SEPARATING ORDERING, CALIBRATION, AND AUDIT EVIDENCE IN TEXT-TO-SQL ACCEPTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-SQL acceptance can improve without changing SQL candidates or the judge body. A SOURCE-trained Qwen residual readout raises disjoint WikiSQL coverage from 2.066% to 3.327% and reduces accepted-error risk from 7.104% to 2.667%. Producer-aware native-prefix controls and matched scalar-spline readouts retain the ordering advantage on the fixed bank. SOURCE-database exclusions preserve fixed-quota gains but reverse CAL-selected correct-mass gains in two of eight conditions. Across 256 paired calibration draws, scalar-plus-residual correct-mass gain over the scalar readout averages 1.264 percentage points and reverses in 50 draws; the residual policy exceeds one-percent risk in 225. Exact acceptance decompositions and without-replacement confidence sequences separate correct-mass gain, risk reduction, and strict dominance. Every-question inspection resolves correct-mass gain earlier; risk reduction and dominance generally require nearly all accepted-union questions in these replays. Calibration-component matching supplies no consistent benefit. Frozen-readout gains on new Mistral answers to exposed questions remain unestablished. Producer-specific adaptation establishes neither incremental ordering gain nor a sufficiently supported one-percent endpoint. These controls support conditional ordering improvement while separating operating-point stability, portability, and claim-specific audit evidence.

1 INTRODUCTION

Text-to-SQL systems can return plausible queries with incorrect results. Selective acceptance offers a useful response: answer only when an estimated error risk is sufficiently low. An acceptance claim combines whether a score ranks candidates usefully, whether a selected threshold transfers, and whether outcome evidence supports the requested risk statement. Different controls answer these questions.

The strongest SOURCE-trained readout using the Qwen judge increases disjoint WikiSQL coverage by 1.261 percentage points and reduces risk by 4.437 points, yet leaves 2.667% risk. The same thresholds yield .996% point risk on unused questions in calibration-pool components, with an interval extending above two percent. Neither relative gain nor a favorable point estimate establishes deployment success. Comparing different roles also cannot establish a calibration-matching cause; the required crossover holds evaluation questions fixed.

Figure 1 separates readout changes, calibration-donor swaps on identical evaluation questions, and label sampling of fixed policies. Changed/shared acceptances connect the first and third interventions; SOURCE refits and new-producer adaptation test stability and transfer.

2 RELATED WORK AND SCOPE

Selective classification already formalizes rejection and the risk–coverage tradeoff (Geifman & El-Yaniv, 2017; 2019; Franc et al., 2023; Gangrade et al., 2021). Confidence calibration, calibrated selection, and evaluation under distribution shift provide complementary tools (Guo et al., 2017; Fisch et al., 2022; Liang et al., 2024). Recent work also emphasizes strong confidence baselines and

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.

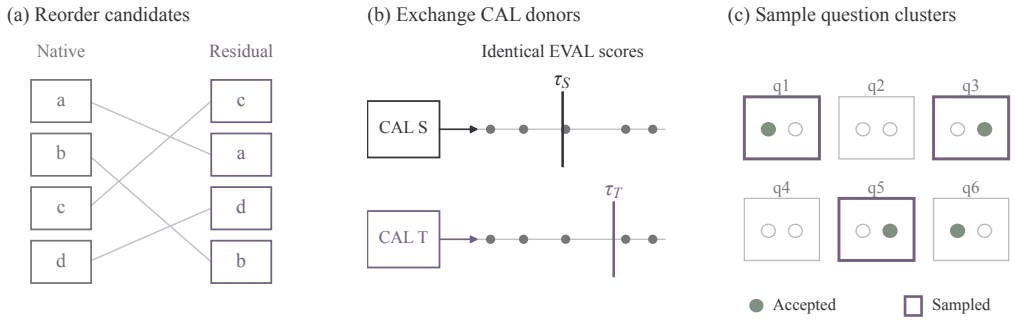

Figure 1: Three interventions on distinct objects. (a) SOURCE changes the ranking of the same frozen candidates. (b) CAL donors provide different thresholds for identical EVAL scores. (c) Sampled question clusters supply both producer outcomes for a fixed acceptance policy; filled dots mark acceptance and plum outlines mark sampled questions. Candidates, cutoffs and clusters are schematic, not measured proportions. Ordering gain, threshold transfer and audit precision remain separate claims. Errors shared by both policies cancel from the difference in wrong-answer mass but remain in absolute risk. No intervention alone establishes one-percent certification.

evaluation choices (Cattelan & Silva, 2024; Rabanser & Papernot, 2025). The present contribution is not a new abstention rule or general risk-control theorem. The contribution is a matched experimental account of *which intervention changes accepted answers, which competing explanation remains possible, and which stronger reliability claim does not follow*.

Relation to reliability and calibration methods. Multicalibration (Detommaso et al., 2024), conformal language modeling (Quach et al., 2024), and penalty-based SQL benchmarks (Lee et al., 2024) address different contracts from fixed-candidate one-percent accepted-error risk. Historical exposure and selection boundaries remain explicit (Cawley & Talbot, 2010); additional fits or replay draws are not fresh populations.

3 A FROZEN-CANDIDATE EXPERIMENTAL CONTRACT

Objects and roles. The original banks have two fixed SQL candidates per question, one per producer family. The later Mistral panel contains one new candidate per selected question and retains a separate component–question denominator. A producer supplies SQL candidates; a judge scores those candidates. SOURCE fits a readout that transforms judge scores and permitted features into an acceptance score. CAL combines that score with a threshold or a fixed mixture of whole score prefixes to define an acceptance policy. EVAL determines reported outcomes, never readout fitting or threshold selection. Later analyses of existing banks are explicitly exposed-data diagnoses, not fresh independent confirmations. Public benchmark train membership does not establish absence of language-model pretraining exposure.

DIRECT supplies the question, schema, and candidate; VALUES additionally supplies a bounded packet of returned values. The corresponding model bodies differ. Consequently, DIRECT–VALUES is a package comparison, not a causal intervention that adds information to an otherwise identical model. The arithmetic-only comparison in Section 7 instead holds captured hidden states and output-head weights fixed.

The benchmarks cover different text-to-SQL regimes (Zhong et al., 2017; Yu et al., 2018; Li et al., 2023); separate rosters prevent inappropriate pooling (Appendix B). The primary frozen WikiSQL cohort contains 28,132 questions: 8,192 CAL questions in 2,048 components and 19,940 EVAL questions in 6,155 component-disjoint units. An additional 3,996 unused questions belong to calibration-pool components. “Pool-known” means membership in the full CAL pool, not that every sampled calibration design has seen every component.

Weights, unknown outcomes, and risk. For a fixed role with D databases and n_d questions in database d, the candidate weight is $w_{dqp} = 1/(2Dn_d)$. The WikiSQL analyses use the declared component/page roster rather than asserting independent production environments. Question blocks retain both producers. Let $A_i \in [0,1]$ be acceptance probability and $E_i \in \{0,1\}$ the execution-error indicator. The finite-bank quantities are

$$C = \sum_i w_i A_i, \quad N = \sum_i w_i A_i E_i, \quad G = C - N, \quad R = N/C, \quad V_\alpha = N - \alpha C, \qquad \alpha = .01. \tag{1}$$

Risk is undefined at $C = 0$. The displayed zero-acceptance decisions mean abstention, not zero-risk successful deployment. Unresolved executions retain their original weights. Completing unresolved outcomes as correct or wrong gives $N^- \le N \le N^+$ and corresponding bounds for risk and violation. These are identification bounds, not confidence intervals. For randomized policies, Equation 1 integrates acceptance probabilities; R is a ratio of expected masses, not the expected realized ratio.

Execution matching tests a stated SQL-result endpoint, not unrestricted semantic equivalence. Test-suite evaluation can strengthen this endpoint (Zhong et al., 2020); the present fixed snapshots cannot retroactively acquire that guarantee. Unresolved outcomes retain their original weights under both correct and incorrect completions.

Three levels of evidence. A relative gain concerns paired changes in G, N, C, or an equal-coverage contrast. Empirical feasibility requires $N^+ \le \alpha C$ on the stated finite bank. Statistical validation additionally requires a valid sampling law and uncertainty procedure. A separate engineering screen of at least 300 accepted questions and 100 components helps avoid vanishing-support operating points, but is neither a theorem nor a certificate. Conformal risk control and learn-then-test provide existing tools under explicit assumptions (Angelopoulos et al., 2024; 2025); hierarchical and betting-based inference emphasize why the sampling unit matters (Lee et al., 2023; Waudby-Smith & Ramdas, 2024). These guarantees are not inferred merely from a point estimate below one percent.

4 A RESIDUAL ORDERING GAIN SURVIVES STRONGER NATIVE CONTROLS

Nested, separately refitted readouts. The readout ladder contains native score alone (N), native plus a public producer feature (NP), native plus an orthogonal-residual hidden-state score (NR), and the full stack (NRP). Each permitted head is refitted on the same 857 eligible SOURCE candidates from eight databases, with common regularization $\lambda = .1$ and four database-held-out folds. The judges are Qwen3-8B with a fixed VALUES adapter and Granite-3.3-8B-Instruct; both expose 4,096-dimensional final-token states. N uses the native YES $-$ NO margin. Appendix B specifies the projection, fit, and CAL rule. Original NRP coefficients and SOURCE/BIRD scores reproduce exactly; maximum WikiSQL projection difference is 2.665×10^{-14}.

The producer feature has a consequential limitation. The SOURCE indicator marks the designated SOURCE producer; both current WikiSQL producer families map to zero. NP and N therefore produce identical target acceptance masks. This is a valid reproduction of the historical feature, but not a test of separate target-producer offsets. NR captures the main gain; the present ablation does not establish that access to actual target-producer identity would be useless. Section 4.1 gives that identity to a stronger hindsight native control instead.

