 query coordinates, six workflow coordinates, eight workflow one-hot indicators, and all $15 \times 6$ interactions: 119 raw coordinates, in task-major order. No DEV utility predictor is used. |
| Rank and scale | For raw matrix $A$, retain singular values above $\max(\mathrm{shape}(A))\,\epsilon_{\mathrm{float64}}\,s_{\max}(A)$. Use the retained $UD$ coordinates divided by their maximum row norm. APPS: rank 78, divisor $15.3671473047$. CodeContests: rank 68, divisor $13.0418928371$. |
| Freeze timing | APPS feature manifest: 2026-09-15 14:13:28 UTC. CodeContests feature manifest: 2026-09-22 07:57:03 UTC. Both precede target workflow execution; CodeContests reuses the builder with its own public texts and fitted transforms. |
| HumanEval+ predictor inputs | 33 public static task descriptors: lengths, word counts, signature/annotation and argument metadata, visible examples, punctuation/requirement counts, and ten keyword groups. The groups cover strings, sequences, dictionaries/sets, numbers, mathematics, parsing, iteration, recursion, combinatorics, and edge cases. Seventeen standardized workflow descriptors encode procedure, model size, call/tool budgets, candidate/repair counts, temperatures, and token limits. No text embedding is used by the selected predictor. |
| DEV fitting and selection | 60 DEV tasks: the original 16 plus 44 chosen by the same ID-hash order, excluding the earlier 30-task TEST split. Four task-level hash folds evaluate 156 feature/model settings. Scaling is fitted within each training fold. Selection minimizes held-out routed regret among candidates passing the DEV gates. The selected ridge predictor uses penalty 100, an intercept, and an SVD solver on workflow indicators, task/workflow descriptors, and their interactions (615 coordinates). It is then refitted on all DEV tasks. |
| Final pair features | Four workflow indicators plus the unclipped DEV-model utility prediction: raw dimension and retained rank 5. All five columns are retained; rank checks use the default float64 singular-value tolerance above. Divide by the maximum selected-row norm on DEV, $1.6145510360$, and keep this divisor for TEST2. |
| TEST2 separation | Predictor frozen on 2026-08-22 at 06:27:06 UTC; TEST2 features at 07:47:54 UTC. TEST2 takes 60 never-executed tasks by a separate ID-hash order, excluding all prior DEV and TEST tasks. Predictor coefficients, scaling, columns, and normalization remain fixed before TEST2 labels. |

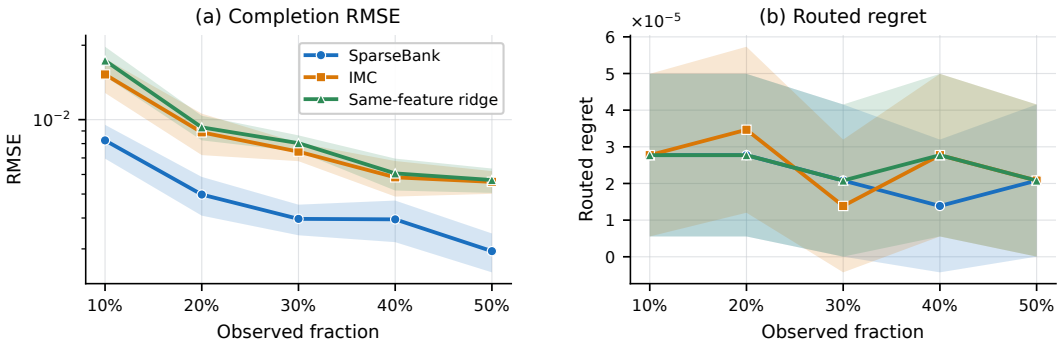



Figure 2: **SparseBank has the lowest mean completion RMSE at every tested budget under matched side information.** SparseBank, IMC, and same-feature ridge use the same frozen descriptors and budget grid. Every plotted budget is empirical-only: $m \leq 360$, while the explicit proof gate requires at least 593 samples. Curves show means over ten matched seeds with pointwise 95% confidence bands.

uses the active-set $G$-optimal backend, while IMC and ridge share uniform masks. SparseBank uses different masks. Thus the original comparison changes acquisition and estimator together.

The canonical protocol freezes query, workflow, and pair features and derives seeds from semantic experiment identifiers. Archived double-run checks record agreement of features, schedules, and metrics.

Figure 2 compares SparseBank with IMC and same-feature ridge on the frozen realizable side-information population. SparseBank has the lowest mean RMSE at every tested budget. At 10%, RMSE is 0.008243 for SparseBank, 0.015235 for IMC, and 0.017343 for ridge; at 50% the values are 0.002937, 0.005601, and 0.005698. Averaged over all five budgets, RMSE is $0.004815/0.008597/0.009289$ and routed regret is $2.22/2.50/2.50 \times 10^{-5}$ for SparseBank/IMC/ridge, respectively. For this population, the explicit proof requires at least 593 samples even when the verified coverage factor is approximately one. All tested fair-comparison budgets use at most 360 observations and fail the canonical gate. No realized-design certificate was established for these archived runs, so their reported status remains empirical-only. Thus the fair realizable comparison favors SparseBank empirically, while theorem-mode certification is not asserted for the current budget grid.

Table 5 adds the realizable and misspecified summary.

Table 5: **The completion ranking reverses under misspecification.** Values are mean RMSE across the five tested budgets; these misspecified rows remain empirical-only; no valid uniform residual bound was supplied.

| Regime | SparseBank | IMC | Same-feature Ridge |
|---|---|---|---|
| Realizable | **0.004815** | 0.008597 | 0.009289 |
| Misspec. 0.05 | 0.063106 | **0.054231** | 0.054508 |
| Misspec. 0.10 | 0.125455 | 0.108385 | **0.108273** |

The completion advantage is therefore conditional on the frozen linear model. Under mild misspecification, IMC and ridge obtain lower RMSE. Nevertheless, SparseBank retains lower routed regret in the audited misspecified regimes: 0.000849 at strength 0.05 and 0.007935 at strength 0.10. This separation between completion ranking and downstream decision ranking is consistent with the decision-transfer focus of SparseBank.

The fail-closed validator also separates empirical behavior from the theorem-valid regime. For this population, $d = 12$, $\delta = 0.05$, and the verified coverage factor is approximately one. The explicit proof gate $\lceil 8\kappa d \log(2d/\delta) \rceil$ requires at least 593 samples. All tested fair-comparison budgets use at

most 360 observations, so every point in the current budget grid retains its historical empirical-only status; being below this gate does not rule out a separate realized-design certificate.

Table 6: **Every tested fair-comparison budget fails closed under the 593-sample proof gate.** Each point remains a valid empirical measurement but carries no theorem-mode claim.

| Budget | $m$ | Required $m$ | Status |
|---|---|---|---|
| 10% | 72 | 593 | Empirical-only (fail closed) |
| 20% | 144 | 593 | Empirical-only (fail closed) |
| 30% | 216 | 593 | Empirical-only (fail closed) |
| 40% | 288 | 593 | Empirical-only (fail closed) |
| 50% | 360 | 593 | Empirical-only (fail closed) |

### B.4 Matched acquisition-by-estimator factorial

Table 7: **Acquisition gains survive matching the estimator under realizability.** Mean RMSE over five budgets and ten matched seeds; all cells are empirical-only.

| Acquisition | OLS | Ridge |
|---|---|---|
| G-opt | 0.004815 | 0.005537 |
| Uniform | 0.008597 | 0.009289 |
| Leverage | 0.006355 | 0.007283 |

The factorial reuses the canonical $60 \times 12$ population, frozen $d = 12$ tensor-product features, five budgets, ten paired seeds, and misspecification strengths $0, 0.05, 0.10$. It crosses three acquisition rules with OLS and ridge at the original fixed $\lambda = 0.01$. Both estimators receive identical rows and observations within each acquisition/seed/budget cell. Across acquisitions, $G$-opt and Uniform retain their distinct original noise streams; Leverage reuses Uniform's noise vector. Matched seeds therefore do not imply identical execution noise across all acquisition rules. The original design is not recomputed.

Table 8: **Full acquisition-by-estimator factorial.** Means across five budgets and ten seeds. Misspecification reverses the RMSE ranking with both estimators.

| Strength | Acquisition | OLS RMSE | Ridge RMSE | OLS routed | Ridge routed |
|---|---|---|---|---|---|
| 0.0 | G-opt | 0.004815 | 0.005537 | $2.218 \times 10^{-5}$ | $2.911 \times 10^{-5}$ |
| 0.0 | Uniform | 0.008597 | 0.009289 | $2.495 \times 10^{-5}$ | $2.495 \times 10^{-5}$ |
| 0.0 | Leverage | 0.006355 | 0.007283 | $2.634 \times 10^{-5}$ | $3.050 \times 10^{-5}$ |
| 0.05 | G-opt | 0.063106 | 0.062862 | $8.488 \times 10^{-4}$ | $8.938 \times 10^{-4}$ |
| 0.05 | Uniform | 0.054231 | 0.054508 | $1.556 \times 10^{-3}$ | $1.563 \times 10^{-3}$ |
| 0.05 | Leverage | 0.054963 | 0.054827 | $9.388 \times 10^{-4}$ | $9.388 \times 10^{-4}$ |
| 0.1 | G-opt | 0.125455 | 0.124989 | $7.935 \times 10^{-3}$ | $7.935 \times 10^{-3}$ |
| 0.1 | Uniform | 0.108385 | 0.108273 | $1.162 \times 10^{-2}$ | $1.175 \times 10^{-2}$ |
| 0.1 | Leverage | 0.110053 | 0.109777 | $8.415 \times 10^{-3}$ | $8.536 \times 10^{-3}$ |

The archive accounts for all 900 cells: 300 historical cells retained unchanged and 600 missing factorial cells fitted in the earlier study. Table 8 averages the five stored budget means, not their confidence limits. At the smallest realizable budget, the paired $G$-opt-minus-Uniform RMSE difference is $-0.006993$ for OLS (95% interval $[-0.010044, -0.004357]$) and $-0.006988$ for ridge ($[-0.010076, -0.003985]$). Lower RMSE under both estimators supports an acquisition contribution in this population; it does not establish a universal interaction pattern. Under misspecification both estimator columns reverse the completion ranking, while $G$-opt retains the lowest mean routed regret. All points retain their historical empirical-only status; no separate realized-design certificate was established.

The artifact directory `results/acquisition_estimator/` contains `summary.csv`, `per_seed.csv`, and `paired_comparisons.csv`, including all budgets, pointwise paired intervals, and interaction contrasts.

### B.5 THEOREM-VALID CERTIFICATION EXPERIMENT

The certification experiment was designed to enter theorem mode without evaluating the full finite table. We used $N = 100$, $K = 20$, $M = NK = 2000$, and $d = 8$. Each row consisted of an intercept and seven normalized Gaussian coordinates assigned to the finite edges, with $\|z_e\|_2 \le 1$. The utility table was generated exactly as $\mu_e = z_e^\top \theta^\star$ before labels were sampled. The design was frozen before labels, observations were sampled with replacement, execution noise was independent Gaussian with $\sigma = 0.002$, and completion used unregularized minimum-norm OLS. No regularization, clipping, adaptive feature construction, or label-dependent design change was used.

An independent finite-population scan verified full rank, design support on all 2000 edges, maximum leverage $8.00000045$, and $\kappa = 1.0000000566$. The design covariance had minimum and maximum eigenvalues $0.07142857$ and $0.5$, respectively. For $\delta = 0.05$, the explicit proof gate from Eq. (20) is

$$\lceil 8\kappa d \log(2d/\delta) \rceil = 370.$$

We tested $m \in \{400, 600, 800, 1200\}$, corresponding to 80%, 70%, 60%, and 40% measurement reduction relative to dense evaluation. Each budget used 100 independent seeds. Both the dense oracle and every completed table use exhaustive search over all banks of sizes $1, \dots, 3$, so $\eta = 0$. The recovered driver and shared bank-search implementation verify this optimization contract. The dense objective was $0.480814$.

Table 9: **The reported completion certificate passes at every tested sparse budget.** Values are means over 100 seeds unless stated otherwise. Exhaustive curation on every completed table gives $\eta = 0$, so the decision bounds are $2r_m(\delta)$.

| $m$ | $m/M$ | $r_m(\delta)$ | max error | pass | bank bound | route bound | route regret |
|---|---|---|---|---|---|---|---|
| 400 | 0.20 | 0.001958 | 0.000742 | 1.00 | 0.003916 | 0.003916 | $4.76\times10^{-7}$ |
| 600 | 0.30 | 0.001599 | 0.000587 | 1.00 | 0.003198 | 0.003198 | $2.11\times10^{-7}$ |
| 800 | 0.40 | 0.001385 | 0.000530 | 1.00 | 0.002769 | 0.002769 | $2.64\times10^{-7}$ |
| 1200 | 0.60 | 0.001131 | 0.000406 | 1.00 | 0.002261 | 0.002261 | $1.32\times10^{-7}$ |

The realized bank regret was zero to numerical precision at every budget, the mean bank overlap was one, and mean route agreement ranged from $0.9982$ to $0.9995$. The largest realized maximum error over the 100 seeds was $0.001293$ at $m = 400$, below the corresponding radius $0.001958$; the certificate and sampled full-rank checks both passed in all 400 runs. Comparing a regret bound with the absolute dense objective does not establish usefulness: adding a constant to every utility changes that objective but leaves regret unchanged. Decision margins and shift-invariant utility ranges are the relevant comparisons. This controlled demonstration does not establish theorem validity for the misspecified HumanEval+ representation.

**Implementation provenance.** The IMC baseline follows Jain–Dhillon's side-information formulation (Jain & Dhillon, 2013); its implementation checks are recorded in the archived published-baseline audit. Same-feature ridge is a controlled side-information baseline receiving the same frozen pair representation as SparseBank; its fixed ridge parameter is $\lambda = 0.01$, chosen in the protocol configuration before the synchronized fair-comparison rerun and not tuned on tested labels. Earlier SoftImpute and adaptive matrix completion comparisons remain historical supplementary evidence, and the matched factorial in Section B.4 separates acquisition from estimator effects.

**Dimension–noise sensitivity.** The headline fair feature-aware comparison fixes one retained feature dimension and noise scale. Figure 3 restores the broader in-model sensitivity sweep from the experiment archive. Across $d \in \{5, 10, 20, 40\}$ and $\sigma \in \{0.02, 0.05, 0.1, 0.2\}$, the maximum finite-edge prediction error decreases systematically with the sample budget. The dashed curves are the corresponding reference scaling curves used in the original sweep. This figure is an in-model scaling diagnostic, not an additional competing-method comparison.

