UNIT-WEIGHT POSTERIOR CORESETS: SHARP KL GEOMETRY AND CERTIFICATION LIMITS

Anonymous authors
Paper under double-blind review

ABSTRACT

We characterize posterior fidelity for literal evidence subsets, with no reweighting and an exact cardinality. Under additive logits, omitted-evidence range equals the maximum posterior log-odds distortion. We derive the exact minimum and maximum forward Kullback–Leibler (KL) divergence at a prescribed range for a fixed full posterior; the minimum depends only on its two smallest masses. The lower envelope turns a range-solver lower bound into a KL optimality certificate. With smallest posterior mass β, exact range initialization followed by KL-decreasing refinement achieves the sharp factor $(1-\beta)/\beta$. A matching binary construction establishes sharpness even with a unique full winner and any fixed exchange radius. For nonadditive consumers, aggregate signed transfer error connects posterior control to decision certification. In 1,800 controlled instances, refinement reduces mean true range by 10% relative to KL-greedy, with mean true KL .3089 versus .3166. RACE, DREAM, and SciFact evaluations with two neural targets identify 172–232 winner-preserving omissions per operating point, while all four calibrated constructions issue zero nontrivial joint winner certificates. Together, the exact envelopes, optimization diagnostics, and source-aware neural evaluations characterize when posterior control supports informative certification.

1 INTRODUCTION

A faithful subset should preserve how the complete evidence supports competing answers, not merely retain relevant statements. Three complementary questions connect an evidence-selection objective to a useful certificate: does the objective control the intended posterior, does the optimizer find a good subset, and is the resulting bound informative enough to certify a decision while deleting evidence? We connect these questions through fixed-posterior geometry, certified optimization, and direct-context evaluation.

We study *unit-weight posterior coresets*: exactly k of n evidence atoms are displayed once, without amplifying their contributions. Under an explicit additive score model, the range of the omitted logit vector is exactly the largest change in pairwise posterior log odds. This is the classical Hilbert projective geometry, with established sharp KL and total-variation (TV) consequences (Dong et al., 2020; Cohen & Fausti, 2024). It makes range minimization interpretable, but a distribution-free upper bound alone does not say how close a selected subset is to the KL-optimal subset.

Our first contribution fills this gap by conditioning on the full posterior. We derive exact lower and upper KL envelopes at fixed residual range. The lower envelope yields a computable certificate of KL suboptimality from a range-solver lower bound. It also gives a positive approximation guarantee when posterior masses are bounded away from zero. We complement it with a construction showing that every fixed exchange neighborhood can be disconnected under the certificate constraint, even with exact scores and exact initialization. The exact worst-case factor is $(1-\beta)/\beta$, where β is the smallest full-posterior mass; a binary construction approaches this factor for each fixed $\beta < 1/2$. This identifies the exact performance guarantee and its optimization boundary within the additive model (Figure 1).

The second contribution concerns the *meaning* of a neural certificate. The direct full-minus-subset residual equals the estimator residual plus a joint mismatch/nonadditivity error. These error components share terms and can cancel arbitrarily; their separate magnitudes do not identify the aggregate-error quantile. We retain signed robust and conditional KL transfer tests, but distinguish their as-

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.

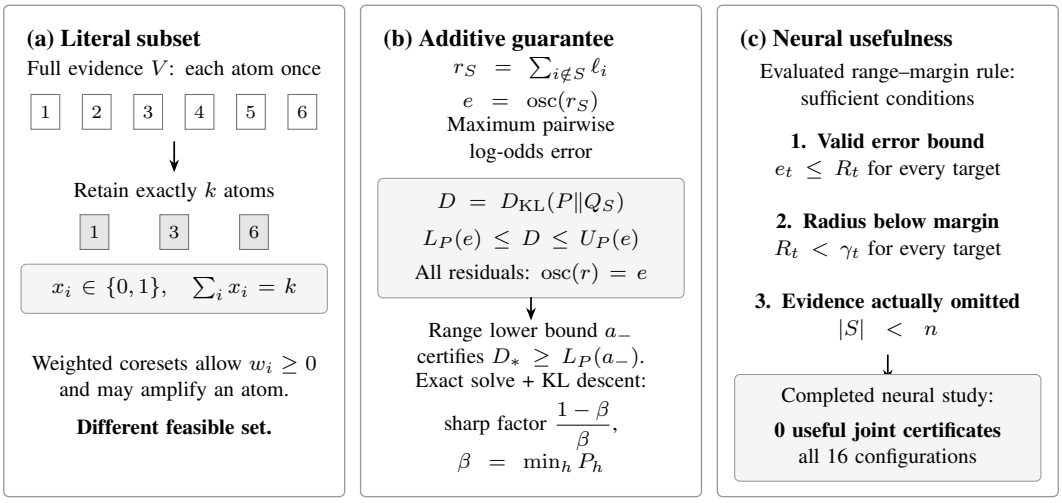

Figure 1: **From evidence selection to an informative certificate.** (a) Literal summaries retain atoms once; weighted coresets use a different feasible set. (b) At a fixed positive additive posterior P, envelopes over all residuals of range e give a solver-based KL lower bound and a sharp conditional performance ratio (Theorem 1, Corollary 2). (c) The evaluated range–margin rule requires valid radii below full-winner margins and actual omission; these conditions are sufficient, not necessary for other signed tests. All 16 evaluated construction/operating-point combinations have zero useful issuance. This is an observed pipeline limitation, not universal impossibility. Panels (a,b) are schematic mathematics, not measurements.

sumptions from properties demonstrated by data. In particular, a high observed coverage fraction need not produce any certificate that both preserves a winner and removes an atom.

The completed neural study makes this distinction concrete. Across RACE, DREAM, and SciFact, two target models, two atomizations, and two item budgets, 172–232 omissions per operating point retain both unique full winners under the recorded argmax rule, and target-wise exact-range oracles certify 92–146. All four calibrated constructions nevertheless issue *zero* joint winner certificates with evidence omitted. Figure 2 separates preserved decisions, oracle-certifiable omissions, and calibrated issuance without mixing full retention into compression. The oracle outcomes are diagnostic, not a deployable method.

The experiments complement the theory by quantifying a range–KL-cost tradeoff and the conditions for informative certification. Controlled results support this specific tradeoff rather than uniform dominance. Exhaustive subset graphs diagnose constrained-search barriers; source-clustered neural comparisons qualify earlier example-level inferences. All empirical tables report the completed study; the supplement includes aggregate outputs, anonymized analysis code, and prepared examples for regeneration (Appendix J). The extremal and approximation results are analytical contributions, not empirical improvements to its pipeline.