The positive result and its boundary. On component-disjoint EVAL, N accepts 2.065953% coverage at 7.103948% risk, whereas NRP accepts 3.326971% at 2.667192% risk (Figure 2). NRP accepts 1,127 unique questions across 876 components; the support screen passes but the empirical risk criterion fails. NR is similar, with 3.344% coverage and 2.653% risk. A fixed-mask paired component bootstrap estimates NRP $-$ NP correct accepted mass at $+1.3190$ percentage points (95% interval $[1.1029, 1.5394]$) and wrong accepted mass at $-.0580$ points ($[-.1105, -.0078]$). These intervals condition on the score and CAL fits. Refitting all readout stages after each of eight SOURCE-database exclusions preserves every disjoint fixed-3%-quota ordering gain, but reverses CAL-selected correct-mass gain in two conditions. NRP risk remains above one percent throughout. The matched scalar-plus-residual control has one empirical sub-one-percent exception, at the cost of correct mass (Appendix B.1). Repeated CAL draws retain a positive mean Qwen gain but not a stable one-percent endpoint (Appendix G).

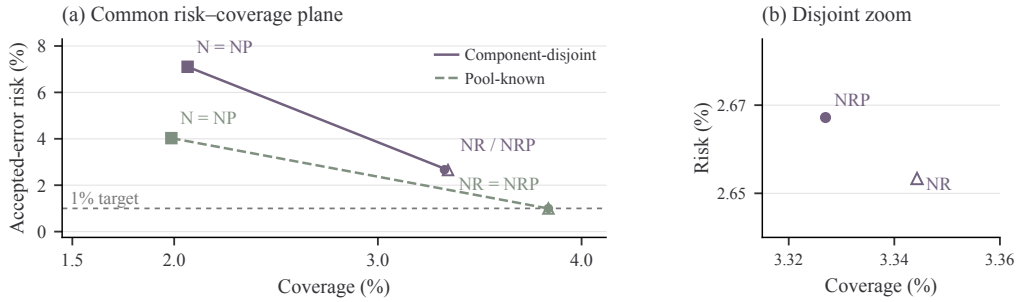

Figure 2: Frozen Qwen VALUES endpoints under the original CAL thresholds. (a) Both EVAL roles share a risk–coverage plane; shading marks risk below the one-percent target. Squares show N = NP, triangles NR and circles NRP; links identify readouts, not a threshold sweep. (b) The disjoint zoom separates NR from NRP; pool-known NR and NRP coincide. Pool-known NRP accepts 250 questions from 201 components; its .996% point risk has a conditional 95% interval of $[.124, 2.296]\%$. Neither endpoint certifies the target.

Table 1: Hindsight native controls at fixed Qwen NRP coverage. All quantities are percentages for coverage and percentage points for masses. The producer-aware family is stronger; both positive gaps are finite-bank ordering evidence, not one-percent certification.

EVAL role	Native family	c (%)	N^-_{oracle}	N^+_{NRP}	Γ
Disjoint	Global mixtures	3.327	.2301	.0887	+.1413
Disjoint	Producer-aware	3.327	.1848	.0887	+.0960
Pool-known	Global mixtures	3.837	.0946	.0382	+.0564
Pool-known	Producer-aware	3.837	.0898	.0382	+.0516

On pool-known questions, NR and NRP coincide at 3.837072% coverage and .995871% risk, accepting 250 questions in 201 components. The original descriptive risk interval, $[.124, 2.296]\%$, crosses one percent. Thus the point estimate is empirically feasible on this role, the 300-question screen fails, and statistical validation remains absent. A favorable relative gain and inadequate certification support coexist.

4.1 A PRODUCER-AWARE HINDSIGHT COUNTEREXPLANATION

Native-score calibration can improve substantially without a residual model. A fair control must therefore admit more than one deterministic threshold. For each frozen native score, every whole-tie prefix, including reject-all, defines a point (C, N). Arbitrary mixtures form a convex feasible region. The lower hull gives the smallest native error mass at a specified coverage. A producer-aware variant permits separate prefix mixtures within each actual target producer, preserving the original weights. This privileged class contains the global class and has target-label hindsight unavailable to a deployable SOURCE/CAL procedure.

Two deterministic thresholds on one score chain have nested acceptance sets: lowering a threshold adds nonnegative correct and wrong mass, while raising the threshold removes both. Such a change cannot produce strict $\Delta G > 0$ and $\Delta N < 0$ simultaneously. Randomized mixtures need not have nested masks, so excluding deterministic thresholding alone leaves a stronger explanation untested. The envelope closes that finite-family gap.

At the *already fixed* NRP coverage c, define the conservative gap

$$\Gamma(c) = N^-_{\text{native oracle}}(c) - N^+_{\text{NRP}}(c). \tag{2}$$

Positive Γ excludes the declared native family as an explanation of that endpoint even under favorable completion of unresolved outcomes for the baseline. Coverage is not selected anew to maximize the gap.

Table 1 rules out even the producer-aware native explanation for the primary Qwen endpoints. The corresponding Granite VALUES gaps are not positive. Consequently the evidence supports an architecture/package-dependent ordering gain, not a universal mechanism. Nor does $\Gamma > 0$ identify semantic reasoning: hidden-state signals can encode many useful correlates. A score ordering that beats this finite class need not meet a stringent deployment constraint.

A stronger supervised scalar control. Native-prefix mixtures preserve the scalar ordering; a nonmonotone scalar function need not. Two matched logistic readouts therefore use the same 857 SOURCE candidates, original out-of-fold residual r, and $\lambda = .1$: $S = g(m)$ and $SR = g(m) + \beta r$, where m is the native margin and g is a five-knot spline with SOURCE-only, label-free quantile knots. Preprocessing, extrapolation, and regularization are frozen before CAL/EVAL. The body and residual direction are unchanged. This control is distinct from the earlier MLP, whose inputs include hidden residual features.

At 3% expected coverage *within each producer*, Qwen SR reduces wrong mass relative to S by .04734 and .06559 percentage points (Table 3). These differences use the full two-producer denominator, not a renormalized accepted subset. Quotas use scores only; 3% coverage is not the 1% risk target. Original-CAL SR has 2.94929% coverage and 1.91421% risk on disjoint EVAL, still above target. Granite's pooled 3% contrast is $+.03741$ points, with 95% interval $[-.01005, .08539]$. The spline family permits nonmonotonicity, but every fitted S segment has positive slope. Thus the observed advantage exceeds this fixed scalar recipe, not every nonmonotone function of the native score. Hidden-state necessity, semantic reasoning, and a universal scalar impossibility remain unestablished.

5 Calibration matching does not close the remaining gap

The .995871% pool-known and 2.667% disjoint risks suggest a calibration-support question, but comparing two different outcome populations cannot answer that question. A reciprocal crossover changes only the CAL donor while keeping each EVAL column fixed. The exposed question census is partitioned, using label-blind component size, schema, and question-length metadata, into component sets S and T. Each set contains 1,024 eligible components. EVAL-S and EVAL-T contain 3,279 and 3,278 questions, respectively; CAL reserves have no question overlap with their EVAL columns. Both producers remain together.

Donors supply 512 or 2,048 unique CAL questions, two per component. Eight overlapping draws are dependent sensitivity analyses, not independent replications. Component–question–producer-uniform weights differ from the preceding full-bank denominator. Every donor policy is evaluated on both columns without retuning.

For readout f, the matching advantage is

$$D_f = \tfrac{1}{2}\{[V_S^+(f,T) - V_S^+(f,S)] + [V_T^+(f,S) - V_T^+(f,T)]\}, \qquad I = D_{\mathrm{NRP}} - D_{\mathrm{NP}}. \tag{3}$$

Positive D_f favors matching; positive I means that NRP benefits more from matching than NP under this specified contrast. The contrast uses V^+, the violation bound obtained by treating unresolved outcomes as errors. This bound differs from a confidence-interval endpoint and from the observed range across CAL draws.

For Qwen VALUES, NRP's mean D is $-.02578$ percentage points at budget 512 and $-.03491$ at 2,048. The eight-draw ranges are $[-.06409, +.03533]$ and $[-.05675, -.01248]$. NP's corresponding means are $-.00187$ and $-.00676$; the mean interactions I are therefore $-.02392$ and $-.02815$ points. No NRP draw improves both evaluation columns at either budget (Figure 3). A donor can simply supply a better threshold for both columns, rather than benefit its matching component set specifically. This result does not prove that calibration support never matters; the result rejects a positive matching account for these fixed rosters and draws.

In the fixed first draw (Table 2), CAL-T increases NRP coverage in both EVAL columns, not only the matched column. The zero-risk NP cell accepts six questions. Neither that cell nor the larger zero-error NRP endpoint meets the full support screen.

Across the full crossover, 640 threshold fits yield 1,280 evaluation cells. The record includes 435 abstentions and 209 positive-coverage cells satisfying the empirical criterion robust to unresolved

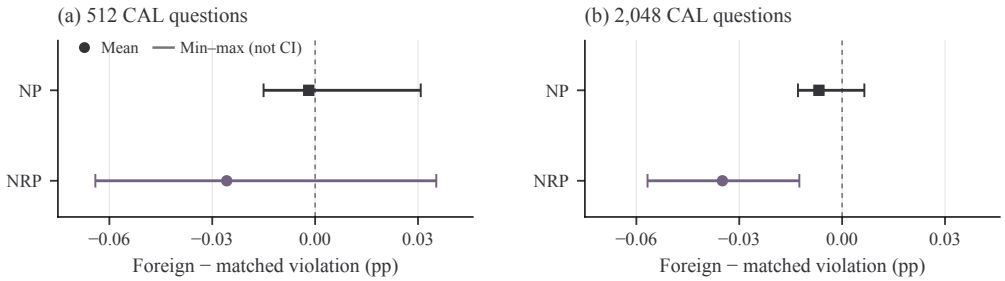

Figure 3: Same-EVAL reciprocal CAL crossover at 512 and 2,048 CAL questions. Points show means; capped horizontal lines show observed minima and maxima across eight dependent CAL draws, not confidence intervals. Positive foreign − matched violation favors the matched donor. NRP has a negative mean at both budgets, and no draw improves both EVAL columns.