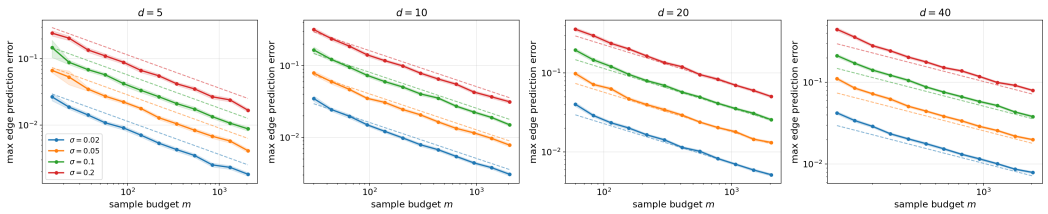



Figure 3: **Maximum edge prediction error falls with sample budget across the realizable dimension–noise grid.** Panels vary retained feature dimension $d$ and curves vary noise scale $\sigma$; shaded bands summarize Monte Carlo variability and dashed curves show the archived reference scaling.

### B.6 Computational scaling and solver validation

The canonical implementation now uses an active-set corrective optimizer for the finite $D$-optimal objective and retains vanilla Frank–Wolfe as a regression backend. Kiefer–Wolfowitz equivalence motivates the candidate design, but every checkpoint below is accepted only after rescanning the original global $G$-optimal certificate over all $NK = 100{,}000$ edges.

Table 10: **Active-set optimization reaches every verified certificate target faster than vanilla Frank–Wolfe.** Times are in seconds for $N = 1000$, $K = 100$, and $d = 12$.

| Target $\kappa$ | Active-set (s) | Vanilla FW (s) |
|---|---|---|
| 1.05 | 1.105 | 7.031 |
| 1.02 | 1.216 | 15.819 |
| 1.01 | 1.427 | 33.969 |
| 1.005 | 1.535 | 65.425 |

At $\kappa \le 1.02$, matched downstream tests give mean RMSE $4.03 \times 10^{-4}$ for active-set versus $3.96 \times 10^{-4}$ for vanilla FW, while routed regret is $3.05 \times 10^{-7}$ versus $5.02 \times 10^{-7}$ and bank regret is zero. Peak process memory is approximately 166 MB and a full certificate scan costs about 0.04 seconds. Thus the speedup is numerical rather than a change in the statistical objective.

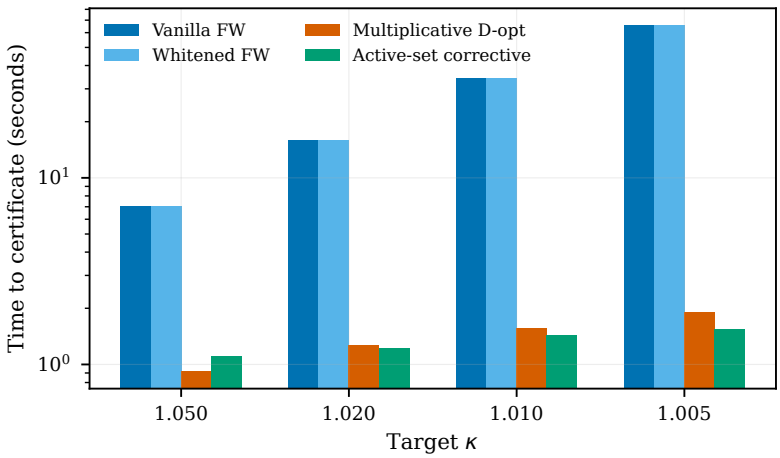



Figure 4: **Active-set optimization reaches the displayed targets about 6–43 times faster than vanilla Frank–Wolfe.** Every point uses the global $G$-optimal certificate on $10^5$ edges; the speedup exceeds tenfold at the three stricter targets.

### B.7 GEOMETRY-STRATIFIED DESIGN QUALITY

The controlled acquisition-geometry study holds the estimator and downstream evaluation fixed while varying acquisition geometry. Under heterogeneous features, the $50\%$-budget mean RMSE is $0.00414$ for $G$-optimal acquisition, $0.00701$ for leverage sampling, and $0.01405$ for uniform sampling. Under homogeneous geometry the corresponding values are $0.01019$, $0.01345$, and $0.01232$. Figure 5 therefore shows a clear gain when the controlled geometry is heterogeneous and a smaller gain when it is already well balanced. This supports an acquisition effect in the controlled population. The real-task evidence is positive on APPS and CodeContests but does not establish a stable acquisition-rule advantage on HumanEval+. Across a broader design-stress sweep, $\rho_{\mathrm{Spearman}}(\log \kappa, \mathrm{RMSE}) = 0.138$ with $p = 0.395$; we use $\kappa$ as a worst-direction certificate, not a universal predictor of average error. Accordingly, $G$-optimal design is most justified for coverage/certification rather than universal routed-regret dominance.

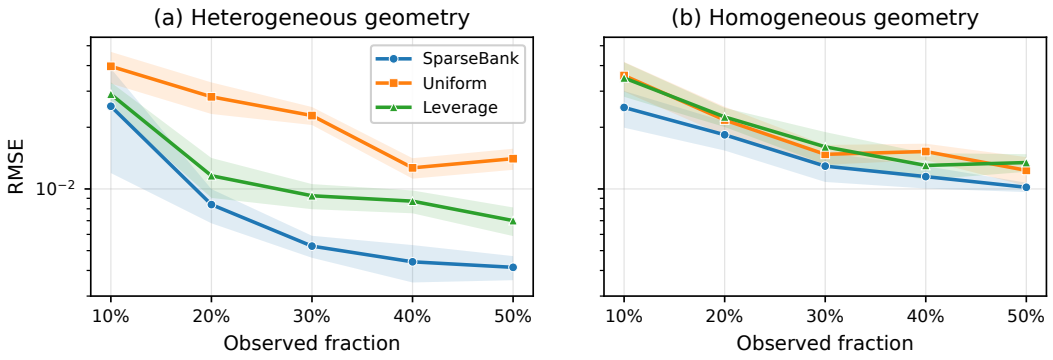



Figure 5: **$G$-optimal acquisition gives its clearest RMSE gain under heterogeneous feature geometry.** Curves compare $G$-optimal, uniform, and leverage acquisition across observation fractions; bands are pointwise 95% confidence intervals over 15 seeds, with a shared logarithmic RMSE scale.

The refreshed acquisition-geometry experiment uses $N = 30$, $K = 8$, $d = 8$, five budgets from $10\%$ to $50\%$, and 15 seeds. Table 11 summarizes the $50\%$ endpoint. The heterogeneous population separates the designs strongly: SparseBank combines a near-optimal leverage certificate with the lowest RMSE. The homogeneous population is less discriminative and all three designs are closer. This supports a geometry-dependent advantage, not the stronger claim that $\kappa$ alone predicts empirical error. Across the broader design-stress suite, the Spearman correlation between $\log \kappa$ and RMSE is $0.138$ with $p = 0.395$.

Table 11: **$G$-optimal acquisition has the lowest 50%-budget RMSE in both geometry regimes.** The certificate is recomputed for the declared acquisition distribution; unique edges are means over 15 seeds.

| Geometry | Method | $\widehat{\kappa}$ | Unique edges | RMSE |
|---|---|---|---|---|
| Heterogeneous | SparseBank | 1.0018 | 8.07 | 0.004144 |
| Heterogeneous | Leverage | 2.6924 | 68.80 | 0.007007 |
| Heterogeneous | Uniform | 14.9276 | 94.07 | 0.014054 |
| Homogeneous | SparseBank | 1.0016 | 18.53 | 0.010188 |
| Homogeneous | Leverage | 1.9570 | 90.40 | 0.013454 |
| Homogeneous | Uniform | 2.5104 | 94.80 | 0.012323 |

**Controlled certificate degradation.** Figure 6 restores a complementary geometry intervention in which the sampling design is continuously moved from a deliberately misaligned distribution toward the certified $G$-optimal design. The verified certificate improves from approximately $4.5$

toward the Kiefer–Wolfowitz optimum 1, realized maximum edge error generally decreases, and the error-to-certified-radius ratio remains below the plotted envelope boundary for all three budgets. This controlled intervention illustrates what the certificate is designed to diagnose; it does not imply that $\kappa$ is a universal cross-population predictor of empirical RMSE, consistent with the weak pooled correlation reported above.

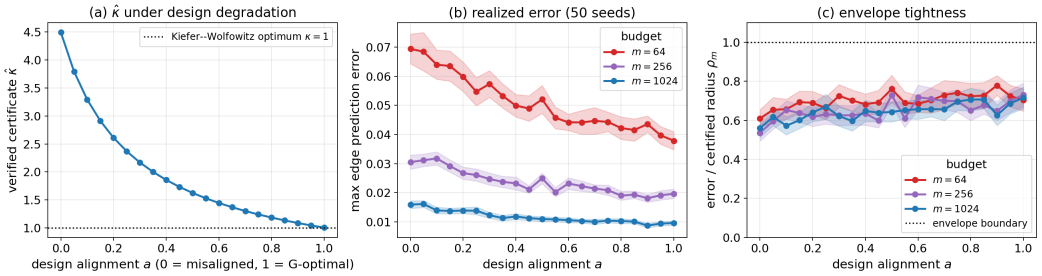



Figure 6: **Controlled alignment improves the verified certificate while realized error remains inside its envelope. Left:** $\widehat{\kappa}$ as the design moves toward the $G$-optimal solution. **Middle:** maximum edge error for three budgets. **Right:** error-to-certified-radius ratio. Bands are archived 95% confidence intervals over 50 seeds per alignment point.

### B.8 Noise-boundary diagnostics

Theorem 1 assumes mean-zero sub-Gaussian noise. This diagnostic intentionally violates that assumption to show where the standard radius ceases to be reliable. Table 12 reports the exported Student-$t$ results. "Diagnostic violation" refers to the auxiliary quantile-scale radius used in the original stress test; it is not a SparseBank theorem guarantee.

Table 12: **The standard sub-Gaussian envelope fails under the heaviest-tailed regimes.** Violation rates are fractions of runs whose maximum edge error exceeds the corresponding reference radius.

| Noise regime | Standard violation | Diagnostic violation | Interpretation |
|---|---|---|---|
| Student-$t_{1.5}$ | 1.00 | 0.26 | Outside finite-variance and theorem mode. |
| Student-$t_{2.0}$ | 0.92 | 0.06 | Boundary case; sub-Gaussian envelope fails. |
| Student-$t_{2.25}$ | 0.04 | 0.02 | Near the target level in this synthetic run. |
| Student-$t_{2.5}$ | 0.06 | 0.02 | Slight standard-envelope excess. |
| Gaussian | 0.02 | – | Theorem-mode light-tailed reference. |

A second stress test contaminates 5% of observations with spikes. The standard-envelope violation rate stays at 0.02 for normalized spike scales $\beta/\sigma \in \{0, 0.1, 0.25, 0.5\}$ and rises to 0.06 at $\beta/\sigma = 1.0$; the auxiliary diagnostic radius has zero violations on the reported grid. These rows are robustness diagnostics only and do not extend the theorem to contaminated noise.

### B.9 Margin-sensitive decision transfer

Theorem 2 controls decision loss, while exact identity also requires margins. At the nominal 30% budget, high-margin and near-tie populations have similar RMSE (0.01943 versus 0.01911), but route agreement is 0.9958 versus 0.8396 and routed regret is $7.56 \times 10^{-5}$ versus $2.72 \times 10^{-3}$. Bank overlap is 1.000 versus 0.933. The reported high-margin route metrics plateau from 30% to 50% while RMSE improves (Appendix B.9); acquisition-hash uniqueness alone cannot verify downstream recomputation.

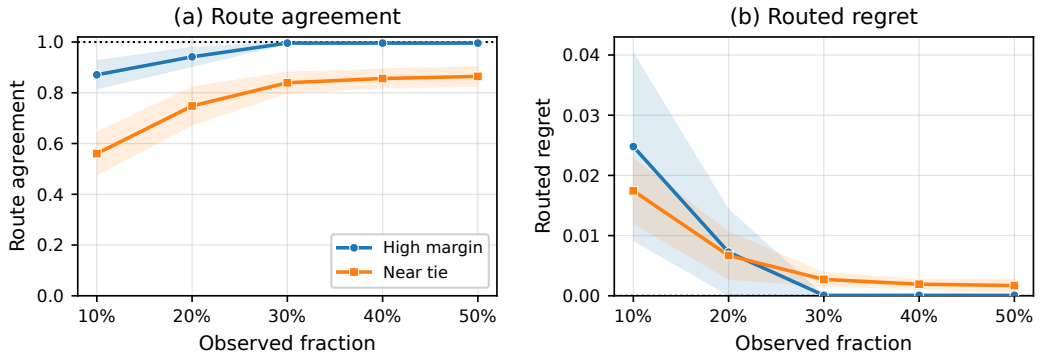



Figure 7: **Similar completion error produces markedly different route stability when decision margins differ.** Route agreement and routed regret are shown for SparseBank under high-margin and near-tie populations. Curves are means over 20 seeds with pointwise 95% confidence bands.

The refreshed margin-transfer study holds the SparseBank completion procedure fixed and changes the downstream decision margins. Table 13 shows that the high-margin population reaches nearly exact route identity by the 30% budget, while the near-tie population remains materially less stable. At 30%, completion RMSE is almost identical in the two conditions (0.01943 versus 0.01911), yet bank overlap is 1.000 versus 0.933 and route agreement is 0.9958 versus 0.8396. This separates completion quality from the additional margin condition needed for exact decision identity.

Table 13: **High-margin routes stabilize substantially earlier than near-tie routes.** Agreement and regret values are means over 20 seeds.

| Budget | High-margin agree | High-margin regret | Near-tie agree | Near-tie regret |
|---|---|---|---|---|
| 10% | 0.8708 | 0.02479 | 0.5604 | 0.01744 |
| 20% | 0.9417 | 0.00724 | 0.7479 | 0.00669 |
| 30% | 0.9958 | 0.000076 | 0.8396 | 0.00272 |
| 40% | 0.9958 | 0.000076 | 0.8562 | 0.00192 |
| 50% | 0.9958 | 0.000076 | 0.8646 | 0.00171 |

**Seed-level saturation audit.** Across all 200 canonical margin-transfer SparseBank records, the reported condition/seed/budget combinations have distinct acquisition hashes. These hashes establish different acquisition records, not downstream recomputation. High-margin RMSE continues to improve from 0.01943 at 30% to 0.01289 at 50%, while the discrete bank/route distributions are already saturated. Figure 8 shows the full seed distributions.

