SCHEMA2BENCH: EVIDENCE-GROUNDED CROSS-DOMAIN TEXT-TO-SQL BENCHMARK MIGRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Schema2Bench formulates cross-domain Text-to-SQL benchmark migration as instrument construction under information, candidate-support and budget constraints. It combines evidence-grounded candidate synthesis with response-blind selection. A support certificate identifies structural deficits, and a sharp release-stability bound budgets changes to a fixed release. On a 2,806-candidate pool with a 2,000-item release budget, cost-aware bounded exchange reduces mean joint test-effect fidelity (TEF) loss by 5.67% versus quota random on a fixed seven-model panel across ten shared-pool seeds. Deficit-guided acquisition reaches the full-pool attainable difficulty discrepancy with 200 additions, versus 400 for source-proportion and 806 for uniform acquisition at the tested budgets. A stratified audit by two control-qualified model families jointly supports 270/300 fixed-release items (weighted estimate 90.07%). In a separate equal-budget construction study, the Schema2Bench deterministic refill variant yields 383 executable, jointly supported items versus 363 for a template baseline, including 109 versus 17 JOIN queries. An exploratory operator-matched BIRD comparison yields 3,835 pairs per method across eight eligible domains. With nine evaluators, domain-mean pairwise-gap error is 0.1724 for the Schema2Bench variant versus 0.2172 for templates (improvement 0.0449; item-bootstrap interval [0.0334, 0.0522]); seven domains improve. The prespecified ten-model measurement gate fails (five of eight domains improve), with Phi-4 producing empty output on 8,779/8,788 prompts. These results concern tested deterministic variants and do not establish general dominance.

1 INTRODUCTION

Text-to-SQL benchmarks bind questions and reference programs to databases. Spider and BIRD broaden schema realism but do not specify how to construct a comparable instrument for a new database (Yu et al., 2018; Li et al., 2023). Migration requires grounded tasks and workload control; reproducing evaluation behavior additionally requires difficulty, accuracy and disagreement fidelity.

We study *response-blind benchmark migration*. A constructor observes schemas, database contents, source SQL structure, and construction evidence, but not the target responses of the evaluator panel. Releases and hashes are fixed before model responses are joined, separating benchmark design from evaluation. This information boundary enables a direct test of whether observable workload control also preserves evaluation behavior.

Figures 1 and 2 show the problem and framework. Two challenges determine the method. **Credible target references** require schema-grounded SQL and questions supported by execution and consistency evidence. **Workload and test-effect mismatch** persists even when individual queries execute: generation can overproduce easy or repetitive tasks. Schema2Bench couples *Evidence-Grounded Candidate Synthesis* with *Response-Blind Benchmark Adaptation*, which selects a fixed-budget target set against the source SQL profile. Construction evidence and downstream evaluation remain separate.

Our contribution is an information-constrained formulation of benchmark instrument migration, implemented by a construction and selection framework. A quota-feasible support certificate identifies candidate deficits; a sharp release-stability bound induces domain-specific exchange costs. Cost-aware selection improves mean joint TEF over random and greedy on the fixed panel, while offline

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.

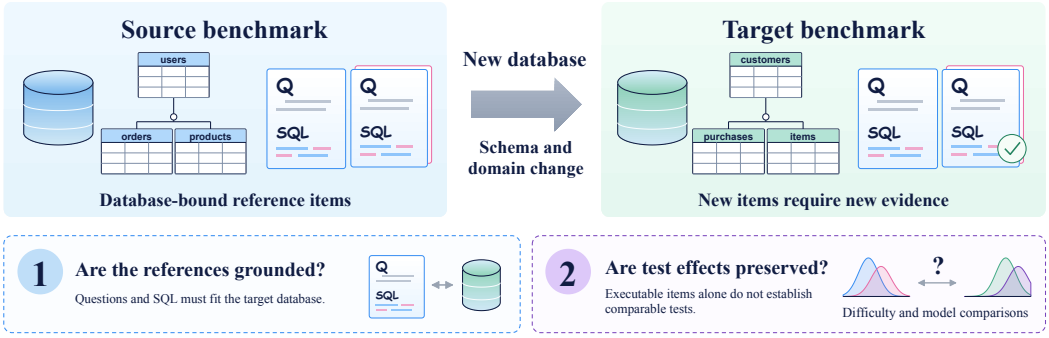


Figure 1: Problem setting: migrating a benchmark to a new database requires new grounded question–SQL references and an assessment of set-level test effects. Executability alone does not establish preserved difficulty or model comparisons. Schemas, items and distribution curves are schematic, not measured results.

deficit-guided acquisition reaches the full-pool difficulty optimum and reproduces the full-reserve release with fewer additions at the tested budgets. A new generation study tests whether deficit-derived demand also acquires useful candidates under equal medium/hard request budgets. Initial blinded model review favors deficit demand with two generators; a stricter precision sensitivity retains a Qwen yield gain but a Llama tie. Ablations isolate repair and duplicate control. A separate stratified audit of the fixed release uses hidden controls to qualify strong-model reviewers and supplies release-level semantic evidence. A further reference-based experiment evaluates whether newly constructed items preserve differences among Text-to-SQL models, connecting candidate quality to measurement behavior.

2 Related work

Text-to-SQL benchmarks broaden schema realism, while controlled diagnostics and evidence-centered design distinguish task construction from measurement (Yu et al., 2018; Li et al., 2023; Lei et al., 2025; Chang et al., 2023; Liu et al., 2024). Cross-benchmark ranking differences motivate evaluating fidelity directly (Sun et al., 2023; Kovatchev & Lease, 2024).

RingSQL uses templates and paraphrasing for synthetic training data (Sterbentz et al., 2026); SING-SQL uses quality-aware synthesis for in-domain training and evaluation (Caferoğlu et al., 2025). Here the objective is source-profile control and measurement fidelity. Our method-inspired baselines are harness adaptations, not official reproductions or trained-model comparisons. OmniSQL/SynSQL-2.5M and EvolSQL scale synthetic training data, the latter through structure-aware evolution with refinement and deduplication (Li et al., 2025; Pan et al., 2026). These are relevant synthesis comparators; we do not claim superior generation quality without matched evaluations. NL2SQL-BUGs evaluates semantic error detection (Liu et al., 2025); we include a separate external-label diagnostic of automatic reviewers in Appendix N.

Data selection, distillation and efficient benchmarking preserve declared properties or exploit response evidence (Xie et al., 2023; Albalak et al., 2024; Wang et al., 2018; Perlitz et al., 2024; Maia Polo et al., 2024; Gupta et al., 2025; Saranathan et al., 2025). Here new database-bound items must first be constructed, with target evaluator responses unavailable to selection. The transferred object is a measurement instrument, rather than a predictor (Ben-David et al., 2010). Benefit–cost greedy selection is classical (Khuller et al., 1999); here the cost is induced by a benchmark measurement-stability bound.

Our formulation connects to herding and minimax design (Bach et al., 2012; Dereziński et al., 2019), but targets a joint evaluator profile on new database-bound items. Appendix A compares coresets, robust optimization and their information regimes.

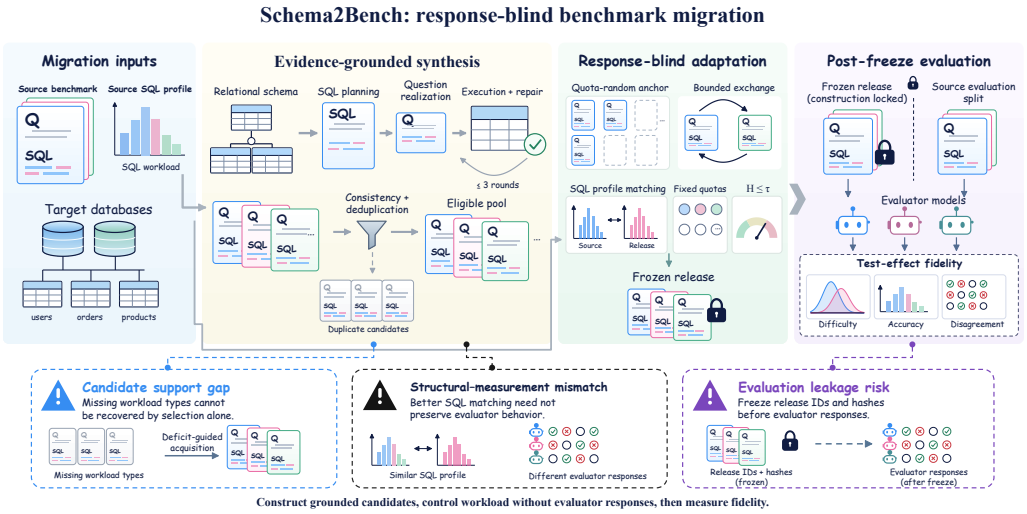


Figure 2: Schema2Bench framework: candidate synthesis, response-blind structural adaptation, and frozen-release evaluation. SQL precedes question realization, execution and bounded repair. The support certificate motivates deficit-guided acquisition, while bounded exchange limits changes from a quota-random anchor. Evaluator responses enter only after release freezing. Schemas, profiles and examples are schematic; structural matching does not establish test-effect fidelity.

3 Method: Cross-Domain Benchmark Migration

3.1 Module 1: Evidence-Grounded Candidate Synthesis

Given source question–SQL pairs B_s and target databases D_t, Module 1 constructs an eligible pool $\mathcal{C}$; Module 2 chooses a fixed-budget instrument. Let $\mathcal{I}$ contain source SQL structure, target schemas, bounded database evidence, candidate provenance, quotas and a response-independent seed. Excluding evaluator responses, the design satisfies

$$z = \pi(\mathcal{I}), \qquad z \in \mathcal{F} = \{z \in \{0,1\}^{|\mathcal{C}|} : Az = q,\ \mathbf{1}^\top z = N\}, \tag{1}$$

where A encodes disjoint domain–provenance strata. Each retained item stores its question, SQL, schema grounding, execution evidence and repair trace. The migrated object is a workload and measurement profile; target entities and domain meaning may change.

Eligibility requires schema grounding, bounded read-only execution, a nonblank answer, automatic consistency, and exact-duplicate control. Each decision retains its execution and repair evidence (Appendix H.1). Semantic examples and audit scope appear in Appendix H.1.

The planner uses tables, keys, values and source operators to specify SQL, followed by LLM question realization, read-only execution and at most three repair rounds. Repair synchronizes LIMIT wording; consistency screening and canonical-signature memory govern admission. These are programmatic roles with an LLM realizer, not separate reasoning agents. Controlled and RingSQL-style variants share SQL templates but use different question prompts; Controlled adds table names. Its realization hook accepts unconditionally, with screening at admission. SING-SQL-style uses schema-partitioned pair generation. Functional comparisons concern these harness variants (Appendix H.1).

3.2 Module 2: Response-Blind Benchmark Adaptation

The structural selector compares source and release distributions over SQL operators, deterministic difficulty, query shapes, join count, and SQL length. Quota random is the stable anchor; greedy and local search modify membership within the same strata. No selector receives response files or response-cache paths. Selected IDs, configuration, and hashes are frozen before evaluation.

Let $\phi(y) = (\phi_{\text{op}}, \phi_{\text{diff}}, \phi_{\text{shape}}, j, \ell)$ denote the deterministic SQL representation. The five equal-coefficient blocks are operator-event total variation, three-level SQL-only difficulty total variation,

six query-shape deviations, relative mean join-count error, and relative mean SQL-length error. Every selector returns $N = 2000$ rows under identical domain and provenance quotas. Quota random is the non-optimized anchor; greedy minimizes marginal structural loss; local search performs quota-preserving swaps from the greedy initialization.

For a selected set $\mathcal{Z}$, the prespecified objective is

$$L_{\text{struct}}(\mathcal{Z}) = \frac{1}{5}\Bigg(\frac{\|\widehat{P}_o - P_o^*\|_1}{2} + \frac{\|\widehat{P}_d - P_d^*\|_1}{2} + \frac{\|\widehat{P}_h - P_h^*\|_1}{6} + \frac{|\widehat{j} - j^*|}{\max(j^*, 10^{-12})} + \frac{|\widehat{\ell} - \ell^*|}{\ell^*}\Bigg), \tag{2}$$

subject to $|\mathcal{Z}| = N$ and $|\mathcal{Z} \cap \mathcal{C}_{d,g}| = q_{d,g}$ for every domain–provenance stratum. The optimizer may change which candidates represent a stratum but not how many it contributes. Four additional selectors remove one structural block at a time; they are sensitivity variants, not alternative primary outcomes. Operator and difficulty terms are total variations in $[0, 1]$. The six shapes (filter, join, grouping, aggregation, ordering and LIMIT) are overlapping occurrence rates, not a simplex; their mean absolute deviation also lies in $[0, 1]$. Relative join/length errors need not be bounded by one, so equal coefficients do not imply equal scales. We retain the frozen scaling. Source means are positive here. Difficulty uses one SQL-only score for both source and target: 0–1 is easy, 2–3 medium, and ≥ 4 hard; it does not merge Spider's official four labels (Appendix M).

The complete construction–selection algorithm is in Appendix B.

3.3 Test-effect fidelity

For $M \geq 2$ evaluators, each item i has a fixed joint correctness law K_i on $Y \in \{0, 1\}^M$. Let g_j enumerate cumulative correct-count indicators $\mathbf{1}\{\sum_m Y_m \leq k\}$ ($k = 0, \ldots, M-1$), accuracies Y_m, and pairwise indicators $|Y_a - Y_b|$. With $r_{i,j} = \mathbb{E}_{K_i} g_j(Y)$, define

$$\Psi_{z,j} = \sum_{i \in \mathcal{C}} \alpha_{z,i}^{c(j)} r_{i,j}, \qquad L_{\text{TEF}}(z) = \sum_j w_j |\Psi_{z,j} - \Psi_{s,j}|. \tag{3}$$

Domain-macro weights pool domain moments before taking absolute differences. Here D is the number of target domains, $N_d > 0$ their fixed release counts, and $c(j) = A$ for accuracy and R otherwise. Accuracy uses domain-macro weights $\alpha_{z,i}^A = z_i/(DN_{d_i})$; difficulty and disagreement use item-micro weights $\alpha_{z,i}^R = z_i/N$. Coordinate weights are $1/M$, $1/M$ and $1/(2\binom{M}{2})$, respectively. Thus $L_{\text{TEF}} = D_{\text{diff}} + D_{\text{acc}} + 0.5D_{\text{dis}}$, with normalized correct-count Wasserstein distance, model-accuracy MAE and pairwise disagreement MAE. Accuracy gives each database equal influence; difficulty and disagreement summarize a uniformly sampled release item. This is a declared mixed estimand, not a claim that the weighting is uniquely optimal. Source moments Ψ_s use the declared evaluation split and weights. Frozen binary responses instantiate point-mass kernels. Item-level laws avoid assuming that a coarse type's conditional mean is unchanged by selection.

3.4 Anchor-constrained instrument design

For a response-law class $\mathfrak{K}(\mathcal{I})$ specified without held-out responses and a frozen quota-random anchor z_0, the extended design problem is

$$\min_{z \in \mathcal{F}} L_{\text{struct}}(z) \quad \text{s.t.} \quad \sup_{K \in \mathfrak{K}(\mathcal{I})} [L_{\text{TEF}}(z; K) - L_{\text{TEF}}(z_0; K)] \leq \tau. \tag{4}$$

For the distribution-free safeguard, $\mathfrak{K}(\mathcal{I})$ contains all collections of fixed item-level joint probability laws on $\{0, 1\}^M$; selection cannot change an item's law. Deterministic cached responses are point-mass members. Here K includes both source and target laws. The same laws evaluate both releases; $\tau \geq 0$ is a prespecified use requirement. The anchor remains feasible. Historical local search optimizes the structural objective; bounded exchange implements a sufficient version of the additional constraint.

Proposition 1 (Sharp release stability)**.** *Fix the panel, binary item responses, source moments, domain counts N_d and normalized domain weights w_d^A, w_d^R for the accuracy and remaining blocks of Eq.* (3)*.*

For r_d within-domain replacements,

$$|L_{\mathrm{TEF}}(z) - L_{\mathrm{TEF}}(z_0)| \le H(z, z_0) := \sum_d (w_d^A + w_d^R) \frac{r_d}{N_d}. \tag{5}$$

The bound is attainable. Under the current mixed weighting, $H = r/N + D^{-1} \sum_d r_d/N_d$, where $r = \sum_d r_d$.