Table 2: Actual crossed policies on one fixed draw (Qwen VALUES, seed 26092300; 2,048 CAL questions). Coverage and risk are percentages; Q counts accepted unique questions. Both EVAL columns are necessary: a better CAL donor is not necessarily a matching effect.

		EVAL-S			EVAL-T		
Readout	CAL	C	R	Q	C	R	Q
NP	S	.3861	13.0664	23	.5404	4.5179	29
NP	T	.1017	16.0000	6	.0899	.0000	6
NRP	S	1.9856	.0000	119	2.8628	4.4062	143
NRP	T	3.1199	.0000	175	3.9429	3.8183	190

outcomes. None of those feasible cells also passes the 300-question/100-component screen. Among the 44,828 whole-tie prefixes, none meets the one-percent risk target with at least 300 accepted questions and 100 components, even when every unresolved outcome is treated as correct. At the required support, minimum NRP risk is 1.2686% on EVAL-S and 3.3106% on EVAL-T. These are finite-library statements, not impossibility results for future scores or new data.

Sparse support is directly observable. Among optimistic one-percent-feasible whole-tie prefixes, the largest accepted-question counts on EVAL-S are 4 for N/NP, 277 for NR, and 271 for NRP; EVAL-T gives 13, 4, and 4. The residual readouts extend the feasible tail on S without reaching 300 questions. Separate 40- and 72-question CAL experiments select identical pooled/worst-group masks because the permitted error mass is below one accepted error. Earlier real/sham grouping controls also fail to repair transfer. These are finite-family and resolution boundaries, not evidence that additional calibration data can never help (Appendix D).

6 Assessment precision is a separate intervention

The next comparison freezes the policy itself. Estimating its error mass does not lower deployment risk. This distinction follows the separation between model evaluation and model improvement in active testing (Kossen et al., 2021). With known full-bank coverage C and unique-question sampling probability $\pi_q > 0$, a Horvitz–Thompson estimator is

$$\widehat{R} = \frac{1}{C} \sum_{q \in S} \frac{z_q}{\pi_q}, \qquad z_q = \sum_{p=1}^{2} w_{qp} A_{qp} E_{qp}. \tag{4}$$

Both producer outcomes are obtained together for a sampled question. The known-coverage formulation avoids replacing the target by a ratio of two estimated totals. The standard unbiasedness argument requires the declared design and inclusion probabilities (Horvitz & Thompson, 1952); no labels are used to choose inclusion probabilities in the compared designs.

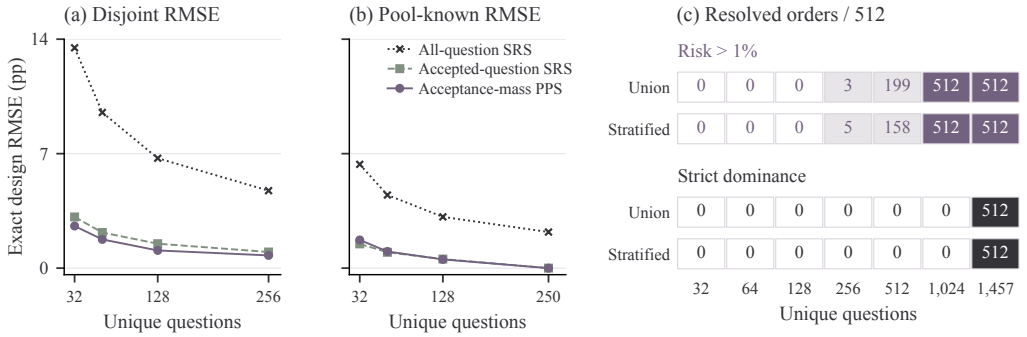

Figure 4: Precision and claim resolution under fixed acceptance policies. (a,b) Exact finite-design NRP risk RMSE at equal unique-question budgets, not confidence intervals. The pool-known accepted-support census caps the budget at 250; zero RMSE for accepted-question SRS and PPS does not mean zero policy risk. (c) Each cell counts resolved orders out of 512 at one reported budget, without interpolation. Union and Stratified denote accepted-union and fixed-stratified SRS without replacement under simultaneous finite-bank confidence sequences. Risk above one percent resolves before strict dominance on this grid. The precision and claim-resolution panels use distinct sampling designs.

Questions with zero acceptance have $z_q = 0$; all remaining questions have positive inclusion probability. The identity $\mathbb{E}[I_q/\pi_q] = 1$ establishes design-unbiasedness, but not an independent-Bernoulli variance formula for systematic sampling. The PPS comparison uses the actual joint inclusion law, certainty units, and label-blind metadata-hash ordering.

The finite-pool replay crosses NP/NR/NRP, two EVAL roles, and budgets 32/64/128/256. All-question SRS, accepted-question SRS, and systematic probability-proportional-to-acceptance-mass sampling (PPS) receive equal unique-question budgets. The accepted-question census caps every pool-known NRP design at 250. Exact design moments accompany 512 simulated draws per cell, not newly acquired human labels.

For disjoint NRP at 128 questions, risk RMSE is 6.7202 percentage points for all-question SRS, 1.4942 for accepted-question SRS, and 1.0817 for PPS. At 256 questions the latter two become .9865 and .7775. PPS is not uniformly best: on pool-known NRP at 32 questions, PPS RMSE is 1.7226 points versus 1.4825 for accepted-question SRS; on disjoint NP at 128, the corresponding values are 2.8183 versus 2.3742. The sampling comparison improves an assessment procedure in specified regimes without modifying the acceptance mask.

No confidence sequence is attached to this PPS replay. The claim-specific experiment below instead uses SRS without replacement; intervals cannot be transferred between sampling laws (Zecchin et al., 2025). Neither replay supplies a universal $1/C$ label saving or measures human acquisition cost.

Changed decisions and absolute risk require different accounting. For two fixed acceptance masks, define $b_i = \min(A_{0i}, A_{1i})$, $u_i = (A_{1i} - A_{0i})_+$, and $v_i = (A_{0i} - A_{1i})_+$. Their error masses obey

$$B = \sum_i w_i b_i E_i, \quad U = \sum_i w_i u_i E_i, \quad D = \sum_i w_i v_i E_i,$$
$$N_0 = B + D, \qquad N_1 = B + U, \qquad \Delta N = U - D. \tag{5}$$

The shared term cancels from ΔN, not necessarily from

$$\Delta R = \sum_i w_i \left(A_{1i}/C_1 - A_{0i}/C_0\right) E_i. \tag{6}$$

Coverage is known from masks without labels. Three comparative claims differ: more correct mass requires $\Delta N < \Delta C$; lower accepted-error risk requires $\Delta R < 0$; *strict acceptance dominance*

requires $\Delta N < \min(0, \Delta C)$, equivalently $\Delta G > 0$ and $\Delta N < 0$. At the headline's fixed positive ΔC, dominance is stronger than risk decrease. At unequal coverage, shared accepted errors can contribute to ΔR. Question-level costs also differ from candidate counts: acquiring one question reveals both producer outcomes, including any unchanged candidate. After every changed question is labeled, let $N_{1,\Delta}$ denote its contribution and M_3 the remaining unchanged accepted mass. The conservative identification interval is $[N_{1,\Delta}/C_1, (N_{1,\Delta} + M_3)/C_1]$, with unresolved execution outcomes propagated separately from unpurchased labels.

For original Qwen N versus NRP, 1,195 changed disjoint questions leave the interval $[2.07101, 23.36281]\%$: comparison labels already refute 1% risk. The pool-known interval after 253 changed questions is $[.99587, 19.95873]\%$ and straddles the target. The fixed S/SR bridge gives the same qualitative distinction: disjoint lower bound 1.14724%, versus pool-known $[.93021, 27.55351]\%$. Thus additional shared-acceptance labels are not universally necessary to refute risk, although the common term can prevent identification in another role.

Equal-budget decisions, not only RMSE. Accepted-union and fixed-stratified SRS-WoR each replay 512 label orders. The Hoeffding-type finite-WoR confidence sequence (Waudby-Smith & Ramdas, 2020) uses deterministic tuning, public unequal-weight bounds and coupled bounds for unresolved outcomes; total error .05 is allocated across roles, designs, strata and endpoints (Appendix E). The joint region for $(\Delta N, N_1)$ also gives intervals for ΔG and ΔR without new endpoints. The guarantee concerns a fixed finite target under sampling randomization, not future databases or all orders jointly.

At 1,024 disjoint questions, both designs support correct-mass gain and risk above 1% in every order. Risk reduction and strict dominance first resolve at the 1,457-question census on the reported grid (Table 10). Inspecting the same sequences after every question gives union mean first-support counts of 613.4 for correct-mass gain, 1,398.3 for risk reduction, 1,419.0 for dominance and 559.8 for absolute violation. All disjoint crossings precede census, but the two relative-risk claims remain near census. These counts describe this confidence sequence, not minimum label requirements. A separate S/SR union audit retains a distinct error family (Appendix E).

7 Controls that limit a broader interpretation

Arithmetic recovery is not semantic recovery. Zero-update arithmetic controls hold captured BF16 hidden states and head rows fixed. DIRECT's maximum margin error relative to FP64 on those operands falls from .123714 to 1.85×10^{-5}; distinct scores rise from 54 to 3,417 (Blanchard et al., 2019). Equal-prefix-mixture opportunity raises exposed-bank feasible coverage from 1.501% to 3.157%, but the locked KaggleDBQA transfer contrast has zero coverage gain. That custom CAL/EVAL split is not the official leaderboard (Lee et al., 2021). Improved arithmetic does not itself identify semantic recovery.