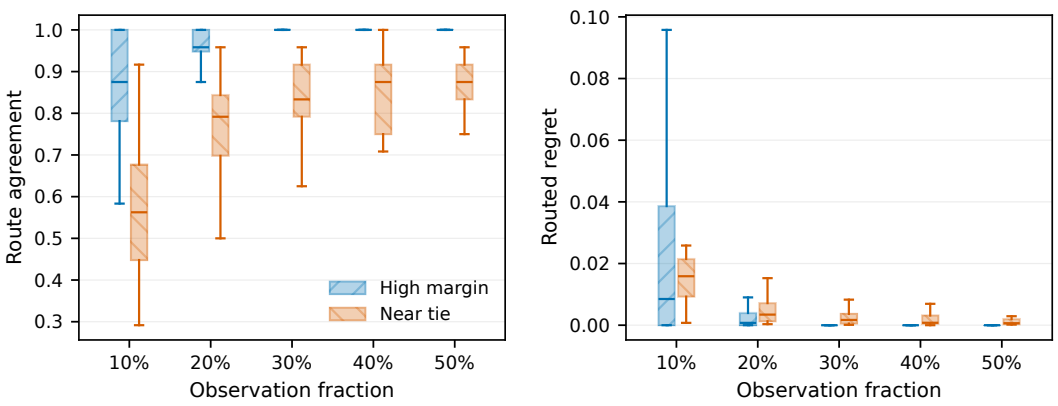



Figure 8: **The reported high-margin decision distribution plateaus after 30% even as completion RMSE improves.** The plot retains the seed-level route-agreement and routed-regret distributions.

### B.10 Prospective HumanEval+ workflow-bank evaluation

**Benchmark and workflow bank.** HumanEval+ augments HumanEval with substantially stronger executable tests (Liu et al., 2023). The confirmatory TEST2 contains 60 tasks that had no prior workflow labels. The frozen candidate set contains four workflows and downstream bank size $B = 2$:

- `he_reasoned`: reasoning-enabled single-pass generation with `Qwen/Qwen3.8-27B`, revision `1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0`;
- `he8_direct`: direct generation with `Qwen/Qwen3-8B`, revision `b968826d9c46dd6066d109eabc6255188de91218`;
- `he8_reasoned`: reasoning-enabled generation with the same `Qwen/Qwen3-8B` revision;
- `he8_multi_visible_select`: four-candidate generation with the same `Qwen/Qwen3-8B` revision, followed by selection using only public docstring examples and syntax/import checks.

Hidden EvalPlus tests and reference solutions were unavailable to every workflow. Utility is the fraction of EvalPlus HumanEval base-plus inputs passed. We set the downstream cost coefficient to $\alpha = 0$, so the real experiment evaluates sparse acquisition of raw task utility rather than heterogeneous execution-cost optimization. The HumanEval+ study is empirical-only and is not reported as a theorem-mode-certified run.

The primary endpoint was specified before TEST2 after an exploratory study on disjoint tasks. All previously observed test tasks were excluded from the confirmatory set. Before any TEST2 utility was observed, the protocol froze the 60 task IDs, the method, pair features, execution seeds, and acquisition masks at $25\%$, $35\%$, and $50\%$. SparseBank outputs were frozen before the remaining labels were revealed. The recovered masks use weighted exponential-race sampling without replacement, unlike theorem-mode i.i.d. sampling. Each sparse construction uses $60$, $84$, or $120$ unique edges; the research ledger ultimately contains all $240$ for evaluation.

The frozen HumanEval+ representation has dimension and rank 5: four workflow indicators plus a DEV-trained utility prediction. Table 4 specifies the predictor, DEV split, and fixed normalization.

**DEV-predictor-only and feature ablations.** The DEV-predictor-only baseline uses the frozen DEV predicted utility as the complete $N \times K$ table, performs zero new TEST2 measurements, and applies the same bank selection, route tie-breaking, and dense-table evaluation used by the other methods. Its selected bank is {`he_reasoned`,`he8_multi_visible_select`}, with bank regret

$0.009045$, routed regret $0.059476$, route agreement $0.2667$, and bank overlap $0.5$. These values are direct post-hoc replay diagnostics.

The one-hot-only ablation removes the DEV predicted-utility coordinate and retains only the four workflow indicators. It reruns the same 100-seed acquisition/completion pipeline at the three finalized budgets using the dense TEST2 table only for replay evaluation. The full-feature Uniform+OLS audit uses the original five-dimensional feature matrix in every 100-seed row; all OLS fits have rank 5 and the feature hash matches the frozen HumanEval+ artifact. No TEST2 label was used to construct that matrix.

Table 14: **Anchor lowers routed regret relative to DEV-only, whereas unshrunk OLS does not.** Anchor has worse bank regret and route agreement than DEV-only. Its coefficient $\lambda = 0.610581$ was selected on DEV; all rows use the same 100 TEST2 masks.

| Method | Draws $m$ | RMSE | Bank regret | Routed regret | Route agree |
|---|---|---|---|---|---|
| DEV predictor only | 0 | 0.403382 | 0.009045 | 0.059476 | 0.2667 |
| Full-feature SparseBank, 25% | 60 | 0.368100 | 0.049018 | 0.072222 | 0.1548 |
| Full-feature SparseBank, 35% | 84 | 0.365179 | 0.048786 | 0.069066 | 0.1313 |
| Full-feature SparseBank, 50% | 120 | 0.362265 | 0.046710 | 0.064050 | 0.1112 |
| SparseBank-Anchor, 25% | 60 | 0.366249 | 0.031163 | 0.057135 | 0.1445 |
| SparseBank-Anchor, 35% | 84 | 0.365941 | 0.032592 | 0.056016 | 0.1395 |
| SparseBank-Anchor, 50% | 120 | 0.365479 | 0.037790 | 0.056430 | 0.1188 |
| One-hot-only SparseBank, 25% | 60 | 0.367793 | 0.072884 | 0.072884 | 0.1350 |
| One-hot-only SparseBank, 35% | 84 | 0.364812 | 0.068487 | 0.068487 | 0.0950 |
| One-hot-only SparseBank, 50% | 120 | 0.362538 | 0.067904 | 0.067904 | 0.0900 |

The unshrunk full-feature rows and one-hot rows are the original negative diagnostics. Relative to DEV-only, unshrunk full-feature SparseBank has paired difference $0.008970$ with 95% bootstrap CI $[0.006374, 0.011792]$, while the one-hot difference is $0.010283$ with CI $[0.007711, 0.013045]$. SparseBank-Anchor instead uses

$$\widehat{U}_{\text{Anchor}} = P_{\text{DEV}} + \lambda \widehat{h}, \qquad \lambda = \text{clip}_{[0,1]} \left( \frac{\sum h^\top \widehat{h}}{\sum \widehat{h}^\top \widehat{h}} \right),$$

where the TEST2 table uses the frozen predictor $P_{\text{DEV}}$ and its fitted TEST2 residual correction $\widehat{h}$. In the coefficient formula, $h$ and $\widehat{h}$ instead denote the true and fitted residual vectors on DEV masks; $\text{clip}_{[0,1]}$ truncates to $[0, 1]$. The sums pool the fixed three acquisition rules, three budgets, and 100 DEV masks. The resulting coefficient is $0.610581$, selected without any TEST2 outcome. This replay is post-hoc and empirical-only; the anchored estimator is not covered by Theorem 1.

Table 15: **SparseBank-Anchor improves over DEV-only across all three acquisition rules in the budget-averaged replay.** Negative $\Delta$ favors the anchored correction; intervals are paired bootstrap intervals over the 100 frozen TEST2 seeds.

| Acquisition | Mean $\Delta$ | 95% CI | Fraction positive |
|---|---|---|---|
| SparseBank | $-0.002949$ | $[-0.003761, -0.002147]$ | 0.06 |
| Uniform | $-0.003424$ | $[-0.004201, -0.002655]$ | 0.06 |
| Leverage | $-0.003764$ | $[-0.004567, -0.002958]$ | 0.04 |

Among these tested estimators, Anchor lowers mean routed regret relative to DEV-only, while unshrunk OLS increases it. This does not imply improvement in bank regret or route identity, or establish acquisition-rule superiority.

Table 16: **Each HumanEval+ sparse mask reveals 50–75% fewer unique edges than the dense table.** Sampling is without replacement. These per-construction counts exclude DEV training and the subsequent dense evaluation; replay makes zero fresh executions.

| Edge fraction | Draws / unique edges | $NK - m$ | Edge-count reduction |
|---|---|---|---|
| 25% | 60 | 180 | 75% |
| 35% | 84 | 156 | 65% |
| 50% | 120 | 120 | 50% |

Table 17: **The preregistered acquisition realization favors SparseBank in budget-averaged routed regret.** The primary endpoint averages task-level regret over the three observation fractions; lower is better, and paired uncertainty remains inconclusive.

| Method | 25% | 35% | 50% | Mean |
|---|---|---|---|---|
| SparseBank | **0.05927** | 0.07512 | **0.05866** | **0.06435** |
| Uniform + OLS | 0.07336 | **0.06626** | 0.06101 | 0.06688 |
| Leverage + OLS | 0.07453 | 0.07234 | 0.06598 | 0.07095 |

SparseBank lowers the budget-averaged routed regret by $3.78\%$ relative to Uniform+OLS and $9.30\%$ relative to Leverage+OLS, and is lower at two of three budgets against each. The task-level paired bootstrap intervals include zero, so the result is directional rather than statistically conclusive.

Table 18: **Both preregistered paired confidence intervals include zero.** $\Delta$ is baseline minus SparseBank budget-averaged task regret, so positive values favor SparseBank.

| Comparison | Mean $\Delta$ | 95% CI | One-sided $p$ | Holm $p$ |
|---|---|---|---|---|
| Uniform+OLS − SparseBank | 0.00253 | $[-0.01501, 0.01913]$ | 0.3873 | 0.3873 |
| Leverage+OLS − SparseBank | 0.00660 | $[-0.00640, 0.02008]$ | 0.1749 | 0.3498 |

The paired task-level comparison contains substantial exact ties. Against both Uniform+OLS and Leverage+OLS, 42 of the 60 TEST2 tasks are exact ties in the budget-averaged task-level endpoint, leaving 18 non-tied tasks; SparseBank has lower regret on 10 of these 18 non-tied tasks. This high tie rate reduces the effective paired information and helps explain why the bootstrap intervals remain wide despite a favorable aggregate direction.

Figure 9 shows the same preregistered acquisition-control comparison budget by budget. Absolute route agreement is secondary: the real experiment is designed around downstream routed regret at matched measurement budgets rather than exact recovery of the dense route identity.

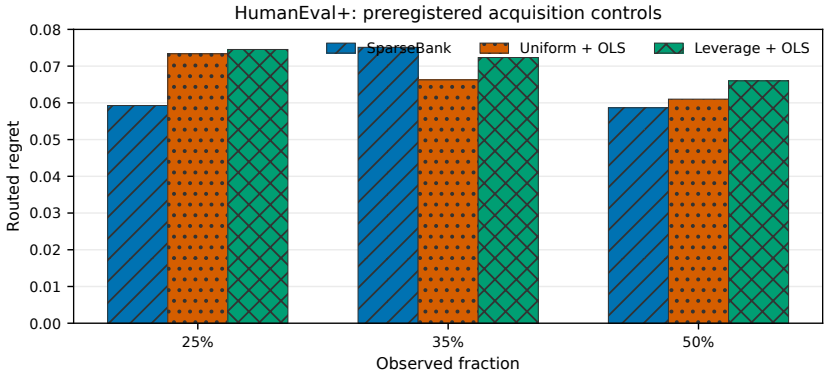



Figure 9: **SparseBank has lower budget-averaged routed regret than both preregistered acquisition controls.** It is lower at two of the three observation fractions against Uniform+OLS and Leverage+OLS; paired uncertainty remains inconclusive.

**Published completion baselines.** After TEST2 was finalized, we performed a frozen-table replay with compatible published methods; no workflow was rerun and SparseBank was unchanged. Official R `softImpute` 1.4-3 used matched uniform masks. AMC used an integrated Krishnamurthy–Singh-style exact-budget compatibility port checked against the staged original-code bundle, revealing labels only through the entries selected by its adaptive policy. IMC is not evaluated here because the frozen real protocol contains compact pair features rather than the separable query/workflow side-feature matrices required by the existing IMC implementation; the fair feature-aware study provides the controlled IMC comparison.

Table 19: **SparseBank has the lowest aggregate regret in the post-hoc published-baseline replay.** SparseBank is the unchanged prospective output; the published baselines are supportive replay analyses, and paired confidence intervals include zero.

| Method | 25% | 35% | 50% | Mean | SB reduction |
|---|---|---|---|---|---|
| SparseBank | **0.05927** | 0.07512 | 0.05866 | **0.06435** | – |
| SoftImpute | 0.09447 | **0.07281** | **0.04784** | 0.07170 | 10.26% |
| AMC | 0.10859 | 0.09941 | 0.07683 | 0.09494 | 32.22% |

SparseBank is lower than AMC at all three budgets and has the lowest budget-averaged regret among the compatible published methods. Against SoftImpute, however, the advantage is aggregate rather than budget-uniform. Paired task-bootstrap 95% intervals for baseline-minus-SparseBank budget-averaged regret are $[-0.0283, 0.0388]$ for SoftImpute and $[-0.0507, 0.1082]$ for AMC; both include zero. These replays therefore support the same directional conclusion without upgrading the confirmatory inference.

At the task level, the post-hoc SoftImpute comparison contains 21 exact ties and 39 non-tied tasks, with SparseBank lower on 33 of the 39 non-tied tasks. The AMC comparison contains 42 exact ties and 18 non-tied tasks, with SparseBank lower on 10 of the 18 non-tied tasks. These counts are descriptive post-hoc diagnostics and do not change the confirmatory status of the preregistered comparisons.

Figure 10 retains the full budget-wise published-baseline replay. It makes explicit that the aggregate ranking does not imply monotone or budget-uniform dominance.

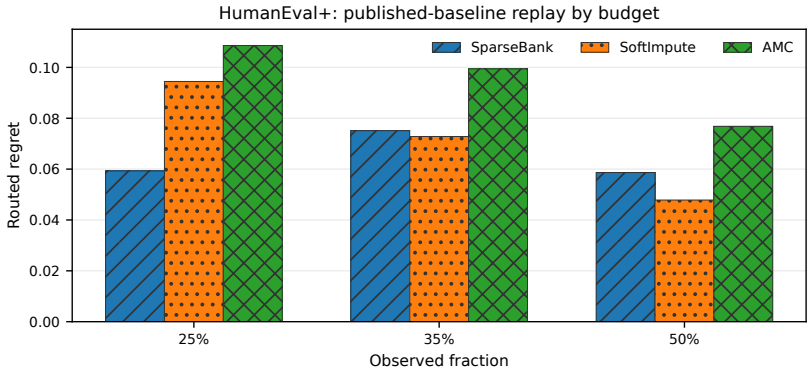



Figure 10: **SparseBank is lower than AMC at every fixed budget, while SoftImpute is lower at 35% and 50%.** Grouped bars show fixed evaluation conditions rather than a monotonicity claim.