2 Posterior geometry at a fixed reference

Let $V = [n] = \{1, \ldots, n\}$, $n \geq 1$, and $k \in \{0, \ldots, n\}$. For finite base logits $b \in \mathbb{R}^m$, atom contributions $\ell_i \in \mathbb{R}^m$, positive probabilities on $m \geq 2$ hypotheses, and natural logarithms, write

$$z = b + \sum_{i \in V} \ell_i, \quad z_S = b + \sum_{i \in S} \ell_i, \quad P = \mathrm{softmax}(z), \quad Q_S = \mathrm{softmax}(z_S), \quad |S| = k. \tag{1}$$

The residual is $r_S = z - z_S$ and $\epsilon(S) = \mathrm{osc}(r_S)$, where $\mathrm{osc}(v) = \max v - \min v$ and $\mathbf{1}$ is the all-ones vector. Pairwise normalizers cancel, giving the exact, common-shift-invariant identity

$$\max_{h,j} \left| \log \frac{P_h}{P_j} - \log \frac{(Q_S)_h}{(Q_S)_j} \right| = \mathrm{osc}(r_S) = 2 \min_c \|r_S - c\mathbf{1}\|_\infty. \tag{2}$$

For any residual r, define the tilt $\mathcal{T}_r(P)_h = P_h e^{-r_h}/\mathbb{E}_P e^{-r}$ and

$$\mathcal{J}(P, r) = P^\top r + \log \mathbb{E}_P e^{-r} = D_{\mathrm{KL}}(P \| \mathcal{T}_r(P)). \tag{3}$$

The known distribution-free bound is $\mathcal{J}(P, r) \le F(\operatorname{osc} r)$, where

$$F(e) = \log \frac{\exp(e) - 1}{e} - 1 + \frac{e}{\exp(e) - 1} \le e^2/8, \qquad F(0) = 0. \tag{4}$$

It also bounds reverse KL; $\mathrm{TV}(P, Q_S) \le \tanh(\epsilon(S)/4)$ with half-ℓ^1 TV (Dong et al., 2020; Cohen & Fausti, 2024). We build on these established comparisons.

2.1 Exact posterior-dependent KL envelopes

The full posterior is fixed when subsets are compared. Exploiting its masses gives more than a worst-case upper bound. Let $\beta_1 \le \beta_2$ be the two smallest masses of P, let $P(A) = \sum_{h \in A} P_h$, and define

$$\begin{aligned} \phi_v(e) &= ve + \log(1 - v + ve^{-e}), \\ L_P(e) &= (\beta_1 + \beta_2)\phi_{\beta_1/(\beta_1+\beta_2)}(e), \\ U_P(e) &= \max_{\varnothing \ne A \subsetneq [m]} \phi_{P(A)}(e). \end{aligned} \tag{5}$$

Theorem 1 (Exact fixed-posterior envelopes)**.** *For every positive P and $e \ge 0$,*

$$\min_{\operatorname{osc}(r)=e} \mathcal{J}(P, r) = L_P(e), \qquad \max_{\operatorname{osc}(r)=e} \mathcal{J}(P, r) = U_P(e). \tag{6}$$

Both extrema are attained over residual vectors of range e; a given subset family need not attain them. L_P is strictly increasing on $(0, \infty)$. For $\beta = \min_h P_h$,

$$2\beta \log\cosh(e/2) \le L_P(e) \le U_P(e) \le \min\{F(e),\, 2(1 - \beta) \log\cosh(e/2)\}. \tag{7}$$

For uniform P, $L_P(e) = (2/m) \log\cosh(e/2)$. Moreover, $U_P(e)/L_P(e) \to (1 - \beta)/\beta$ as $e \to \infty$.

Why it holds. Shift r into $[0, e]^m$. Convexity puts the maximum at a binary vertex $e\mathbf{1}_A$. For the minimum, fix one coordinate a at zero and another b at e. All other coordinates equal

$$t_{ab} = -\log \frac{P_a + P_b e^{-e}}{P_a + P_b}.$$

This is the stationary point on the endpoint face and hence its minimum. The minimizing endpoints belong to the two least likely hypotheses, with the smaller mass at e; optimizing over faces proves the formula (Appendix B). The upper-vertex argument is consistent with established Hilbert-polytope geometry (Cohen & Fausti, 2024); the lower-envelope calculation and its optimization consequences are the focus here.

Local geometry and computation. The lower envelope also quantifies small residuals:

$$L_P(e) = \frac{\beta_1 \beta_2}{2(\beta_1 + \beta_2)} e^2 + O(e^3) \qquad (e \to 0).$$

Its coefficient depends only on the two smallest posterior masses. Both the exact lower envelope and its local quadratic scale require one $O(m)$ scan; the displayed exact evaluation of U_P enumerates subset masses. For a positive reference and $e > 0$, the envelopes coincide only for two equally likely hypotheses. Otherwise residual orientation or interior coordinates affect KL (Appendix B).

3 Optimization guarantees and sharpness

3.1 Robust initialization and uncertainty

Maximum Log-Odds Coresets (MLOC) minimize $\widehat{\epsilon}(S) = \operatorname{osc}(\sum_{i \notin S} \widehat{\ell}_i)$ over $|S| = k$. Given pairwise radii $\rho_{ihj} \ge 0$ for $h < j$, Robust Log-Odds Coresets (RLOC) minimize

$$C(S) = \max_{h<j} \left(\left| \sum_{i \notin S} (\widehat{\ell}_{ih} - \widehat{\ell}_{ij}) \right| + \sum_{i \notin S} \rho_{ihj} \right). \tag{8}$$

Both have mixed-integer linear programming (MILP) formulations (Appendix C). A feasible incumbent has a certificate even when optimality is unresolved. On *simultaneous* atomwise interval validity, $\epsilon(S) \leq C(S)$ for every subset, including an adaptively selected one. Marginal interval coverage does not establish this event.

If $\rho_{ihj} = \rho$, fixed cardinality gives

$$C(S) = \widehat{\epsilon}(S) + (n-k)\rho. \tag{9}$$

Exact MLOC and RLOC minimizers therefore coincide. This uncertainty model changes the bound, not the optimal selected set. Certificate-Preserving Coreset (CPC) refinement starts from a RLOC incumbent S_0 and, for $\tau \geq 0$, makes deterministic KL-decreasing single exchanges under $C(S) \leq (1+\tau)C(S_0)$. Under constant radii the nominal threshold is $(1+\tau)\widehat{\epsilon}(S_0) + \tau(n-k)\rho$: larger uncertainty also allows more nominal slack. The controlled and exhaustive implementations use at most 20 passes; the neural implementation continues until no improving feasible single exchange remains. Neither rule guarantees global KL optimality (Appendix G).