Judge and endpoint dependence remain material. Earlier external VALUES panels improve Qwen risk on BIRD and WikiSQL but worsen Granite risk on both. Different rosters and denominators preclude pooling these endpoints. Such judge/task dependence also motivates prior text-to-SQL verification work (Somov & Tutubalina, 2025; Alrashed et al., 2026; Richardson, 2026). Column-permutation tolerance changes some execution labels without making the alternative readouts meet one-percent risk. Result equivalence remains a proxy, not proof of intended semantics.

Matched value replacement does not establish a semantic mechanism. A label-blind matched donor replaces returned values while holding the candidate, body, readouts, and thresholds fixed. The common subset contains 419 questions/28 databases, or 7.532% of original weighted mass. The original-versus-donor readout interactions are $-.0689$ and $-.0094$ points for Granite and Qwen, with joint bands $[-.2471, .1093]$ and $[-.1876, .1688]$. Both include zero. A key-order sham has identical tokens. Neither semantic reasoning nor transfer to unmatched questions follows.

A new producer tests conditional transfer. Frozen Mistral-v0.3 generates SQL from question and schema on 4,000 WikiSQL and 2,000 BIRD questions. Original candidates remain unchanged; each new panel has its own component–question weights and preserves unresolved outcomes. No new judge or SOURCE readout is selected for favorable target performance. The Qwen 3% SR −

Table 3: Matched SR − S wrong-mass contrasts using the Qwen judge at 3% expected coverage. Lower is better; all entries are percentage points. Original producer rows use within-producer quotas and the full two-producer denominator, with 97.5% conditional intervals; new-producer rows use separate panel weights and 95% intervals. These rows are not pooled or population certificates.

Panel	Readout fit	ΔN	Conditional interval
Original producers: full two-producer mass; 97.5% intervals			
Qwen-generated candidates	Original	$-.04734$	$[-.09297, -.00603]$
Granite-generated candidates	Original	$-.06559$	$[-.10146, -.03516]$
New producer: separate panel mass; 95% intervals			
Mistral WikiSQL target	Original	$+.02162$	$[-.12970, +.17294]$
Mistral BIRD target	Original	$+.16961$	$[-.33782, +.63036]$
Mistral WikiSQL target	Adapted	.00000	$[-.12970, +.12970]$

S contrasts are $+.02162$ and $+.16961$ percentage points, respectively, with intervals crossing zero (Table 3). The gain is therefore not established on this producer. On WikiSQL, SR with the original numerical threshold accepts 56 target questions with zero observed errors; this different-coverage, low-support endpoint does not reverse the ordering result. Granite remains an adverse boundary. New answers to previously exposed questions are not new independent question or database populations.

Additional controls. Literal/schema, agreement, two-judge, and support/capacity variants do not repair one-percent acceptance (Appendix F). Two-judge risks of 4.3094%/3.9339% at 31–33% coverage also fail the strongest producer-matched NRP comparison. Agreement can share nonempty wrong results.

Producer-specific adaptation: risk and acceptance support. A frozen follow-up generates Mistral answers for 1,024 SOURCE and 2,048 CAL questions, component-disjoint from the previously scored target but from the exposed question population. Four S/SR fits use 397 resolved eligible SOURCE questions from 190 components; body, adapter, and residual direction remain fixed. Old/new SOURCE readouts are crossed with old/new producer answers to the *same* 2,048 CAL questions. Each readout receives its own threshold; original numeric cutoffs remain references. Adaptation changes answer distribution and effective fit support.

On the fixed WikiSQL target, adapted SR matches S at 2.16169% risk under 3% coverage, compared with 2.88226% for old SR; the paired adapted SR − S interval still spans zero. None of 16 CAL operating points reaches the 300-question/100-component screen (maximum 117 questions). Adapted Qwen SR with old-producer CAL yields 1.54021% target coverage, 1.40351% risk, and 45 accepted questions; new-producer CAL yields .04323% coverage, zero observed errors, and one question. Across 56 target endpoints, no endpoint passes the support screen. All fitted scalar components remain monotone. Adaptation establishes neither incremental ordering gain nor supported one-percent acceptance; target reuse also precludes independent confirmation.

8 Conclusion

A useful readout improvement and an unmet risk target coexist. Qwen gains survive producer-aware prefix controls and the fixed scalar-spline recipe. These controls do not establish superiority over arbitrary scalar functions of the native score. Exact changed/shared acceptance accounting distinguishes mass dominance, risk reduction, and absolute risk. Every-question audits resolve correct-mass gain earlier than risk reduction and strict dominance; the latter remain near census. Matching CAL components does not explain recovery, and a new producer plus matched SOURCE/CAL adaptation does not establish additional gain or supported one-percent acceptance. SOURCE refits preserve quota ordering but can reverse CAL-selected correct-mass gains. Repeated CAL draws preserve a positive mean Qwen gain but retain reversals and frequent risk violations; Granite contrasts remain negative. The supported result is conditional improvement with estimand-specific evidence requirements, not universal verifier repair.

AI USE STATEMENT

Generative AI tools assisted hypothesis refinement, experimental design and implementation, mathematical checking, numerical analysis, literature search and reference verification, figure preparation, and manuscript drafting and editing. Language models supplied SQL candidate answers for existing benchmark questions; no new questions, databases, or gold labels were generated. The authors are responsible for the final content.

ETHICS STATEMENT

SQL execution and returned values can expose sensitive database contents. This study uses the declared benchmark snapshots and bounded result packets; applying these procedures to private data would require separate access, privacy, and execution-safety controls. Abstention and the one-percent experimental target do not certify an accepted query's semantic correctness or safety.

REPRODUCIBILITY STATEMENT

Appendix B specifies frozen readouts, data roles, weights, and uncertainty conventions. Appendices C, D, and E give the finite-family controls, same-evaluation crossover, and fixed-policy sampling details; Appendix F retains additional and adverse comparisons. Appendix G specifies repeated calibration draws and their conditional interpretation.

REFERENCES

Tarfah Alrashed, Madhup Sukoon, David R. Karger, and Natasha Noy. Developing and benchmarking verification algorithms to improve text-to-SQL generation. *Proceedings of the VLDB Endowment*, 19(11):3732–3744, 2026. URL https://www.vldb.org/pvldb/vol19/p3732-alrashed.pdf.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2208.02814.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *Annals of Applied Statistics*, 19(2), 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Pierre Blanchard, Desmond J. Higham, and Nicholas J. Higham. Accurate computation of the log-sum-exp and softmax functions, 2019. URL https://arxiv.org/abs/1909.03469. arXiv:1909.03469.

Luís Felipe Cattelan and Danilo Silva. How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks. In *Conference on Uncertainty in Artificial Intelligence*, volume 244, pp. 547–584. PMLR, 2024. URL https://proceedings.mlr.press/v244/cattelan24a.html.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107, 2010. URL https://jmlr.org/papers/v11/cawley10a.html.

Gianluca Detommaso, Martin Andres Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in LLMs. In *International Conference on Machine Learning*, volume 235, pp. 10624–10641. PMLR, 2024. URL https://proceedings.mlr.press/v235/detommaso24a.html.

Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Calibrated selective classification. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2208.12084.

Vojtech Franc, Daniel Prusa, and Vaclav Voracek. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11):1–49, 2023. URL https://jmlr.org/papers/v24/21-0048.html.

Aditya Gangrade, Anil Kag, and Venkatesh Saligrama. Selective classification via one-sided prediction. In *International Conference on Artificial Intelligence and Statistics*, volume 130, pp. 2179–2187. PMLR, 2021. URL https://proceedings.mlr.press/v130/gangrade21a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, volume 97, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, volume 70, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446. URL https://doi.org/10.1080/01621459.1952.10483446.

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

Chia-Hsuan Lee, Oleksandr Polozov, and Matthew Richardson. KaggleDBQA: Realistic evaluation of text-to-SQL parsers. In *Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing*, pp. 2261–2273, 2021. URL https://aclanthology.org/2021.acl-long.176/.

Gyubok Lee, Woosog Chay, Seonhee Cho, and Edward Choi. TrustSQL: Benchmarking text-to-SQL reliability with penalty-based scoring, 2024. URL https://arxiv.org/abs/2403.15879. arXiv:2403.15879.

Yonghoon Lee, Rina Foygel Barber, and Rebecca Willett. Distribution-free inference with hierarchical data, 2023. URL https://arxiv.org/abs/2306.06342. arXiv:2306.06342; revised 2025.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, 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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.03111.

Hengyue Liang, Le Peng, and Ju Sun. Selective classification under distribution shifts. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.05160.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.10193.

Stephan Rabanser and Nicolas Papernot. What does it take to build a performant selective classifier?, 2025. URL https://arxiv.org/abs/2510.20242. arXiv:2510.20242.

Robert Richardson. What predicts correctness in text-to-SQL? a selective-prediction study, 2026. URL https://arxiv.org/abs/2607.06799. arXiv:2607.06799.

Oleg Somov and Elena Tutubalina. Confidence estimation for error detection in text-to-SQL systems. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 25137–25145, 2025. URL https://ojs.aaai.org/index.php/AAAI/article/view/34699.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/e96c7de8f6390b1e6c71556e4e0a4959-Abstract.html.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009. URL https://arxiv.org/abs/2010.09686.

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 *Conference on Empirical Methods in Natural Language Processing*, pp. 3911–3921, 2018. URL https://aclanthology.org/D18-1425/.

Matteo Zecchin, Sangwoo Park, and Osvaldo Simeone. Adaptive learn-then-test: Statistically valid and efficient hyperparameter selection. In *International Conference on Machine Learning*, volume 267, pp. 74018–74036. PMLR, 2025. URL https://proceedings.mlr.press/v267/zecchin25a.html.