**Post-hoc acquisition-rule robustness.** The confirmatory SparseBank run used one preregistered acquisition realization. After TEST2 was finalized, we replayed SparseBank, Uniform+OLS, and Leverage+OLS over seeds 0–99 on the finalized dense table without any model calls or workflow reruns. Each method used its own acquisition distribution, with the shared integer seed used only for paired seed-level comparisons. This post-hoc robustness analysis preserves the frozen features, OLS completion, bank size, tie-breaking, and metric definitions from the original evaluation.

Table 20: **The unshrunk OLS diagnostic does not show a stable acquisition-rule advantage.** Values are seed-level budget-averaged routed regret on the finalized TEST2 table; these rows are retained to diagnose the unregularized OLS estimator, not to define SparseBank-Anchor.

| Method | Mean | 2.5% quantile | 97.5% quantile | Median |
|---|---|---|---|---|
| Full-feature OLS / SparseBank | 0.068446 | 0.054374 | 0.109513 | 0.064146 |
| Uniform+OLS | 0.067444 | 0.053228 | 0.104137 | 0.064146 |
| Leverage+OLS | 0.067263 | 0.054859 | 0.112454 | 0.059167 |

Table 21: **Both paired seed-level intervals cross zero in the post-hoc acquisition replay.** $\Delta$ is baseline minus SparseBank budget-averaged routed regret.

| Comparison | Mean $\Delta$ | 95% bootstrap CI | SB lower | Tied |
|---|---|---|---|---|
| Uniform+OLS − SparseBank | -0.001002 | $[-0.002804, 0.000596]$ | 0.31 | 0.31 |
| Leverage+OLS − SparseBank | -0.001183 | $[-0.002755, 0.000337]$ | 0.25 | 0.37 |

Budget-specific mean routed regrets for the unshrunk full-feature OLS diagnostic at $25\%$, $35\%$, and $50\%$ are $0.072222/0.069066/0.064050$ for SparseBank, $0.069910/0.068035/0.064386$ for Uniform+OLS, and $0.069230/0.067827/0.064731$ for Leverage+OLS. The paired means are negative and the bootstrap intervals cross zero, so the replay does not establish a stable acquisition-rule advantage for SparseBank.

The earlier SparseBank-only seed-sensitivity diagnostic helps explain the original non-monotone $25\% \to 35\% \to 50\%$ routed-regret trajectory. It replayed the frozen SparseBank design over 100 fixed diagnostic acquisition seeds on the finalized dense TEST2 table. Mean routed regret across seeds was $0.07553$, $0.06706$, and $0.06509$ at the three budgets, with 95% seed-distribution intervals $[0.04622, 0.14988]$, $[0.04961, 0.12086]$, and $[0.04237, 0.09235]$. The original realization lies at the 47th, 78th, and 29.5th percentiles, respectively. Only $7\%$ of seeds exhibit a local $35\%$ peak, while $64\%$ are non-increasing across all three budgets. Thus finite-sample acquisition randomness contributes to the observed spike, although the broad seed intervals and weak geometry correlation (largest audited Spearman $\rho = -0.17$) do not support attributing it to a single geometric statistic.

The selected bank itself remains unchanged across the original three budgets. For HumanEval/108, the selected route changes as `he8_multi_visible_select` $\rightarrow$ `he8_direct` $\rightarrow$ `he8_multi_visible_select` across the 25%, 35%, and 50% budgets, with realized utility $1.0 \rightarrow 0.032738 \rightarrow 1.0$. The $25\% \rightarrow 35\%$ switch accounts for nearly all of the observed regret increase. This is consistent with routed regret being a discontinuous downstream quantity under feature misspecification: additional observations can perturb completed utilities enough to change a route even when the selected bank is unchanged. Figure 11 visualizes the diagnostic. The preregistered single-seed trajectory remains the reported confirmatory result.

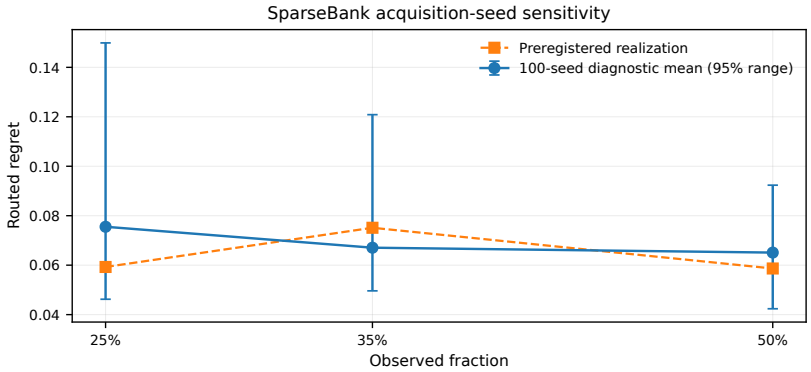



Figure 11: **The seed-averaged regret curve decreases with budget, while the preregistered 35% realization is relatively high.** Error bars show the central 95% range across 100 fixed post-hoc replay seeds; the dashed curve is the original preregistered realization. This diagnostic does not replace the preregistered result.

### B.11 Canonical APPS workflow-bank evaluation

This subsection reports the second real-task study for the canonical method: $G$-optimal acquisition with unregularized minimum-norm OLS, followed by the unchanged bank-selection and routing code. The study has no SparseBank-Anchor variant and no DEV utility predictor. It is an empirical study, not a theorem-mode-certified run.

**Protocol and provenance.** The run is `20260915_canonical_apps_workflow_bank_v2`. It uses snapshot `21e74ddf8de1a21436da12e3e653065c5213e9d1` of the `codeparrot/apps` test set, filtered to stdin/stdout tasks (Hendrycks et al., 2021), with $N = 100$ tasks, $K = 8$ workflows, $B = 3$, $\alpha = 0$, and $NK = 800$ task–workflow edges. We set $\alpha = 0$ deliberately to isolate acquisition/completion from heterogeneous token, latency, or API-cost estimation; cost-aware curation is therefore outside the empirical claims of this run. The workflows, task IDs, raw pair features, retained rank coordinates, design probabilities, budgets, and masks were frozen before target utility labels. The raw feature dimension is 119 and the retained rank is 78. The design and OLS operate only on the frozen exact-rank coordinates; the exact raw feature matrix, rank transform, and feature hash are retained as immutable run artifacts, with no label-dependent feature selection or post-hoc coordinate regrouping. The budgets are $m \in \{200, 280, 400\}$, corresponding to 25%, 35%, and 50% of $NK$. The artifact provides the frozen protocol, features, dense table, and result CSVs with file hashes.

Table 22: **The APPS workflow pool was frozen before target labels.** The first four workflows use Qwen3-8B; the last four use Qwen3.8-27B.

| Workflow ID | Procedure |
|---|---|
| `direct_greedy` | Direct single-pass generation |
| `plan_then_code` | Plan, then generate |
| `public_repair` | Draft, inspect public tests, repair |
| `candidate_public_select` | Generate candidates, select by public tests |
| `direct_27b` | 27B direct single-pass generation |
| `critique_revise_27b` | 27B draft, critique, and revision |
| `direct_27b_long` | 27B direct generation with longer output budget |
| `plan_27b_long` | 27B plan, then generation with longer output budget |

For scale, applying the theorem's sufficient sample gate to this empirical retained rank would not certify savings: with $d = 78$, $\delta = 0.05$, and even $\kappa = 1$, $\lceil 8d \log(2d/\delta) \rceil = 5021 > 800$. This is a sufficient-gate diagnostic, not evidence that 5021 samples are necessary. APPS remains empirical-only because the mean/noise assumptions and a valid residual bound are not established; the realized-design corollary does not remove these requirements.

The dense oracle is the exhaustive best bank of size three, {`critique_revise_27b`, `direct_27b_long`, `plan_27b_long`}, with objective $0.7444152132$. The gap between the best and second-best bank is only $0.0010459184$; the dense route margin is zero because some task-level routes are tied. The maximum-budget sparse union executed $555$ unique edges, and dense reveal executed the remaining $245$ edges. The merged table therefore contains $800$ unique physical edges, with zero failed edges and zero external paid cost. All $100$ task records passed the APPS schema validator using case-level all-stdin inference; hidden inputs, private expected outputs, and reference solutions were not exposed to the workflows.

**Formal sparse-first realization.** Table 23 reports the formal seed-0 realization. These rows are retained for reproducibility; the replay average below is the primary estimate for the acquisition comparison.

Table 23: **Formal APPS sparse-first results.** Lower is better for RMSE and regret; higher is better for route agreement and bank overlap.

| Method | $m$ | RMSE | Routed regret | Bank regret | Route agree. | Bank overlap |
|---|---|---|---|---|---|---|
| G-opt | 200 | 0.524081 | 0.230149 | 0.059307 | 0.380 | 0.666667 |
| G-opt | 280 | 0.478453 | 0.272649 | 0.059307 | 0.270 | 0.666667 |
| G-opt | 400 | 0.443913 | 0.174124 | 0.060942 | 0.410 | 0.666667 |
| Uniform | 200 | 0.677997 | 0.206666 | 0.017057 | 0.280 | 0.666667 |
| Uniform | 280 | 0.477229 | 0.112504 | 0.000000 | 0.330 | 1.000000 |
| Uniform | 400 | 0.467666 | 0.149157 | 0.000000 | 0.380 | 1.000000 |
| Leverage | 200 | 0.526186 | 0.175504 | 0.000000 | 0.310 | 1.000000 |
| Leverage | 280 | 0.457724 | 0.116830 | 0.000000 | 0.270 | 1.000000 |
| Leverage | 400 | 0.429447 | 0.060163 | 0.000000 | 0.290 | 1.000000 |

**Frozen-table replay.** We replayed the frozen acquisition distributions for 100 seeds using with-replacement sampling, the same budgets, minimum-norm unregularized OLS, exhaustive bank search, and deterministic tie-breaking. G-optimal acquisition has the highest mean bank overlap and lowest mean bank regret at every budget. Its exact-bank recovery rates are $42\%/57\%/57\%$, compared with $8\%/16\%/20\%$ for Uniform and $19\%/25\%/24\%$ for Leverage. The routed ordering is different: G-opt is better than Uniform at $m = 200$ but worse at $m = 280$ and $400$, and it is slightly worse than Leverage at $m = 200$ and worse at the two larger budgets.

Table 24: **One-hundred-seed frozen-table replay.** Values are replay means; lower is better for RMSE and regret, while higher is better for route agreement and bank overlap.

| Method | $m$ | RMSE | Routed regret | Bank regret | Route agree. | Bank overlap |
|---|---|---|---|---|---|---|
| G-opt | 200 | 0.519894 | 0.186836 | 0.016994 | 0.2581 | 0.780000 |
| G-opt | 280 | 0.467073 | 0.160058 | 0.006632 | 0.2740 | 0.843333 |
| G-opt | 400 | 0.446558 | 0.141100 | 0.004737 | 0.2731 | 0.853333 |
| Uniform | 200 | 0.578876 | 0.200395 | 0.032062 | 0.2113 | 0.590000 |
| Uniform | 280 | 0.481014 | 0.141193 | 0.012760 | 0.2166 | 0.666667 |
| Uniform | 400 | 0.439883 | 0.108517 | 0.008284 | 0.2245 | 0.706667 |
| Leverage | 200 | 0.506905 | 0.185187 | 0.021171 | 0.2178 | 0.670000 |
| Leverage | 280 | 0.454156 | 0.147464 | 0.012186 | 0.2226 | 0.710000 |
| Leverage | 400 | 0.428297 | 0.123238 | 0.009455 | 0.2195 | 0.723333 |

**Bank–route decomposition.** For every replay cell, routed regret equals bank-selection loss plus within-bank routing loss. Table 25 shows this decomposition for G-opt. The bank-selection term is small and decreases with budget, while the within-bank term remains the dominant contribution. Thus exact or near-exact bank recovery does not imply route recovery.

Table 25: **G-optimal routed regret is dominated by within-bank routing loss.** The two loss columns add to routed regret up to displayed precision.

| $m$ | Routed regret | Bank-selection loss | Within-bank routing loss |
|---|---|---|---|
| 200 | 0.186836 | 0.016994 | 0.169842 |
| 280 | 0.160058 | 0.006632 | 0.153426 |
| 400 | 0.141100 | 0.004737 | 0.136364 |

**Descriptive task-coverage diagnostics.** G-opt samples fewer distinct workflows per task and revisits more edges than the controls. This coexists with concentrated routed loss. These associations do not identify coverage as the cause or exclude misspecification:

Table 26: **G-optimal acquisition has narrower realized task coverage.** Values are means over the 100 replay seeds.

| $m$ | G-opt | | Uniform | | Leverage | |
|---|---|---|---|---|---|---|
| | Unique WF/task | Visit rate | Unique WF/task | Visit rate | Unique WF/task | Visit rate |
| 200 | 1.314 | 0.337 | 1.773 | 0.867 | 1.718 | 0.825 |
| 280 | 1.607 | 0.345 | 2.368 | 0.941 | 2.272 | 0.903 |
| 400 | 1.911 | 0.350 | 3.152 | 0.984 | 2.997 | 0.957 |

For G-opt, the top five/top ten task contributions account for $21.2\%/38.3\%$, $25.1\%/45.1\%$, and $29.0\%/50.4\%$ of routed regret at $m = 200, 280, 400$, respectively. The union of comparator-specific catastrophic tasks receives only $0.321/0.146/0.792$ mean G-opt draws, versus $2.621/3.733/4.958$ for the non-catastrophic tasks. Completion error is not a single explanation: G-opt has lower RMSE than Uniform at $m = 200, 280$ but not $400$, and Leverage has lower RMSE at all three budgets.

**Pre-label geometry.** Using only frozen pre-label features and acquisition probabilities, the Spearman association between G-opt task probability mass and excess routed regret versus Uniform is $-0.371/-0.386/-0.319$ at the three budgets. The association between G-opt probability concentration and unique-workflow coverage is approximately $0.872$ at all three budgets. These are descriptive associations from a frozen task set, not causal estimates. The formal seed-0 rows remain separate because that realization selects a non-dense bank at all three budgets.

**Invalid predecessor run.** The earlier `20260915_canonical_apps_workflow_bank_v1` run is retained only as invalid reproducibility history: an APPS adapter schema mismatch prevented a valid edge result. It contributes no evidence, metrics, or claims here.

### B.12 APPS ACQUISITION INTERVENTIONS AND RETAINED PROTOCOL

All follow-ups use the same frozen $100 \times 8$ observed utility table, 800-by-78 feature matrix (from 119 raw coordinates), minimum-norm unregularized OLS, $B = 3$, $\alpha = 0$, uniform task weights, and deterministic exhaustive bank/routing code as the canonical study. No clipping, regularization, feature refitting, or scoring change is introduced. The dense bank is $[5, 6, 7]$ with objective $0.7444152131782945$. Seeds are $0, \ldots, 99$ throughout. Canonical acquisition repeats remain repeated rows in OLS with the same cached score.