Thus $H \le \tau$ is a response-free sufficient condition for the constraint in Eq. (4), over fixed item laws with these weights. It yields an implementable safeguard for structural swaps, without assuming invariant coarse-type responses. The sharpness construction and proof are in Appendix F; historical local search did not enforce this safeguard.

Cost-aware bounded exchange. For a feasible within-stratum swap e in domain $d(e)$, let $\Delta_e = L_{\mathrm{struct}}(z) - L_{\mathrm{struct}}(z \oplus e)$ and $c_d = N^{-1} + (DN_d)^{-1}$. We choose

$$e^* \in \arg \max_{e:\, \Delta_e > 0,\ H + c_{d(e)} \le \tau} \frac{\Delta_e}{c_{d(e)}}. \tag{6}$$

Swaps replace original-anchor items by original-complement items, preserve stratum quotas and use deterministic ID ties. This allocates the release perturbation budget without evaluator responses; absolute-gain exchange maximizes Δ_e instead. Experiments fix $\tau = 0.01$ and at most 25 swaps as a computational setting (Appendix E).

4 Experimental protocol

Construction and selection. Spider train supplies 7,000 SQL queries for the structural source profile; its disjoint development split supplies 1,034 source-response items only for post-selection evaluation. The seven target databases are Chinook, IMDb, Instacart, SFDFFD, TPC-DS, TPC-H, and USAccidents. The aligned experiment begins with 3,360 candidates, applies two response-independent screens, and retains 2,806 before each selector chooses 2,000 under identical quotas. The primary local-search configuration and seed 20260811 were declared before response evaluation; ten seeds measure path stability over one shared universe. Here cross-domain denotes construction across target databases, not demonstrated generalization to a population of independently sampled schemas.

Construction costs, fixed domain quotas, and optimizer settings are retained in Appendix H.1.

Panels. Old-13 development, fresh-four confirmation, a separate seven-model selector comparison, and a two-model radius reserve retain their frozen protocols and distinct information regimes (Appendix H).

Uncertainty. Schema clusters are the replication units in the expert-reference study. Shared-pool seeds measure path stability. Historical single-model and joint-panel estimands retain separate intervals; the radius study clusters six method families. Frozen criteria, outcomes and full inference details appear in Appendices H.4 and H.

Fresh-construction measurement. We compare deterministic refill with templates under execution and dual-model semantic review. The earlier two- and four-domain protocols appear in Appendix N. The later exploratory BIRD expansion pairs candidates by SQL operator signature before review: of 11 attempted domains, eight meet the frozen eligibility gate, yielding 3,835 pairs per method and 1,118 public reference items. We score both arms and the reference with one extraction and execution policy, measuring domain-mean error in signed pairwise model-accuracy gaps. The ten-model panel is primary; a prespecified nine-model panel excludes Phi-4. This expansion reuses the frozen raw pool after an earlier STOP and is not an independent-domain confirmation.

5 Release optimization and construction results

Cost-aware bounded exchange. Table 1 compares the same ten anchors and quotas. In this exploratory extension, cost-aware exchange lowers mean TEF by 5.67% versus random (10/10 pairs)

Table 1: Release optimization: ten-seed means on the same pool and panel. Joint TEF is mean ± seed SD. Both bounded rules use $H \leq 0.01$; lower is better.

Selector	Structural loss	Joint TEF
Quota random	0.269519	0.050071 ± 0.003030
Structural greedy	0.213055	0.048692 ± 0.002120
Unconstrained local	0.189298	0.055945 ± 0.003501
Bounded, absolute gain	0.261163	0.047747 ± 0.002284
Bounded, cost-aware	0.260790	0.047229 ± 0.002392

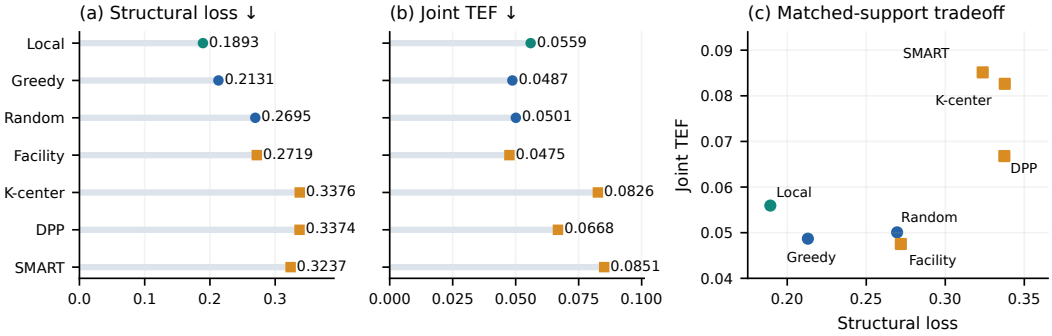


Figure 3: Original matched-support selectors: (a) five-block structural loss, (b) joint TEF, and (c) their tradeoff. All methods use the same 2,806 candidates, 2,000-item budget and domain/provenance quotas. Circles are ten historical seed means; squares are single fixed-seed alternatives. Lower is better on both axes; local has the strongest structural match and facility the lowest TEF among these selectors. Bounded extensions appear in Table 1.

and 3.00% versus greedy (7/10), while improving structure over random. The absolute cost-aware minus random difference is -0.002841 (paired-seed 95% bootstrap interval $[-0.003379, -0.002337]$); against greedy it is -0.001463 with interval $[-0.003170, 0.000032]$. These post-hoc intervals describe fixed-pool path stability, not independent schema replication; the greedy comparison remains uncertain. Against absolute-gain exchange it lowers TEF in 8/10 pairs, with no higher actual H in any pair. Greedy retains lower structural loss. An exact-cost random-swap control has TEF 0.049857 (cost-aware wins 10/10); restricting absolute-gain swaps to the largest quota domain reproduces the cost-aware releases. Protocol and decision-preservation checks appear in Appendix E.

Automatic screens retain 2,806 executable, nonblank candidates; the release has no within-domain exact SQL or question duplicates under the canonicalizer.

Figure 3 separates structural and response outcomes. Its supplementary Facility, K-center, DPP and SMART-style methods represent coverage, geometric diversity, determinant diversity and a benchmark-selection adaptation, respectively; their fixed-run scope is in Appendix H.3. Quota random has mean five-block distance 0.2695, greedy reaches 0.2131, and local search 0.1893: a 29.76% reduction from random and 11.15% from greedy. All ten local seeds beat paired random and greedy sets on the aggregate objective, with mixed block-level gains. Under the BIRD structural source, local improves 14.73% over random but only 0.00000541 absolute over greedy (Appendix H.3); this is structural sensitivity.

Functional ablations and parameter scope. Appendix Figure 5 summarizes 147 cells (seven methods, seven domains, three seeds; 500 attempts each). Repair raises controlled synthesis yield from 396.00 to 417.90 within-cell unique SQL signatures per 1,000 attempts (5.53% relative): 20 of 21 domain–seed pairs improve and one ties. These are shared-domain comparisons. Memory rejects repeated signatures: removing it nearly doubles raw acceptances but leaves unique yield at 418.19. RingSQL-style and template yields are 431.24 and 420.67. These comparisons identify functional contributions within the construction pipeline; their automatic qualification does not establish semantic accuracy. Across all seven domains, controlled synthesis has lower profile error but higher token cost and lower yield than RingSQL-style: an observed tradeoff, not dominance.

Block, optimizer and radius sensitivity, full costs and question-realization diagnostics appear in Appendices H and H.1.

Fresh SQL quality and measurement fidelity. Under 1,080 attempts per method across six databases, the deterministic Schema2Bench refill variant produces 383 executable items jointly supported by two qualified model reviewers, compared with 363 for the template baseline. The respective pools contain 109 and 17 JOIN queries. This comparison extends the construction ablation from automatic admission to reviewed semantic support and structural coverage.

Figure 4 shows the eight eligible domains under the prespecified nine-model sensitivity panel. Domain-mean error falls from 0.2172 to 0.1724 ($+0.0449$ improvement); seven domains improve, while California schools reverses. The 95% item-bootstrap interval for the item-weighted improvement is $[0.0334, 0.0522]$. Equal-skeleton resampling instead gives $+0.0358$ with interval $[0.0229, 0.0447]$. Under the prespecified ten-model primary panel, improvement is only $+0.0197$ and five of eight domains improve, below the frozen six-domain gate. Phi-4 has empty raw output on 8,779 of 8,788 prompts; we retain its scores and report the nine-model result as sensitivity, not a passed primary result. The earlier four-domain comparison and full attrition appear in Appendix N.

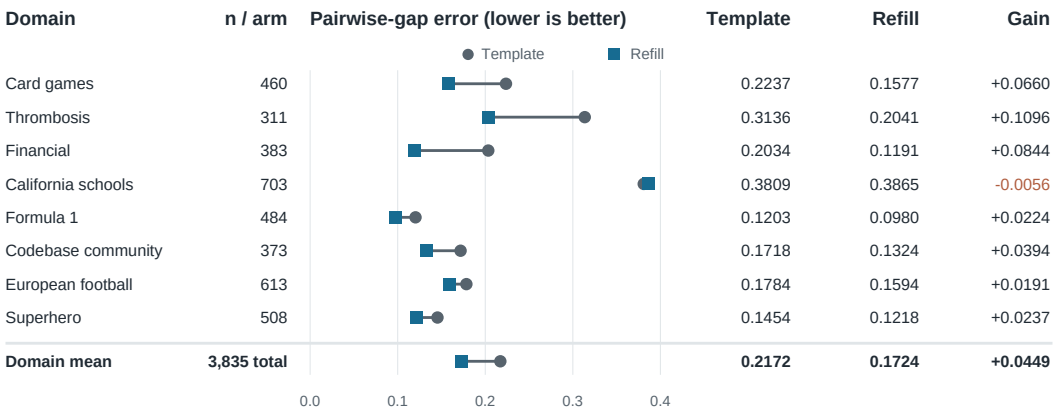


Figure 4: Exploratory eight-domain BIRD comparison with nine evaluators. Each row shows both methods' reference-based pairwise-gap errors, paired synthetic items per method, and Template-minus-Schema2Bench improvement. Lower error is better; the negative California schools row remains in the domain mean.

6 Support certificates and bounded adaptation

Proposition 2 (Quota-feasible support certificate)**.** *For finite structural types T, one fixed weighting channel, a nonempty feasible release set $\mathcal{F}$, and source probability vector p, let $\mathcal{Q} = \mathrm{conv}\{q_z : z \in \mathcal{F}\}$. Then*

$$\delta_{\sup} = \min_{q \in \mathcal{Q}} \mathrm{TV}(p, q) = \max_{h \in [0,1]^{|T|}} \left[h^\top p - \max_{z \in \mathcal{F}} h^\top q_z \right]. \tag{7}$$

Every feasible integer release satisfies $\mathrm{TV}(p, q_z) \geq \delta_{\sup}$.

For $D_T(z) = \mathrm{TV}(p, q_z)$, define the integer optimum $D_T^* = \min_{z \in \mathcal{F}} D_T(z)$. The certificate separates three quantities:

$$D_T(z) = \delta_{\sup} + \underbrace{(D_T^* - \delta_{\sup})}_{\text{integer feasibility gap}} + \underbrace{(D_T(z) - D_T^*)}_{\text{selection headroom}}. \tag{8}$$

All terms are nonnegative. If a feasible integer witness attains the dual value, the middle term vanishes and the remaining selection headroom is exact. A partition oracle supplies the witness (Appendix A). This is an optimality accounting for the declared structural distance, not an additive decomposition or lower bound for TEF.

Application to the historical candidate pool. A post-hoc CPU audit fixes the three SQL difficulty levels, 2,806 candidates and the primary 2,000-item release's quotas. An integer count witness is feasible in every domain–provenance stratum and attains the dual bound exactly:

$$D_T(z_{\text{primary}}) = 0.2725 = \underbrace{0.257}_{D_T^* = \delta_{\sup}} + \underbrace{0.0155}_{\text{remaining headroom}} . \tag{9}$$

The witness $h = (0, 1, 1)$ exposes medium/hard demand of 53.6% against maximum capacity of 27.9% (558 items). Thus selection alone cannot remove the 25.7-point support deficit. The witness also attains domain-equal pooled-TV 0.395546, versus 0.401924 for the release. Other weighting channels and counts appear in Appendix C and Appendix Figure 6; the displayed certificate concerns difficulty. Extending the audit to all five structural blocks gives a valid joint loss lower bound of 0.14523. For example, the minimum feasible LIMIT frequency is 46.1%, versus a source target of 15.81%, identifying a concrete acquisition mismatch beyond difficulty (Appendix L).

From diagnosis to acquisition. With fixed quotas and item-micro weighting, adding b eligible items leaves $D_T^*(\mathcal{C}^+) \geq \max\{0, \delta_{\sup}(\mathcal{C}) - b/N\}$. Thus exact difficulty matching requires at least 514 new candidates: at least 384 medium and 130 hard. This is a necessary support-capacity bound, not a generation guarantee (Appendix D).

Offline candidate acquisition. Starting from each 2,000-item anchor and its 806-item reserve, we compare uniform, source-proportion and largest-difficulty-deficit-first acquisition. With SQL-derived types and fixed quotas, deficit-guided acquisition reaches the full-pool integer optimum 0.257 at 200 additions, versus 400 and 806 for the other policies among tested budgets. Passing the acquired candidates to cost-aware exchange, deficit-guided acquisition at 200 additions reproduces the full-806 release exactly in 10/10 seeds, including TEF 0.047229; the other policies reach this endpoint at 400 and 806 tested additions. On the BIRD structural target, cost-aware and absolute-gain exchange improve structure by 1.358% and 1.070%.

Certificate-guided generation. We next generate new SQL rather than acquire existing reserve items. On seven development databases, each policy receives 24 slots per database (168 total) under matched prompts, scoped schema contexts, execution checks and repair budgets. Source-conditional demand allocates 109/59 medium/hard requests; the certificate deficits induce 126/42. Two blinded model reviewers assess support and unnecessary complexity; the precision stage rechecks only the initially retained medium/hard subset under stricter criteria. These are model judgments, not semantic gold.

Table 2 separates the initial endpoint from the precision sensitivity. Deficit demand improves the initial Deficit versus Source-conditional medium/hard count for Qwen (19/14) and Llama (12/10), so the observed direction repeats under a changed generator. Precision review retains the Qwen gain (12/10) but gives a Llama tie (8/8). At these precision endpoints, deficit uses 17.66% fewer generation tokens per retained item for Qwen and 2.07% fewer for Llama, excluding review and startup. Llama's unspecified arm retains 15 items, more than either targeted arm. Hence this study supports generator-conditioned construction utility, not stable cross-generator yield superiority. The two generators share the same seven databases and are not 14 independent schemas. Protocol, format failures, reviewer agreement and paired uncertainty appear in Appendix J.

Release-level semantic audit. We audit a proportional stratified sample of 300 fixed-release items using scoped schema and execution evidence plus 60 hidden controls. Qualified Qwen and ordinary Hy3 jointly support 270/300 targets, a stratum-weighted 90.07% (95% interval [86.77%, 93.06%]). A same-family Hy3 SQL-agent sensitivity agrees on 290/300 labels. DeepSeek fails the frozen gate. Appendix K gives the protocol and provenance; these are model judgments, not human annotation or item-level gold.

Decision preservation. Across ten fixed releases, every selector retains the source-reference leader and top-three set. Random, greedy and both bounded rules preserve 210/210 model-pair orders; unconstrained local preserves 208/210. Cost-aware exchange has zero source selection regret. Its TEF gain therefore improves profile fidelity without demonstrating a model-choice advantage (Appendix E).

Table 2: Generation endpoints under 168 slots per policy. Counts are nonempty, unique M/H items jointly supported and judged nonredundant by two model reviewers. Precision rechecks the initial retained subset. Cost includes failed generation attempts and repairs, but excludes review and startup.