Ruiqi Zhong, Tao Yu, and Dan Klein. Semantic evaluation for text-to-SQL with distilled test suites. In *Conference on Empirical Methods in Natural Language Processing*, pp. 396–411, 2020. URL https://aclanthology.org/2020.emnlp-main.29/.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2SQL: Generating structured queries from natural language using reinforcement learning, 2017. URL https://arxiv.org/abs/1709.00103. arXiv:1709.00103.

Table 4: Three interventions test different claims. The third column states the supported conclusion, not a proposed future success.

Bottleneck	Matched intervention	Evidence and limitation
Score ordering	Nested and scalar/residual readouts; prefix mixtures	Original Qwen gain survives fixed controls; Mistral transfer/adaptation gain remains unestablished.
CAL transfer	Same EVAL; reciprocal CAL donors; equal budget	No consistent matching benefit or positive joint interaction.
Audit support	Fixed policy; equal question budgets and finite-design intervals	Disjoint violation resolves before strict acceptance dominance on the reported grid; no general ordering of audit difficulty.

A Questions, controls, and interpretation

B Readout, role, and uncertainty details

Frozen model and scoring card. Qwen3-8B uses checkpoint revision `b968826d9c46dd6066d109eabc6255188de91218` and the fixed VALUES adapter selected at epoch 1, support 24, seed 20260909. Granite-3.3-8B-Instruct uses fixed model weights identified by file hashes and does not use the Qwen adapter. Both expose the 4,096-dimensional last non-padding-token state after the final transformer layer. The native score m is the stored YES $-$ NO logit difference; both verbalizers are single tokens. Body operands are BF16 and the two-row output projection is accumulated in FP32. Granite logits include the model's division by 16; Qwen uses scale 1. Same-operand arithmetic controls do not replace this body with an FP64 forward pass.

Let $v = (W_{\mathrm{YES}} - W_{\mathrm{NO}})/s$ be the frozen verbalizer direction, μ the training-fold weighted hidden mean, and $d = 4096$. The implemented orthogonal feature and scale are

$$u(h) = (h - \mu) - v\frac{v^\top(h-\mu)}{v^\top v}, \qquad \sigma^2 = \sum_{i\in\mathrm{train}} w_i \|u(h_i)\|_2^2/d. \tag{7}$$

The base score is $b(h, m) = e^a m + c + \theta^\top u(h)/\sigma$, requiring $v \neq 0$ and $\sigma > 0$. Weighted logistic loss uses correctness labels and penalty $\lambda[(e^a - 1)^2 + c^2 + \|\theta\|^2]/2$, with $\lambda = .1$. Each of four folds excludes two SOURCE databases; fold-only means and RMS scales produce out-of-fold scores for the meta-fit. A full-SOURCE base fit then supplies target predictions. The stacked residual input is $r = b - m$, not the raw hidden vector. N, NP, NR, and NRP refit a positive-native- slope logistic head using respectively $\varnothing$, centered producer indicator, r, or both as auxiliary inputs, with the same penalty recipe. Removing a feature and refitting is distinct from zeroing a fitted coefficient.

CAL maximizes correct accepted mass, then coverage, over whole-tie prefixes satisfying $N^+ \leq .01C$. If no nonempty prefix satisfies the constraint, the policy rejects all candidates. Unavailable or ineligible candidates abstain without leaving the denominator. Execution eligibility requires executable, untruncated candidates. The common availability mask and outcome convention are identical within each contrast. CAL labels choose only thresholds, never SOURCE coefficients.

The SOURCE reconstruction uses the same 857 eligible rows, eight databases, regularization .1, feature scaling, and four-fold out-of-fold assignments for N, NP, NR, and NRP. Original unavailable rows remain unavailable rather than receiving synthetic labels. The literal SOURCE producer indicator has no contrast on the current two target producers. Both map to zero, explaining the identical NP/N target chain. A different indicator that explicitly distinguishes target producers would be a new feature intervention. The producer-aware prefix oracle answers the narrower question of whether privileged target-producer allocation of native scores explains the observed endpoint; that control does not require retraining a target-aware head.

The earlier external BIRD panel has 1,860 CAL questions across 13 databases and 5,470 evaluation questions across 41; the earlier WikiSQL panel has 1,536 CAL and 11,624 evaluation questions. These are not the later primary cohort. The component-disjoint cohort was frozen separately from the earlier external-bank diagnoses. The later crossover, native-prefix comparisons, fixed-mask un-

Table 5: Primary Qwen VALUES endpoints at frozen original CAL thresholds. Coverage and risk are percentages. Question/component counts describe actual support, not independent candidate rows.

Role	Readout	Coverage	Risk	Questions	Components
Disjoint	N/NP	2.065953	7.103948	746	640
Disjoint	NR	3.344	2.653370	1,131	881
Disjoint	NRP	3.326971	2.667192	1,127	876
Pool-known	N/NP	1.986116	4.028243	140	117
Pool-known	NR/NRP	3.837072	.995871	250	201

certainty analysis, and bounded alternatives reuse existing data and are exposed diagnoses. A study-internal CAL/EVAL separation does not erase earlier project exposure. Neither previously unused question identifiers inside a known component nor new random seeds establish a new population. Shared benchmark training provenance also leaves language-model pretraining exposure unresolved.

The full-bank headline weights are component/database balanced. The reciprocal crossover instead constructs two explicitly balanced rosters and uses component–question–producer-uniform weights within each. Consequently, different coverage numbers across these panels need not imply that the same acceptance mask changed. All matched comparisons use one common denominator within their own panel. Unresolved-outcome completions are applied after preserving that denominator.

For a fixed policy, a component bootstrap resamples whole components while retaining paired candidate outcomes and score-arm pairing. The 2,000-draw fixed-mask analysis gives disjoint NR $-$ NP $\Delta G = +1.3364$ percentage points with interval $[1.1200, 1.5612]$; NRP $-$ NP gives $+1.3190$ $[1.1029, 1.5394]$. Both yield $\Delta N = -.0580$ $[-.1105, -.0078]$. On pool-known questions, NR $-$ NP and NRP $-$ NP both give $\Delta G = +1.8927$ $[1.3057, 2.5207]$ and $\Delta N = -.0418$ $[-.1389, +.0413]$. These are marginal descriptive 95% intervals conditional on frozen models, selection, and role construction. The independently reproduced fixed-mask pool-known risk interval $[.1187, 2.3410]\%$ differs slightly from the original descriptive interval $[.124, 2.296]\%$ because the bootstrap realization and recorded protocol differ. Both cross one percent; neither is selected as a favorable certification result.

Scalar-control fitting. The S/SR basis contains an intercept, the SOURCE-standardized native margin, and positive-part linear terms at the .1/.3/.5/.7/.9 SOURCE quantiles. Coefficients are unconstrained: nonmonotone scalar functions are allowed, not imposed. SR adds the original out-of-fold residual. Fold-specific preprocessing and database exclusions follow the original residual contract; target scores never determine knots or regularization. Four fitted models have maximum KKT gradient 1.159×10^{-8}. Although the scalar family permits nonmonotone functions, every fitted S slope is positive, so the fitted ordering matches the native ordering.

B.1 Sensitivity to SOURCE databases

Each of eight leave-one-database-out conditions refits every retained fold's hidden mean, residual scale, orthogonal residual coefficients and out-of-fold predictions, then refits the full base, producer centering and N/NRP meta heads. Judge, adapter, regularization .1, positive native slope, weighting and availability rules remain fixed. Both heads use the same retained SOURCE rows. Fits precede score freezing and the original CAL selector. No database exclusion or threshold is chosen using EVAL. Full-SOURCE reconstruction reproduces all sixteen original CAL/quota masks and base parameters; target score error is at most 2.71×10^{-14}.

All eight disjoint fixed-quota gains persist, including within each producer, but CAL-selected correct mass decreases in two conditions and strict dominance holds in only five. Pool-known ordering improves in seven conditions and is unchanged in one; two sub-one-percent CAL endpoints accept only 66 and 145 questions. No NRP exclusion combines the risk target with the support screen. These results measure frozen-body readout sensitivity on exposed banks, not backbone robustness.

The same native-prefix oracle families are evaluated at each already selected NRP coverage, without selecting a new point on the curve. Original weights and producer masses remain fixed. The

Table 6: All SOURCE exclusions on disjoint EVAL. Each NRP head has its own CAL threshold; N remains at 2.065953% coverage and 7.103948% risk with 746 questions/640 components. The last column instead fixes a pooled 3% quota. Differences are mass percentage points. These overlapping sensitivities are not independent replications or confidence intervals.

Excluded database	C (%)	R (%)	ΔG_{CAL}	Q / Comp.	$\Delta N_{3\%}$
None: primary	3.327	2.667	+1.3190	1127 / 876	-.130666
apartment_rentals	2.394	6.131	+.3280	848 / 716	-.003791
behavior_monitoring	.987	1.138	-.9431	337 / 296	-.134241
game_injury	1.999	2.689	+.0258	705 / 583	-.106293
hospital_1	.509	1.715	-1.4190	181 / 166	-.124299
music_1	2.961	2.604	+.9652	994 / 795	-.131469
network_2	3.904	2.946	+1.8698	1344 / 1017	-.127948
solvency_ii	3.349	2.626	+1.3418	1121 / 880	-.133625
voter_2	4.098	3.217	+2.0475	1385 / 1032	-.133740

producer-aware witness is positive in seven disjoint exclusions and five pool-known exclusions. Nonpositive gaps leave the native explanation open; positive gaps at the behavior-monitoring and hospital exclusions do not undo their negative CAL-selected correct-mass changes.

Table 7: Native-oracle gaps at the fixed residual endpoints, in mass percentage points. The global and producer-aware families remain distinct. None is a population interval.