Pure Task-balanced acquisition cycles through randomized task rounds, choosing distinct workflows within each task. The $q = 1$ and $q = 2$ hybrids first observe that many distinct workflows per task, then append the unchanged with-replacement $G$-opt stream. The initialization is outcome-independent. Table 27 preserves the full fixed-draw allocation grid, and Table 28 gives its downstream outcomes. The plain deduplicated 400-edge row includes numerical tails and is a sensitivity result.

Table 27: **Fixed-draw acquisition counts and task coverage.** Every row uses exactly $m$ draws. TB denotes Task-balanced; TB-$G$ denotes its warm-start hybrids. WF/task includes unvisited tasks. Plain deduplication at 400 draws includes 112 tail edges.

| $m$ | Method | Unique edges | Duplicate fraction | WF/task | Visit rate |
|---|---|---|---|---|---|
| 200 | G-opt | 131.40 | 0.3430 | 1.3140 | 0.3372 |
| 200 | Plain dedup. | 200.00 | 0.0000 | 2.0000 | 0.3517 |
| 200 | Uniform | 177.34 | 0.1133 | 1.7734 | 0.8671 |
| 200 | Leverage | 171.83 | 0.1409 | 1.7183 | 0.8245 |
| 200 | TB | 200.00 | 0.0000 | 2.0000 | 1.0000 |
| 200 | TB-$G$ ($q = 1$) | 169.84 | 0.1508 | 1.6984 | 1.0000 |
| 200 | TB-$G$ ($q = 2$) | 200.00 | 0.0000 | 2.0000 | 1.0000 |
| 280 | G-opt | 160.67 | 0.4262 | 1.6067 | 0.3454 |
| 280 | Plain dedup. | 280.00 | 0.0000 | 2.8000 | 0.3598 |
| 280 | Uniform | 236.81 | 0.1542 | 2.3681 | 0.9413 |
| 280 | Leverage | 227.19 | 0.1886 | 2.2719 | 0.9030 |
| 280 | TB | 280.00 | 0.0000 | 2.8000 | 1.0000 |
| 280 | TB-$G$ ($q = 1$) | 207.27 | 0.2597 | 2.0727 | 1.0000 |
| 280 | TB-$G$ ($q = 2$) | 249.62 | 0.1085 | 2.4962 | 1.0000 |
| 400 | G-opt | 191.07 | 0.5223 | 1.9107 | 0.3495 |
| 400 | Plain dedup. | 400.00 | 0.0000 | 4.0000 | 0.5101 |
| 400 | Uniform | 315.16 | 0.2121 | 3.1516 | 0.9839 |
| 400 | Leverage | 299.73 | 0.2507 | 2.9973 | 0.9568 |
| 400 | TB | 400.00 | 0.0000 | 4.0000 | 1.0000 |
| 400 | TB-$G$ ($q = 1$) | 245.81 | 0.3855 | 2.4581 | 1.0000 |
| 400 | TB-$G$ ($q = 2$) | 298.23 | 0.2544 | 2.9823 | 1.0000 |

Table 28: **Full fixed-draw APPS follow-up outcomes.** All are 100-seed means. The plain 400-edge deduplicated result is support-sensitive; the 200/280 results equal the support-aware variant exactly.

| $m$ | Method | RMSE | Bank regret | Overlap | Routed regret | Within-bank | Agreement |
|---|---|---|---|---|---|---|---|
| 200 | G-opt | 0.519894 | 0.016994 | 0.780000 | 0.186836 | 0.169842 | 0.258100 |
| 200 | Plain dedup. | 0.435564 | 0.004821 | 0.833333 | 0.124653 | 0.119832 | 0.259900 |
| 200 | Uniform | 0.578876 | 0.032062 | 0.590000 | 0.200395 | 0.168333 | 0.211300 |
| 200 | Leverage | 0.506905 | 0.021171 | 0.670000 | 0.185187 | 0.164016 | 0.217800 |
| 200 | TB | 0.523912 | 0.022160 | 0.643333 | 0.169052 | 0.146891 | 0.223700 |
| 200 | TB-$G$ ($q = 1$) | 0.503334 | 0.020686 | 0.703333 | 0.180991 | 0.160305 | 0.243600 |
| 200 | TB-$G$ ($q = 2$) | 0.523912 | 0.022160 | 0.643333 | 0.169052 | 0.146891 | 0.223700 |
| 280 | G-opt | 0.467073 | 0.006632 | 0.843333 | 0.160058 | 0.153426 | 0.274000 |
| 280 | Plain dedup. | 0.416668 | 0.003517 | 0.666667 | 0.092049 | 0.088532 | 0.238100 |
| 280 | Uniform | 0.481014 | 0.012760 | 0.666667 | 0.141193 | 0.128433 | 0.216600 |
| 280 | Leverage | 0.454156 | 0.012186 | 0.710000 | 0.147464 | 0.135278 | 0.222600 |
| 280 | TB | 0.442844 | 0.009828 | 0.703333 | 0.121028 | 0.111200 | 0.222600 |
| 280 | TB-$G$ ($q = 1$) | 0.452257 | 0.009760 | 0.760000 | 0.146501 | 0.136741 | 0.253700 |
| 280 | TB-$G$ ($q = 2$) | 0.446190 | 0.009642 | 0.763333 | 0.137468 | 0.127826 | 0.245800 |
| 400 | G-opt | 0.446558 | 0.004737 | 0.853333 | 0.141100 | 0.136364 | 0.273100 |
| 400 | Plain dedup. | 0.403010 | 0.003333 | 0.666667 | 0.057152 | 0.053819 | 0.243700 |
| 400 | Uniform | 0.439883 | 0.008284 | 0.706667 | 0.108517 | 0.100234 | 0.224500 |
| 400 | Leverage | 0.428297 | 0.009455 | 0.723333 | 0.123238 | 0.113783 | 0.219500 |
| 400 | TB | 0.409777 | 0.005910 | 0.743333 | 0.090763 | 0.084852 | 0.226500 |
| 400 | TB-$G$ ($q = 1$) | 0.435035 | 0.004981 | 0.813333 | 0.132785 | 0.127804 | 0.263200 |
| 400 | TB-$G$ ($q = 2$) | 0.423502 | 0.004839 | 0.806667 | 0.115995 | 0.111156 | 0.247700 |

Within-bank loss is $F(\widehat{S}) - R(\widehat{S}, \widehat{r})$, so it plus bank regret equals routed regret. All displayed follow-up means use 100 matched seeds. Paired intervals reuse the stored 5000-resample percentile bootstrap with seed 2026091903. They are exploratory pointwise 95% intervals without multiplicity adjustment, conditional on this one observed table. Replaying these scores does not estimate the utility distribution of fresh executions.

### B.13 APPS EQUAL-UNIQUE CONTROLS

Table 29: **Matching unique observations reverses Task-balanced's apparent advantage.** Mean routed regret and paired Task-balanced-minus-$G$-opt differences with pointwise 95% bootstrap intervals; 100 matched seeds.

| $U$ | $G$-opt | Task-balanced | Uniform | Difference [95% interval] |
|---|---|---|---|---|
| 120 | 0.20663 | 0.28250 | 0.28142 | $0.07587\ [0.05555, 0.09656]$ |
| 160 | 0.16160 | 0.21734 | 0.22095 | $0.05574\ [0.03985, 0.07131]$ |
| 190 | 0.14314 | 0.18546 | 0.18262 | $0.04232\ [0.02860, 0.05644]$ |

Each unchanged acquisition stream stops on the draw that reveals its $U$-th unique edge. Repeats count as draws and OLS rows but not as new observations. There is no draw cap and no renormalization onto unseen edges. The primary targets are $U = 120, 160, 190$. The $q = 2$ hybrid has not finished its 200-edge initialization at these targets and is exactly Task-balanced there; these rows do not test a completed hybrid. The supplemental $U = 220$ target was specified before outcomes to examine $q = 2$ after initialization.

Table 30: **Full equal-unique control, including hybrids.** Unique edges equal $U$ exactly in every seed and WF/task equals $U/100$. The $q = 2$ rows at $U < 200$ are initialization-only; $U = 220$ is supplemental.

| $U$ | Method | Draws | Visit rate | RMSE | Bank regret | Overlap | Routed regret |
|---|---|---|---|---|---|---|---|
| 120 | G-opt | 174.55 | 0.3328 | 0.55736 | 0.02430 | 0.73667 | 0.20663 |
| 120 | Uniform | 130.11 | 0.7257 | 1.04236 | 0.06617 | 0.52333 | 0.28142 |
| 120 | Leverage | 133.30 | 0.6928 | 0.68592 | 0.03893 | 0.59333 | 0.23591 |
| 120 | TB | 120.00 | 1.0000 | 1.02191 | 0.06684 | 0.54333 | 0.28250 |
| 120 | TB-$G$ ($q = 1$) | 124.28 | 1.0000 | 0.87235 | 0.05842 | 0.52000 | 0.27610 |
| 120 | TB-$G$ ($q = 2$) | 120.00 | 1.0000 | 1.02191 | 0.06684 | 0.54333 | 0.28250 |
| 160 | G-opt | 276.51 | 0.3457 | 0.46724 | 0.00662 | 0.84000 | 0.16160 |
| 160 | Uniform | 178.16 | 0.8322 | 0.63858 | 0.03981 | 0.56667 | 0.22095 |
| 160 | Leverage | 184.53 | 0.8005 | 0.52100 | 0.02367 | 0.66000 | 0.19295 |
| 160 | TB | 160.00 | 1.0000 | 0.63267 | 0.03535 | 0.60667 | 0.21734 |
| 160 | TB-$G$ ($q = 1$) | 183.28 | 1.0000 | 0.53305 | 0.02460 | 0.68667 | 0.19284 |
| 160 | TB-$G$ ($q = 2$) | 160.00 | 1.0000 | 0.63267 | 0.03535 | 0.60667 | 0.21734 |
| 190 | G-opt | 394.58 | 0.3494 | 0.44756 | 0.00498 | 0.86333 | 0.14314 |
| 190 | Uniform | 216.51 | 0.8871 | 0.54852 | 0.02474 | 0.63667 | 0.18262 |
| 190 | Leverage | 224.96 | 0.8554 | 0.48042 | 0.01664 | 0.69000 | 0.16590 |
| 190 | TB | 190.00 | 1.0000 | 0.54177 | 0.02684 | 0.63667 | 0.18546 |
| 190 | TB-$G$ ($q = 1$) | 240.52 | 1.0000 | 0.46883 | 0.01482 | 0.75333 | 0.16104 |
| 190 | TB-$G$ ($q = 2$) | 190.00 | 1.0000 | 0.54177 | 0.02684 | 0.63667 | 0.18546 |
| 220 | G-opt | 580.82 | 0.3531 | 0.43691 | 0.00298 | 0.88333 | 0.13142 |
| 220 | Uniform | 256.84 | 0.9235 | 0.49851 | 0.01672 | 0.67000 | 0.15425 |
| 220 | Leverage | 268.91 | 0.8961 | 0.45731 | 0.01066 | 0.71333 | 0.14836 |
| 220 | TB | 220.00 | 1.0000 | 0.49609 | 0.01737 | 0.65667 | 0.15648 |
| 220 | TB-$G$ ($q = 1$) | 313.87 | 1.0000 | 0.44445 | 0.00776 | 0.79000 | 0.13968 |
| 220 | TB-$G$ ($q = 2$) | 229.09 | 1.0000 | 0.48015 | 0.01540 | 0.67667 | 0.15189 |

At every displayed target, canonical $G$-opt has the lowest mean bank regret and highest mean overlap of these six methods. At $U = 190$, Task-balanced-minus-$G$-opt bank regret is $+0.021854$ ($[0.015056, 0.029334]$) and overlap is $-0.226667$ ($[-0.273333, -0.180000]$). Its routed-regret differences from Uniform at $U = 120, 160, 190$ are $+0.001086$ ($[-0.021787, 0.023630]$), $-0.003611$ ($[-0.024397, 0.016915]$), and $+0.002837$ ($[-0.014471, 0.020639]$). These intervals neither support an advantage nor establish equivalence.

The seedwise bridge holds each archived canonical mask and metric fixed and stops Task-balanced at exactly that mask's unique count. Defining benefit as canonical minus Task-balanced routed regret,

the earlier equal-draw benefits at $m = 200, 280, 400$ are $0.017784, 0.039030, 0.050338$; unique matching changes them to $-0.070022, -0.055467, -0.036670$. The entire apparent mean benefit disappears and reverses. These are changes in policy contrasts, not percentages causally mediated by duplicate avoidance.

The $q = 1$ hybrid offers a conditional tradeoff: at $U = 190$ it requires 240.52 draws rather than canonical's 394.58, but has higher routed regret by $0.017901$ ($[0.006429, 0.030019]$). It improves over Task-balanced and Uniform there. At $U = 220$ its difference from canonical is $0.008258$ ($[-0.001642, 0.018461]$), which does not establish noninferiority. The $q = 2$ routed-regret difference from $q = 1$ at $U = 220$ is $0.012213$ ($[0.001409, 0.023313]$). Neither broad visitation nor unique count alone determines the ranking; edge identities, geometry, and duplicate OLS weights also differ.

### B.14 APPS DEDUPLICATED REALIZATION

Table 31: **APPS regret advantages remain at matched draws and unique observations.** Support-aware deduplicated $G$-opt minus comparator; negative favors the variant. The Uniform-WOR rows match both counts. Exploratory pointwise 95% seed-bootstrap intervals condition on the frozen table, without multiplicity adjustment.

| $m$ | Comparator | Routed-regret difference | Bank-regret difference |
|---|---|---|---|
| 200 | Uniform-WOR | $-0.04545$ $[-0.05858, -0.03288]$ | $-0.01449$ $[-0.02011, -0.00942]$ |
| 280 | Uniform-WOR | $-0.03241$ $[-0.04007, -0.02504]$ | $-0.00719$ $[-0.01057, -0.00412]$ |
| 280 | $G$-opt-WR | $-0.06801$ $[-0.07623, -0.06041]$ | $-0.00312$ $[-0.00678, -0.00029]$ |
| 280 | Task-balanced | $-0.02898$ $[-0.03614, -0.02219]$ | $-0.00631$ $[-0.00851, -0.00416]$ |
| 280 | Uniform-WR | $-0.04914$ $[-0.05868, -0.04029]$ | $-0.00924$ $[-0.01347, -0.00576]$ |

The canonical active-set solver is called once and its probabilities $\pi$ are frozen. The deduplicated sampler excludes observed edges and samples proportionally to the remaining original positive weights. For numerical stability it divides eligible weights by their maximum and then their sum. It performs no design reoptimization, rejection sampling, uniform mixture, or outcome-dependent adaptation. Thus each actual draw reveals one new edge. This changes sampling and OLS row multiplicities, while reusing the exact estimator and decision functions.