3.2 A CONSTRUCTIVE KL OPTIMALITY CERTIFICATE

Corollary 2 (From a range lower bound to KL quality)**.** *Fix a full posterior P and residuals from the same model. Let $\mathcal{F}$ be a nonempty subset family, $a_- \geq 0$ a valid lower bound on $\min_{S\in\mathcal{F}} \operatorname{osc}(r_S)$, and $S \in \mathcal{F}$ have loss d. Then*

$$L_P(a_-) \leq D_* := \min_{T\in\mathcal{F}} \mathcal{J}(P, r_T) \leq d, \quad d - D_* \leq d - L_P(a_-). \tag{10}$$

If $a_- > 0$, $d/D_* \leq d/L_P(a_-)$. If S_0 globally minimizes range at e_* and refinement never increases KL, then, for $e_* > 0$,

$$\frac{\mathcal{J}(P, r_S)}{D_*} \leq \frac{U_P(e_*)}{L_P(e_*)} \leq \frac{1-\beta}{\beta}, \qquad \beta = \min_h P_h. \tag{11}$$

At $e_* = 0$, both the initialization and any KL-decreasing refinement have zero loss.

For every $T \in \mathcal{F}$, Theorem 1 gives $\mathcal{J}(P, r_T) \geq L_P(\operatorname{osc}(r_T)) \geq L_P(a_-)$. At exact range initialization, KL descent yields $\mathcal{J}(P, r_S) \leq \mathcal{J}(P, r_{S_0}) \leq U_P(e_*)$, whereas $D_* \geq L_P(e_*)$. Dividing establishes the ratio. For two equally likely hypotheses, $L_P(e) = U_P(e) = \log\cosh(e/2)$, so range minimization is exactly forward-KL optimal. This is a conditional performance ratio, not a polynomial-time approximation algorithm; it holds even when the certificate-feasible graph is disconnected. A verified MILP lower bound suffices for Eq. (10); an unresolved incumbent alone does not. When evaluated using $\widehat{P}, \widehat{r}$, it certifies *nominal* KL quality only. The completed experiments did not measure these envelopes or this certificate; no retrospective tightness claim is made. Appendix C.2 gives an endpoint-error correction when computed KL descent is not exact.

3.3 THE POSTERIOR-MASS DEPENDENCE IS NECESSARY

Theorem 3 (Sharp factor despite bounded exchanges)**.** *Fix any $\beta \in (0, 1/2)$, integer $q \geq 1$, $\tau > 0$, and $0 < \zeta < (1-\beta)/\beta - 1$. There is an exact-score, zero-uncertainty instance with $m = 2$, $P = (1-\beta, \beta)$, $k = q+1$, $n = 2k$, a unique range-minimizing initialization, and exactly two subsets satisfying $C \leq (1+\tau)C(S_0)$. They differ by k exchanges. The initialization is isolated under every feasible exchange of at most q atoms and its KL divided by the constrained optimum is greater than $(1-\beta)/\beta - \zeta$.*

The construction assigns opposite residual orientations to two complete atom blocks; mixed subsets violate the certificate threshold. Appendix C.3 gives the vectors and the finite-parameter argument. Since the unconstrained optimum is no larger, no smaller β-only constant can replace Eq. (11) for such refinements. An additional three-hypothesis construction keeps atom vectors fixed while $\beta \downarrow 0$ and still has ratio $\Theta(1/\beta)$. These are worst-case limits, not a claim that typical neural instances attain the factor.

4 FROM ADDITIVE FIDELITY TO A USEFUL NEURAL CERTIFICATE

4.1 THE RELEVANT TRANSFER ERROR IS AGGREGATE AND SIGNED

Let $f_T(A)$ be a target's direct logits under a fixed context renderer, and $\ell_{T,i} = f_T(\{i\}) - f_T(\varnothing)$. Define $g_T(A) = f_T(\varnothing) + \sum_{i \in A} \ell_{T,i}$ and $\Delta_T(A) = f_T(A) - g_T(A)$. For estimator increments $\ell_{E,i}$ and $r_E = \sum_{i \notin S} \ell_{E,i}$,

$$r_T := f_T(V) - f_T(S) = r_E + \underbrace{\sum_{i \notin S} (\ell_{T,i} - \ell_{E,i})}_{M(S)} + \underbrace{\Delta_T(V) - \Delta_T(S)}_{N(S)}. \tag{12}$$

Expanding $E_S = M + N$ cancels every target singleton and target empty-context term. Separate $\mathrm{osc}(M)$ and $\mathrm{osc}(N)$ can diverge while $r_T = r_E = 0$; their marginal laws can be identical in two models with radically different aggregate-error quantiles (Appendix E.2). Therefore large separate diagnostics do not identify a direct-error floor or the causal benefit of repairing one component.

For a supplied nonempty compact aggregate set U, subset logits $s = f_T(S)$, and nominal residual $\widehat{r}$, a candidate a is the unique full winner for all admissible errors exactly when

$$s_a - s_j + \widehat{r}_a - \widehat{r}_j > h_U(-v_{aj}) \quad (j \neq a), \quad v_{aj} = \mathbf{e}_a - \mathbf{e}_j. \tag{13}$$

Here $\mathbf{e}_j$ is the jth coordinate vector and $h_U(v) = \sup_{u \in U} v^\top u$ is the support function. This standard robust-counterpart principle (Ben-Tal & Nemirovski, 1999) preserves signed contrasts and joint error structure. It does not provide a valid fitted U automatically.

4.2 CONDITIONAL KL TRANSFER

Corollary 4 (Reference and residual transfer)**.** *Let $D = \mathcal{J}(P, r)$ and $\widehat{D} = \mathcal{J}(\widehat{P}, \widehat{r})$, with positive references. If $\mathrm{osc}(r - \widehat{r}) \leq \delta$ and $\mathrm{osc}(\log P - \log \widehat{P}) \leq \eta$, then*

$$e^{-\eta} \left[\sqrt{\widehat{D}} - \delta/\sqrt{8} \right]_+^2 \leq D \leq \left(e^{\eta/2} \sqrt{\widehat{D}} + \delta/\sqrt{8} \right)^2. \tag{14}$$

Here $[x]_+ = \max\{x, 0\}$. This composes established reference-weight/Jensen comparison and nonnegative-smooth square-root stability (Dragomir, 2006; Glaeser, 1963; Srebro et al., 2010); Appendix E proves it, including the sharp residual constant. It is not a new general divergence inequality. Computing η, δ from actual target scores makes an *oracle audit*, not a deployable bound. In the completed neural audit these intervals separate no CPC-versus-baseline ordering, even with oracle allowances.