	Disjoint		Pool-known	
Excluded database	Global	Producer-aware	Global	Producer-aware
None: primary	+.1413	+.0960	+.0564	+.0516
apartment_rentals	+.0102	-.0201	-.0243	-.0403
behavior_monitoring	+.0476	+.0346	+.0019	-.0049
game_injury	+.0724	+.0502	+.0258	+.0084
hospital_1	+.0200	+.0121	-.0060	-.0085
music_1	+.1243	+.0844	+.0244	+.0156
network_2	+.1603	+.1083	+.0469	+.0286
solvency_ii	+.1438	+.0983	+.0498	+.0457
voter_2	+.1587	+.1060	+.0499	+.0373

Matched scalar sensitivity. A separate S/SR analysis refits both scalar models after the same exclusions using identical retained rows, weights, five hinge knots and regularization .1. SR uses the corresponding fully refitted residual and out-of-fold predictions, not a truncated full-SOURCE residual. The eighteen scalar fits do not retrain the judge or adapter. All nine fitted S functions have positive slopes on nonzero-width intervals, although coefficients are unconstrained. Both roles retain

Table 8: Matched scalar-only S versus scalar-plus-residual SR sensitivities. Each pair gives own-CAL correct-mass change / pooled-3%-quota wrong-mass change, in percentage points. All exclusions and both evaluation roles are retained.

Excluded database	Disjoint $\Delta G_{\mathrm{CAL}}/\Delta N_{3\%}$	Pool-known $\Delta G_{\mathrm{CAL}}/\Delta N_{3\%}$
None: primary	+.9736 / -.1521	+1.5125 / -.0571
apartment_rentals	+.3439 / -.0588	+.5984 / -.0521
behavior_monitoring	-.7842 / -.1336	-.8122 / -.0800
game_injury	+.0398 / -.1063	+.3001 / -.0528
hospital_1	-1.3069 / -.1311	-1.3064 / -.0571
music_1	+1.1144 / -.1246	+1.5395 / -.0348
network_2	+1.3668 / -.1203	+2.1324 / -.0702
solvency_ii	+1.1759 / -.1248	+1.8868 / -.0580
voter_2	+1.7203 / -.1334	+2.2291 / -.0754

eight quota-level improvements but two negative CAL-selected correct-mass changes. Disjoint SR

risk spans .980–3.984%; pool-known risk spans .500–1.726%. Excluding the behavior-monitoring database yields disjoint SR risk .980354% at 1.146267% coverage, with 392 accepted questions, 341 components and seven wrong candidates. This passes the empirical target and support screen but loses .7842 points of correct mass relative to S. No confidence certificate or selection of this exclusion as a replacement primary follows. Full-SOURCE SR risk remains 1.914%.

C FINITE-FAMILY CONTROLS AND SHORT DERIVATIONS

Bounds for unresolved outcomes. Write K for accepted candidates with observed outcomes and U for accepted candidates with unresolved outcomes. Then

$$N^- = \sum_{i \in K} w_i A_i E_i, \qquad N^+ = N^- + \sum_{i \in U} w_i A_i, \quad G^- = C - N^+, \quad G^+ = C - N^-. \tag{8}$$

Each allowed completion lies in this interval. Such bounds identify the finite-bank quantity under missing labels; sampling uncertainty is separate. Using optimistic baseline bounds and pessimistic NRP bounds in Equation 2 makes positive gaps robust to independent endpoint completion, at the cost of potential conservatism for a shared completion.

Whole-prefix mixtures. Let (C_j, N_j) denote every whole-tie prefix, including $(0, 0)$. A mixture with coefficients $\theta_j \geq 0$, $\sum_j \theta_j = 1$, satisfies $(C, N) = \sum_j \theta_j (C_j, N_j)$. Hence the minimum attainable N at fixed $C = c$ is the lower convex envelope of the prefix points. A linear program or an equivalent lower-hull interpolation computes the envelope. An interior point on one hull edge needs at most two supporting prefixes. If every nonempty constituent has risk greater than α, then every nonempty mixture also does, because $R = \sum_j \theta_j C_j R_j / \sum_j \theta_j C_j$ is coverage weighted.

For producer-aware mixtures, each producer receives a separate normalized prefix mixture while its original mass in the bank remains fixed. The resulting feasible set contains the global native mixture, because copying the same random threshold law into both producer components reproduces the global policy. Allowing independent producer allocations may lower the error envelope, making exclusion more demanding. The primary Qwen gap remains positive under this stronger family (Table 1). The statement applies only to the enumerated score/producer family, fixed stored scores, tie convention, bank, and endpoint; arbitrary verifier functions are not bounded by this hull.

Family expansion versus threshold transfer. For nested families $\Pi_K \subseteq \Pi_{K+1}$, the hindsight maximum

$$H_B(K) = \max_{\pi \in \Pi_K} \{C_B(\pi) : N_B^+(\pi) \leq \alpha C_B(\pi)\} \tag{9}$$

cannot decrease, since every previously feasible policy remains available. This monotonicity does not hold for the EVAL coverage or risk of a policy selected on CAL. Moreover, an infeasible transferred policy can have coverage larger than H_{EVAL}, so $H_{\mathrm{EVAL}} - C_{\mathrm{transferred}}$ is not a nonnegative regret. Reporting that difference as recoverable coverage would mix feasibility with optimization.

D SAME-EVALUATION CALIBRATION CROSSOVER

The label-blind census contains 32,128 unique questions and 8,203 components; 2,630 components have at least five questions. Metadata matching retains 2,048 components and 14,285 questions. Component sets S and T each contain 1,024 components. EVAL-S contains 3,279 questions/6,558 candidates, and EVAL-T contains 3,278/6,556. CAL reserves and EVAL question identifiers are disjoint. Component size, schema, and question length determine the frozen rosters before the crossed outcome comparison; observed correctness does not define the partition.

The experiment crosses four judge/input packages, five frozen readouts, two unique-question budgets, eight draws, and two CAL donors, producing 640 threshold fits and 1,280 EVAL cells. Unresolved-outcome counts on the two EVAL rosters are 64 and 73; the original full outcome set contains 593. The comparison does not drop these outcomes. Every donor threshold is evaluated on both columns with the same readout scores and role weights. Threshold ties follow the original deterministic convention; no EVAL-dependent boundary selection is added.

The main-text draw in Table 2 illustrates why both columns matter: a CAL-T policy can improve both S and T rather than benefit only its matching column. The eight-draw summaries, not this single row, determine the stated matching contrast. The mean interaction is negative at both budgets; no positive joint readout–matching claim is supported. Because the draws overlap and the bank was previously exposed, their range is descriptive sensitivity, not a confidence interval or evidence from eight independent populations.

Requiring 300 questions and 100 components raises the minimum possible risk to 4.7201%/1.2773%/1.2686% for native/NR/NRP on S and 5.3992%/3.3170%/3.3106% on T. These descriptive screens locate the finite-family limitation without converting a support count into a theorem.

Granularity and finite error resolution. Earlier $K = 1, 2, 4, 8$ group policies retain the coarse policy, so hindsight feasible coverage cannot decrease with family expansion. Numeric transfer does not inherit that property: Granite DIRECT NRP's held risk rises from .741% to 14.250%, while a matched sham reaches 4.479% at $K = 8$. Qwen VALUES still has 11.843% risk at $K = 8$. With 40 balanced two-producer CAL questions, a single erroneous candidate has 1.25% total mass, requiring zero accepted errors at the 1% risk target. The 72-question follow-up permits one error only at coverage at least 69.44%; selected coverage remains below that point. Pooled and worst-group selectors consequently coincide on these draws. This discrete-budget coincidence is not a general equivalence of calibration rules.

E FIXED-POLICY LABEL-SAMPLING DETAILS

For each question define acceptance mass $a_q = \sum_p w_{qp} A_{qp}$ and error mass $z_q = \sum_p w_{qp} A_{qp} E_{qp}$. The policy, score, calibration threshold, weights, and total coverage $C = \sum_q a_q$ are fixed before sampling. Under all-question SRS, every question in the role has equal inclusion probability. Under accepted-question SRS, only $a_q > 0$ questions are sampled; excluded questions have $z_q = 0$, so no error mass is omitted. Under systematic PPS, inclusion is based on public acceptance mass, with certainty units handled explicitly. Sampling order is fixed by metadata hashes, not errors.

The inclusion-law argument establishes design-unbiasedness of Equation 4; Table 9 retains the exact finite-design RMSE values behind Figure 4.

Table 9: NRP finite-design risk RMSE (percentage points). The actual budget is unique questions with both producer outcomes. Requested 256 is capped at 250 for every pool-known design.

Role	Actual Q	All-Q SRS	Accepted-Q SRS	Acceptance PPS
Disjoint	32	13.4729	3.1286	2.5755
	64	9.5191	2.1797	1.7562
	128	6.7202	1.4942	1.0817
	256	4.7365	.9865	.7775
Pool-known	32	6.3463	1.4825	1.7226
	64	4.4694	.9683	1.0181
	128	3.1345	.5545	.5321
	250	2.2072	.0000	.0000

The finite census of 250 accepted questions recovers that bank's accepted error mass exactly. Future questions or databases still require a sampling model. No interval procedure is attached to this PPS replay; the later claim-specific SRS-WoR experiment has a separate sampling law. A simulation with existing labels does not measure the latency or labor of obtaining new labels. PPS setup and known coverage also consume information and computation; neither is silently assigned zero deployment cost.

Claim-specific without-replacement audit. Question stratum H1 changes an acceptance and retains some acceptance under A_1; H2 changes acceptance but rejects both candidates under A_1; H3 has unchanged positive acceptance. H0 rejects everything and contributes exactly zero. Disjoint H1/H2/H3 sizes are 865/330/262; pool-known sizes are 197/56/53. Fixed-stratified sampling starts

from allocation .5/.25/.25; empty or censused strata redistribute remaining budget by public sizes. Uniform random orders within strata implement SRS-WoR. Unique question counts, not candidate counts or repeated draws, define cost.