The stored validation reproduced all 300 canonical cells within $2.23 \times 10^{-16}$ in reported metrics and all 1800 prior control masks exactly. The subsequent tail study reproduced all 300 plain-deduplicated masks, predictions, banks, routes, and metrics exactly before threshold evaluation.

At $m = 280$, canonical-minus-Task-balanced routed regret is $0.039030$, whereas deduplicated-minus-Task-balanced is $-0.028979$. The $0.068009$ improvement exceeds the original deficit; it is not a causal mediation estimate. Both latter methods use exactly 280 unique rows. Deduplication also beats the retained with-replacement Uniform control. The Uniform-WOR follow-up in Appendix B.16 matches duplicate avoidance and retains a regret advantage for support-aware $G$-opt.

Bank identity worsens despite lower bank regret. At $m = 280$ and 400, canonical recovers the dense bank in 57/100 seeds and plain deduplication in none. Deduplication chooses $[3, 6, 7]$ in 98 seeds at $m = 280$ and $[2, 6, 7]$ in the other two; all seeds choose $[3, 6, 7]$ at $m = 400$. The former bank's observed-table regret is $0.003333$. At $m = 280$, overlap decreases by $0.176667$ ($[0.140000, 0.213333]$ for canonical minus deduplicated). The mean bank-regret improvement at $m = 400$ has an interval crossing zero.

Task coverage remains a residual limitation. At $m = 280$, visit rates are $0.3454$ for canonical, $0.3598$ for deduplicated, and $1.0000$ for Task-balanced. Deduplicated mean per-task loss is $0.030683$ on visited tasks and $0.126538$ on unvisited tasks. These groups differ in task identity and difficulty, so this association does not identify a coverage effect. The large improvement with little change in visitation, and the advantage over Task-balanced at identical unique count, support useful allocation structure beyond merely visiting more tasks.

### B.15 APPS NUMERICAL-TAIL AND SUPPORT SENSITIVITY

The original design has 19 exactly zero weights, 781 positive weights, and 493 extremely small positives at or below $10^{-12}$. Their total mass is $1.23100725 \times 10^{-31}$ and their largest weight is $4.47063387 \times 10^{-32}$. The smallest of the other 288 weights is $7.68178284 \times 10^{-5}$. This gap motivates $\tau = 10^{-10}$ before thresholded outcomes; $10^{-8}$ and $10^{-6}$ yield exactly the same support. The $10^{-4}$ sensitivity removes three additional non-tail edges. It is not preferred on the basis of regret.

Table 32: **Original canonical design-weight distribution.** Statistics include all 800 weights; quantiles use linear interpolation. Values are copied from the stored weight audit.

| Statistic | Value |
|---|---|
| Minimum | 0.00000000e+00 |
| Minimum positive | 4.94065646e-324 |
| Median | 6.63658544e-249 |
| Maximum | 1.12108038e-02 |
| 1% quantile | 0.00000000e+00 |
| 5% quantile | 4.94065646e-324 |
| 10% quantile | 4.94065646e-324 |
| 25% quantile | 4.94065646e-324 |
| 50% quantile | 6.63658544e-249 |
| 75% quantile | 2.08380189e-03 |
| 90% quantile | 4.54743607e-03 |
| 95% quantile | 5.81410845e-03 |
| 99% quantile | 8.08698342e-03 |
| Exactly zero | 19 |
| Positive | 781 |
| Positive and $\leq 10^{-12}$ | 493 |

Thresholding applies to the original frozen probabilities, not the changing conditional probabilities. The wrapper zeros $\pi_e \leq \tau$ and calls the same deduplicated sampler up to the smaller of the requested budget and eligible support size. It never fills the budget with excluded edges. Table 33 distinguishes requested and actual counts; workflows per task is actual count divided by 100.

Table 33: **Support-aware outcomes distinguish requested budgets from actual observations.** The three smaller positive thresholds are identical. Only the unthresholded row reaches 400 edges.

| $\tau$ | Cap | Actual | Routed | Bank regret | Overlap | RMSE | Visit | Agreement |
|---|---|---|---|---|---|---|---|---|
| None | 200 | 200 | 0.124653 | 0.004821 | 0.833333 | 0.435564 | 0.3517 | 0.2599 |
| None | 280 | 280 | 0.092049 | 0.003517 | 0.666667 | 0.416668 | 0.3598 | 0.2381 |
| None | 400 | 400 | 0.057152 | 0.003333 | 0.666667 | 0.403010 | 0.5101 | 0.2437 |
| $10^{-10}, 10^{-8}, 10^{-6}$ | 200 | 200 | 0.124653 | 0.004821 | 0.833333 | 0.435564 | 0.3517 | 0.2599 |
| $10^{-10}, 10^{-8}, 10^{-6}$ | 280 | 280 | 0.092049 | 0.003517 | 0.666667 | 0.416668 | 0.3598 | 0.2381 |
| $10^{-10}, 10^{-8}, 10^{-6}$ | 400 | 288 | 0.094252 | 0.003333 | 0.666667 | 0.416261 | 0.3600 | 0.2400 |
| $10^{-4}$ | 200 | 200 | 0.120865 | 0.004030 | 0.830000 | 0.435179 | 0.3500 | 0.2623 |
| $10^{-4}$ | 280 | 280 | 0.092156 | 0.003700 | 0.666667 | 0.416928 | 0.3600 | 0.2382 |
| $10^{-4}$ | 400 | 285 | 0.094252 | 0.003333 | 0.666667 | 0.416445 | 0.3600 | 0.2400 |

Table 34: **Reachability and selected numerical-tail counts.** Remaining counts include only eligible unseen edges. Counts below $10^{-12}, 10^{-10}, 10^{-8}, 10^{-6}$ coincide and share one column. No selected weight is exactly zero.

| $\tau$ | Cap | Actual | Reached? | Remaining | Below small scales | Below $10^{-4}$ |
|---|---|---|---|---|---|---|
| None | 200 | 200 | Yes | 581 | 0.00 | 0.11 |
| None | 280 | 280 | Yes | 501 | 0.00 | 0.68 |
| None | 400 | 400 | Yes | 381 | 112.00 | 115.00 |
| $10^{-10}, 10^{-8}, 10^{-6}$ | 200 | 200 | Yes | 88 | 0.00 | 0.11 |
| $10^{-10}, 10^{-8}, 10^{-6}$ | 280 | 280 | Yes | 8 | 0.00 | 0.68 |
| $10^{-10}, 10^{-8}, 10^{-6}$ | 400 | 288 | No | 0 | 0.00 | 3.00 |
| $10^{-4}$ | 200 | 200 | Yes | 85 | 0.00 | 0.00 |
| $10^{-4}$ | 280 | 280 | Yes | 5 | 0.00 | 0.00 |
| $10^{-4}$ | 400 | 285 | No | 0 | 0.00 | 0.00 |

The three small positive thresholds preserve all 100 ordered masks, predictions, banks, and routes exactly at $m = 200/280$. They are support-equivalence checks, not independent replications. The paired comparisons in Table 35 retain clear regret improvements at those budgets. At a requested cap of 400, their 288-edge endpoint has routed regret $0.094252$, versus $0.057152$ for plain 400-edge deduplication: a difference of $0.037100$ ($[0.036200, 0.038000]$). Bank regret and overlap are unchanged at $0.003333$ and $0.666667$, so this loss is within-bank routing loss in the retained outputs.

Table 35: **Support-aware paired differences.** The threshold is $\tau = 10^{-10}$. Variant minus comparator with pointwise 95% intervals. At cap 400 the variant uses 288 observations and controls use 400 draws; those rows are not matched-draw comparisons.

| Cap | Actual | Comparator | Routed-regret difference | Bank-regret difference |
|---|---|---|---|---|
| 200 | 200 | G-opt | $-0.06218$ $[-0.07558, -0.04873]$ | $-0.01217$ $[-0.01881, -0.00632]$ |
| 200 | 200 | Task-balanced | $-0.04440$ $[-0.05806, -0.03122]$ | $-0.01734$ $[-0.02403, -0.01147]$ |
| 200 | 200 | Uniform | $-0.07574$ $[-0.08990, -0.06182]$ | $-0.02724$ $[-0.03623, -0.01941]$ |
| 280 | 280 | G-opt | $-0.06801$ $[-0.07623, -0.06041]$ | $-0.00312$ $[-0.00678, -0.00029]$ |
| 280 | 280 | Task-balanced | $-0.02898$ $[-0.03614, -0.02219]$ | $-0.00631$ $[-0.00851, -0.00416]$ |
| 280 | 280 | Uniform | $-0.04914$ $[-0.05868, -0.04029]$ | $-0.00924$ $[-0.01347, -0.00576]$ |
| 400 | 288 | G-opt | $-0.04685$ $[-0.05405, -0.03955]$ | $-0.00140$ $[-0.00334, 0.00027]$ |
| 400 | 288 | Task-balanced | $0.00349$ $[-0.00208, 0.00860]$ | $-0.00258$ $[-0.00429, -0.00102]$ |
| 400 | 288 | Uniform | $-0.01427$ $[-0.02153, -0.00769]$ | $-0.00495$ $[-0.00847, -0.00212]$ |

The endpoint still beats the 400-draw canonical and Uniform controls, but not clearly Task-balanced. This is a 288-versus-400-draw comparison, not a completed 400-edge thresholded result. Every endpoint seed observes the same support set and returns the same bank, route, and routed regret; only row order differs. Endpoint intervals reflect comparator-seed variability, not 100 distinct information sets. Removing the tail simultaneously reduces the attainable unique count, so the sensitivity cannot isolate the value of tiny weights at a matched 400-edge budget.

The more aggressive $10^{-4}$ cutoff has routed differences from plain deduplication of $-0.003788$ ($[-0.011469, 0.004139]$) at $m = 200$ and $+0.000107$ ($[-0.001564, 0.001748]$) at $m = 280$. No equivalence margin was specified. We retain the simpler original-weight threshold $10^{-10}$ and explicit support exhaustion, without claiming a universally appropriate cutoff or empirical certification under Corollary 4.

**Follow-up provenance.** The artifact directory `results/apps_controls/` retains the Task-balanced, equal-unique, deduplicated, and support-threshold summaries and paired contrasts. The central replay inputs and commands are mapped in the artifact `README.md`. All conclusions remain conditional on the frozen APPS table, its representation, and its passing-prefix score.

### B.16 APPS Uniform-without-replacement control

Uniform-WOR provides a direct comparison with support-aware deduplicated $G$-opt when both methods acquire exactly $m$ distinct observations. It uses the unchanged successive-sampling function with 800 equal weights: after $t$ selections, each unseen edge has probability $1/(800 - t)$ and selected edges have probability zero. The follow-up uses seeds 0–99, the same `default_rng(seed)` coupling and nested prefixes as deduplicated $G$-opt, and budgets $m = 200, 280$. Acquisition masks were saved before evaluating the new outcomes.

The support-aware arm reuses the stored $\tau = 10^{-10}$ results exactly. It preserves relative original weights on unseen eligible edges without recomputing the design. Its 288-edge support leaves 88 and 8 eligible edges unseen at the two budgets; Uniform-WOR samples from all 800 edges and leaves 600 and 520 unseen. These comparisons match draws and unique observations, not eligible support. The weights and support restriction together define the tested $G$-opt allocation. No 400-edge comparison was run because the support-aware arm cannot reach that count.

Only the Uniform-WOR arm was fitted in the archived follow-up: 200 fits across the two budgets and 100 seeds. The 600 support-aware and with-replacement control cells were reused without refitting. Every arm uses the same frozen $100 \times 8$ passing-prefix utility table, $800 \times 78$ features, minimum-norm unregularized OLS with `rcond=None` and no clipping, uniform task weights, $\alpha = 0$, and bank size at most three. Exhaustive bank search retains the size-first lexicographic strict-

improvement rule; routing retains smallest-index ties. The deterministic dense bank is $[5, 6, 7]$, with objective $0.7444152131782945$.

Table 36: **Uniform-WOR narrows the APPS regret gap, but support-aware $G$-opt retains lower regret.** Means over 100 matched seeds. Overlap and agreement compare bank and route identities with the deterministic dense reference. WR denotes sampling with replacement; both other arms observe exactly $m$ unique edges.

| $m$ | Method | Routed regret | Bank regret | Bank overlap | RMSE | Route agreement |
|---|---|---|---|---|---|---|
| 200 | Support-aware $G$-opt | 0.124653 | 0.004821 | 0.833333 | 0.435564 | 0.259900 |
| 200 | Uniform-WOR | 0.170105 | 0.019313 | 0.690000 | 0.530699 | 0.244900 |
| 200 | Canonical $G$-opt-WR | 0.186836 | 0.016994 | 0.780000 | 0.519894 | 0.258100 |
| 200 | Uniform-WR | 0.200395 | 0.032062 | 0.590000 | 0.578876 | 0.211300 |
| 280 | Support-aware $G$-opt | 0.092049 | 0.003517 | 0.666667 | 0.416668 | 0.238100 |
| 280 | Uniform-WOR | 0.124461 | 0.010708 | 0.723333 | 0.451113 | 0.234400 |
| 280 | Canonical $G$-opt-WR | 0.160058 | 0.006632 | 0.843333 | 0.467073 | 0.274000 |
| 280 | Uniform-WR | 0.141193 | 0.012760 | 0.666667 | 0.481014 | 0.216600 |

Table 37: **Matching unique observations does not match task visitation.** Every row uses $m$ draws. WF/task counts distinct workflows and includes unvisited tasks. Different task and workflow identities remain part of each allocation.

| $m$ | Method | Unique edges | Duplicate fraction | WF/task | Task visit rate |
|---|---|---|---|---|---|
| 200 | Support-aware $G$-opt | 200.00 | 0.0000 | 2.0000 | 0.3517 |
| 200 | Uniform-WOR | 200.00 | 0.0000 | 2.0000 | 0.9049 |
| 200 | Canonical $G$-opt-WR | 131.40 | 0.3430 | 1.3140 | 0.3372 |
| 200 | Uniform-WR | 177.34 | 0.1133 | 1.7734 | 0.8671 |
| 280 | Support-aware $G$-opt | 280.00 | 0.0000 | 2.8000 | 0.3598 |
| 280 | Uniform-WOR | 280.00 | 0.0000 | 2.8000 | 0.9697 |
| 280 | Canonical $G$-opt-WR | 160.67 | 0.4262 | 1.6067 | 0.3454 |
| 280 | Uniform-WR | 236.81 | 0.1542 | 2.3681 | 0.9413 |

**Paired regret comparisons.** Table 38 reports the new matched-unique contrast and Uniform's own replacement contrast. Intervals use the stored 5000-resample percentile bootstrap with seed 2026091903, resampling matched acquisition seeds. They are exploratory pointwise 95% intervals without multiplicity adjustment, conditional on this frozen table; they do not quantify population or fresh-execution uncertainty. The support-aware arm lowers both routed and bank regret at both budgets. Uniform-WOR lowers routed regret relative to Uniform-WR at both budgets, but its bank-regret interval at 280 includes zero. The retained comparisons against canonical $G$-opt, Task-balanced, and Uniform-WR appear in Table 35.