4.3 CERTIFYING DECISION-PRESERVING OMISSION

Proposition 5 (Three distinct decision events)**.** *For a selected subset and targets t, let $e_t = \mathrm{osc}(r_t)$, let $R_t \geq 0$ be reported radii, and let γ_t be the full winner margin, zero if not unique. Define $B = \{R_t \geq e_t \ \forall t\}$, $I = \{|S| < n,\ R_t < \gamma_t \ \forall t\}$, $O = \{|S| < n,\ e_t < \gamma_t \ \forall t\}$, and W as agreement of each subset's deterministic argmax with its unique full winner, with $|S| < n$. Then*

$$I \cap B \subseteq O \subseteq W. \tag{15}$$

If $\Pr(B) \geq 1 - \alpha$ holds for the frozen selected pipeline, $\Pr(I \setminus W) \leq \alpha$. This is not conditional precision among issued certificates.

A common calibrated radius is the special case $R_t = R$. The exact-range oracle O is target-wise and may exploit information unavailable to a common-radius construction. Equation (15) distinguishes conservative bounds, conservatism of the range–margin test, and genuinely changed winners. The condition $R_t < \gamma_t$ is sufficient, not necessary for preservation or for a signed certificate.

Why signed information helps. For $a > 0$, take full logits $z = (a, 0)$ and subset logits $z_S = (2a, 0)$. Both have the same unique winner, yet $r = (-a, 0)$ has range a, equal to the full margin, so the strict range–margin test fails. The signed test with exact residual and $U = \{0\}$ certifies the positive full margin $2a - a = a$. This is a literal two-atom example with zero base logits, $\ell_1 = (2a, 0)$ retained and $\ell_2 = (-a, 0)$ omitted (Appendix D), not an untested neural improvement.

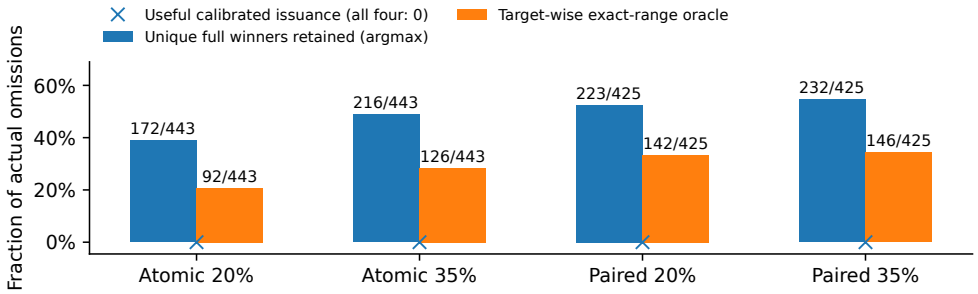

Figure 2: **Decision preservation and certification.** Completed neural evaluation, restricted to actual omissions: 443 atomic or 425 paired bases per operating point. Counts and percentages are descriptive, with both targets required. Agreement uses unique full winners and the recorded deterministic subset-argmax rule; subset ties are not excluded. The exact-range oracle uses each target's own error and margin; calibrated pipelines use a common radius and are not granted oracle information. All four constructions have zero useful issuance. Full retention (57 atomic, 75 paired bases) is excluded. No new model calls are used.

Setting	CPC KL	LLR KL	CPC minus LLR, 95% CI	Coverage	Median R
Atomic .20	1.7109	1.3480	$+.3629\ [+.1471, +.6052]$	.956	17.625
Atomic .35	1.1695	1.0207	$+.1488\ [-.0044, +.3051]$	.958	14.875
Paired .20	1.1091	1.0662	$+.0429\ [-.0578, +.1522]$	.956	13.875
Paired .35	1.0155	.9633	$+.0522\ [-.0443, +.1542]$	.958	14.188

Table 1: Completed direct-context evaluation. LLR denotes log-likelihood ratio; CI denotes confidence interval. KL means average two targets per base; pointwise supplementary intervals use 5,000 source-cluster resamples within dataset. Positive differences disfavor CPC. Coverage and R refer to additive-nominal joint calibration at $\alpha = .05$; they are observed values, not validated source-exchangeability guarantees. All four calibrated constructions issue zero winner certificates with any atoms omitted.

5 Neural posterior fidelity and certification

Protocol and scope. Prepared RACE, DREAM, and SciFact examples (Lai et al., 2017; Sun et al., 2019; Wadden et al., 2020) use Qwen2.5-1.5B-Instruct as estimator and Qwen2.5-3B-Instruct and Phi-3.5-mini-instruct as targets (Qwen Team, 2024a;b; Microsoft, 2024). CUDA scoring sums answer-label continuation log probabilities and normalizes over fixed alternatives. Prompt tokens and space-prefixed label tokens are encoded separately and then concatenated; the resulting token-sequence convention, rather than re-tokenization of the combined string, defines the evaluated scores. These are discriminative answer-score effects, not verified independent likelihood factors or unrestricted-generation accuracy. Atomic and adjacent-paired representations share base examples. For fraction $f \in \{0.20, 0.35\}$,

$$k = \min\{n, \max\{2, \min\{10, \mathrm{round}(nf)\}\}\}. \tag{16}$$

The study is item-budget matched, not token-budget or runtime matched. Separate base-ID roles contain 150 interval-fitting, 300 development, 300 residual-calibration, and 500 evaluation examples. Evaluation has 240 RACE, 160 DREAM, and 100 SciFact bases. The 6,400 held-out target bundles yield 83,200 method rows, including 52,000 evaluation rows. Methods, targets, atomizations, and budgets repeat bases; they are not independent replicates. All 13 selector rows appear in Table 3. Tuning rules and calibration results are detailed in Appendix G and Table 2; checkpoint revisions accompany the supplementary code.

Coverage and decision-preserving compression. For frozen CPC selections, let $e(x) = \max_t e_t(x)$ and a nominal error estimate be $a(x)$. Dataset-stratified split conformal calibration uses $[e(x) - a(x)]_+$ and $R = a + q_{.95}$. Nominals are additive range, zero, factorwise robust range, and a development-fitted predictor. Primary observed coverage is .956–.958 with median radii 13.875–17.625 (Table 1); no joint winner certificate is issued. Factorwise radii cover every evaluated error

Nominal	Atomic .20	Atomic .35	Paired .20	Paired .35
Additive	.956 / 17.625	.958 / 14.875	.956 / 13.875	.958 / 14.188
Zero	.956 / 16.250	.952 / 14.875	.936 / 14.250	.940 / 14.250
Factorwise	1.000 / 221.000	1.000 / 175.750	1.000 / 88.625	1.000 / 88.625
Development predictor	.956 / 17.625	.958 / 14.875	.956 / 13.875	.958 / 14.188

Table 2: All four calibration constructions: each entry is observed joint coverage / median radius R at $\alpha = .05$, on the same 500 evaluation bases per operating point. All 16 combinations issue zero joint winner certificates with evidence omitted. Factorwise issuance occurs only with full retention; empirical coverage is not a source-exchangeability guarantee.