For each union or stratum, public weights and masks bound the per-question contributions to ΔN and N_1. The audit applies the Hoeffding-type confidence sequence of Waudby-Smith & Ramdas (2020, Theorem 3.1), with their weighted without-replacement mean, Hoeffding penalty and deterministic tuning sequence (Equations 3.2, 3.3 and 3.7). The four scalar populations are the lower and upper completions for each of ΔN and N_1; these are distinct from the four claims derived below. Coupled completion bounds retain unresolved execution outcomes separately from unobserved audit labels.

The interval construction. Let $M > 0$ denote the number of questions in one union or stratum. In a uniform random order, x_i is a scalar weighted contribution with public bounds $\ell_i \le x_i \le u_i$. The contributions already include the original role-level candidate weights. Set

$$y_i = x_i - \ell_i, \qquad W = \max_{i \le M}(u_i - \ell_i), \qquad O = \sum_{i=1}^{M} \ell_i, \qquad \alpha_* = .05/48.$$

For $W > 0$, the two-sided mean interval uses

$$\lambda_t = \min\left\{ \sqrt{\frac{8\log(2/\alpha_*)}{t\log(t+1)W^2}}, \frac{1}{W} \right\}, \qquad D_t = \sum_{i=1}^{t} \lambda_i \frac{M}{M-i+1},$$

$$\widehat{\mu}_t = \frac{\sum_{i=1}^{t} \lambda_i \left(y_i + \frac{\sum_{j<i} y_j}{M-i+1} \right)}{D_t}, \qquad r_t = \frac{\sum_{i=1}^{t} \lambda_i^2 W^2/8 + \log(2/\alpha_*)}{D_t}.$$

All logarithms are natural. The tuning depends only on t, W, and α_* and is predictable without adaptation to empirical variance. The common width W enters directly; values are not divided by their individual widths or by W. Multiplication by M and restoration of the offset give total bounds $\widetilde{L}_t = M(\widehat{\mu}_t - r_t) + O$ and $\widetilde{U}_t = M(\widehat{\mu}_t + r_t) + O$. The accepted subset is not renormalized. Logical bounds for the remaining labels and running intersection then give

$$A_t = \sum_{i \le t} x_i + \sum_{i > t} \ell_i, \qquad B_t = \sum_{i \le t} x_i + \sum_{i > t} u_i,$$

$$L_t = \max_{s \le t} \max(\widetilde{L}_s, A_s), \qquad U_t = \min_{s \le t} \min(\widetilde{U}_s, B_s).$$

At census, each scalar interval is explicitly set to the exact total $L_M = U_M = \sum_i x_i$. Distinct lower and upper completion populations can still leave identification width for unresolved executions. If $W = 0$, public bounds fix the total to O at every budget; empty strata contribute zero without evaluating the formula. At budget zero, the interval is the public total range $[\sum_i \ell_i, \sum_i u_i]$.

The lower-completion sequence supplies the lower endpoint and the upper-completion sequence supplies the upper endpoint for each target. Error probability .05 is allocated over two roles, two designs, at most three strata and four scalar populations, with two-sided error α_* per population. Independent uniform orders within strata and the label-blind allocation above yield stratum total bounds that are summed, not unweighted stratum means. The finite target and masks remain fixed; validity concerns sampling randomization, not future databases or a simultaneous event over all 512 orders.

Reported budgets are 32/64/128/256/512/1,024 and census, capped by actual support. The 306-question pool-known census resolves both finite claims without certifying future risk. First support on this grid is the first reported budget whose valid interval implies the stated inequality; unresolved orders remain in the denominator. The separate S/SR comparison and inspection after every question are described below.

Derived claims and a separate scalar audit. For positive coverages, let $a = 1/C_1 - 1/C_0$ and $b = 1/C_0$. The joint intervals $I_{\Delta N} = [L_d, U_d]$ and $I_{N_1} = [L_1, U_1]$ imply

$$I_{\Delta G} = [\Delta C - U_d, \Delta C - L_d], \quad I_{\Delta R} = [\min(aL_1, aU_1) + bL_d, \max(aL_1, aU_1) + bU_d].$$

Table 10: Original N–NRP disjoint audit: percentage of 512 replays supporting strict dominance / risk above 1% at equal unique-question budgets. No policy changes.

Design	512 Q	1,024 Q	1,457 Q (census)
Accepted-union SRS	0 / 38.87	0 / 100	100 / 100
Fixed-stratified SRS	0 / 30.86	0 / 100	100 / 100

Both intervals share the same event without independence or further error allocation. Coupled completion bounds and undefined zero-coverage risks retain their original conventions.

A separate audit fixes the original S/SR scores and own-CAL thresholds. Disjoint coverage is $2.065953\%/2.949294\%$, with risk $7.103948\%/1.914205\%$ and 746/1,000 accepted questions. The difference is $+.973649$ points of correct mass and $-.090309$ points of wrong mass. Accepted unions contain 1,311 disjoint and 274 pool-known questions. Public subset filtering of the original 512 uniform orders preserves SRS without replacement; changed strata prevent reuse of the stratified allocation. The four elementary endpoints keep error allocation $.05/48$ each. Across two roles and one design, the separate family has error at most $8(.05/48) \leq .05$; no combined guarantee across the N/NRP and S/SR families or all orders is asserted.

On the reported S/SR grid, 510 of 512 disjoint orders support risk above 1% by 1,024 questions. Every order supports correct-mass gain there, while risk reduction and dominance first resolve at the 1,311-question census. Pool-known SR risk is $.930208\%$, so above-one-percent risk is false at census. Its 222 accepted questions/184 components fail the 300-question support screen. No new-producer or Granite confidence-sequence audit is included.

Inspection after every question. The same populations, weights, CAL thresholds, completion rules, public bounds, original error allocations and 512 orders are inspected at every integer budget through census. First support is the first valid rectangle implying the named inequality, whether or not support persists at every later budget. Time-uniform validity permits this inspection change without adding endpoints. These counts are not minimum label requirements or optimal stopping times for other confidence sequences. Every disjoint order supports each true inequality before census.

Table 11: Disjoint first-support counts under every-question inspection. Means and ranges describe 512 orders, not confidence intervals. Grid delay is the paired difference between the reported-grid and integer first-support counts (grid − integer). Census sizes are 1,457 for N/NRP and 1,311 for S/SR.

Pair/design	Claim	Integer mean	Range	Grid mean	Delay
N/NRP union	$\Delta G > 0$	613.4	533–688	1024.0	410.6
	$\Delta R < 0$	1398.3	1369–1423	1457.0	58.7
	Strict dominance	1419.0	1397–1438	1457.0	38.0
	$R_1 > 1\%$	559.8	202–1017	823.5	263.7
N/NRP strata	$\Delta G > 0$	716.4	625–781	1024.0	307.6
	$\Delta R < 0$	1420.7	1404–1433	1457.0	36.3
	Strict dominance	1433.8	1419–1447	1457.0	23.2
	$R_1 > 1\%$	573.5	185–945	863.5	290.0
S/SR union	$\Delta G > 0$	631.7	519–731	1024.0	392.3
	$\Delta R < 0$	1236.2	1196–1268	1311.0	74.8
	Strict dominance	1247.2	1212–1278	1311.0	63.8
	$R_1 > 1\%$	692.3	273–1053	952.1	259.8

Nevertheless, union risk-reduction/dominance means remain $96.0\%/97.4\%$ of the N/NRP census and $94.3\%/95.1\%$ of the S/SR census. Dense inspection reveals earlier crossings without changing intervals or readout quality. Pool-known union risk-reduction/dominance means are $299.0/301.0$ of 306 questions and $265.9/267.6$ of 274; five and six of 512 orders, respectively, first support each claim only at census. Absolute violation is not supported in these replays; the event is false at census.

Near census, 524 union-prefix rectangles have numerical endpoint inversions, at most 6.60×10^{-17} wide; none occur under stratified sampling. An inverted rectangle supports none of the four claims,

without swapping endpoints or using a tolerance to restore support. The empirical containment check uses a separate 10^{-12} tolerance, so zero observed noncoverage is not a claim of no exact endpoint inversions. Applying that tolerance to support changes only four pool-known crossing entries by one question; disjoint first crossings are unchanged. Per-budget fractions measure current support, not whether support occurred earlier. Zero observed wrong support does not imply zero theoretical error.

Earlier betting resolution control. A distinct Qwen full-stack NRP policy has 2.351% coverage and 7.595% observed risk. Under its 616-question law, observed evidence is .822, versus 1.028 after replacing accepted outcomes with all-correct labels and 3.255 for all-accept/all-correct. A different 13-database law gives .980/1.003/1.140. None reaches evidence 20 at level .05. These artificial completions test whether the audit reaches its evidence threshold; no natural-performance result or impossibility for other audits follows. Smaller RMSE and larger betting evidence are different targets.

F Additional matched comparisons and adverse results

Numerical and evaluator controls. The arithmetic census contains 3,644 SOURCE candidates from 32 databases, with 3,638 available computations. FP64 comparisons use the same stored operands, not an FP64 language-model forward pass. For the native-producer subset, DIRECT whole-tie coverage decreases from .565% to .414%, even though aggregate randomized coverage improves. The KaggleDBQA evaluator ignores row ordering and duplicate multiplicity in its bounded SQLite snapshot; 50 correct, 101 incorrect, and 63 unresolved EVAL candidates remain in the denominator. Alternative execution-metadata/returned-value score mixtures accept some mass but exceed one-percent risk even under optimistic completion of unresolved outcomes. New score distinctions are therefore insufficient evidence of useful cross-database recovery.

Semantic proxies and transfer. The bounded alternative panel retains 857 SOURCE fit candidates across eight databases, 300 CAL candidates across four, and a 370-question/740-candidate evaluation panel across eight. Pair-derived and agreement features use candidate and execution information available at scoring, not gold correctness. NR, NRP, pair, and agreement readouts have evaluation risks 5.184%, 3.685%, 3.994%, and 2.372%, respectively. The agreement policy reaches 26.952% coverage. At common producer quotas of 1%, 2%, and 5%, the strongest comparisons have no incremental gain; a favorable isolated operating point is not promoted over those controls.