Table 38: **The allocation advantage remains after matching duplicate avoidance.** Differences are left minus right; negative regret favors the left method. "Support-aware" denotes support-aware deduplicated $G$-opt. Intervals condition on the frozen table.

| $m$ | Left − right | Routed difference [95% interval] | Bank difference [95% interval] |
|---|---|---|---|
| 200 | Support-aware − Uniform-WOR | $-0.045452$ $[-0.058583, -0.032879]$ | $-0.014492$ $[-0.020110, -0.009423]$ |
| 280 | Support-aware − Uniform-WOR | $-0.032412$ $[-0.040072, -0.025038]$ | $-0.007192$ $[-0.010574, -0.004116]$ |
| 200 | Uniform-WOR − Uniform-WR | $-0.030290$ $[-0.047222, -0.012930]$ | $-0.012749$ $[-0.022704, -0.003322]$ |
| 280 | Uniform-WOR − Uniform-WR | $-0.016732$ $[-0.028641, -0.004939]$ | $-0.002052$ $[-0.007060, +0.002683]$ |

**Descriptive gap accounting.** For either regret metric, let $W$, $U$, and $G$ denote the regrets of Uniform-WR, Uniform-WOR, and support-aware $G$-opt. The identity $(W - G) = (W - U) + (U - G)$ holds for every matched seed. Table 39 reports mean gains and the ratio of mean $W - U$ to mean $W - G$. Replacing Uniform-WR removes 40.0%/34.0% of the previous routed-regret gap at 200/280 edges, leaving 60.0%/66.0%. These ratios describe a change of baseline; they do not identify a causal mediator or the fraction of $G$-opt's own deduplication gain caused by duplicate avoidance.

Table 39: **Generic duplicate avoidance explains part of the earlier baseline gap.** Entries are mean regret gains; the final column is a ratio of means, not a causal mediation estimate. $W$, $U$, and $G$ are defined in the text.

| $m$ | Metric | Old gap $W - G$ | Uniform gain $W - U$ | Remaining $U - G$ | Gap removed |
|---|---|---|---|---|---|
| 200 | Routed regret | 0.075742 | 0.030290 | 0.045452 | 40.0% |
| 200 | Bank regret | 0.027241 | 0.012749 | 0.014492 | 46.8% |
| 280 | Routed regret | 0.049143 | 0.016732 | 0.032412 | 34.0% |
| 280 | Bank regret | 0.009243 | 0.002052 | 0.007192 | 22.2% |

**Identity, coverage, and scope.** Lower regret does not imply better identity recovery. At 200 edges, support-aware $G$-opt has higher bank overlap than Uniform-WOR ($0.833333$ versus $0.690000$); at 280, overlap is lower ($0.666667$ versus $0.723333$), with paired difference $-0.056667$ ($[-0.096667, -0.016667]$). Route-agreement intervals include zero at both budgets. The loss of exact-bank recovery relative to canonical $G$-opt in Appendix B.14 is unchanged.

Uniform-WOR visits 90.49%/96.97% of tasks, versus 35.17%/35.98% for support-aware $G$-opt, despite identical mean workflows per task. Broad visitation alone therefore does not order routed regret here, but the comparison is not a randomized mediation analysis of coverage. Equal counts and one OLS row per edge rule out unequal unique-edge counts as the sole explanation. They test the full frozen allocation, including its support, rather than isolating a feature direction or showing that the classical $G$-optimal criterion is the unique mechanism.

All findings retain the passing-prefix score and one observed table. Its stored full-table projection RMSE of $0.384241$ indicates substantial lack of fit to that replay target, not measured misspecification of expected fresh-execution utilities. Repeated cached draws add no score information, while independent fresh executions could. The empirical sampler is outside the canonical IID completion theorem; no universal acquisition superiority, optimality of deduplication, cross-population generalization, or robust 400-unique-edge result follows. HumanEval+ boundary results remain unchanged.

**Stored validation and provenance.** The artifact directory `results/apps/` contains `summary.csv`, `paired_comparisons.csv`, `gain_accounting.csv`, and per-seed outcomes; `data/apps/` supplies the frozen table, features, and masks. Stored validation reports exact agreement for all 600 reused cells and all 200 new masks' unique counts, with zero discrepancy between saved predictions and reported metrics.

### B.17 Prospective CodeContests: protocol and complete results

**Population and scoring.** The study uses the pinned CodeContests TEST revision `802411c3010cb00d1b05bad57ca77365a3c699d6`. Technical eligibility requires a noninteractive, deterministic stdin/stdout task with no required files, network, external assets, or third-party packages. Each test group must contain paired input/output strings; public examples and at least one private or generated test are required. Statement screening excludes tasks needing a semantic checker for noncanonical outputs. The retained comparator uses exact tokens, case-normalized yes/no, or the two declared floating-point rules. Of 121 technically eligible tasks, 11 had prior workflow-artifact matches, leaving 110 artifact-free candidates. Sorting by the pinned numeric TEST row ID and taking the first 100 fixes the population without outcomes or difficulty-based selection. The filename/archive-entry inventory found zero prior workflow artifacts for every selected task. This claim concerns recorded project history, not model pretraining or unrecorded external use. The eight workflows preserve the APPS procedures; $B = 3$, $\alpha = 0$, and task weights are $1/100$. Public-text features and workflow descriptors give an $800 \times 119$ raw matrix reduced to rank 68 before target execution.

Utility is passed records divided by all provided public, private, and generated records. The 100 tasks contain 20882 records. Every valid-source program is graded on the complete ordered sequence, preserving repeated records and continuing after wrong answers, runtime errors, and per-input 30-second timeouts. One record is one entire stdin/stdout pair, which can contain multiple contest cases. The frozen comparator is exact tokens except for case-normalized yes/no answers and the two statement-specific floating-point rules. Public workflow feedback uses only the original first at-most-three public tests; private/generated fixtures are available only after the final program has

been saved. Policy rejection assigns zero; none occurred. This full-test fraction differs from APPS's passing-prefix score, whose original values and qualifications are retained.

**Preregistration and sparse-first chronology.** $G$ is the original support-aware deduplicated sampler from Section 4.4, with $\pi_e > 10^{-10}$ and no design refit. $M$ samples without replacement from $0.9\pi_e + 0.1/800$, and $U$ does so from uniform weights. The preregistered primary comparison is $M - U$; $G - U$ is the secondary unchanged-method replication, and $M - G$ tests the exploration floor. All 600 arm/seed/budget masks were frozen before execution: seeds 0–99, budgets 200 and 280, nested within each arm and seed. The prospective acquisition seed is 0.

The predeclared feasibility rule retains support-aware acquisition as the primary arm only if meaningful support contains at least 280 edges and covers at least 50 of the 100 tasks. Original support has 288 edges on 36 tasks, so both budgets are reachable but the task-coverage condition fails. The resulting 90/10 mixture is a preregistered coverage intervention; the secondary $G$ arm still tests the original sampler at both reachable budgets. It is not retrospectively promoted to the primary arm.

Table 40: **All sparse decisions precede dense execution.** UTC on 2026-09-22. Completion/start times are from committed edge records; separate commit and launch events corroborate the ordering.

| Milestone | UTC timestamp |
|---|---|
| Final sparse execution completion | 21:06:18.544584 |
| Final sparse commit event | 21:06:18.581124 |
| All six estimator/bank/route artifacts frozen | 21:06:19.058028 |
| Decision hash recorded in chronology | 21:06:19.136061 |
| First dense worker launch | 21:06:19.272269 |
| First dense execution | 21:06:19.351750 |
| Final dense execution completion | 23:52:19.309995 |
| Final dense commit event | 23:52:19.465093 |

The 492-edge sparse union was completed before six separate minimum-norm OLS fits, each using only its own frozen mask. Predictions, banks and routes were hashed before any of the 308 remaining edges began execution. There is no separate fit-finish timestamp; all six fits were complete by the decision freeze. The final ledger has 800 unique committed population edges, with no duplicates, omissions, or out-of-population edges. Local timestamps, immutable artifact hashes and the event hash chain establish record order; they are not external notarization.

**Prospective realization and bank identity.** Table 41 evaluates the actual frozen seed-0 decisions. The dense-optimal bank is $[3, 5, 7]$, corresponding to `candidate_public_select`, `critique_revise_27b`, and `plan_27b_long`, with objective $0.640459928416$. Indices 0, 1 and 2 denote `direct_greedy`, `plan_then_code`, and `public_repair`, respectively. Exhaustive bank search visits sizes up to three in size-first lexicographic order and updates only on strict improvement. Within-bank ties choose the smallest workflow index.

Table 41: **The single prospective realization is distinct from replay means.** Acquisition seed 0; all six predictions, banks and routes were frozen before dense execution. $M$ versus $U$ is primary, and unchanged $G$ is secondary.

| $m$ | Arm | Routed regret | Bank regret | RMSE | Overlap | Agreement | Bank |
|---|---|---|---|---|---|---|---|
| 200 | $M$ | 0.197956 | 0.052657 | 0.430115 | 0.333333 | 0.200000 | $[0, 2, 7]$ |
| 200 | $U$ | 0.276309 | 0.158227 | 0.447476 | 0.666667 | 0.310000 | $[1, 3, 5]$ |
| 200 | $G$ | 0.160358 | 0.035709 | 0.431574 | 0.666667 | 0.300000 | $[0, 3, 7]$ |
| 280 | $M$ | 0.147861 | 0.010461 | 0.414193 | 0.666667 | 0.310000 | $[2, 3, 7]$ |
| 280 | $U$ | 0.230364 | 0.000000 | 0.436600 | 1.000000 | 0.370000 | $[3, 5, 7]$ |
| 280 | $G$ | 0.142361 | 0.010461 | 0.413023 | 0.666667 | 0.330000 | $[2, 3, 7]$ |

At $m = 200$, the $M$ and $U$ banks are disjoint and their routes differ on all 100 tasks; at $m = 280$, their routes differ on 67 tasks. At the latter budget, $U$ recovers the exact dense-optimal bank with zero bank regret and overlap one but incurs routed regret $0.230364$. $G$ has bank regret $0.010461$ and overlap $0.666667$, yet routed regret $0.142361$. $M$ and $G$ select the same bank at $m = 280$ but differ on 15 routes. Dense-bank routing has ties on 34 tasks, so agreement also depends on the declared

tie rule. Bank identity and routing utility must both be measured; lower overlap is not intrinsically desirable.

**Conditional 100-seed replay.** The completed table is reused with the pre-execution masks and identical unregularized minimum-norm OLS, without clipping, and unchanged decisions. The replay has 600 arm/seed/budget results; seed 0 reproduces all six prospective banks, routes and metrics. Tables below keep these means separate from the prospective realization. Paired differences use the already frozen 5000 paired seed-bootstrap resamples, bootstrap seed 2026091903, with pointwise 95% percentile intervals and no multiplicity adjustment. These quantify acquisition-seed mean stability conditional on one realized table, not execution-noise, population, or simultaneous uncertainty.

Table 42: **Frozen-table replay means over 100 acquisition seeds.** Every row uses exactly $m$ unique edges. Means condition on the completed observed table; seed 0 reproduces the prospective decisions.

| $m$ | Arm | Routed regret | Bank regret | RMSE | Overlap | Agreement |
|---|---|---|---|---|---|---|
| 200 | $M$ | 0.203203 | 0.027684 | 0.437632 | 0.616667 | 0.266400 |
| 200 | $U$ | 0.220449 | 0.034369 | 0.473382 | 0.710000 | 0.304000 |
| 200 | $G$ | 0.188634 | 0.027810 | 0.436654 | 0.593333 | 0.276400 |
| 280 | $M$ | 0.172516 | 0.025248 | 0.415913 | 0.613333 | 0.297900 |
| 280 | $U$ | 0.188827 | 0.023684 | 0.431975 | 0.736667 | 0.334300 |
| 280 | $G$ | 0.141710 | 0.010461 | 0.411276 | 0.666667 | 0.333900 |

Table 43: $M - U$**: preregistered primary comparison.** Paired mean differences with pointwise 95% seed-bootstrap intervals. Negative favors the left arm for regret and RMSE; positive favors it for overlap and agreement.

| $m$ | Metric | Mean difference | Pointwise 95% interval |
|---|---|---|---|
| 200 | Routed regret | $-0.017245$ | $[-0.028662, -0.006113]$ |
| 200 | Bank regret | $-0.006685$ | $[-0.016890, +0.003082]$ |
| 200 | RMSE | $-0.035750$ | $[-0.040634, -0.030916]$ |
| 200 | Bank overlap | $-0.093333$ | $[-0.143417, -0.040000]$ |
| 200 | Route agreement | $-0.037600$ | $[-0.060902, -0.013200]$ |
| 280 | Routed regret | $-0.016311$ | $[-0.025003, -0.008125]$ |
| 280 | Bank regret | $+0.001563$ | $[-0.005337, +0.007706]$ |
| 280 | RMSE | $-0.016062$ | $[-0.018085, -0.014002]$ |
| 280 | Bank overlap | $-0.123333$ | $[-0.166667, -0.080000]$ |
| 280 | Route agreement | $-0.036400$ | $[-0.053600, -0.018200]$ |

Table 44: $G - U$**: unchanged-method secondary replication.** Paired mean differences with pointwise 95% seed-bootstrap intervals. Negative favors the left arm for regret and RMSE; positive favors it for overlap and agreement.

| $m$ | Metric | Mean difference | Pointwise 95% interval |
|---|---|---|---|
| 200 | Routed regret | $-0.031814$ | $[-0.041989, -0.021652]$ |
| 200 | Bank regret | $-0.006559$ | $[-0.016256, +0.002294]$ |
| 200 | RMSE | $-0.036728$ | $[-0.041612, -0.031890]$ |
| 200 | Bank overlap | $-0.116667$ | $[-0.166667, -0.070000]$ |
| 200 | Route agreement | $-0.027600$ | $[-0.049100, -0.006000]$ |
| 280 | Routed regret | $-0.047118$ | $[-0.054429, -0.040587]$ |
| 280 | Bank regret | $-0.013223$ | $[-0.020327, -0.007488]$ |
| 280 | RMSE | $-0.020698$ | $[-0.022534, -0.018860]$ |
| 280 | Bank overlap | $-0.070000$ | $[-0.103333, -0.036667]$ |
| 280 | Route agreement | $-0.000400$ | $[-0.014700, +0.014000]$ |

Table 45: $M - G$**: exploration-floor secondary comparison.** Paired mean differences with pointwise 95% seed-bootstrap intervals. Negative favors the left arm for regret and RMSE; positive favors it for overlap and agreement.

| $m$ | Metric | Mean difference | Pointwise 95% interval |
|---|---|---|---|
| 200 | Routed regret | $+0.014569$ | $[+0.006960, +0.022481]$ |
| 200 | Bank regret | $-0.000126$ | $[-0.005194, +0.005183]$ |
| 200 | RMSE | $+0.000978$ | $[-0.001268, +0.003187]$ |
| 200 | Bank overlap | $+0.023333$ | $[-0.023333, +0.073333]$ |
| 200 | Route agreement | $-0.010000$ | $[-0.030400, +0.010600]$ |
| 280 | Routed regret | $+0.030807$ | $[+0.025769, +0.035930]$ |
| 280 | Bank regret | $+0.014786$ | $[+0.011679, +0.017924]$ |
| 280 | RMSE | $+0.004637$ | $[+0.003803, +0.005465]$ |
| 280 | Bank overlap | $-0.053333$ | $[-0.083333, -0.026667]$ |
| 280 | Route agreement | $-0.036000$ | $[-0.048300, -0.023900]$ |

Both $M - U$ routed-regret intervals lie below zero, and $G - U$ independently replicates the direction seen on APPS. Both $M - G$ intervals lie above zero, rejecting an improvement claim for the exploration floor. This does not imply a win for every seed: $M$ has lower routed regret than $U$ in 60% and 67% of matched seeds at $m = 200/280$; $G$ does so in 74% and 92%. The secondary outcomes are mixed. Both $M$ and $G$ improve RMSE relative to $U$, but $M$ has lower bank overlap and route agreement at both budgets, and neither of its bank-regret intervals excludes zero. At $m = 280$, $G$ also improves bank regret relative to $U$ while its route-agreement interval includes zero.