Generator	Demand	Initial	Precision	Tokens/precision
Qwen	Unspecified	11	8	21,427
	Source-cond.	14	10	18,332
	Deficit	**19**	**12**	**15,095**
Llama	Unspecified	18	15	11,622
	Source-cond.	10	8	22,555
	Deficit	**12**	8	**22,089**

7 Discussion and limitations

Why structure can misalign with measurement. Local search reduces structural loss by 0.080221 relative to random, but raises TEF by 0.005874: difficulty contributes 0.003337, accuracy 0.001291, and weighted disagreement 0.001246. Thus the mismatch is not confined to one response coordinate. The five SQL blocks omit question meaning and schema-linking ambiguity; even exact feature matching leaves response components outside their span unidentified (Appendix A). The decomposition locates observed changes, not their semantic cause.

Semantic uncertainty. The 90.07% estimate quantifies agreement of qualified model reviewers on the fixed release, rather than gold semantic accuracy. Our external-label diagnostic identifies shared reviewer errors; close model comparisons should therefore be interpreted with the audit protocol in view (Appendix N).

Use and evidence scope. The release supports workload-controlled experiments and descriptive model comparisons on the evaluated panel. The stability bound limits changes to an anchor; the fresh BIRD studies test whether constructed items reproduce model-to-model differences on public reference sets. The eight-domain expansion is positive in seven domains under the nine-model sensitivity but fails its ten-model primary gate. Public development schemas, automatic semantic review, easier constructed items and reuse of a diagnosed raw pool limit generalization. The earlier external absolute-fidelity and worst-model accuracy-deviation gates remain failed. Fixed-pool TEF and reference-based pairwise-gap error are distinct endpoints.

What the fresh endpoint identifies. Operator pairing controls SQL syntax, not question meaning or reference difficulty. Item and SQL-skeleton bootstrap intervals describe variation within fixed public schemas, not replication across independent datasets. The California schools reversal remains in the eight-domain mean. Replacing a reference benchmark requires new schemas and stronger semantic adjudication; these results support only the tested deterministic variant and pairwise-gap endpoint.

8 Conclusion

Schema2Bench makes benchmark migration actionable through support diagnosis and stability-constrained selection. Cost-aware exchange improves fixed-panel TEF, while deficit-guided acquisition reaches the pool's difficulty optimum at smaller tested budgets. Qwen retains a generation gain after precision review, and a qualified Qwen–Hy3 audit jointly supports 270/300 stratified release items. Equal-budget construction increases reviewed JOIN coverage. In an exploratory eight-domain BIRD expansion, the deterministic variant improves reference-based pairwise-gap fidelity in the nine-model sensitivity, with one negative domain; the ten-model primary gate fails. Together these results connect structural adaptation with evaluator differentiation in the tested settings.

Reproducibility statement

Claims bind artifacts, hashes and decision gates. Shared-universe seeds measure stability, not independent replication. Full freeze, reproduction and routing details appear in Appendix H.4. An anonymous code and experimental-evidence archive accompanies this submission; full reruns also require the source datasets and model weights under their respective licenses.

AI use statement

Generative AI assisted part of the coding work, synthetic question realization, part of the experiment check, automated semantic review, analysis and interpretation of results, and part of manuscript drafting, translation. The authors checked the reported claims against the retained artifacts and take responsibility for the final text, code and results. Model judgments are identified as such and are not human annotations.

Ethics statement

No human annotation or human-subject study was conducted. Schema/data use and redistribution remain subject to source licenses, access and privacy rules. Executability and model checks are not semantic gold labels. Frozen failures and privileged diagnostic status are retained.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haejun Jeong, Colin Raffel, Shuaichen Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models. *Transactions on Machine Learning Research*, 2024.

Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms, 2012. URL https://arxiv.org/abs/1203.4523.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

Hasan Alp Caferoğlu, Mehmet Serhat Çelik, and Özgür Ulusoy. SING-SQL: A synthetic data generation framework for in-domain text-to-sql translation. *arXiv preprint arXiv:2509.25672*, 2025. URL https://arxiv.org/abs/2509.25672.

Shuaichen Chang, Jun Wang, Mingwen Dong, Lin Pan, Henghui Zhu, Alexander Hanbo Li, Wuwei Lan, Sheng Zhang, Jiarong Jiang, J. Lilien, Steve Ash, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, and Bing Xiang. Dr.spider: A diagnostic evaluation benchmark towards text-to-sql robustness. In *International Conference on Learning Representations*, 2023.

Michał Dereziński, Kenneth L. Clarkson, Michael W. Mahoney, and Manfred K. Warmuth. Minimax experimental design: Bridging the gap between statistical and worst-case approaches to least squares regression, 2019. URL https://proceedings.mlr.press/v99/derezinski19b.html.

Vipul Gupta, Candace Ross, David Pantoja, Rebecca J. Passonneau, Megan Ung, and Adina Williams. Improving model evaluation using SMART filtering of benchmark datasets. In *Proceedings of NAACL*, pp. 4595–4615, 2025. doi: 10.18653/v1/2025.naacl-long.235.

Samir Khuller, Anna Moss, and Joseph Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999. doi: 10.1016/S0020-0190(99)00031-9.

Venelin Kovatchev and Matthew Lease. Benchmark transparency: Measuring the impact of data on evaluation. In *Proceedings of NAACL*, pp. 1536–1551, 2024. doi: 10.18653/v1/2024.naacl-long.86.

Fangyu Lei, Jixuan Chen, Yiming Ye, Ruisheng Cao, Dongchan Shin, Hongjin Su, Zhaoyang Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, Victor Zhong, Caiming Xiong, Ruoxi Sun, Qian Liu, S. I. Wang, and Tao Yu. Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. In *International Conference on Learning Representations*, 2025.

Haoyang Li, Shang Wu, Xiaokang Zhang, Xinmei Huang, Jing Zhang, Fuxin Jiang, Shuai Wang, Tieying Zhang, Jianjun Chen, Rui Shi, Hong Chen, and Cuiping Li. Omnisql: Synthesizing high-quality text-to-sql data at scale. *Proceedings of the VLDB Endowment*, 18(11):4695–4709, 2025.

Jinyang Li, Binyuan Hui, Ge Qu, Binhua Li, Jiaxi Yang, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Cao, Rui Geng, Nan Huo, Chenhao Ma, Kevin C.-C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *arXiv preprint arXiv:2305.03111*, 2023.

Xinyu Liu, Shuyu Shen, Boyan Li, Nan Tang, and Yuyu Luo. NL2SQL-BUGs: A benchmark for detecting semantic errors in NL2SQL translation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 5662–5673, 2025. doi: 10.1145/3711896.3737427.

Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. ECBD: Evidence-centered benchmark design for NLP. In *Proceedings of ACL*, pp. 16349–16365, 2024. doi: 10.18653/v1/2024.acl-long.861.

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, 2024.

Xuanguang Pan, Chongyang Tao, Jiayuan Bai, Jianling Gao, Zhengwei Tao, Xiansheng Zhou, Gavin Cheung, and Shuai Ma. EvolSQL: Structure-aware evolution for scalable text-to-sql data synthesis, 2026. URL https://arxiv.org/abs/2601.04875.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking (of language models). In *Proceedings of NAACL*, pp. 2519–2536, 2024. doi: 10.18653/v1/2024.naacl-long.139.

Gayathri Saranathan, Cong Xu, Mahammad Parwez Alam, Tarun Kumar, Martin Foltin, Soon Yee Wong, and Suparna Bhattacharya. SubLIME: Subset selection via rank correlation prediction for data-efficient LLM evaluation. In *Proceedings of ACL*, pp. 30572–30593, 2025. doi: 10.18653/v1/2025.acl-long.1477.

Marko Sterbentz, Kevin Cushing, Cameron Barrie, and Kristian J. Hammond. RingSQL: Generating synthetic data with schema-independent templates for text-to-sql reasoning models. *arXiv preprint arXiv:2601.05451v1*, 2026. URL https://arxiv.org/abs/2601.05451v1.

Kaiser Sun, Adina Williams, and Dieuwke Hupkes. The validity of evaluation results: Assessing concurrence across compositionality benchmarks. In *Proceedings of CoNLL*, pp. 274–293, 2023. doi: 10.18653/v1/2023.conll-1.19.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems*, 2023.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of EMNLP*, pp. 3911–3921, 2018.

A INFORMATION-CONSTRAINED INSTRUMENT MIGRATION: DEFINITIONS AND SUPPORT AUDIT

This appendix defines the extended design problem and its local support audit. Historical selection experiments retain their original algorithms; the anchor-constrained selector is prospective. Source manifests, aggregate counts, code and mathematical checks are in the released V2 artifact.

A.1 THE OBJECT BEING MIGRATED

A benchmark is a measurement instrument, rather than only a collection of executable queries. Migration changes target entities and reference programs while seeking a declared workload and measurement profile. The task is to construct an instrument under restricted information, limited candidate support, and fixed release resources.

Let B_s be the structural source benchmark, D_t the target databases, and $\mathcal{I}$ the information available before evaluation:

$$\mathcal{I} = \sigma(B_s, \mathrm{schema}(D_t), E_t, \mathcal{C}, A, q, N, \phi, \xi), \quad z = \pi(\mathcal{I}), \quad z \text{ is } \mathcal{I}\text{-measurable.} \tag{10}$$

Here E_t is bounded database evidence, $\mathcal{C}$ the eligible pool, A a stratum-membership matrix, q quota counts, N release size, ϕ a frozen structural representation, and ξ a response-independent seed. Eligibility is operational, not a semantic-correctness certificate. The information excludes held-out evaluator responses and their cache paths; source response items may belong to a separate evaluation split. Candidate synthesis $\mathcal{C} = \mathcal{G}(B_s, D_t, E_t; \xi_G)$ and release selection are separate decisions.

For n eligible candidates, the release indicator has feasible set

$$\mathcal{Z}(\mathcal{I}) = \{z \in \{0,1\}^n : Az = q,\ \mathbf{1}^\top z = N\}. \tag{11}$$

With disjoint exhaustive strata, $\mathcal{Z} \neq \varnothing$ iff $0 \le q_g \le |\mathcal{C}_g|$ for every g and $\sum_g q_g = N$. This does not ensure structural representativeness or valid natural-language references.

Item-level measurement, type-level structure. Let S be the held-out source evaluation items and C the target candidate items. Each item has an immutable identity and fixed prompt, reference and evaluation protocol. Let $K_{s,i}$ and $K_{t,i}$ be joint correctness laws on $\{0,1\}^M$, with $M \ge 2$ evaluators. The target law of a particular item does not depend on which other items are released; contextual evaluation would require a different model. These item laws may be unknown before freezing. For channel $c \in \{A, R\}$ define normalized item weights

$$\alpha^A_{z,i} = \frac{z_i}{DN_{d_i}}, \qquad \alpha^R_{z,i} = \frac{z_i}{N}, \qquad \sum_i \alpha^c_{z,i} = 1. \tag{12}$$

All domain quotas $N_d > 0$ are fixed, D counts target domains, and source weights β^c_i use the source evaluation protocol. Observable types $t(i) \in T$ are used only for structural summaries $q^c_z(t) = \sum_{i:t(i)=t} \alpha^c_{z,i}$. No invariant coarse-type response law is needed to define measurement.

A.2 A JOINT-RESPONSE MEASUREMENT OPERATOR

For $k = 0, \dots, M-1$ and pairs (a, b), use coordinates

$$h_k(y) = \mathbf{1}\left\{\sum_m y_m \le k\right\}, \quad a_m(y) = y_m, \quad d_{ab}(y) = |y_a - y_b|. \tag{13}$$

Let g_j enumerate these cumulative-count, accuracy and disagreement coordinates and set $r_{u,i,j} = \mathbb{E}_{K_{u,i}} g_j(Y)$. The item-level operator is

$$\Psi_{z,j} = \sum_{i \in C} \alpha^{c(j)}_{z,i} r_{t,i,j}, \quad \Psi_{s,j} = \sum_{i \in S} \beta^{c(j)}_i r_{s,i,j}, \quad c(j) = \begin{cases} A & \text{accuracy,} \\ R & \text{otherwise,} \end{cases}$$
$$L(z; K_s, K_t) = \sum_j w_j |\Psi_{z,j} - \Psi_{s,j}|. \tag{14}$$

where $w_j = 1/M$ for each cumulative-count or accuracy coordinate and $w_j = 1/(2\binom{M}{2})$ for disagreement. The cumulative-count sum is exactly the Wasserstein-1 distance on $\{0, 1/M, \ldots, 1\}$; thus L equals the current Schema2Bench three-block TEF under the checked mapping: accuracy is domain-macro, difficulty and disagreement are item-micro, with block coefficients $1, 1, 1/2$. For frozen binary responses the kernels are point masses. It is a population functional for random kernels, not the expectation of a finite-sample absolute-loss estimator. The released V2 artifact supplies the paper and scoring-code provenance; other historical configurations are not asserted identical. Joint K enforces compatibility between coordinates. Allowing every coordinate to vary independently is only an outer relaxation.

A.3 Information-constrained design, not response optimization

The response-blind structural objective is the historical five-block proxy

$$L_{\mathrm{struct}}(z) = \frac{1}{5}\Bigg[\tfrac{1}{2}\|P_o(z) - P_o^*\|_1 + \tfrac{1}{2}\|P_d(z) - P_d^*\|_1 + \tfrac{1}{6}\|P_h(z) - P_h^*\|_1 + \frac{|\bar{j}(z) - j^*|}{\max(j^*, 10^{-12})} + \frac{|\bar{\ell}(z) - \ell^*|}{\ell^*}\Bigg]. \tag{15}$$

P_o normalizes counts over the five declared operator events (an item may contribute multiple events), P_d is the three-level deterministic SQL difficulty distribution, and P_h contains the six query-shape frequencies. $\bar{j}, \bar{\ell}$ are mean join and SQL token counts. Stars are the frozen source structural profile, independent of source response evaluation. The token source mean must be positive; the historical token division has no zero-denominator fallback. Source means in the current experiment are positive, so the join guard is inactive. Operator event normalization uses a denominator of at least one, as in code. This is not a linear-moment objective in every block. The concrete normalization follows the historical code; it is not re-tuned here.

Let $\mathfrak{K}(\mathcal{I})$ be a nonempty declared class of joint source/target response laws consistent with assumptions made without held-out responses. Relative to a frozen feasible anchor z_0, define

$$\mathcal{R}_{\mathcal{I}}(z) = \sup_{(K_s,K_t)\in\mathfrak{K}(\mathcal{I})} \{L(z; K_s, K_t) - L(z_0; K_s, K_t)\}, \qquad \min_{z\in\mathcal{Z}(\mathcal{I})} L_{\mathrm{struct}}(z) \text{ s.t. } \mathcal{R}_{\mathcal{I}}(z) \le \tau. \tag{16}$$

The same law is used for both releases inside the supremum. The noninferiority tolerance $\tau \ge 0$ must come from a use requirement, not observed results. The anchor has regret zero and is feasible. Regret need not be nonnegative for every release in a restricted response class; no nonnegativity premise is needed. If a requested structural improvement cannot coexist with the constraint, return the anchor or abstain, as specified in advance.

Equation (16) is the extended problem definition, not a claim that historical local search solved a minimax problem. Historical local search optimizes a five-block proxy under (11), without a verified robust-utility constraint. Generic robust optimization and ambiguity sets are established tools[1]; any contribution must reside in this migration-specific information regime, measurement operator, support conditions and justified solution, not the min–max notation itself.

A.4 What the observable representation can identify

Let $\Phi \in \mathbb{R}^{|T|\times p}$ contain frozen structural features. For this section the observable profile is the linear moment $\Phi^\top q$; this is not an assertion that every existing event-normalized structural block is linear in item frequencies.

Proposition 3 (Observable and unidentified response components)**.** *Over the full probability simplex on finite T (which admits a strictly positive distribution), for a fixed response coordinate $f \in \mathbb{R}^{|T|}$, its expectation is identical under every pair of probability distributions with identical Φ moments if and only if*

$$f \in \mathrm{col}([\mathbf{1}, \Phi]) \quad \Longleftrightarrow \quad v^\top f = 0 \quad \forall v \in \ker([\mathbf{1}, \Phi]^\top). \tag{17}$$

On a restricted support or quota-feasible family, membership remains sufficient but need not be necessary.

[1] See, for example, https://arxiv.org/abs/2006.06981 and https://doi.org/10.1007/s10915-022-02063-8.