Method	Atomic .20	Atomic .35	Paired .20	Paired .35
KL-greedy	1.7350	1.3469	1.2273	1.1462
CPC	1.7109	1.1695	1.1091	1.0155
Lead	1.5978	1.3055	1.2674	1.1200
Lexical	0.9461	0.7504	0.9676	0.8324
LLR norm	1.3480	1.0207	1.0662	0.9633
MLOC	1.4527	1.1425	1.1114	1.0133
MMR lexical	1.0637	0.8289	0.9070	0.8293
Random (seed 113)	1.4414	1.2043	1.2716	1.1256
Random (seed 17)	1.5569	1.2887	1.2141	1.1561
Random (seed 29)	1.4956	1.1937	1.0317	1.0734
Random (seed 43)	1.4699	1.1764	1.1770	1.1013
Random (seed 71)	1.5685	1.2061	1.2178	1.1727
RLOC	1.5263	1.1949	1.1209	1.0092

Table 3: All neural selectors: mean direct forward KL over 1,000 target rows (500 bases, two targets) per operating point. Lower is better. All five random seeds are retained. Columns share bases and are not independent samples. These means are descriptive; the source-cluster intervals in Table 1 apply only to CPC versus LLR norm.

and issue certificates on .114 of atomic and .150 of paired bases, but *every issuance retains all atoms*. All 16 construction/operating-point combinations have zero useful issuance. Table 2 reports every construction, separating radius size from observed coverage.

Among actual omissions, joint agreement with the unique full winners is 172/443, 216/443, 223/425, and 232/425. Target-wise exact-range oracle success is 92/443, 126/443, 142/425, and 146/425 (Figure 2). Hence some existing selections preserve decisions but are uncertified; others preserve decisions even when the scalar exact-range test fails. These are observed properties of the selected subsets, not evidence that another calibrator will attain the oracle. The omission-only signed-ball oracle counts 59, 88, 86, and 89 use a different uncertainty description and do not establish dominance over the exact-range test.

Signed transfer diagnostics. At atomic $f = .20$, mean mismatch and nonadditivity-difference ranges are 53.106 and 53.855, versus direct range 5.506. The supplied score audit reports that the vector identity reconstructs scores within 2.84×10^{-14}. The components cannot be interpreted as additive shares of this scalar error; their coupling is essential. Across the four operating points, median oracle $\log_{10}(U/D)$ is 3.75–4.72 on positive-loss omitted target rows. None of the eight CPC-versus-KL-greedy/LLR operating-point comparisons has an interval-separated ordering in either direction. These diagnostics identify aggregate-error control as the relevant certification quantity; they do not establish whether nonadditivity or factor mismatch is the dominant causal source. Appendix F preserves all denominators and diagnostics.

Baseline context. Table 3 includes every evaluated selector, including maximal marginal relevance (MMR) lexical selection (Carbonell & Goldstein, 1998). CPC has lower displayed direct KL than additive KL-greedy at all four operating points, but lexical selection has the lowest means for atomic evidence and MMR lexical for paired evidence. These descriptive comparisons distinguish the value

Method	True KL	True range	Est. KL	Nominal range
Random	1.0324	10.9008	1.2076	11.8701
Position	0.2493	8.7164	0.5296	10.2412
Lexical	0.2493	8.7164	0.5296	10.2412
Coverage	0.5520	9.7112	0.7793	10.9090
LLR norm	0.6384	7.9954	0.7014	8.1465
KL-greedy	0.3166	7.5994	0.0207	6.6519
MLOC	0.3120	6.5370	0.0041	4.3604
RLOC	0.3102	6.5390	0.0041	4.3618
CPC	0.3089	6.8394	0.0009	4.7276

Table 4: Completed additive controlled evaluation: means over 1,800 paired instances, with equal condition weights. All columns are errors (lower is better). Stable true KL uses float64 log-softmax; estimated KL and nominal range use estimator factors. Methods are not runtime matched.

of certificate-constrained refinement within the additive-selector family from superiority over all baselines. They do not establish statistical equivalence or an adjusted ranking of methods.

Source-aware inference and calibration. The 500 evaluation bases form 278 source components under exact full-text, RACE source-ID, and SciFact document-ID matching. Across all 1,250 bases, 62 components span roles. No direct residual-calibration/evaluation source overlap was detected, but fitted-nuisance overlap prevents an unqualified source-exchangeability claim. Near duplicates can remain unidentified. We keep all records rather than remove unfavorable overlaps retrospectively. Source-clustered intervals make atomic $f = .35$ inconclusive where the example-level interval had excluded zero; atomic $f = .20$ remains unfavorable to CPC. Neither paired comparison resolves the sign. Cluster bootstrapping does not validate conformal coverage (Barber et al., 2023), and an interval containing zero does not establish equivalence.

6 Controlled fidelity and exact optimization diagnostics

The completed controlled study has $n = 40, m = 5$, four generators, noise levels $\sigma \in \{0, .25, .50\}$, retained fractions $\{.20, .35, .50\}$, 15 development and 50 evaluation instances per condition: 1,800 evaluations across 36 conditions. Development selects τ from $\{0, .02, .05, .10, .20\}$. All methods share each instance and exact cardinality; each MILP call has a 120-second limit and requested zero relative gap. Constant pairwise radii 2.77σ are approximately 95% *marginal* Gaussian pairwise radii, not simultaneous coverage. Table 4 includes every method and the generator-induced lexical/position tie.

CPC reduces mean true range by 10.00% relative to KL-greedy, with observed true KL .3089 versus .3166. Supplementary paired stratified inference supports the 1% range-improvement threshold and a permitted 10% KL-cost threshold; approximate bootstrap p-values with conservative Monte Carlo (MC) adjustment are .001237 and .003994. This establishes the stated range–KL-cost tradeoff, rather than KL superiority. Position and lexical mean true KL .2493, below CPC, and MLOC/RLOC attain smaller true ranges, so the benefit is specific to the comparison with KL-greedy. Stable-metric auditing changes no evaluation KL by more than 10^{-9}; all 4,500 CPC development/evaluation records satisfy their incumbent-relative constraints. Constant-radius nominal objectives agree in 931 evaluation cases with both solves optimal.

In the separate exhaustive $n = 16, m = 5$ panel, CPC reaches 74/96 constrained estimated-KL optima. Eighteen optima are unreachable from its terminal component and four are reachable but blocked by a local minimum. No terminal node has an improving feasible single exchange. Component separation accounts for 92.5% of summed estimated-KL regret. A frozen dimension-change panel ($n = 18, m = 7, k = 9$) yields 18 optimal, 18 unreachable-optimum, and four connected-local-minimum cells. These panels share tensors within cells and are descriptive structural tests, not independent failure-rate estimates or neural confirmation (Figure 3). They illustrate the single-exchange mechanism characterized by Theorem 3; the $q > 1$ construction is established analytically.