**Representation and acquisition diagnostics.** The full-table projection RMSE is $0.387952$, with $R^2 = 0.149229$. These residuals concern the realized table and do not separately identify execution noise or misspecification of expected utilities. All replay fits have rank 68. The original design has 734 floating-point-positive weights, but only 288 meaningful edges; the excluded tail mass is $3.814931855039833 \times 10^{-30}$. The frozen solver reached its 4000-iteration cap without meeting the requested $\kappa - 1$ tolerance $10^{-7}$. Its nominal geometry alone does not certify either without-replacement run. Corollary 4 would require the actual Gram matrix and valid model/noise assumptions, which are not established for this replay target.

Table 46: **The exploration floor broadens nominal support but does not improve routed regret.** Meaningful support is $\pi_e > 10^{-10}$; effective edge support is $1/\sum_e \pi_e^2$. These nominal-weight diagnostics do not certify the WOR sampler.

| Design | Edges | Tasks | Effective support | $\kappa$ | Top 10/20/30 task mass |
|---|---|---|---|---|---|
| Original $G$-opt | 288 | 36 | 181.139278 | 1.000033 | 51.8531% / 83.8723% / 98.6200% |
| $M$ | 800 | 100 | 212.350420 | 1.075528 | 47.6678% / 77.4850% / 91.7580% |
| $U$ | 800 | 100 | 800.000000 | 3.960740 | 10.0000% / 20.0000% / 30.0000% |

Table 47: **Equal unique-edge budgets do not imply equal task coverage.** Task counts are out of 100; support edges refer to the original meaningful $G$-opt support. The minimum fit rank is over all 100 replay seeds.

| $m$ | Arm | Prospective tasks | Mean replay tasks | Prospective support edges | Min. rank |
|---|---|---|---|---|---|
| 200 | $M$ | 57 | 53.79 | 177 | 68 |
| 200 | $U$ | 92 | 90.49 | 73 | 68 |
| 200 | $G$ | 34 | 33.51 | 200 | 68 |
| 280 | $M$ | 75 | 72.80 | 226 | 68 |
| 280 | $U$ | 97 | 96.97 | 98 | 68 |
| 280 | $G$ | 36 | 36.00 | 280 | 68 |

Original $G$ uses 200/288 and 280/288 meaningful support edges at the two budgets. Mixture support spans all 800 edges and all 100 tasks, but its actual visitation is lower than Uniform-WOR's. The support restriction and weights together define the tested allocation; the study is not a causal decomposition of coverage. Saved per-seed results retain sampled rank, visitation, support use, selected banks and all routes.

Table 48: **CodeContests execution accounting.** Unique edges count committed workflow observations. Recorded model calls are used in the post-hoc cost-sensitivity analysis.

| Phase | Unique edges | Recorded model calls |
|---|---|---|
| Sparse | 492 | 859 |
| Dense | 308 | 541 |
| Total | 800 | 1400 |

**Execution accounting and prospective chronology.** All 492 sparse-union edges were completed before the six bank-and-route decisions were frozen. The remaining 308 edges were evaluated only afterward, completing the 800-edge observed table. Each edge contributed one committed score. One interrupted, uncommitted request was retried under the preregistered same-seed transport policy; no duplicate committed observation was used.

**Frozen artifacts and validation.** The artifact directory `results/codecontests/` preserves prospective results, replay means, paired comparisons, and seed-level outputs. `data/codecontests/` contains the frozen inputs and anonymized protocol records, including the frozen sparse decisions and execution chronology. Stored validation verified frozen hashes, the sparse-first chronology, and exact seed-0 agreement.

### B.18 Post-hoc CodeContests robustness

These are post-hoc robustness analyses on the same observed table. The original $\alpha = 0$ $M - U$ primary comparison and $G$ secondary arm retain their preregistered roles. Task-balanced-WOR ($T$) follows the recovered APPS rule: randomize task order each round and use a separate workflow permutation per task, without repeated edges. Its new masks visit all tasks twice at $m = 200$ and twice or three times at $m = 280$. The 100 matched seed indices reuse the original $G/M/U$ masks, frozen rank-68 features, unregularized OLS, exhaustive curation, and tie rules. Replay reproduces all 600 original cells exactly; no new workflow was executed. Both analyses use the saved 5000 paired seed-bootstrap resamples (seed 2026091903) and pointwise 95% percentile intervals, without multiplicity adjustment. These intervals condition on one observed table, not fresh executions or a task population.

Table 49: **Post-hoc Task-balanced comparisons on the CodeContests observed table.** Paired differences over 100 seeds; negative favors the left method. $T$'s mean routed regret is $0.23624/0.19472$ and bank regret is $0.04622/0.02023$ at $m = 200/280$.

| $m$ | Contrast | Routed-regret difference [95% CI] | Bank-regret difference [95% CI] |
|---|---|---|---|
| 200 | $G - T$ | $-0.04761$ $[-0.05933, -0.03623]$ | $-0.01841$ $[-0.03123, -0.00613]$ |
| 200 | $M - T$ | $-0.03304$ $[-0.04582, -0.02043]$ | $-0.01854$ $[-0.03130, -0.00608]$ |
| 200 | $U - T$ | $-0.01579$ $[-0.02971, -0.00138]$ | $-0.01185$ $[-0.02751, +0.00377]$ |
| 280 | $G - T$ | $-0.05301$ $[-0.06035, -0.04591]$ | $-0.00977$ $[-0.01582, -0.00475]$ |
| 280 | $M - T$ | $-0.02220$ $[-0.03054, -0.01378]$ | $+0.00502$ $[-0.00215, +0.01140]$ |
| 280 | $U - T$ | $-0.00589$ $[-0.01628, +0.00435]$ | $+0.00346$ $[-0.00516, +0.01240]$ |

$G$ and $M$ have lower routed regret than $T$ at both budgets (Table 49). At $m = 280$, the $U - T$ routed-regret contrast and $M - T$ bank-regret contrast remain unresolved. $T$ visits every task and has higher mean bank overlap than $G/M$ at that budget, yet worse routed regret; this does not isolate a coverage effect.

**Downstream call-count sensitivity.** Let $c_k$ be the mean recorded completed logical calls for workflow $k$ across its 100 executions. The unnormalized vector is $c = [1, 2, 2, 3, 1, 2, 1, 2]$ in workflow-index order; counts are constant across tasks within each workflow. Set $S = 100^{-1} \sum_i (\max_k U_{ik}^{\text{obs}} - \min_k U_{ik}^{\text{obs}}) = 0.6385884828101276$ and $\Delta c = 2$. The grid $\alpha = fS/\Delta c$, $f \in \{0, 0.1, 0.5, 1\}$, was fixed before these method comparisons, after the original outcomes existed. Alpha has units of utility per logical call; exact values are archived and Table 50 rounds

them. At each alpha, subtract $\alpha c_k$ from both the observed and completed tables and recompute their optimal banks and routes, keeping acquisition masks and OLS completions fixed. Regret uses the corresponding cost-adjusted dense optimum. Recorded realized calls equal the fixed workflow costs here.

Table 50: **Post-hoc downstream call-count sensitivity.** Entries show mean routed regret / realized logical calls per routed task over 100 matched seeds. Banks and routes change with $\alpha$; masks and OLS completions do not. Mean rankings alone do not establish pairwise superiority.

| $m$ | $\alpha$ | $G$ | $M$ | $U$ | $T$ |
|---|---|---|---|---|---|
| 200 | 0 | 0.1886 / 2.054 | 0.2032 / 2.021 | 0.2204 / 2.124 | 0.2362 / 2.120 |
| 200 | 0.03193 | 0.1886 / 1.956 | 0.2049 / 1.894 | 0.2222 / 2.035 | 0.2353 / 2.051 |
| 200 | 0.15965 | 0.2397 / 1.608 | 0.2481 / 1.519 | 0.2713 / 1.705 | 0.2740 / 1.712 |
| 200 | 0.31929 | 0.2417 / 1.264 | 0.2369 / 1.216 | 0.2688 / 1.354 | 0.2644 / 1.351 |
| 280 | 0 | 0.1417 / 2.119 | 0.1725 / 2.051 | 0.1888 / 2.179 | 0.1947 / 2.128 |
| 280 | 0.03193 | 0.1415 / 2.060 | 0.1768 / 1.933 | 0.1894 / 2.084 | 0.1956 / 2.054 |
| 280 | 0.15965 | 0.2098 / 1.658 | 0.2241 / 1.551 | 0.2366 / 1.655 | 0.2390 / 1.670 |
| 280 | 0.31929 | 0.2273 / 1.215 | 0.2226 / 1.187 | 0.2345 / 1.267 | 0.2313 / 1.281 |

Table 51: **Post-hoc paired routed-regret contrasts at nonzero call penalties.** Mean differences and pointwise 95% paired intervals; negative favors $M$.

| $m$ | $\alpha$ | $M - U$ [95% CI] | $M - T$ [95% CI] |
|---|---|---|---|
| 200 | 0.03193 | $-0.017275$ $[-0.028777, -0.006434]$ | $-0.030367$ $[-0.042622, -0.018556]$ |
| 200 | 0.15965 | $-0.023249$ $[-0.034961, -0.011549]$ | $-0.025858$ $[-0.037905, -0.013366]$ |
| 200 | 0.31929 | $-0.031936$ $[-0.044127, -0.020023]$ | $-0.027476$ $[-0.038219, -0.016550]$ |
| 280 | 0.03193 | $-0.012624$ $[-0.021704, -0.003969]$ | $-0.018807$ $[-0.027449, -0.009975]$ |
| 280 | 0.15965 | $-0.012517$ $[-0.021189, -0.003718]$ | $-0.014975$ $[-0.022763, -0.007071]$ |
| 280 | 0.31929 | $-0.011932$ $[-0.020506, -0.003435]$ | $-0.008746$ $[-0.016083, -0.001570]$ |

$M$'s routed-regret differences from $U$ and $T$ have intervals below zero throughout the grid (Table 51). At the highest alpha and $m = 280$, $G - T$ is $-0.004075$ $[-0.010354, +0.002309]$ and $G - U$ is $-0.007261$ $[-0.015017, +0.000303]$. $M$ has the lowest mean regret there, but the high-alpha $G - M$ intervals include zero at both budgets. Logical calls across different models do not measure equal GPU work, latency, or money; this sensitivity does not measure acquisition or total-project cost.

The artifact directory `results/codecontests_robustness/` retains metrics, seed differences and tie counts, routes, bank identities, and utility-only and adjusted scores; `data/codecontests/` contains the masks and recorded logical-call costs. No compatible saved CodeContests stronger-estimator predictions were available, so estimator replacement remains untested.

## C Fail-Closed Audit Checklist

The companion `tools/realized_design_diagnostic.py` scans the full frozen population using the actual retained rows and a Cholesky solve. It records rank and conditioning tolerances, abstains on unreliable geometry, and reports a radius only with an explicitly declared valid noise scale and model assumptions. Its floating-point diagnostics do not issue a statistical certificate or reclassify historical runs. Deterministic algebra checks cover full rank, lost span, repeated rows, actual-count accounting, and zero noise.

Table 52: **Certification requires an applicable statistical and geometric contract.** The canonical and conditional paths have distinct design gates; if neither applies, the run remains empirical-only.

| Check | Required condition |
|---|---|
| Finite population | Queries and workflow IDs are fixed before design and equal the downstream decision population. |
| Feature provenance | Raw features, normalization, and rank-span coordinates are frozen before target utilities; separately collected DEV information is declared. |
| Model validity | Exact frozen-span realizability must hold; alternatively use Proposition 15 with a valid uniform residual bound and its larger radius. |
| Noise validity | A valid $\sigma$ and the relevant conditional independent centered sub-Gaussian model are required; replayed copies of one noisy label are not fresh noise. |
| Sample sufficiency | The canonical path requires (2); the conditional path uses actual retained rows and full rank without that threshold. |
| Gram rank | The actual sampled $X$ has population rank $d$; use $V = X^\top X$. Rank failure requires abstention, not a pseudoinverse certificate. |
| Design certificate | Scan all population edges in frozen OLS coordinates: verify $\Sigma(\pi)$ and $\kappa$ for the canonical path, or $L(V)$ for the conditional path. |
| Sampling integrity | Canonical: i.i.d. with replacement and fresh repeated executions. Conditional: outcome-independent acquisition/stopping, including deduplication. Geometry-only acceptance is allowed. |
| Estimator semantics | Completion uses unregularized minimum-norm OLS without clipping, ridge regularization, or label-dependent feature changes. |
| Downstream reuse | Costs, weights, $B$, objective, optimizer certificate $\eta$, workflow IDs, routing, and tie-breaking are identical in dense and sparse pipelines. |
| Target and confidence | State mean versus observed-table target, residual bound if used, and pre-outcome $\delta$. Do not select the smaller of two nominal-$\delta$ radii without a valid joint guarantee. |