Proof. The forward direction follows from equality of moments. Conversely, if f is outside the column space, there exists v in the orthogonal complement with $v^\top f \neq 0$. It has $\mathbf{1}^\top v = 0$. Perturb a strictly positive probability vector by sufficiently small $\pm\epsilon v$ to obtain two distributions with equal moments and different expectations. This is a full-simplex statement; restricted feasible supports can weaken the necessary condition. □

Equivalently, for orthogonal projector P onto that column space,

$$\langle q - p, f\rangle = \langle P(q-p), Pf\rangle + \langle (I-P)(q-p), (I-P)f\rangle. \tag{18}$$

Matching observed moments removes the first term, not necessarily the second. This standard null-space argument makes the information missing from the benchmark-migration objective explicit; it is not a new general linear-algebra theorem.

A.5 A RESPONSE-BLIND SUPPORT CERTIFICATE

Compute separate certificates for A and R; they can have different optimizers and cannot be silently combined into a jointly feasible optimum. For one weighting channel let $\mathcal{Q} = \mathrm{conv}\{q_z : z \in \mathcal{Z}\}$, assuming nonempty $\mathcal{Z}$, and source measure p. The best fractional-support discrepancy and its dual witness are

$$\delta_{\sup} = \min_{q\in\mathcal{Q}} \mathrm{TV}(p,q) = \max_{h\in[0,1]^{|T|}} \left\{ h^\top p - \max_{z\in\mathcal{Z}} h^\top q_z \right\} \leq \min_{z\in\mathcal{Z}} \mathrm{TV}(p, q_z). \tag{19}$$

Proof. For probability vectors, $\mathrm{TV}(p,q) = \max_{0\leq h\leq 1} h^\top(p-q)$. Bilinearity and compact convex domains allow minimax interchange. A linear functional over $\mathcal{Q}$ attains its maximum at a feasible-release distribution. The final inequality follows by relaxation. □

A positive witness proves workload support deficiency at the chosen resolution even for randomized mixtures of releases. It is not a TEF lower bound, because a response law can be constant. A zero relaxed deficit does not guarantee an integer release exactly matches the profile. The dual is a standard convex separation construction; reporting such a witness is a proposed operational addition.

Scalable partition oracle. With disjoint strata g, let n_{gt} count eligible items of type t, x_{gt} be selected fractional counts, and b^c_g their channel weight. Then $0 \leq x_{gt} \leq n_{gt}$, $\sum_t x_{gt} = q_g$, and $q^c(t) = \sum_g b^c_g x_{gt}$. This capped partition polytope is the projection of the convex hull of quota-feasible releases. Its TV minimization is a small linear program, without enumerating releases. The inner support functional is computed by filling each stratum's quota in decreasing $b^c_g h_t$ order. This standard partition optimization supplies both a primal discrepancy and a dual witness. Additional overlapping or semantic constraints would require a different oracle.

A.6 CONDITIONAL BRIDGE FROM MOMENTS TO MEASUREMENT

Conditional type-level bridge only. Assume a common finite type space with normalized nonnegative measures, paired primal/dual norms, and response functions $f_{u,j}(t)$ that represent the channel-specific item averages for every compared release. One sufficient condition is homogeneous item response coordinates inside each type, for source and target separately. Otherwise the following type-level bridge must include within-type mixture error. All radii are justified independently and frozen; they are not calibrated by these synthetic checks.

Under this additional assumption $\Psi_{z,j} = \langle q_z^{c(j)}, f_{t,j}\rangle$ and $\Psi_{s,j} = \langle p_s^{c(j)}, f_{s,j}\rangle$. Suppose, for each coordinate, the response class admits

$$f_{s,j} = \Phi\theta_j + e_j, \quad \|\theta_j\| \leq R_j, \quad \|e_j\|_\infty \leq \epsilon_j, \quad \|f_{t,j} - f_{s,j}\|_\infty \leq \delta_j, \tag{20}$$

with a compatible joint law and coordinate range $[0,1]$. For $v_j = q_z^{c(j)} - p_s^{c(j)}$, exact decomposition yields

$$\Psi_{z,j} - \Psi_{s,j} = \langle \Phi^\top v_j, \theta_j\rangle + \langle v_j, e_j\rangle + \langle q_z^{c(j)}, f_{t,j} - f_{s,j}\rangle, \tag{21}$$

$$L(z; K_s, K_t) \leq \sum_j w_j \left[R_j \|\Phi^\top v_j\|_* + \epsilon_j \|v_j\|_1 + \delta_j \right]. \tag{22}$$

The proof is Holder's inequality and normalization of q_z. The terms respectively express feature-controlled discrepancy, unmodeled within-feature response structure, and conditional domain shift. Candidate support restricts the achievable v_j; it is not an additional independent additive error term. Adding it again would double count.

Unconditional item-level safeguard. Without any coarse-type homogeneity assumption, set $v_i^c = \alpha_{z,i}^c - \alpha_{0,i}^c$. Common fixed item laws and reverse triangle inequality imply

$$\mathcal{R}_{\mathcal{I}}(z) \le U_{\text{item}}(z, z_0) = \sum_j w_j \tfrac{1}{2} \sum_i |v_i^{c(j)}|. \tag{23}$$

This needs no unknown functional radius. It can be conservative; $U \le \tau$ is sufficient, not necessary. The initial demo implements precisely this branch when channel weights coincide, giving $2.5\,\text{TV}$; it did not test the general function-class minimum below. We do not replace the existing sharper joint-response replacement bound.

For target type functions satisfying the boxed assumption and analogous class with radii R_j^t, ϵ_j^t, reverse triangle inequality gives a usable but conservative regret bound: Writing $v_j = q_z^{c(j)} - q_0^{c(j)}$, the conditional bound is

$$\mathcal{R}_{\mathcal{I}}(z) \le U(z, z_0) := \sum_j w_j \min\left\{ R_j^t \|\Phi^\top v_j\|_* + \epsilon_j^t \|v_j\|_1,\ \tfrac{1}{2}\|v_j\|_1 \right\}. \tag{24}$$

Coarse-type negative control. One coarse type contains two items with deterministic joint responses $(0,0)$ and $(1,1)$. The source and anchor have $(0,0)$ and the new release selects $(1,1)$, with size one. Both coarse-type distributions are (1), so a misapplied type-TV bound is zero. Actual TEF increases by two (one difficulty plus one accuracy); item-level $U_{\text{item}} = 2.5$ remains valid. The failure is a release-dependent conditional mean, not a vector outside the one-dimensional coarse feature span. The released test preserves this counterexample.

The TV term follows from $f_{t,j} \in [0,1]$. This bound assumes a common fixed target response function across the compared releases, obtainable on full item types; it may fail on coarse types whose within-type composition changes. Without justified functional radii, only the TV branch remains valid. Summing coordinate bounds can be looser than the existing joint-response replacement bound; it does not supersede that sharper result.

A.7 Algorithm and evidence correspondence

A proposed safe structural selector initializes at z_0 and accepts quota-preserving swaps only if they strictly reduce the fixed structural loss and satisfy the item-level $U_{\text{item}}(z, z_0) \le \tau$ (or the conditional type bound only when its assumptions are justified). Finite feasible support and strict improvement imply termination at a feasible restricted local optimum. No global-optimality or positive-TEF-gain guarantee follows. If no swap is admitted, return z_0 with that decision recorded.

Three levels must remain distinct: (i) historical measured structural selection; (ii) the operator and problem reformulation of what is being measured; (iii) the prospective constrained algorithm. Local synthetic checks can validate arithmetic and constraints, not the response-class assumptions on real databases. A new empirical test must freeze types, radii, tolerance, quota rules and failure criteria before independent evaluator responses are accessed. Semantic reference quality needs its own evidence.

A.8 Local application of the support certificate

A new post-hoc CPU diagnostic fixes the existing easy/medium/hard SQL bins, the 2,806-item historical pool, and the 2,000-item primary release's domain/provenance quotas before solving the certificate. No evaluator responses or database queries are used. The structural source profile is $(0.464, 0.348, 0.188)$ from Spider train. The same structural target is considered separately under target item-micro and domain-macro weights; these are workload certificates, not response statistics.

Channel	Relaxed minimum TV	Dual witness value	Primary release TV
Item-micro	0.257000	0.257000	0.272500
Domain-macro	0.395546	0.395546	0.401924

In both channels the witness is $h = (0, 1, 1)$: the quota-feasible pool cannot supply enough medium-or-hard workload mass. In the item-micro channel the source requests 53.6%, while the maximum feasible mass is 27.9%, leaving a 25.7 percentage-point shortage. This is a genuine support obstruction for these fixed bins and quotas, even after fractional relaxation. It is not a lower bound on TEF, does not establish reference semantics, and is not a comparison of all five structural blocks. It does not authorize changing historical bins, quotas or response outcomes. The diagnostic motivates acquiring candidates in certified deficient strata rather than further tuning the same selector on the same pool.

A.9 Relation to the closest formulations

Line	Object and available information	Decision	Guarantee / distinction
Kernel herding [1,2]	Target feature moments	Representative points	Moment discrepancy; response-free selection is not unique here.
Coresets [3]	Dataset and declared loss family	Weighted subset	Uniform loss approximation; differs from preserving a source instrument's joint evaluator profile.
Experimental design [4]	Design vectors before hidden responses	Measurement subset	Estimator risk under a model; hidden responses and minimax objectives already occur.
DRO [5,6]	Declared ambiguity class	Robust decision	Worst-case objective; this formulation does not add statistical calibration theory for radii.
Benchmark validity [7]	Construct and benchmark scope	Evaluation design	Instrument interpretation predates this work.
This formulation	New-schema candidates, quotas, structural source; withheld panel responses	Unweighted target release, anchor or abstention	Joint measurement fidelity under explicit information/support constraints; no unconditional improvement.

The candidate contribution is the specific cross-schema instrument-design problem and its operational support diagnosis. Neither response blindness, an instrument metaphor, quotas, nor minimax notation alone establishes novelty. This targeted comparison does not establish literature-wide priority.

1. Chen, Welling and Smola. Super-Samples from Kernel Herding (UAI 2010). https://arxiv.org/abs/1203.3472.
2. Bach, Lacoste-Julien and Obozinski. On the Equivalence between Herding and Conditional Gradient Algorithms (ICML 2012). https://arxiv.org/abs/1203.4523.
3. Tolochinsky, Jubran and Feldman. Generic Coreset for Scalable Learning of Monotonic Kernels: Logistic Regression, Sigmoid and more (ICML 2022). https://proceedings.mlr.press/v162/tolochinksy22a.html.
4. Derezinski, Clarkson, Mahoney and Warmuth. Minimax experimental design (COLT 2019). https://proceedings.mlr.press/v99/derezinski19b.html.
5. Zhu et al. Kernel Distributionally Robust Optimization (AISTATS 2021). https://arxiv.org/abs/2006.06981.
6. Nie et al. Distributionally Robust Optimization with Moment Ambiguity Sets (2023). https://doi.org/10.1007/s10915-022-02063-8.
7. Raji et al. AI and the Everything in the Whole Wide World Benchmark (2021). https://arxiv.org/abs/2111.15366.

Table 3: Shared-pool, fixed-panel comparison: absolute fidelity on seven models, averaged over ten selection seeds. Lower TEF/accuracy MAE is better; higher rank correlation is better. These are fixed-panel descriptive means, not new-schema noninferiority.

Selector	Joint TEF	Accuracy MAE (pp)	Spearman ρ
Quota random	0.05007	2.392	1.000
Structural greedy	0.04869	2.468	1.000
Structural local search	0.05594	2.521	0.993

A.10 Prior transport results and versioned context

Let X be an observable item type, q_s, q_0, q_1 the source, random, and selected type distributions, and $p_E(x) = \Pr(Y = 1 \mid X = x, E)$ an evaluator panel's conditional correctness law. For $\ell_E(q) = |\langle q - q_s, p_E \rangle|$, selected-minus-random utility is $\Gamma_E = \ell_E(q_0) - \ell_E(q_1)$.

Proposition 4 (Non-identifiability)**.** *For an observable state space with at least three values and unrestricted conditional Bernoulli response laws, fixed q_s, q_0, q_1 can have positive utility under one evaluator law and negative utility under another. Structural observations alone therefore do not identify the utility sign.*

The three-state witness and its opposite-sign response laws are retained in Appendix G.

Uniform non-worsening and the complete TEF bridge. For common finite-type weights and unrestricted response laws, universal non-worsening holds exactly when $q_1 - q_s = c(q_0 - q_s)$ for $|c| \leq 1$. TEF is a weighted ℓ_1 norm of cumulative correct-count, model-accuracy and pairwise-disagreement expectations. Common-measure affine contraction scales all blocks by $|c|$; coupled responses supply the converse. For arbitrary mixed measures, simultaneous blockwise contraction is sufficient, without a claimed converse. This characterizes a boundary, not an implementable positive-gain rule.

Approximate guarantees and the missing quantities. With blockwise contraction factor at most $\rho \leq 1$, a full-type residual E, and conditional-response shift errors Δ_j, Appendix F.1 derives

$$L_1 - L_0 \leq -(1 - \rho)L_0 + E + \Delta_1 + \rho\Delta_0. \tag{25}$$

Five structural marginals do not certify E or the response-shift errors. Thus the inequality identifies sufficient conditions, not a guarantee for historical local search. Separately, fixed-response replacement obeys the sharp bound $|L_1 - L_0| \leq \sum_d (w_d^A + w_d^R) r_d / N_d$; it reduces to $2r/N$ only with uniform item weights in every block. Full proofs and historical phase diagrams remain in Appendices G and F.

A separate frozen replacement-radius study finds attenuation of TEF harm for smaller changes (reserve-panel contrast 0.11297, cluster interval $[0.04044, 0.18661]$). The small-radius release still has negative gain, so this supports intervention-size sensitivity rather than a positive-transfer rule (Appendix H.2).

An earlier 2,470-item release reports conditional TEF gain $+0.036516$ (18.55%) under a different measurement setting, not combinable with the current structural reduction. Versioned generation feasibility studies are reported separately in Appendix H.1.

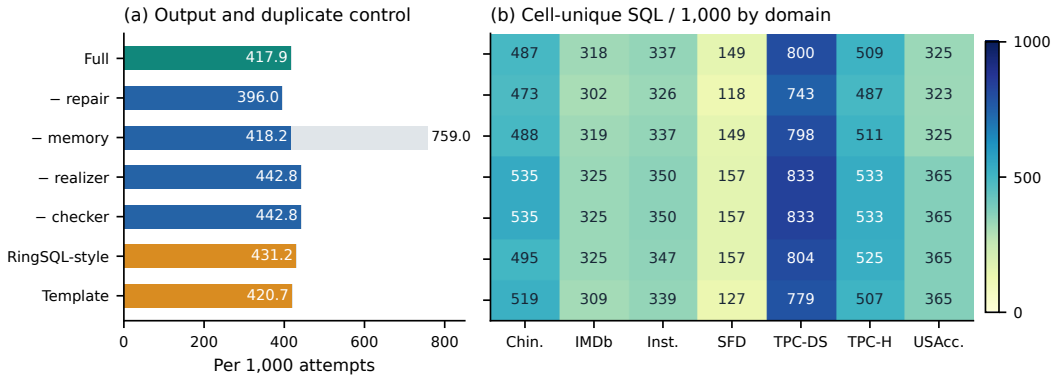


Figure 5: Construction ablations. (a) Within-cell unique SQL signatures per 1,000 attempts; the gray extension denotes repeated acceptances without memory. (b) Domain means over three seeds, with 500 attempts per cell. Removing checks or realization can raise throughput; these counts measure automatic admission, not semantic validity. Method names denote harness adaptations.

Table 4: Same-schema external reference and post-hoc three-axis diagnostic at budget 50. Lower TEF is better; gaps are improvements over the fully constrained cell.

Quantity	Estimate	95% interval / gate
Primary external TEF	0.448675	fails absolute gate
Rank Spearman	0.890110	passes rank gate
Worst-model accuracy deviation	0.296667	fails deviation gate
Candidate-universe gap	0.345532	[0.219698, 0.455442]
Response-access gap	0.162269	[0.088655, 0.247618]
Panel gap	0.023426	[-0.094120, 0.098195]

B Main-result numerical details and algorithm

Schema2Bench: two-module benchmark migration.

Require: Source SQL profile, target schemas and data, attempt and repair budgets, release size N, and fixed quotas $q_{d,g}$.