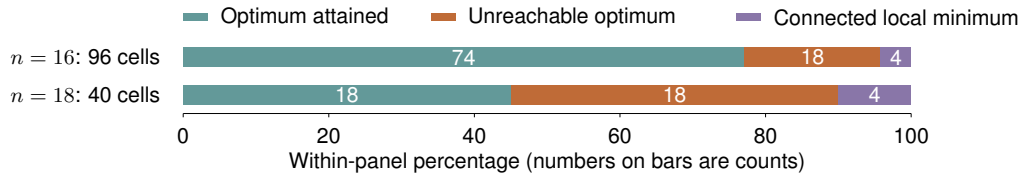

Figure 3: **Exhaustive optimization diagnostics.** The 96-cell panel shares 12 latent tensors; the 40-cell dimension-change panel shares 20. Counts are descriptive. Unreachable means no constrained optimum is in CPC's terminal single-exchange component. The panels evaluate single exchanges; larger exchange radii are covered by Theorem 3, not by these measurements.

7 Related work

Positioning. Hilbert geometry and sharp range-to-KL/TV inequalities are established (Cohen & Fausti, 2024; Dong et al., 2020). Our fixed-reference lower envelope, sharp approximation constant, and matching bounded-exchange obstruction supplement, rather than replace, these tools. This is a finite-support exact-range characterization, not a new general divergence inequality or a distribution-free compression guarantee. Reference-weight and square-root transfer use classical Jensen and smoothness arguments (Dragomir, 2006; Glaeser, 1963); broader Bayesian perturbation and coreset bounds address different assumptions (Sprungk, 2020; Campbell, 2024). Greedy Iterative Geodesic Ascent (GIGA) reweights representatives geometrically and sparse variational inference (SparseVI) optimizes sparse nonnegative weights with reverse KL (Campbell & Broderick, 2018; Campbell & Beronov, 2019). Their feasible sets differ from literal fixed-cardinality subsets. Discrepancy-based coreset theory already studies how signed balancing controls approximation (Karnin & Liberty, 2019); our claims concern the fixed-posterior KL envelope and certificate-constrained search, not novelty of vector balancing itself. Robust support functions are classical (Ben-Tal & Nemirovski, 1999). Finally, token-compression successes such as LLMLingua-2 (Pan et al., 2024) address token-efficiency objectives complementary to our certificate analysis; we make no aligned token-efficiency or method-superiority comparison.

Limitations. The guarantees require positive reference masses and the stated score and uncertainty assumptions. The $(1 - \beta)/\beta$ factor can be weak for peaked posteriors and requires exact range initialization; nominal guarantees need transfer control before becoming target guarantees. The real study covers fixed-option scores on three prepared benchmarks, rather than independently verified evidence, human beliefs, or factual accuracy. Source dependence precludes an unconditional conformal-coverage claim; oracle scores are diagnostic, supplementary inference is post-outcome, and total compute is not matched.

8 Conclusion

Exact fixed-posterior envelopes turn residual range into computable KL-quality certificates and establish the sharp posterior-mass dependence of constrained refinement. Controlled evaluation demonstrates a 10% range reduction relative to KL-greedy within the tested KL-cost allowance, and exhaustive graphs distinguish component barriers from connected local minima. Neural evaluation identifies winner-preserving omissions while all evaluated calibrated constructions issue no useful joint certificates. This separates successful decision preservation from calibrated issuance; signed information can also certify preservation beyond the sufficient range–margin rule. Together, these results connect posterior geometry, optimization quality, and the conditions for informative certification.

Reproducibility Statement

We provide anonymized source code for the results reported in this paper in the supplementary material.

Artificial Intelligence (AI) Use Statement

The author developed the research idea and proofs. AI tools provided coding, debugging, writing, and editing support, including polishing the exposition and presentation of the proofs. Assistance also included literature retrieval, suggestions on mathematical derivations and proof steps, experimental planning, synthetic-benchmark construction, and analysis of the supplied results. The author reviewed and verified the resulting material and retains responsibility for the research questions, methods, proofs, experiments, interpretation, and final content.

References

Rohit Agrawal and Thibaut Horel. Optimal bounds between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 22(128):1–59, 2021. URL https://jmlr.org/papers/v22/20-867.html.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276. URL https://arxiv.org/abs/2202.13415.

Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999. doi: 10.1016/S0167-6377(99)00016-4. URL https://www2.isye.gatech.edu/~nemirovs/stablpn.pdf.

Trevor Campbell. General bounds on the quality of Bayesian coresets. In *Advances in Neural Information Processing Systems*, volume 37, pp. 126946–126976, 2024. doi: 10.52202/079017-4031. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e589022774244df75b97eae18bb3628d-Paper-Conference.pdf.

Trevor Campbell and Boyan Beronov. Sparse variational inference: Bayesian coresets from scratch. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11461–11472, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/7bec7e63a493e2d61891b1e4051ef75a-Abstract.html.

Trevor Campbell and Tamara Broderick. Bayesian coreset construction via greedy iterative geodesic ascent. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 698–706, 2018. URL https://proceedings.mlr.press/v80/campbell18a.html.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998. doi: 10.1145/290941.291025. URL https://www.cs.cmu.edu/~jgc/publication/MMR_DiversityBased_Reranking_SIGIR_1998.pdf.

Samuel N. Cohen and Eliana Fausti. Hyperbolic contractivity and the Hilbert metric on probability measures. *arXiv preprint arXiv:2309.02413v2*, 2024. doi: 10.48550/arXiv.2309.02413. URL https://arxiv.org/abs/2309.02413v2.

Jinshuo Dong, David Durfee, and Ryan Rogers. Optimal differential privacy composition for exponential mechanisms. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2597–2606, 2020. URL https://proceedings.mlr.press/v119/dong20a.html.

Sever S. Dragomir. Bounds for the normalised Jensen functional. *Bulletin of the Australian Mathematical Society*, 74(3):471–478, 2006. doi: 10.1017/S000497270004051X. URL https://rgmia.org/papers/v9e/BoundingJensen.pdf.

Georges Glaeser. Racine carrée d'une fonction différentiable. *Annales de l'Institut Fourier*, 13 (2):203–210, 1963. doi: 10.5802/aif.146. URL https://www.numdam.org/item/AIF_1963__13_2_203_0/.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1975–1993, 2019. URL https://proceedings.mlr.press/v99/karnin19a.html.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, 2017. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082/.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116. URL https://www.stat.cmu.edu/~ryantibs/papers/conformal.pdf.

Microsoft. Phi-3.5-mini-instruct, 2024. URL https://huggingface.co/microsoft/Phi-3.5-mini-instruct. Model card.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981, 2024. doi: 10.18653/v1/2024.findings-acl.57. URL https://aclanthology.org/2024.findings-acl.57/.

Qwen Team. Qwen2.5-1.5B-Instruct, 2024a. URL https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct. Model card.