A fixed public-literal/schema extension reduces the initial agreement-scope point from 24.6026% coverage/1.3041% risk to 23.6296%/1.0187%, accepting 98 questions across eight databases. Rounding 1.0187% to one percent would misstate feasibility. On a frozen remaining-eight-database role of 496 questions, the literal variant has 27.8710% coverage and 2.8889% risk, compared with 29.0924%/2.2561% for agreement alone. This additional role is historically exposed, not a fresh external confirmation.

Two-judge stacking uses paired judge inputs on the same 857 SOURCE rows. The resulting coverage/risk is 30.982%/4.3094% on this evaluation panel and 32.894%/3.9339% on the remaining role. The producer-aware NRP control is better at matched coverage on both roles. Extra judge-forward cost is not a free feature. A utility analysis at a looser error price finds some positive acceptance utility, but no new advantage over the strongest matched residual baseline and no repair of the one-percent target.

Earlier external and capacity endpoints. On the earlier BIRD VALUES panel, native/residual coverage and risk are .565%/8.816% versus 1.012%/3.823% for Qwen, and 2.701%/8.858% versus 3.314%/11.917% for Granite. Earlier WikiSQL gives .255%/11.425% versus 4.766%/4.748% for Qwen and 1.450%/9.416% versus .232%/14.317% for Granite. These roles differ from the later component-disjoint cohort and from the Mistral-only panel.

New-producer protocol and adaptation. Mistral-7B-Instruct-v0.3, revision `c170c708c41dac9275d15a8fff4eca08d52bab71`, uses greedy decoding (one beam, no sampling, repetition penalty 1), at most 384 new tokens and a fixed 4,096-token prompt bound. Prompts request SQL from the question and allowed schema; neither gold SQL nor correctness

Table 12: Support and capacity comparison at the original one-percent CAL target. This exposed discovery panel is not the primary WikiSQL cohort. Zero errors on four to seven questions do not establish useful reliability.

Frozen readout	Coverage (%)	Risk (%)	Accepted Q
Native	.487031	10.227403	22
NR/NRP	.045371	.000000	5
SOURCE linear	.062788	.000000	7
SOURCE MLP	.033317	.000000	4
Augmented linear	.401111	9.323094	32
Augmented MLP	.158200	7.964329	14

enters generation. The recipe and metadata-hash roster precede outcomes. WikiSQL has 3,000 target and 1,000 pool-known questions; BIRD has 2,000 target questions across 41 databases. Generation uses question and allowed schema, never gold SQL or execution correctness. BIRD excludes external knowledge from the new producer prompt; historical producer answers with a different knowledge interface are reference-only. The 240 BIRD prompts exceeding the fixed context bound remain unresolved in the full denominator. Comparisons S versus SR use the same new candidates, judge packets, weights, and unresolved outcomes. WikiSQL retains 471 correct, 2,525 incorrect, and 1,004 unresolved outcomes. Frozen old numeric thresholds and score-only 1%/3% coverage quotas remain separate endpoints.

The later SOURCE/CAL adaptation is a separate exposed-data follow-up, not retroactive replacement of the frozen-transfer primary. SOURCE, CAL, and target component/question rosters are disjoint within that protocol; the target predictions and outcomes are retained from the prior Mistral study. Source fitting uses 125 correct and 272 incorrect eligible answers; unavailable or unresolved rows are not converted into correct labels. The old/new producer CAL crossover keeps the same 2,048 questions, using two old candidates or one new candidate with their declared weights. No target labels select coefficients, thresholds, or quotas. All 16 CAL points and 56 target endpoints are retained, including reject-all. The final accounting finds 15 target endpoints with observed one-percent violation, 14 empirically feasible but insufficiently supported endpoints, 11 abstentions, and 16 fixed-coverage ordering-only endpoints. The adapted readouts still provide no one-percent-risk endpoint with the required question and component counts.

Endpoint sensitivity. A separate endpoint-sensitivity census classifies 3,644 SOURCE candidates: 2,268 strict matches, 91 matches only after a global column permutation, 489 shape failures, 250 value failures, and 546 unresolved outcomes. Under the relaxed endpoint, NR/NRP/pair/agreement risks become 3.851%/2.548%/2.762%/2.059%, all above one percent. Original labels and primary results remain unchanged. An alternative endpoint can explain part of a numerical difference without establishing that a generated query expresses the requested semantics.

Interpretive scope. These bounded alternatives were motivated after earlier outcomes and retain historical exposure. A single unsuccessful architecture does not prove that larger models, improved objectives, or genuinely new data cannot help. The supported statement is narrower: the tested interventions do not supply a reliable one-percent policy with adequate support, while the primary Qwen residual ordering improvement remains measurable. Future independent confirmation would require a newly declared population and sampling law, not another seed on the same existing bank.

G REPEATED CALIBRATION DRAWS WITH FROZEN READOUTS

This sensitivity analysis fixes the original SOURCE-fitted Qwen and Granite VALUES readouts and the EVAL banks. The CAL reserve contains 2,048 components, four questions per component, and two producer answers per question: 8,192 unique questions and 16,384 candidate rows. A question-label budget B therefore corresponds to $2B$ candidate rows, not B rows. At each of 256 paired draws, budgets $B \in \{512, 1024, 2048\}$ are nested. Breadth sampling takes one question from each of B components; depth sampling takes four questions from each of $B/4$ components. The designs share component order. At breadth $B = 2048$, every component is present and only the question within each component changes.

Each CAL candidate has weight $1/(2B)$. Unavailable candidates retain denominator mass, unresolved outcomes count as errors, and a whole-tie prefix maximizes correct mass, then coverage, subject to empirical risk at most .01, with numerical tolerance 10^{-12}. An empty feasible set gives abstention. EVAL weights and outcomes remain fixed: the disjoint bank has 19,940 questions in 6,155 components (39,880 candidate rows), and the pool-known bank has 3,996 questions in 1,363 components (7,992 rows). No EVAL outcome chooses a draw, budget, threshold, or SOURCE refit. Historical exposure remains unchanged. N and the matched scalar readout S have identical contrasts in these draws.

The primary contrast is SR $-$ S for Qwen on disjoint EVAL at breadth $B = 2048$. Mean correct-mass gain is $+1.26398$pp, with 206 positive and 50 negative draws. Mean coverage and wrong-mass differences are $+1.26201$pp and $-.001969$pp. The conditional 2.5th and 97.5th draw quantiles are $[-.57181, 4.95129]$pp. NRP $-$ S averages $+1.26046$pp, with 200 positive and 56 negative draws and quantiles $[-.98526, 4.96325]$pp. These overlapping draws are dependent sensitivity analyses; the quantiles are not confidence intervals, population guarantees, or a simultaneous risk certificate.

Table 13: Qwen disjoint EVAL across 256 breadth draws at $B = 2048$. Risk is the arithmetic mean of each nonempty draw's ratio, not a pooled ratio. Support requires at least 300 accepted unique questions and 100 components. All arms have nonempty, empirically feasible CAL prefixes in all draws.

Arm	Mean coverage (%)	Mean risk (%)	Risk $\leq 1\%$	Support	Both
S (=N)	.57915	8.46776	0/256	31/256	0/256
SR	1.84116	1.63762	31/256	127/256	1/256
NRP	1.84343	1.87577	4/256	114/256	4/256

SR exceeds one-percent EVAL risk in 225 of 256 draws; 30 of the 31 risk-feasible draws lack sufficient support. NRP exceeds the target in 252 draws. Mean weighted acceptance-set Jaccard similarities to the original policy are .25966, .49276, and .43836 for S, SR, and NRP. Thus repeated calibration does not establish stable acceptance sets or recover a supported one-percent policy. The original NRP operating point, with 3.326971% coverage and 2.667192% risk, remains a separate result.

Table 14: All repeated-CAL panels. Each gain entry gives mean correct-mass differences SR − S / NRP − S in percentage points. Each success entry counts draws satisfying both one-percent risk and minimum support, again SR / NRP out of 256. Pool-known membership refers to the full reserve, not every sampled CAL design.

Judge	B	Design	Disjoint ΔG	Both	Pool-known ΔG	Both
Qwen	512	Breadth	$1.95677/2.21145$	$10/10$	$2.36111/2.65643$	$0/0$
Qwen	512	Depth	$1.97427/2.08287$	$8/8$	$2.38751/2.52112$	$0/0$
Qwen	1024	Breadth	$1.55823/1.67171$	$15/14$	$1.85833/1.99730$	$0/0$
Qwen	1024	Depth	$1.49355/1.55006$	$11/11$	$1.79872/1.87161$	$0/0$
Qwen	2048	Breadth	$1.26398/1.26046$	$1/4$	$1.48369/1.51490$	$0/0$
Qwen	2048	Depth	$1.04682/1.08145$	$1/3$	$1.23719/1.30844$	$0/0$
Granite	512	Breadth	$-.00399/-.50793$	$0/0$	$-.12979/-.81362$	$0/0$
Granite	512	Depth	$-.05481/-.48737$	$0/0$	$-.17806/-.79039$	$0/0$
Granite	1024	Breadth	$-.09884/-.36577$	$0/0$	$-.26633/-.62701$	$0/0$
Granite	1024	Depth	$-.11925/-.36279$	$0/0$	$-.26745/-.61371$	$0/0$
Granite	2048	Breadth	$-.14024/-.23683$	$0/0$	$-.28844/-.40997$	$0/0$
Granite	2048	Depth	$-.15762/-.24807$	$0/0$	$-.30414/-.41514$	$0/0$

Every Granite mean contrast is negative across budgets, designs, and EVAL roles, and no Granite draw passes both risk and support screens. Higher mean coverage at a smaller CAL budget does not imply safer acceptance. These results separate conditional Qwen ordering gains from draw-uniform superiority, judge portability, and SOURCE-fitting uncertainty.