1: **Synthesis:** ground a workload plan in schema evidence; generate SQL and realize its question.
2: Execute read-only and repair within budget; synchronize LIMIT wording, then apply automatic consistency and exact duplicate checks. Store eligible candidates in $\mathcal{C}$ with their traces.
3: If any required stratum lacks its quota, stop without refilling, shrinking the release, or dropping the stratum.
4: **Adaptation:** compute the shared SQL representation ϕ; initialize quota-feasible greedy selection and improve L_{struct} with within-stratum swaps, retaining the best set B.
5: Freeze B, its quota-random comparator B_0, and the evaluation specification before a separate scorer joins model responses.

Ensure: A target reference set and construction evidence; test-effect transfer is evaluated separately and is not guaranteed by structural fit.

C Integer-attainable migration certificate and frozen-release bounds

This post-hoc audit reads only historical structural counts and frozen release memberships. It does not access evaluator responses, change selections, or redefine types or quotas. The exact-rational verifier and input hashes are in `migration_certificate_strengthening_20260913`.

Integer optimality witness. For each domain–provenance stratum, the following selected counts are integers, within candidate capacities, and sum to the frozen quota:

Domain	Provenance	Easy	Medium	Hard	Quota
Chinook	core	165	12	0	177
Chinook	refill	49	74	39	162
IMDb	core	165	0	0	165
Instacart	core	90	0	0	90
SFDFFD	core	35	0	0	35
TPC-DS	core	204	23	0	227
TPC-DS	refill	242	164	189	595
TPC-H	core	225	17	0	242
TPC-H	refill	81	22	18	121
USAccidents	core	186	0	0	186

Disjoint strata and type cells allow selecting these counts from distinct eligible items, so the witness is realizable, not merely a fractional plan. Its item-micro profile is $(0.721, 0.156, 0.123)$. Against $(0.464, 0.348, 0.188)$, TV equals $257/1000$. The dual witness $h = (0, 1, 1)$ attains the same value. Weak duality therefore proves integer optimality with zero integrality gap. The same counts attain the domain-macro lower bound $1944960557/4917152625 = 0.395546102659\ldots$ under the original positive quotas. This simultaneous witness is a property of this instance; separate channel optima do not generally share an optimizer.

The main-text accounting follows by adding and subtracting the integer optimum. Both gaps are nonnegative by feasible-set inclusion and optimality. This identity is not itself a new optimization theorem; its operational content is the computable certificate and realizable witness. For the primary release, remaining item-micro difficulty headroom is 0.0155 and domain-macro headroom is approximately 0.006378. No optimality claim is made for other structural blocks, joint TEF, or candidate semantic validity.

How informative is the response-free stability bound? Matching each historical seed to its quota-random anchor yields:

Selector	Replacements (range)	Mixed-weight H (range)
Greedy	564–579	0.564273–0.582641
Local	556–584	0.561020–0.595226

All ten seed pairs preserve domain counts; removed and added counts match in every domain. These distribution-free bounds are broad and cannot establish useful empirical noninferiority of the historical releases. They remain valid safeguards for a prospective bounded-change algorithm, not evidence that the historical optimizer enforced such a constraint. The audited catalog contains these historical methods and anchors; no missing alternatives are inferred.

Aggregation-order sensitivity. For domain profiles q_d and source profile p, distinguish $\mathrm{TV}(D^{-1}\sum_d q_d, p)$ from $D^{-1}\sum_d \mathrm{TV}(q_d, p)$. The former is the domain-equal pooled quantity used above; the latter averages domain losses and is at least as large by convexity. Neither aggregation order changes the historical TEF endpoint.

Difficulty statistic	Integer optimum	Primary release	Headroom
Item-micro TV	0.257000	0.272500	0.015500
Domain-equal pooled TV	0.395546	0.401924	0.006378
Mean domain TV	0.401536	0.405133	0.003597

For mean domain TV, the exact optimum is $1974412199/4917152625$, and the primary release has $3984201773/9834305250$. The same integer witness attains all three optima in this instance. For each domain, choosing $h_t = \mathbf{1}\{p_t > q_{d,t}\}$ and maximizing its support over that domain's original stratum capacities attains the witness's TV. Summing these independent domain dual bounds certifies

the mean-domain optimum. The portable review package includes typed candidate IDs, primary and historical memberships, pinned hashes and `verify_membership.py`; it independently recomputes these values and replacement counts without overwriting the expected results. It does not independently validate SQL type assignment or semantic correctness.

D A NECESSARY ACQUISITION BUDGET FOR STRUCTURAL MIGRATION

The migration framework can translate a support witness into a necessary candidate-acquisition budget. The result holds under fixed release size, disjoint fixed quotas, unchanged types, and item-micro weights. Eligible means admitted under the existing protocol, not independently proven semantic truth.

Proposition 5 (Acquisition lower bound)**.** *Let $\mathcal{C}^+$ add at most b distinct eligible items to $\mathcal{C}$, without removing old items or changing quotas. If the original feasible set is nonempty, then*

$$D_T^*(\mathcal{C}^+) \geq \max\{0, D_T^*(\mathcal{C}) - b/N\} \geq \max\{0, \delta_{\sup}(\mathcal{C}) - b/N\}. \tag{26}$$

Consequently, achieving TV at most $\epsilon < \delta_{\sup}$ requires at least $\lceil N(\delta_{\sup} - \epsilon)\rceil$ added items.

Proof. Any new feasible release selects $k \leq b$ added items. Within each stratum, replace these by unselected old items. There are enough old items because the original stratum capacity is at least its fixed quota. The resulting old-pool release changes at most k of N items, so its type distribution changes by TV at most k/N. The triangle inequality gives $D_T^*(\mathcal{C}) \leq D_T(z^+) + b/N$. Minimize over new releases and use nonnegativity and the relaxation bound. □

This is a capacity sensitivity result, not a TEF guarantee. Changing the source profile, quotas, type definitions, weighting, or eligibility contract falls outside this claim. It does not bound generation attempts or token cost.

Concrete lower bound and idealized tightness. The source requires 928 easy, 696 medium and 376 hard items at $N = 2000$. The existing eligible pool has only 312 medium and 246 hard items, all of which fit the support-optimal integer witness. Exact matching therefore needs at least 384 additional medium and 130 additional hard items, 514 in total. A count-space construction adds 204 medium to TPC-DS core, 180 medium to TPC-H core and 130 hard to TPC-DS refill, replacing the corresponding easy selections. All quotas remain fixed. This construction is conditional on the availability of distinct eligible items in those cells; it is not a newly constructed dataset.

For every integer $0 \leq b \leq 514$, a prefix of those acquisitions attains $0.257 - b/2000$ in count space. Exact-fraction verification checks all 515 budgets and their capacity constraints. A separate exhaustive two-type, two-stratum check verifies 2,352 small acquisition cases. These checks support the algebra and implementation; they do not provide empirical evidence of reference quality, generator capacity, or downstream measurement improvement.

The audit is available as `acquisition_certificate_20260913/check.py` with input hash and results. Historical inputs and response caches remain unchanged. The lower bound guides acquisition requirements without authorizing post-hoc tuning or claiming successful end-to-end migration.

E BOUNDED RELEASE OPTIMIZATION AND OFFLINE CANDIDATE ACQUISITION

This diagnostic reuses the historical 2,806-item universe, 2,000-item quota release, 1,034 source items and seven-model cached panel. It evaluates new memberships without new model calls. All results are developmental: the primary seed was evaluated first, then all remaining historical seeds under the unchanged rule. The same-budget random control was designed after those results were observed. Each stage froze memberships before scoring them; this does not make the reused evidence an independent confirmation set.

Algorithm and fixed computational budget. Starting from each quota-random anchor, select the greatest five-block structural loss reduction among swaps from original-anchor items to the original complement, preserving domain–provenance counts. Admit a swap only if total mixed-weight $H \leq 0.01$. Use deterministic ID ties and at most 25 swaps, without undoing previous changes. The budget is a development setting, not a use-calibrated noninferiority margin. This is a standard best-improvement exchange implementation with a budget safeguard, not a new generic optimizer.

Method	Structural loss	Joint TEF
Quota random	0.269519	0.050071
Structural greedy	0.213055	0.048692
Unconstrained local	0.189298	0.055945
Budget-matched random exchange	0.269159	0.049822
Bounded structural exchange	0.261163	0.047747

Values average the same ten historical seeds. Bounded exchange improves structure by 3.10% and TEF by 4.64% relative to its random anchors, with 10/10 positive pairs on both metrics. Against greedy its mean TEF is 1.94% lower (7/10 positive pairs) but its structural loss is higher in every pair. Relative to unconstrained local its TEF is lower in every pair, at the cost of weaker structural matching. No dominance over all alternatives is claimed. The measured H ranges from 0.00943309 to 0.00999737; every observed anchor-relative TEF change satisfies the bound.

Budget-matched control. For each seed, draw one uniform random set of removals from the original anchor and additions from its complement in each domain–provenance stratum, using exactly the bounded method's replacement counts. A fixed seed offset 20260913 is used for all controls. Consequently counts and H match exactly. Bounded exchange has lower structure loss and TEF in all ten paired controls. This distinguishes the observed structured-swap outcome from one set of same-budget random paths; it is not a population causal effect or exhaustive comparison to every random path. No alternate draws were selected for reporting.

Reproduction and scope. The original three same-seed baselines reproduce their historical metrics to 10^{-12}. A separate exact-fraction implementation recomputes all ten bounded TEFs from cached binary responses, matching the scoring implementation to 10^{-12}. Plans, memberships, input hashes and all seed/control outcomes are retained in `bounded_cached_stability_20260913`; the original seed artifacts remain in `bounded_method_feasibility_20260913` and `bounded_cached_utility_20260913`. Shared seeds and evaluator models are not independent schemas. These outcomes establish an observed tradeoff on reused evidence, not semantic reference truth, statistical superiority, new-schema generalization or externally justified noninferiority.

E.1 Cost-aware exchange and module-level evaluation

The extension uses the same anchors, maximum 25 monotone swaps and $H \leq 0.01$ cap. It maximizes structural improvement per unit c_d rather than absolute improvement. All ten new memberships were frozen before their panel scores were computed. This extension was designed after the earlier results; it uses the same response records and is not an independent confirmation study.

Mean structural loss is 0.260790 and joint TEF 0.047229. Relative to absolute-gain exchange, structure improves in 6/10 pairs and TEF in 8/10. Accuracy MAE is 0.023648 versus 0.024043; difficulty Wasserstein distance is 0.012975 versus 0.012932; disagreement MAE is 0.021213 versus 0.021546. All ten runs make 14 swaps in TPC-DS, consuming $H = 0.0094330900$: less than absolute-gain exchange in nine cases and equal in one. The domain allocation reflects the lower bound-induced cost in a larger domain; it is not evidence of improvement within every domain. A new control matches the cost-aware rule's exact domain–provenance replacement counts and H. Its mean structural loss is 0.269239 and TEF 0.049857; cost-aware exchange improves TEF in all ten pairs. Restricting absolute-gain exchange to the largest-quota domain produces exactly the same ten releases as cost-aware exchange. Thus this pool supports an allocation mechanism, not a distinct within-domain search advantage. Neither greedy optimality nor a calibrated noninferiority margin is claimed.

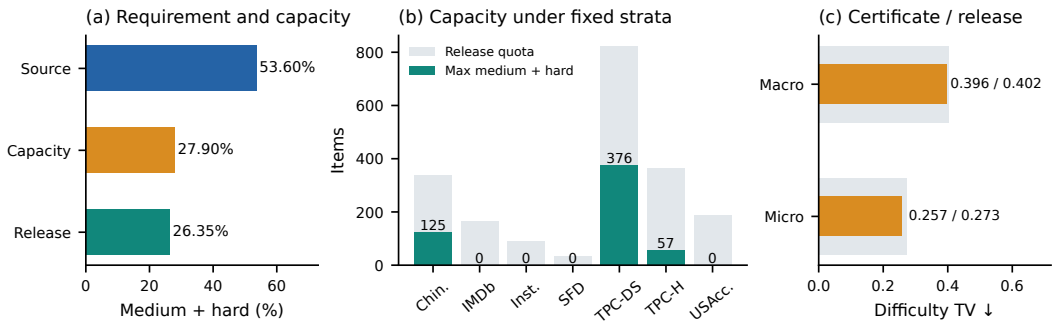


Figure 6: Response-free support diagnosis of the historical pool. (a) Medium/hard source demand, quota-feasible capacity and the primary release. (b) Domain quotas and eligible medium/hard capacity under fixed strata. (c) Certified integer optimum (orange) versus released (gray) difficulty TV under item-micro and domain-equal pooled weighting (TV after pooling). These certificates do not bound TEF.

Offline acquisition protocol. For each of the ten anchors, the 806 other candidates are the reserve. All policies have access to the same SQL-derived three-class metadata. Uniform acquisition shuffles reserve IDs with seed offset 20260913. Source-proportion acquisition selects the available type with the largest shortfall relative to its cumulative source-proportion allocation; deficit acquisition selects the type with the largest source-demand minus pool-count shortfall. Deterministic type and ID ties are used. After each of the fixed budgets $0, 50, 100, 200, 400, 806$, an integer min-cost flow computes the best 2,000-item difficulty match preserving the original domain–provenance quotas. This is acquisition from an existing, typed reserve, not new generation; metadata extraction costs are not part of the addition budget.

Additions	Uniform	Source proportion	Deficit guided
0	0.33575	0.33575	0.33575
50	0.33115	0.32225	0.31075
100	0.32580	0.30875	0.28575
200	0.31635	0.28225	0.25700
400	0.29820	0.25700	0.25700
806	0.25700	0.25700	0.25700

These are ten-initial-set means, not independent-schema estimates. The tested budget thresholds are not exact minimum acquisition costs. The full-pool value 0.257 is the difficulty optimum, not a TEF or semantic-quality bound. A separate maximum-overlap flow implementation reproduces all 180 cells exactly; 15 small integer instances were also checked by exhaustive enumeration.

Alternative structural target and reproducibility. With the existing BIRD train structural profile, absolute-gain and cost-aware exchange reduce mean structure loss from 0.415485 to 0.411037 and 0.409843, respectively. No BIRD source-response utility is inferred. Source hashes, plans, memberships and all outcomes are in `method_extension_development_20260913`; the independent arithmetic and membership audit is in `method_extension_review_20260913`.

E.2 Acquisition-to-release chain

We apply both unchanged exchange rules to each acquired pool at the six fixed budgets above: 360 seed–policy–budget–rule cells, plus 20 mechanism controls. All 380 memberships were fixed before scoring; 195 unique releases were scored once. Exact-fraction replay reproduces their TEFs within 10^{-12}. Every release preserves 2,000 items, fixed strata and the H cap.

Additions	Uniform	Source proportion	Deficit guided
0	0.050071	0.050071	0.050071
50	0.048594	0.049156	0.047093
100	0.048015	0.047408	0.047034
200	0.047686	0.047409	0.047229
400	0.047304	0.047229	0.047229
806	0.047229	0.047229	0.047229

Values are mean final-release TEF for cost-aware exchange. At 200 additions, deficit-guided acquisition reproduces the full-reserve memberships in all ten seeds. Source-proportion and uniform acquisition reach that endpoint at 400 and 806 tested additions. TEF is not monotone in acquisition budget; these are endpoint-reproduction thresholds on the tested grid, not selected TEF-optimal budgets. The chain uses existing typed candidates and does not establish generation-to-semantic-validation performance. All cells and controls are retained in `acquisition_release_chain_20260913`.

E.3 Model-selection decision preservation

For the five selectors in Table 1, we fix all 50 releases and compute exact rational domain-macro accuracies. Let a_m^S be source reference accuracy and $\hat{m}_T \in \arg\max_m a_m^T$ the release-selected model, with lexicographic model-ID ties. Source decision regret is $\max_m a_m^S - a_{\hat{m}_T}^S$; it evaluates source-decision preservation, not identification of a target-optimal model. We also compare top-three sets and the signs of all 21 non-tied source model pairs.

All five methods retain the source leader and all three leading models in every seed, yielding zero source decision regret. Random, greedy and both bounded rules preserve 210/210 pair orders across ten seeds; unconstrained local preserves 208/210. All source pair gaps exceed one percentage point, so the prespecified descriptive gap filter gives the same counts. These repeated pairs are correlated, not 210 independent observations. The leader and top-three results support decision preservation but do not distinguish bounded exchange from random or greedy. Exact accuracy replay agrees with the historical implementation to 1.12×10^{-16}. Plans, memberships and per-model results are retained in `model_selection_utility_20260913`.