Qwen Team. Qwen2.5-3B-Instruct, 2024b. URL https://huggingface.co/Qwen/Qwen2.5-3B-Instruct. Model card.

Björn Sprungk. On the local Lipschitz stability of Bayesian inverse problems. *Inverse Problems*, 36: 055015, 2020. doi: 10.1088/1361-6420/ab6f43. URL https://arxiv.org/abs/1906.07120.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, volume 23, pp. 2199–2207, 2010. URL https://proceedings.neurips.cc/paper/2010/file/76cf99d3614e23eabab16fb27e944bf9-Paper.pdf.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, 2019. doi: 10.1162/tacl_a_00264. URL https://aclanthology.org/Q19-1014/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

A SCOPE, NOTATION, AND ESTABLISHED POSTERIOR IDENTITIES

All statements use a finite common hypothesis support, finite logits, natural logarithms, and strictly positive probability vectors unless explicitly stated otherwise. Subsets have *exactly* k atoms, with $V = [n]$ and $0 \le k \le n$; empty sums are zero. The theoretical boundary cases $k = 0, n$ are well defined. The archived selector interface requires $1 \le k \le n$, and the completed study uses Eq. (16). The base logits are included once. Contributions may be signed; conditional independence is sufficient for a likelihood interpretation but is not required for the score identities.

A.1 RANGE, TILT, AND THE KNOWN DIVERGENCE BOUNDS

Subtracting normalized log probabilities gives $\log P_h - \log(Q_S)_h = r_{S,h} + c$, with the same constant c for all coordinates. Taking coordinate differences proves Eq. (2). The minimizer of $\max_h |r_h - c|$ is the midpoint between $\min r$ and $\max r$, so the quotient-norm formula follows. Also

$$Q_{S,h} = \frac{P_h e^{-r_{S,h}}}{\sum_j P_j e^{-r_{S,j}}}, \qquad D_{\mathrm{KL}}(P\|Q_S) = \mathbb{E}_P r_S + \log \mathbb{E}_P e^{-r_S}.$$

Common shifts in either logits or residuals leave all these expressions unchanged.

For completeness, shift r to $[0, e]$ with $e > 0$ and put $v = \mathbb{E}_P r/e$. Convexity of the exponential gives its endpoint chord,

$$e^{-r_h} \le 1 - \frac{r_h}{e} + \frac{r_h}{e} e^{-e}, \qquad \mathcal{J}(P, r) \le ve + \log(1 - v + ve^{-e}) = \phi_v(e).$$

The derivative with respect to v is $e - (1 - e^{-e})/(1 - v + ve^{-e})$. The unique maximizer is $v_* = (1 - e^{-e})^{-1} - 1/e \in (0, 1)$. Substitution gives Eq. (4); a binary residual taking values $0, e$ with mass v_* at e attains it. This is the established sharp calculation of Dong et al. (2020, Section 7 and supplementary Lemma E.3). Hoeffding's lemma (Hoeffding, 1963) applied to $-r + \mathbb{E}_P r$ gives $\mathcal{J}(P, r) \le e^2/8$. Its continuous value is zero at $e = 0$, with $F(e) = e^2/8 + O(e^4)$ near zero. Reversing P, Q leaves the log-ratio range unchanged, so the same distribution-free bound applies to reverse KL.

Here is also a direct TV calculation. For $X_h = Q_h/P_h$, let $a = \min X_h$, $b = \max X_h$. Normalization gives $a \le 1 \le b$, and the Hilbert bound implies $b/a \le e^e$. For $a < b$, the endpoint chord for the convex function $(x - 1)_+$ gives

$$\mathrm{TV}(P, Q) = \mathbb{E}_P(X - 1)_+ \le \frac{(b - 1)(1 - a)}{b - a}.$$

At a fixed ratio $b/a = R \ge 1$, this expression equals $(R + 1 - Ra - 1/a)/(R - 1)$. It is maximized at $a = R^{-1/2}$ and has value $(\sqrt{R} - 1)/(\sqrt{R} + 1)$. This is increasing in R, yielding $\tanh(e/4)$. The constant case is immediate. Endpoint likelihood ratios attain the bound. We include this standard comparison for completeness and attribute the sharp Hilbert-metric form to Cohen & Fausti (2024).

A.2 WINNER PRESERVATION AND VARIANCE GEOMETRY

If a is the unique full winner, the selected a-versus-j margin is

$$z_{S,a} - z_{S,j} = z_a - z_j - (r_{S,a} - r_{S,j}).$$

It is positive for all $j \ne a$ exactly when a remains the unique winner. In particular, $\epsilon(S) < \gamma$ is sufficient. Equality allows a tie and is not a strict certificate.

For $A(t) = \log \mathbb{E}_P e^{-tr}$ and $P_t = \mathcal{T}_{tr}(P)$, differentiation of the finite sum gives $A(0) = 0$, $A'(0) = -\mathbb{E}_P r$, and $A''(t) = \mathrm{Var}_{P_t}(r)$. Taylor's integral identity proves

$$\mathcal{J}(P, r) = \int_0^1 (1 - t)\,\mathrm{Var}_{P_t}(r)\,dt, \qquad D_{\mathrm{KL}}(\mathcal{T}_r(P)\|P) = \int_0^1 t\,\mathrm{Var}_{P_t}(r)\,dt. \tag{17}$$

Thus range ignores the probability mass carrying the residual, whereas KL weights its variation along the tilted path. This identity motivates conditioning the extremal calculation on P; it is a standard log-partition argument, not a new general divergence representation.

B PROOF OF THE EXACT FIXED-POSTERIOR ENVELOPES

B.1 EXTREMA AND THEIR ATTAINABILITY

Proof of Theorem 1, extrema. Shift invariance lets us restrict to the compact set $\{r \in [0,e]^m : \min r = 0, \max r = e\}$. Continuity ensures extrema. For $e = 0$, every residual is constant and all formulas are zero. Assume $e > 0$.

The map $r \mapsto \mathcal{J}(P, r)$ is convex: its linear term is convex and log-sum-exp is convex. Every point of $[0,e]^m$ is a convex combination of its binary vertices. Hence the maximum over the cube is attained at a vertex $r = e\mathbf{1}_A$. There the loss is $\phi_{P(A)}(e)$. Vertices with $A = \varnothing$ or $A = [m]$ give zero, while strict Jensen inequality makes every nonconstant vertex positive. Therefore a maximizing vertex has exact range e, proving the upper formula and its attainment. This is a finite convex-vertex argument, aligned with the polytope perspective in Cohen & Fausti (2024).