F Sharp joint-feature replacement bound

The historical $2.5r/N$ bound is conservative when every TEF block uses uniform item mass. That condition does not hold for the seven-model twelve-method endpoint: its accuracy is domain-macro, while difficulty and disagreement are item-micro. We retain the endpoint and derive its bound explicitly.

Assume binary responses on a fixed K-model panel, uniform model and unordered-pair weights, normalized correctness-count Wasserstein distance, and disagreement coefficient $1/2$. For response vectors a, b at Hamming distance h, the accuracy-block diameter is $h/K \le 1$. The combined difficulty and weighted-disagreement feature distance is

$$\delta_R(a,b) = \frac{|\sum_m a_m - \sum_m b_m|}{K} + \frac{h(K-h)}{K(K-1)} \le 1 \quad (K \ge 2). \tag{27}$$

Exactly $h(K-h)$ unordered pairs change disagreement. For integer $h < K$, bounding the first term by h/K leaves the nonnegative gap $(K-h)(K-1-h)/(K(K-1))$ from 1; for $h = K$, the distance is at most 1. Equality is attained by the all-zero and all-one vectors; complementary vectors need not attain it. For $K = 1$, disagreement is absent and the difficulty diameter is 1. All-zero and all-one vectors jointly attain both block diameters.

Block-specific domain weights. Let each release contain the same positive count N_d in each of D domains. Accuracy uses nonnegative normalized domain weights w_d^A, and both difficulty and disagreement use weights w_d^R. Source moments and all weights remain fixed. For r_d within-domain replacements, coupling removed and added responses, the triangle inequality for feature means, and

the reverse triangle inequality for distance to the source give

$$|L_{\mathrm{TEF}}(B_1) - L_{\mathrm{TEF}}(B_0)| \le H := \sum_d (w_d^A + w_d^R)\frac{r_d}{N_d}. \tag{28}$$

This bound is sharp for every feasible vector (r_d): take an all-zero source and anchor, then replace r_d rows by all-one vectors within each domain. Accuracy changes by $\sum_d w_d^A r_d/N_d$, difficulty by $\sum_d w_d^R r_d/N_d$, and disagreement remains zero. The construction preserves quotas and attains H. Coupling and convexity also bound changes in expected feature laws; they do not remove response-sampling uncertainty.

For the historical mixed endpoint, $w_d^A = 1/D$ and $w_d^R = N_d/N$, hence

$$H_{\mathrm{mixed}} = \frac{r}{N} + \frac{1}{D}\sum_d \frac{r_d}{N_d}, \qquad r = \sum_d r_d. \tag{29}$$

If all blocks instead share w_d, the bound becomes $2\sum_d w_d r_d/N_d$; the same bound holds when full within-domain losses are averaged after their norms. Only when every block has $w_d = N_d/N$ does this reduce to

$$|L_{\mathrm{TEF}}(B_1) - L_{\mathrm{TEF}}(B_0)| \le 2r/N. \tag{30}$$

Why the distinction matters. In a synthetic all-zero anchor with the historical seven-domain quotas, one all-one replacement in the 35-item stratum changes the mixed loss by $1/2000 + 1/(7 \cdot 35) = 0.00458163$. This exceeds both $2/2000$ and $2.5/2000$. It disproves an unconditional radius-only certificate; it is not an observed violation by a historical release. All seven strata are included in the CPU witness report. Exhaustive checks of 87,380 ordered binary-vector pairs for $K = 1, \ldots, 8$ corroborate the two diameter bounds. The algebra above, not those finite checks, proves the result for arbitrary K.

The future rule clips the fixed path at its longest prefix with $H \le \epsilon$. The shortcut $\lfloor \epsilon N/2 \rfloor$ is restricted to uniform item mass in every block. Arbitrary model/pair weights, different difficulty and disagreement stratum weights, or an added pattern-TV endpoint need separate bounds. Historical radii, metrics, budget snapshots, and failed gates are unchanged. Neither the weighted theorem nor the old illustrative budget table retroactively certifies a historical release or guarantees positive gain.

F.1 Approximate contraction: separating observable residual from response shift

The exact contraction characterization can be relaxed quantitatively. The following proposition is a consequence of linear expectation maps and the triangle inequality, rather than a claim of a new general contraction theorem. Its purpose is to expose which additional assumptions would connect structural control to evaluation fidelity.

For each TEF block b, let $q_{j,b}$ be its probability measure on a common finite type space, for source $j = s$, anchor $j = 0$, and selection $j = 1$. Different blocks may use different domain weights. Let $F_{j,b}(x)$ be the conditional expected response feature given type x, and $\lambda_b \ge 0$. Use cumulative correctness-count indicators for the difficulty block, model correctness for accuracy, and pairwise disagreement for disagreement, with normalization absorbed into each block norm and coefficient. Then full TEF is

$$L_j = \sum_b \lambda_b \left\| \sum_x q_{j,b}(x) F_{j,b}(x) - \sum_x q_{s,b}(x) F_{s,b}(x) \right\|_b . \tag{31}$$

This representation retains joint response information; it does not replace TEF by averaged single-model effects. Suppose a common reference map T_b has feature diameter at most D_b and satisfies $\sup_x \|F_{j,b}(x) - T_b(x)\|_b \le \eta_{j,b}$ for every j. These are explicit assumptions about conditional response laws, not consequences of SQL execution or matching five structural marginals.

Proposition 6 (Residual and response-shift decomposition)**.** *Choose coefficients $|c_b| \le \rho \le 1$, and define*

$$e_b = q_{1,b} - q_{s,b} - c_b(q_{0,b} - q_{s,b}), \qquad E = \sum_b \lambda_b D_b \frac{\|e_b\|_1}{2}, \qquad \Delta_j = \sum_b \lambda_b(\eta_{j,b} + \eta_{s,b}). \tag{32}$$

Then

$$L_1 - L_0 \le -(1-\rho)L_0 + E + \Delta_1 + \rho\Delta_0. \tag{33}$$

Proof. Each e_b has total mass zero. Its positive and negative parts have the same mass $\|e_b\|_1/2$; their normalized feature expectations lie in the convex hull of $T_b(x)$. Thus $\|\sum_x e_b(x)T_b(x)\|_b \le D_b\|e_b\|_1/2$. For the reference-map loss L_j^T, linearity gives $L_1^T \le \rho L_0^T + E$. Each probability-weighted conditional-feature error is at most its uniform bound, so $|L_j - L_j^T| \le \Delta_j$. Combining $L_1 \le L_1^T + \Delta_1$ and $L_0^T \le L_0 + \Delta_0$ proves the result. □

What can and cannot be checked without responses. On a declared full finite type space, E is a distributional residual that can be computed without evaluator responses once the coefficients and weights are fixed. Five marginal SQL summaries do not in general determine it. The response-shift terms $\eta_{j,b}$ and anchor loss L_0 are not certified by those summaries. Under a genuinely shared conditional response map all η terms vanish; when $E = 0$, the exact sufficient contraction result is recovered. A sufficient non-worsening condition is $E + \Delta_1 + \rho\Delta_0 \le (1-\rho)L_0$. The present experiments do not certify this condition. In particular, source and target databases can induce different responses for the same SQL type. The proposition identifies the missing quantities rather than asserting that historical local search already controls them.

Feasible replacement budgets. For a shared pool with M_d distinct items and two releases of size N_d in domain d, set cardinality implies $0 \le r_d \le \min\{N_d, M_d - N_d\}$. Combining this with the sharp replacement theorem yields the observable sufficient harm-budget condition

$$\sum_d \left(\frac{1}{N} + \frac{1}{DN_d}\right) r_d \le \epsilon \quad \Longrightarrow \quad |L_1 - L_0| \le \epsilon \tag{34}$$

for the historical mixed endpoint, with fixed responses and fixed domain counts. Unlike approximate contraction, this bound requires no shared conditional response law, but guarantees only a bounded change, not an improvement. At fixed N, replacing an item in a smaller domain consumes more of this budget. Pool expansion changes feasibility, not the meaning of this bound. No budget, coefficient, or threshold is fitted to rescue a historical result.

G THEORY DETAILS

Let q_s, q_0, q_1 be distributions on a finite observable type space and $p_E \in [0,1]^d$ the evaluator-panel conditional correctness law. Define $\ell_E(q) = |\langle q - q_s, p_E\rangle|$ and $\Gamma_E = \ell_E(q_0) - \ell_E(q_1)$.

Proposition 7 (Response-blind non-identifiability)**.** *If the state space has at least three values and evaluator response laws are unrestricted, fixed observable distributions can have positive transport utility under one law and negative utility under another.*

Proof. Take $q_s = (0.4, 0.3, 0.3)$, $q_0 = (0.6, 0.2, 0.2)$, and $q_1 = (0.3, 0.5, 0.2)$. For $p_{E+} = (1,0,0)$, source, random, and selected response rates are 0.4, 0.6, and 0.3, so $\Gamma_{E+} = 0.1$. For $p_{E-} = (0,1,0)$, the rates are 0.3, 0.2, and 0.5, so $\Gamma_{E-} = -0.1$. Observables are unchanged; only the response law changes. □

Proposition 8 (Uniform non-worsening)**.** *For $q_0 \ne q_s$, $|\langle q_1 - q_s, p\rangle| \le |\langle q_0 - q_s, p\rangle|$ for every $p \in [0,1]^d$ iff $q_1 - q_s = c(q_0 - q_s)$ for some $|c| \le 1$. If $q_0 = q_s$, uniform non-worsening requires $q_1 = q_s$.*

Proof. Sufficiency follows by scaling. For necessity, write $a = q_0 - q_s$ and $b = q_1 - q_s$. Both sum to zero. For any $v \in \mathbb{R}^d$, choose small enough $\epsilon > 0$ that $p = \frac{1}{2}\mathbf{1} + \epsilon v \in [0,1]^d$. The inequality gives $|\langle b, v\rangle| \le |\langle a, v\rangle|$, hence $\ker(a) \subseteq \ker(b)$ and $b = ca$. Applying the inequality away from $\ker(a)$ yields $|c| \le 1$; when $a = 0$, it forces $b = 0$. □

Proposition 9 (Lift to complete TEF)**.** *Let $R = (R_1, \ldots, R_K) \in \{0,1\}^K$ be the response vector for a panel and let $S = \sum_m R_m$. Under common finite-type weights across blocks, define*

$$\psi(R) = \big(\mathbf{1}\{S \le j\}_{j=0}^{K-1}, R_1, \ldots, R_K, (|R_m - R_{m'}|)_{m<m'}\big).$$

Then complete TEF under these common weights is a weighted ℓ_1 norm of the source–release difference in $\mathbb{E}[\psi(R)]$. Consequently, $q_1 - q_s = c(q_0 - q_s)$ with $|c| \leq 1$ is uniformly non-worsening for complete TEF. Under unrestricted joint response laws, complete-TEF utility is not identifiable from q_s, q_0, q_1 alone, and uniform non-worsening still requires the same affine contraction.

Proof. On the ordered correctness-count support, Wasserstein-1 distance equals the weighted ℓ_1 distance between cumulative distribution functions. Its cumulative coordinates are the expectations of $\mathbf{1}\{S \leq j\}$. Model-macro accuracy and pairwise disagreement are the remaining coordinates of $\mathbb{E}[\psi(R)]$; the TEF coefficient 0.5 changes a nonnegative coordinate weight. If $q_1 - q_s = c(q_0 - q_s)$, every source–release coordinate difference is multiplied by c, so TEF is multiplied by $|c|$.

For necessity and non-identifiability, restrict the admissible joint laws to $R_1 = \cdots = R_K = Y$ with $\Pr(Y = 1 \mid X = x, E) = p_E(x)$. Pairwise disagreement is then zero, while both the difficulty and accuracy terms are positive multiples of $|\langle q - q_s, p_E \rangle|$. The scalar witness and uniform-non-worsening characterization therefore embed in complete TEF. This lift adds no empirical assumption and does not turn the radius bound into a positive-gain guarantee. □

The historical mixed endpoint uses domain-macro accuracy but item-micro difficulty and disagreement. Write q_i^A, q_i^R for the corresponding finite-type measures, $i \in \{s, 0, 1\}$. Simultaneous $q_1^b - q_s^b = c(q_0^b - q_s^b)$ for both blocks with the same $|c| \leq 1$ is sufficient, since the concatenated block differences scale by c. The scalar argument does not prove a converse for arbitrary mixed measures. The unrestricted class still contains the common-measure non-identifiability witness; a condition specific to the observed historical panel is not inferred.

For scalar-witness panels E, E', the gain shift is bounded by

$$|\Gamma_E - \Gamma_{E'}| \leq (\|q_0 - q_s\|_1 + \|q_1 - q_s\|_1)\|p_E - p_{E'}\|_\infty.$$

For equal-size, equally item-weighted releases differing by r quota-preserving replacements out of N, $|L_{\mathrm{TEF}}(B_1) - L_{\mathrm{TEF}}(B_0)| \leq 2.5r/N$. The first bound explains why a development sign need not transfer; the second limits worst-case harm without certifying benefit. Appendix F derives the sharp weighted bound for the actual mixed endpoint; neither total-radius expression is applied to it unconditionally.

The controlled phase diagram crosses panel-shift strength, family correlation, structural alignment, panel size, intervention radius, and domain mixture. It contains 1,728 cells and 5,184 method–cell rows. Mean sign-flip probability for the response-blind proxy rises from zero at shift strength 0 to 0.2855, 0.4828, and 0.5414 at strengths 0.5, 1.0, and 1.5. Because response laws are known by construction, this is explanatory evidence, not another empirical panel.

H Construction evidence, experimental scope and reproducibility

H.1 Construction protocol and semantic scope

SQL-first construction uses schema metadata and bounded read-only execution, at most three repair rounds, question realization, consistency screening and canonical-signature duplicate rejection. These are programmatic stages with an LLM realizer, not independently validated reasoning agents. Comparators share the harness: RingSQL-style and SING-SQL-style are adaptations, not claims of official system reproduction. The current ablation has 147 cells: seven methods, seven domains, three seeds and 500 attempts per cell. Repair increases unique admitted SQL per 1,000 attempts from 396.00 to 417.90; 20 domain–seed pairs improve and one ties. Removing memory leaves unique yield at 418.19 while nearly doubling raw acceptance. RingSQL-style and Template yield 431.24 and 420.67. These are admission and duplication effects, not semantic-quality superiority. Controlled synthesis has lower profile error but higher token cost and lower yield than RingSQL-style.

A separate model audit reports 237/245 supported items (96.73%), with Fleiss' $\kappa = 0.331$ and none of three judge-level noninferiority gates passing. The conjunctive qualification therefore failed. No human semantic annotation was conducted. Neither this sample nor SQL executability certifies all 2,000 released questions. Earlier five-domain generation studies retained 4.70–6.00 consistent, nonduplicate references per 1,000 attempts despite execution above 80%. These protocols differ

from the current SQL-signature ablation and their counts must not be pooled. Historical schema-link examples also include questionable identifier mappings; they are not reference gold.

H.2 Confirmation and diagnostic interpretation

The earlier fresh-four confirmation reports joint gain -0.006573, with interval $[-0.049714, 0.020661]$, and fails its criterion. In the expert-reference study, seven schemas each supply 40 reference tasks and 20 selected tasks; gain is 0.004728, interval $[-0.014473, 0.025102]$, with 4/7 positive schemas. These tests do not establish general fidelity and are not confirmation of the subsequently designed bounded rule. Full external-reference estimates and matched-support comparisons follow below. Shared-pool selection seeds are correlated paths, not independent schema replications. Radius and panel sensitivity studies retain their original protocols in the separate technical record; no additional independent evidence is inferred from repeated summaries.

H.3 Matched support and alternative structural target

The matched comparison fixes the original 2,806-item universe, 2,000-item budget and domain/provenance counts. Added k-center, DPP and SMART-style alternatives each have one fixed-seed run. Local improves both losses against those runs, while greedy and facility location have lower TEF. This does not establish dominance over the baseline suite. The BIRD train structural profile gives local 14.73% improvement over random, but only 0.00000541 absolute improvement over greedy. This is a structural target sensitivity result, not evaluation on BIRD response references. The numerical source and original sensitivity tables remain in the separate technical record.

H.4 Measurement convention and reproducibility

Selection uses the 7,000-item Spider training structural profile. The declared source-response reference contains 1,034 development items. Accuracy uses domain-macro weighting; difficulty and disagreement use item-micro weighting. The historical single-model loss and the current joint-panel TEF are different estimands and are not pooled. Freeze selected IDs, quotas, configuration and hashes before joining evaluator responses. New bounded analyses reuse the fixed response table and are developmental; this per-stage freeze does not make them fresh confirmation. Exact-fraction replay, membership checks and all seed outcomes support the reported arithmetic, not population generalization.

The accompanying technical record preserves previous complete protocols, model rosters, detailed costs, radius analyses and audit tables. It is an archival reference, not an additional source of independent results. The current appendix retains the proofs, principal comparisons and material limitations needed to assess the main claims.

I Panel comparisons and external checks

I.1 Frozen alternatives and panel sensitivity

Figure 3 reports the seven-model comparison; accuracy MAE and rank correlations appear in Appendix B. A post-hoc matched-support comparison restricts four additional response-free selectors to the original 2,806 candidates and exact domain/provenance quotas before selection. Local search has the lowest structural loss and lies on the observed two-objective frontier with greedy and facility location (Figure 3); it is not TEF-optimal. The unequal-support union and old-13/fresh-four panel results remain separate (Appendices H.3 and H.2).

Local improves both losses over the fixed k-center, DPP and SMART-style runs; greedy and facility location offer lower TEF. Added methods each have one fixed-seed run, not ten independent replications (Figure 3).

I.2 Expert-reference support stress test

A fixed pool of expert-authored tasks isolates selection from generated-item support. Seven schemas each provide 40 references; three selectors release 20 items per schema and are evaluated by seven

models. Local search's random-minus-local gain is $+0.004728$, with a schema-cluster interval of $[-0.014473, +0.025102]$ and $4/7$ positive schemas. The result does not meet the frozen confirmation criterion. This study controls reference support, rather than certifying generated tasks; complete results appear in Appendix H.2.

I.3 External-reference evaluation and diagnostic contrasts

We next use 100 expert-authored BIRD items on each of three schemas as a same-schema external criterion. Two generators, three generation seeds, eight response-free baselines, 30 selection seeds, and 25 evaluators yield 111,000 frozen responses. The replication units are three schemas, not 25 evaluators. On the 13-model confirmatory core, primary TEF is 0.448675, rank Spearman is 0.890110, and worst-model accuracy deviation is 0.296667. Ranking passes its frozen criterion; absolute fidelity and worst-model deviation do not. A post-hoc budget-50 diagnostic reduces TEF by 0.345532 when substituting same-schema expert support, 0.162269 with response access, and 0.023426 with a seen panel. Only the first two intervals exclude zero. These are privileged diagnostic contrasts, not deployable methods or causal ablations; complete results appear in Appendix H.3.

J Certificate-guided generation: protocol and evidence

Each of seven development databases receives 24 slots per policy, split into three batches of eight. B specifies no difficulty. C conditions the source medium/hard proportions $0.348/0.188$, rounding to 109/59 requests; D rounds the 384/130 deficits to 126/42 requests. This tests difficulty composition, not adaptive allocation across databases. All 504 requests are interleaved before collection. Every policy uses Qwen2.5-Coder-7B-Instruct, temperature 0.7 and seed $20260914 + \text{slot}$. Scoped contexts are identical within each slot: a root table cycling in sorted order, at most two declared foreign-key parents, complete scoped DDL and up to three sample rows per table. This is not a test of whole-schema planning superiority. The model generates new SQL, without a shared template generator preassigning it.

Execution and qualification. Only SQL errors permit one repair; changed question text is rejected. Read-only execution is limited to supplied tables, five seconds and 1,000 rows. Initial/repair output limits are 1,536/768 tokens and complete inputs are capped at 6,000 tokens. All costs include failures and repairs. All 504 slots completed, with 535 distinct completion identifiers (31 repairs) and no unresolved calls. There were 443 nonempty executions, 22 timeouts, 10 row-cap outcomes, 27 SQL errors, one format error and one empty result. Lexical SQL deduplication preserves string literals; it does not certify semantic equivalence. The historical classifier reproduces all 2,806 original labels; this tests implementation parity, not difficulty validity on new SQL.

Blinded model review. The 345 distinct review items include parseable initial, repaired and final outputs. Reviewers receive question, SQL and evidence, but no policy identity, requested/classified difficulty or outcome summaries. The separately submitted reviews are user-identified as hy3 and doubao2.1pro; returned identities are WorkBuddy and unknown, so underlying versions are not independently verified. Both cover all 345 items. A final item counts only if it is nonempty, new, unique within its policy, supported by both reviewers, and judged free of unnecessary complexity by both. These are model judgments, not human gold labels.

Review A assigns 320 supported, 23 incorrect and two insufficient labels; review B assigns 250, 87 and eight. Agreement is 273/345 (79.13%), with Cohen's $\kappa = 0.3288$. Joint support on this review universe is 250/345, not final-release accuracy. On automatically eligible unique candidates, reviewer-A-only qualified medium/hard counts are 16/29/32 for B/C/D; reviewer-B-only and joint counts are both 11/14/19. Each reviewer separately preserves D>C, but the joint result is constrained by the stricter review.

Effects and uncertainty. D–C medium/hard count differences are $0, 0, 3, -1, 1, 0, 2$ for chinook, imdb, instacart, sfdffd, tpcds, tpch and usaccidents. The mean per-slot difference is 0.02976. A fixed-seed 10,000-resample database bootstrap gives the descriptive interval $[-0.00595, 0.07143]$, crossing zero. It does not establish population generalization from seven selected databases. Every leave-one-database-out mean remains positive. D–B yields five wins, one loss and one tie, also with

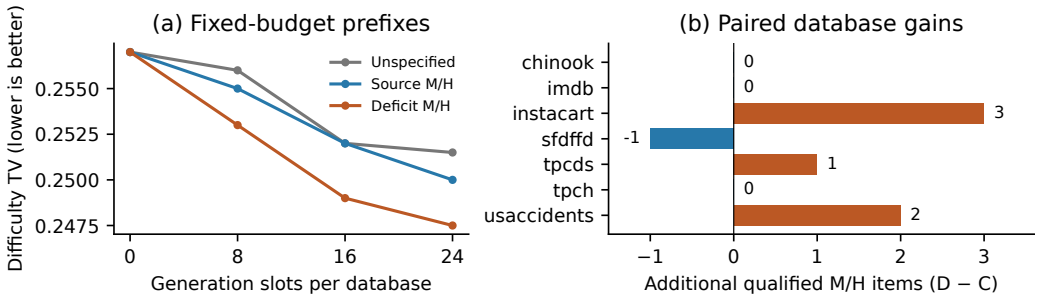


Figure 7: Real candidate acquisition under the initial joint review. (a) Difficulty TV after adding jointly model-supported candidates at fixed slot prefixes. (b) Paired medium/hard candidate gains of deficit versus source-conditional demand. Prefixes share one run; databases are development domains.

an interval crossing zero. The frozen development continuation rule requires complete collection, no complete SQL convergence and positive D–C macro yield and token efficiency. It is met, but is not a statistical superiority gate.

C–D lexical SQL is identical in 101/167 parseable slot pairs (60.48%); SQL-set Jaccard overlap is 43.63%. The manipulation changes some SQL but substantial convergence remains. Adding qualified candidates reduces difficulty TV from 0.257 to 0.2515/0.2500/0.2475 for B/C/D. These are exact integer optima on operational expanded pools; the original pool is not globally semantically certified. Prefixes in Figure 7 share one run and are not independent replications. No downstream TEF or ranking benefit is established.

Precision recheck and mechanism sensitivity. A follow-up review covers all 34 distinct contents contributing the initial 44 qualified medium/hard items across policies, without selecting by policy or dispute. Reviewers explicitly check population scope, aggregation unit, time boundaries, ordering effects, code meanings and task redundancy. Applying the same joint-support and nonredundancy requirement retains 8/10/12 items for B/C/D, compared with 11/14/19 initially. This is a precision analysis of the previously accepted subset, not an independent experiment or full-pool reannotation; neither review is treated as semantic gold. Original labels and counts remain available. The direction is also positive under either follow-up reviewer alone (B/C/D: 9/10/12 and 9/11/13).

The initial 35.71% yield gain and 37.35% token-efficiency gain describe the frozen first review; the follow-up retains a two-item gain (20%) and 21.45% higher aggregate yield per token. Its database differences are $0, 0, 1, -1, 1, -1, 2$, with descriptive interval $[-0.01786, 0.04167]$ per slot. Removing USAccidents makes the follow-up mean difference zero. Both review stages therefore support a positive observed direction, with limited precision. Recheck concerns include an ineffective ORDER BY on a scalar AVG, an unstated prior-split restriction, and an hour-truncated time boundary. These are concrete reasons to retain sensitivity results alongside the initial model labels.

A paired decomposition of the initial result places four of the five additional items in slots where the requested difficulty changed, and one in a slot with unchanged demand. This is descriptive, not causal identification: outputs may differ under the same demand and within-policy uniqueness depends on earlier outputs. Literal-masked SQL grouping leaves initial counts at 11/14/19; removing unreferenced output aliases gives 10/13/19. These checks do not constitute complete semantic deduplication. The acquisition benefit thus extends beyond simple constant changes, while the precision recheck bounds how strongly the observed gain should be interpreted.

Changed-generator reproduction and precision sensitivity. We repeat the same 7-database, three-policy, 504-slot prompt and budget protocol with the server-resident `Meta-Llama-3.1-8B-Instruct` directory as generator. Recorded file hashes bind the run, but the directory name does not independently authenticate publisher identity. All 504 slots and 546 requests, including 42 repairs, are settled with no pending request. Under the frozen parser, 262/504

Table 5: Semantic audit of a 300-item stratified release sample. The gate uses 60 hidden controls. Weighted rates extrapolate only to the fixed 2,000-item release. Hy3+SQL is a same-family configuration sensitivity.

Reviewer/configuration	Controls	Supported	Weighted
Qwen3.8-Max	59/60 pass	276/300	92.05%
Hy3 ordinary	54/60 pass	285/300	95.03%
Hy3+SQL expert	56/60 pass	289/300	96.33%
DeepSeek-V4-Pro	53/60 fail	265/300	88.39%

slots (51.98%) fail JSON formatting; we neither recover them post hoc nor rewrite the primary outcome.

Two blinded model reviewers each cover all 197 review items. User-provided session labels are HY3 and doubao2.1pro, while both returned `reviewer_id` fields are unknown; model versions are therefore not independently authenticated. Initial jointly supported, nonredundant M/H counts are 18/10/12 for unspecified/source-conditional/deficit demand. D–C is 12/10 (+20%) and uses 18.39% fewer generation tokens per item; database outcomes are three wins, two losses and two ties, with a descriptive per-slot bootstrap interval $[-0.02976, 0.05357]$. Unspecified is higher at 18, precluding an overall-policy dominance claim. Label agreement is 175/197 (88.83%), with Cohen's $\kappa = 0.7808$.

The precision sensitivity covers all 34 unique contents corresponding to the 40 initially retained M/H arm records. Requiring both new reviews to mark an item supported and without unnecessary complexity retains 15/8/8 items. Thus D–C is 8/8; deficit uses 2.07% fewer generation tokens per retained item. Across databases this is two wins, two losses and three ties, with interval $[-0.03571, 0.03571]$; IMDb remains in the analysis with zero output. Reviewer A alone gives C/D 10/10 and reviewer B 8/8, so the tie is not created only by intersection. Precision-review label agreement is 29/34 (85.29%), with $\kappa = 0.4426$; the selected subset differs from the 197-item universe, so the two agreement coefficients do not measure a change in reviewer ability.

Across generators, initial D/C counts are Qwen 19/14 and Llama 12/10; precision counts are 12/10 and 8/8. The initial model-review direction therefore repeats, but precision sensitivity does not establish a stable cross-generator yield advantage. Both runs use the same seven databases, and neither review is human annotation, semantic gold, downstream TEF validation, or new-schema evidence.

Per-policy Llama format attrition and conditional rates appear in Appendix M.

K STRONG-MODEL SEMANTIC AUDIT OF THE FIXED RELEASE

Frozen sample and evidence. We audit a deterministic 300-item sample from the 2,000-item cost-aware release. Sampling is proportional within domain–difficulty–provenance strata. Each blinded item contains the question, SQL, scoped SQLite DDL and stored-value samples for referenced tables, plus an execution preview. Execution status is evidence, not the review label. Sixty hidden controls calibrate each reviewer: 30 executable Spider development question–gold-SQL pairs and 30 executable within-schema question–SQL mismatches. The frozen qualification gate requires at least 54/60 exact control decisions, at most four false rejections among valid controls, and at most two false acceptances among mismatches.

Review configurations. Qwen3.8-Max and DeepSeek-V4-Pro are called through their provider APIs. Two WorkBuddy runs use isolated workers for the same twelve 30-item batches. The first uses the ordinary review prompt; the second uses a user-designated SQL expert-agent configuration. The expert run's receipt binds its workers to Hy3. For the ordinary run, Hy3 is user-reported because the platform receipt exposes only the WorkBuddy runner. The two Hy3 configurations share a model family and are not independent replications. All final outputs cover 360/360 IDs with valid schemas; raw replies, receipts and hashes are retained.

Qualification and target judgments. Qwen passes 59/60 controls, ordinary Hy3 passes at the 54/60 boundary, and the SQL-agent configuration passes 56/60. DeepSeek scores 53/60 and fails

because it rejects seven valid controls, although it rejects all 30 mismatches. We keep DeepSeek as sensitivity evidence and do not replace it or change the gate. On the 300 release targets, the three qualified configurations support 276, 285 and 289 items, respectively. Qwen and ordinary Hy3 jointly support 270/300; the stratum-weighted release estimate is 0.9007 with a stratified bootstrap 95% interval $[0.8677, 0.9306]$. Qwen and the SQL-agent configuration jointly support 273/300, with weighted estimate 0.9104. Requiring Qwen and both Hy3 configurations gives 269/300 and weighted estimate 0.8973. Ordinary and SQL-agent Hy3 agree on 290/300 target labels (96.67%).

Scope. This audit adds semantic-validity evidence for the fixed release, but it is a model panel rather than human annotation or semantic gold. The controls test question–SQL discrimination on Spider schemas; they do not certify all target domain meanings. Sampling uncertainty does not cover model dependence or evidence omissions. The audit also does not test new-schema generalization, repair the failed external absolute-fidelity gate, or turn the generation study into a stable policy-superiority result. Because DeepSeek fails qualification, the originally frozen all-reviewers-qualified endpoint is not met.

Qualitative control-case inspection in Appendix M separates possible reference ambiguity from SQLite-specific reviewer errors; qualification remains defined by the original frozen control key.

L MULTI-BLOCK SUPPORT AND CONSTRUCTION COST

Notation guide. N is release size, N_d is its domain quota, D is the number of domains, r_d counts replacements, and H bounds response-profile change. L_{struct} is the five-block structural loss; TEF measures response-profile discrepancy. δ_{sup} is a convex support certificate and D_T^* the integer optimum for the declared structural distance. These quantities answer different questions: feasible support, selection quality, and measurement.

Beyond difficulty. We retain the original feature extractor, 7,000 source SQL queries, 2,806 candidates and the original domain–provenance quotas totaling 2,000 items. For each coordinate, sorting within strata gives exact feasible extrema. Operator shares normalize by total operator events; their extrema use a linear-fractional residual rather than per-item normalization. We independently replay all 32 extremal releases, their quotas, and within-stratum exchange optimality. This audit uses SQL features only.

Coordinate	Source	Feasible minimum	Feasible maximum
Easy fraction	.4640	.7210	.9140
Filter fraction	.5003	.5905	.8930
Join fraction	.3959	.1560	.3570
Group-by fraction	.2536	.0245	.2035
Aggregation fraction	.4671	.0895	.3640
Order-by fraction	.2326	.0165	.1925
Limit fraction	.1581	.4610	.6660
Mean join count	.5419	.1560	.4325
Mean SQL tokens	15.8799	10.4450	14.0010

Summing valid block lower bounds under the original equal-block metric gives $L_{\mathrm{struct}} \geq 0.1452323$. All 34 checked historical releases satisfy it. Individual extrema are exact; their combination need not be jointly attainable, so this is not a claim of a tight joint optimum or a TEF bound. The resulting acquisition directions include fewer LIMIT-bearing candidates and greater support for meaningful aggregation and joins, subject to task requirements.

Construction expense under precision sensitivity. The table uses all frozen prefixes and includes failed attempts and repairs. Items are retained by the follow-up model-review sensitivity analysis; token costs exclude external reviewing and server startup.

Policy	Slots/domain	Retained M/H	Generation tokens
Unspecified	8	1	55,802
	16	7	109,898
	24	8	171,417
Source-conditional	8	4	60,716
	16	7	118,627
	24	10	183,318
Deficit	8	6	61,800
	16	12	118,013
	24	12	181,135

At 24 slots, deficit demand uses 17.66% fewer tokens per retained item than source-conditional demand and 29.55% fewer than unspecified demand. These are descriptive construction costs on the same reviewed run. The plateau after 16 slots motivates a prospective stopping-policy test; it does not validate an adaptive stopping rule retrospectively.

M Additional diagnostics and sensitivity analyses

Paired uncertainty. This post-hoc analysis retains all ten frozen seed pairs. We bootstrap paired seed differences with replacement 50,000 times (seed 20260919), using percentile endpoints. It conditions on the shared pool and panel; it is not a prospective test or a confidence interval over new schemas. No seed, threshold or method is selected using these intervals.

Cost-aware minus	Mean difference	95% paired-seed interval
Random	-0.002841	$[-0.003379, -0.002337]$
Greedy	-0.001463	$[-0.003170, 0.000032]$
Local	-0.008715	$[-0.011254, -0.005982]$
Absolute gain	-0.000518	$[-0.000817, -0.000193]$

Llama attrition by demand. All counts retain the original 168 slots per policy, including parsing failures. No parser repair or generation rerun is introduced.

Stage	Unspecified	Source-cond.	Deficit
Format failure	69 (41.07%)	101 (60.12%)	92 (54.76%)
Parse success	99	67	76
Nonempty execution	58	34	37
Nonempty medium/hard	34	29	29
Initial joint retention	18	10	12
Precision retention	15	8	8
Precision / parse success	15.15%	11.94%	10.53%

Uneven parse attrition is a concrete confounder of semantic-capacity comparisons. The unspecified arm still retains more items conditional on parse success, but that conditioning selects different subsets and is not a causal adjustment. Targeted arms show no stable end-to-end precision-yield advantage. All generation and repair tokens remain in the cost denominator.

SQL-only difficulty and planner. The shared score adds join count, twice the additional SELECT count, twice the UNION/INTERSECT/EXCEPT count, and indicators for GROUP BY, HAVING, aggregation, ORDER BY, CASE, DISTINCT and at least two AND/OR tokens. The thresholds are ≤ 1, 2–3, and ≥ 4. This transparent lexical proxy does not implement the official Spider hardness evaluator. The programmatic planner samples one of filter, aggregation, join, grouping or top-k using the configured operator weights, multiplied by $1 + \max(0, \text{coverage gap})$ when enabled. Seeded templates choose schema columns, observed predicate values and available join edges. Numeric columns support aggregates; grouping and top-k add grouping/order clauses. Template fallbacks can simplify infeasible plans, so final SQL is rescored independently of requested difficulty. LLM realization then verbalizes the SQL. Execution and admission checks do not make these templates a novel general SQL synthesis algorithm.

Scale and workflow limits. Read-only SQLite execution uses a progress-handler time/step budget and a result-row cap. These bound operational exposure rather than eliminate expensive scans or prove latency scaling on dirty BIRD-scale data. Timeouts remain failures, not successful validation. The item-law formulation assumes standalone queries: stateful multi-turn enterprise workflows require a trajectory-level instrument and are outside the demonstrated scope.

Audit follow-up scope. The existing qualified-pair intersection and the stricter Qwen–both-Hy3 intersection are retained. Reusing hidden controls can diagnose disputed cases but cannot retroactively change the original qualification gate. Any additional channel review is a separately frozen, post-hoc diagnostic; no pending review is counted as a result.

Domain substitution diagnostic. For each seed, substitute only one domain's local-search membership into the random anchor, retaining all domain quotas. The resulting contrasts isolate membership changes, not semantic causes. Their averages are:

Domain	Replacements	One-domain TEF change
chinook	102.9	-0.001139
imdb	44.3	0.000491
instacart	27.3	0.000080
sfdffd	11.3	-0.000130
tpcds	220.9	-0.002945
tpch	108.1	-0.002577
usaccidents	57.3	0.000535

Despite several negative one-domain contrasts, applying all local memberships together raises TEF by 0.005874. Absolute deviations after pooling make these contrasts nonadditive: partial shifts can reduce a discrepancy whereas their combined shift can overshoot it. Direct coordinate inspection finds 3–6 coordinates per seed that cross the source value and end farther from it than the random anchor. In contrast, cost-aware exchange changes only 14 TPC-DS items per seed, with full TEF change -0.002841. This describes this pool's intervention pattern, not a universal explanation of semantic quality.

Disputed control cases. Inspection of the seven DeepSeek-rejected valid-control labels identifies heterogeneous issues. A "youngest dog" reference uses MAX(age), an apparent reference-direction problem under the ordinary age interpretation. Another asks for years with cars below 4000 and cars above 3000, while its SQL requires both conditions on one row; these readings are not generally equivalent. Conversely, rejecting a bare language column next to a single built-in MAX(Percentage) overlooks SQLite's documented argmax-row behavior (ties may select any maximizer).[2] Other objections concern zero-concert singers, area-code/state mapping, "popular" language semantics and padded airport codes. These require scope or data interpretation rather than an automatic verdict. Three in-memory SQLite checks confirm the MAX(age) counterexample, the single-MAX behavior and the non-equivalence of the two car readings; they do not relabel the original controls. Thus frozen qualification is operational agreement with the control key, not proof that every control is indisputable or that all rejections are reasoning failures. All original gates remain unchanged.

N Fresh construction and reference-based measurement

Construction comparisons. The fresh-SQL study compares a template generator with a deterministic schema-aware refill implementation of the Schema2Bench planner. With 1,080 attempts per method across IMDb, Instacart, SFDFFD, TPC-DS, TPC-H and USAccidents, the static-and-dual-review supported counts are 363 and 383, respectively. The corresponding JOIN counts are 17 and 109. These are attempt-denominated construction outcomes, not a comparison of two selected equal-size releases. The dual-review label is model support for a question–SQL pair after execution and static screening.

[2] `https://sqlite.org/lang_select.html`

An additional component comparison uses two generators, seven domains and 168 slots per generation strategy. Explicit planning reduces the mean three-level difficulty-support TV by 0.03050 for Qwen and 0.01925 for Granite; an OmniSQL prompt-component adaptation obtains 0.01200 and 0.01225 under the same harness. Generation-token use is 23.28% and 24.57% lower, respectively, for the explicit-planning arm. This compares planning and prompt components rather than the complete external synthesis systems.

Reference comparison. The measurement experiment uses BIRD card_games and thrombosis_prediction. The reference has 171 and 161 question–SQL items, respectively. Each construction method uses three generation seeds and 100 attempts per seed–domain cell. Before reference questions or evaluator responses are opened, each method supplies an 80-item provisional pool per domain (60 JOIN, ten subquery and ten other queries). Dual-reviewed items then fill a 50-item release per method and domain (40 JOIN, five subquery and five other). All 200 selected constructed items receive support from both version-identified semantic reviewers (DeepSeek-V4-Pro-0813 and GLM-5.2) and have executable, nonempty reference SQL. The two methods share the subquery-generation control, with disjoint selected SQL; results that exclude this common control are reported below.

The fixed eight-model panel is Qwen3-4B, Granite-3.3-8B, Qwen2.5-Coder-7B, Qwen2.5-7B, Llama-3.1-8B, Phi-4, Mistral-7B and Qwen2.5-14B. All arms use the first complete SELECT/WITH statement from the saved model output and the same execution scorer. Let $a^r_{d,m}$ and $a^c_{d,m}$ be model m's execution accuracies on reference and constructed items in domain d. The reported loss is

$$G_d(c) = \binom{M}{2}^{-1} \sum_{m<n} \left| (a^r_{d,m} - a^r_{d,n}) - (a^c_{d,m} - a^c_{d,n}) \right|. \tag{35}$$

The improvement is $G_d(\mathrm{Template}) - G_d(\mathrm{Schema2Bench})$, averaged equally over the two domains. It assesses model-to-model discrimination, not absolute mean difficulty. The 2,000-draw item and SQL-skeleton bootstraps resample each arm within domain and report percentile intervals of the domain-mean improvement.

Count-unit correction. The first version of the thrombosis JOIN pool exposed a count-unit error: `COUNT(*)` over a one-to-many Laboratory or Examination join counts joined rows when the question asks for Patient records. After this semantic diagnosis, 39 affected SQL statements were changed to `COUNT(DISTINCT Patient.ID)`, re-executed and rejudged by both models before the reference questions and evaluator scores were opened. The original semantic-gate result is retained; the main-text measurement result belongs to this corrected construction variant. No reference-based score was used to select or edit the corrected items.

Table 6: Sensitivity of the domain-mean pairwise-gap improvement (Template minus Schema2Bench). Positive values favor Schema2Bench. Intervals use 2,000 resamples within each reference and constructed arm.

Panel and item set	Improvement	Item 95% interval	Skeleton 95% interval
Eight models, all items	+0.0650	[0.0159, 0.1122]	[0.0143, 0.1063]
Six-model sensitivity	+0.0554	[0.0038, 0.1136]	[0.0005, 0.1102]
Eight models, native items	+0.0665	[0.0122, 0.1177]	[0.0120, 0.1190]

The six-model sensitivity excludes Granite and Phi-4; its domain effects are -0.0078 on card games and $+0.1187$ on thrombosis prediction. The native-item comparison excludes the shared EXISTS/subquery control and remains positive in both domains. Mean eight-model accuracy differs from the reference: card games has 0.2478 on reference items, 0.4625 on templates and 0.5950 on Schema2Bench items; thrombosis prediction has 0.2399, 0.3950 and 0.3100. Thus the pairwise-gap result does not imply equal absolute difficulty. The reference comprises two public BIRD databases; the three construction seeds do not create three independent database samples.

Later four-database operator-matched comparison. A separate, exploratory version uses card games, thrombosis prediction, financial and California schools. Its raw pool contains 7,634 generated records. After response-blind static exclusions, template and deterministic refill candidates are paired

before review on SQL family, WHERE presence and aggregate function, with nearest SQL-token complexity and a skeleton cap. The 316 provisional pairs (632 items) receive fresh judgments from local Qwen2.5-32B and Llama-3.3-70B; previous labels are diagnostic only. The questions in both arms uniformly clarify SQL sort directions, DATE meaning where relevant and IN-subquery membership. Both judges pass their frozen 22-per-label control gate, and 619/632 items receive joint support. Each arm then fills 30 JOIN+LIMIT, five subquery and three other items per domain; all semantic, diversity and complexity gates pass before opening the public BIRD reference or evaluator responses. The four reference sets contain 171, 161, 101 and 85 items, respectively.

The nine-model domain-mean pairwise-gap errors are 0.2912 for template and 0.2150 for refill, an improvement of 0.0762. The item-bootstrap interval is $[0.0372, 0.1166]$ and the SQL-skeleton interval is $[0.0300, 0.1148]$ over 2,000 draws. Card games, thrombosis and financial contribute $+0.0518$, $+0.0808$ and $+0.1832$; California schools contributes -0.0112. The original seven-model sensitivity gives $+0.0831$. The earlier two-database experiment above uses a different item set, quota and model panel; its scores are not pooled with this version. This version followed a frozen semantic STOP in the immediately preceding development attempt, and is not an untouched-domain confirmation. Operator pairing does not make questions semantically identical. The constructed sets are easier in mean evaluator accuracy than their public references; model-judge agreement is not human semantic gold. The result supports a bounded, descriptive comparison of two deterministic variants, not the full historical multi-agent or official generation systems.

Table 7: Earlier four-domain operator-matched comparison. Each method contributes 38 items per database; nine models evaluate both methods and the reference. Improvement is Template minus Schema2Bench.

Reference database	Template	Schema2Bench	Improvement
Card games ($n = 171$)	0.2351	0.1832	+0.0518
Thrombosis ($n = 161$)	0.3494	0.2686	+0.0808
Financial ($n = 101$)	0.2648	0.0816	+0.1832
California schools ($n = 85$)	0.3154	0.3266	-0.0112
Domain mean	0.2912	0.2150	+0.0762

Exploratory eight-domain max-pool expansion. This later version reuses the frozen raw candidate pool after a previous 200-item-per-domain STOP. It attempts all 11 available BIRD dev domains, pairs candidates before semantic review by the same SQL operator signature, and applies the frozen skeleton cap. Eight domains pass the prespecified eligibility gates. Student club (199/200 supported pairs), toxicology (51/200) and debit card specializing (84/200) fail before evaluator scoring; they are excluded by eligibility, not counted as measurement reversals. The eligible domains supply 3,835 pairs per method and 1,118 public reference items. Ten models score all 8,788 items (87,880 score records).

Table 8: Eight-domain nine-model sensitivity. The n column counts paired items per constructed method. Errors compare pairwise model-accuracy gaps against the public BIRD reference; improvement is Template minus refill.

Reference database	n/arm	Template	Refill	Improvement
Card games	460	0.2237	0.1577	+0.0660
Thrombosis	311	0.3136	0.2041	+0.1096
Financial	383	0.2034	0.1191	+0.0844
California schools	703	0.3809	0.3865	-0.0056
Formula 1	484	0.1203	0.0980	+0.0224
Codebase community	373	0.1718	0.1324	+0.0394
European football	613	0.1784	0.1594	+0.0191
Superhero	508	0.1454	0.1218	+0.0237
Equal-domain mean	3,835	0.2172	0.1724	+0.0449

The nine-model item-bootstrap interval for the item-weighted improvement is $[0.0334, 0.0522]$. Equal-skeleton resampling weights each distinct SQL skeleton once rather than by its item count; its improvement is $+0.0358$, with percentile interval $[0.0229, 0.0447]$. Seven domains improve. With all ten prespecified models, equal-domain errors are 0.2369 and 0.2173 ($+0.0197$), with an item-bootstrap interval $[0.0097, 0.0267]$ and an equal-skeleton resampling interval $[0.0033, 0.0225]$. Only five of eight domains improve, below the frozen six-domain gate, so the primary descriptive measurement gate fails. Phi-4 returns empty raw output for 8,779/8,788 prompts. Its original scores remain in the ten-model analysis; the nine-model panel is a sensitivity analysis rather than a replacement primary endpoint. The cause of the empty outputs is unresolved. Both arms here are deterministic variants, and the experiment provides no independent-dataset or human semantic validation.

Earlier measurement diagnostics. An earlier three-domain BIRD comparison produced a domain-mean improvement of $+0.0181$ with an item-bootstrap interval $[-0.0385, 0.0665]$; the subsequent two-domain experiment above has a distinct item set and corrected construction protocol. A six-domain 99-item-per-method comparison initially showed a 5.81-point raw accuracy difference, concentrated in USAccidents. Swapping the projection columns in a 16-item-per-source crossover reduced the corresponding matched-set difference to approximately 0.25 points. These diagnostics motivate evaluating reference-based model gaps, rather than using unadjusted synthetic-set accuracy as a construction-quality measure. Historical external absolute-fidelity and worst-model-deviation outcomes remain separate endpoints.

External-label reviewer diagnostic. On all 2,018 released NL2SQL-BUGs examples, Qwen3-4B and Granite-3.3-8B classify the published correctness labels with 71.31% and 61.30% accuracy, respectively, counting invalid responses in the full denominator. A separate 220-item diagnostic across 11 databases gives 83.18% for a Hy3 channel. These reviewers and prompts are not identical to the fixed-release Qwen3.8-Max–Hy3 audit or the two reviewers in the BIRD measurement experiment. The diagnostic therefore measures an important source of automatic-review uncertainty without converting model agreement on new items into semantic gold.