For the lower formula, the exact-range set is the finite union of endpoint faces

$$H_{ab} = \{r \in [0,e]^m : r_a = 0, r_b = e\}, \qquad a \neq b.$$

Fix one face, and set $p = P_a$, $q = P_b$, $w = p + q$, and

$$Z_* = (p + qe^{-e})/w, \qquad t_* = -\log Z_*.$$

Positivity gives $e^{-e} < Z_* < 1$, hence $0 < t_* < e$. Take $r_a = 0, r_b = e$ and all other coordinates t_*. Its exponential normalizer is

$$Z = p + qe^{-e} + (1-w)e^{-t_*} = wZ_* + (1-w)Z_* = Z_*.$$

For each free coordinate $h \notin \{a, b\}$,

$$\frac{\partial \mathcal{J}}{\partial r_h} = P_h - \frac{P_h e^{-r_h}}{Z} = 0.$$

Convexity makes this stationary point a global minimizer on the face. When $m = 2$, there are no free coordinates and the same evaluation applies. The attained value is

$$qe + (1-w)t_* + \log Z_* = qe + w \log Z_* = w\phi_{q/w}(e).$$

Taking the minimum over ordered endpoint pairs gives an explicit attaining residual. The next calculation reduces that minimum to the closed form in Eq. (5). This is exact for the specified P; a two-point probability distribution is not required, because interior coordinates absorb the remaining mass. □

B.2 MONOTONICITY AND THE MINIMUM-MASS BOUND

For a pair with masses $p, q > 0$, define

$$K_{p,q}(e) = qe + (p+q)\log\frac{p + qe^{-e}}{p+q}.$$

Its derivative is

$$K'_{p,q}(e) = \frac{pq(1-e^{-e})}{p + qe^{-e}} > 0 \quad (e > 0).$$

A finite minimum of these strictly increasing functions is strictly increasing: if $e_2 > e_1$, choose a pair minimizing at e_2; its value at e_2 is larger than its value at e_1, which is at least the minimum at e_1.

For fixed $e > 0$, let $Z = (p + qe^{-e})/(p+q)$. Direct differentiation yields

$$\frac{\partial K}{\partial p} = \log Z + Z^{-1} - 1 \geq 0, \qquad \frac{\partial K}{\partial q} = e + \log Z + e^{-e}Z^{-1} - 1 \geq 0.$$

Both inequalities follow from $\log x + x^{-1} - 1 \geq 0$ for $x > 0$; for the second take $x = e^e Z$. Thus $p, q \geq \beta$ implies

$$K_{p,q}(e) \geq K_{\beta,\beta}(e) = 2\beta \log\cosh(e/2).$$

To identify the minimizing pair, first orient its masses so $p \geq q$. This cannot increase the objective. Indeed, with $u = e/2$ and $d = (p-q)/(p+q) \in [0,1)$,

$$K_{p,q}(e) - K_{q,p}(e) = 2(p+q)\{\operatorname{arctanh}(d\tanh u) - du\} \leq 0.$$

The inequality follows from convexity of $\operatorname{arctanh}$ on $[0,1)$: its value at $d\tanh u$ is at most $d\operatorname{arctanh}(\tanh u) = du$. Among any two distinct coordinates, the larger mass is at least β_2 and the smaller at least β_1. Coordinatewise monotonicity of K therefore gives

$$K_{p,q}(e) \geq K_{\beta_2,\beta_1}(e) = (\beta_1+\beta_2)\phi_{\beta_1/(\beta_1+\beta_2)}(e).$$

Equality is attained by placing the second-smallest-mass coordinate at zero and the smallest-mass coordinate at e, with every other coordinate at the stationary value above. This proves the stated closed form. Tied masses allow multiple equivalent extremizers. For uniform P it becomes $(2/m)\log\cosh(e/2)$. We already know $U_P(e) \leq F(e)$.

For the upper bound in Eq. (7), apply the independent reference comparison in Lemma 6 to a binary reference $(1/2, 1/2)$ and residual $(e, 0)$:

$$\phi_v(e) \leq 2\max\{v, 1-v\}\log\cosh(e/2).$$

Every nonempty proper subset mass $v = P(A)$ belongs to $[\beta, 1-\beta]$. Consequently

$$U_P(e) \leq 2(1-\beta)\log\cosh(e/2),$$

which gives the stated sharper upper bound when combined with $U_P \leq F$. There is no circular dependence: Lemma 6 is proved directly from concavity in the reference weights.

The asymptotic claim follows from finite positive masses. For each fixed $v \in (0,1)$, $\phi_v(e) = ve + \log(1-v) + o(1)$ as $e \to \infty$. The largest attainable proper subset mass is $1-\beta$, obtained by excluding a minimum-mass coordinate. Since there are finitely many subset masses,

$$U_P(e) = (1-\beta)e + O(1), \qquad L_P(e) = \beta e + (\beta_1+\beta_2)\log\frac{\beta_2}{\beta_1+\beta_2} + o(1).$$

Thus $U_P(e)/L_P(e) \to (1-\beta)/\beta$ for every fixed positive reference. The envelope ratio cannot have a smaller uniform-in-e bound depending on β alone. This completes Theorem 1.

Small-range and computational interpretation. Expansion for each ordered pair gives

$$K_{p,q}(e) = \frac{pq}{2(p+q)}e^2 + O(e^3), \qquad L_P(e) = \left(\min_{a\neq b}\frac{P_aP_b}{2(P_a+P_b)}\right)e^2 + O(e^3).$$

The constants in the remainder may depend on the fixed reference. Finding the two smallest masses requires a single $O(m)$ scan; the exact lower envelope then needs constant-time scalar arithmetic. The exact upper envelope has up to $2^m - 2$ subset-mass candidates; no polynomial-time claim in m is made. Numerical implementation should use log-sum-exp or `logaddexp` rather than subtracting nearly equal exponentials.

When range alone determines KL. For $m = 2$ and $P = (1/2, 1/2)$, both nonconstant vertex masses equal $1/2$, and the extremal envelopes coincide at $\log\cosh(e/2)$. Thus range minimization is exactly KL minimization. In fact, for positive P and $e > 0$, this is the only nontrivial finite-support case in which $L_P(e) = U_P(e)$. For two unequal masses the strict convexity of $\operatorname{arctanh}$ in the orientation calculation gives different losses for the two orientations. For $m \geq 3$, the minimizing endpoint face has a free coordinate strictly between 0 and e. Varying that coordinate preserves exact range, and its second derivative is $Q_h(1-Q_h) > 0$. Moving it away from the face minimizer strictly increases KL. Hence the maximum exceeds the minimum. This is a loss-geometry characterization, not a claim that the empirical pipeline was tested in each such regime.

C Optimization consequences and fixed-exchange obstructions

C.1 Exact-cardinality robust optimization

Let $x_i \in \{0,1\}$ indicate retained atoms and $\widehat{d}_{ihj} = \widehat{\ell}_{ih} - \widehat{\ell}_{ij}$. RLOC is the mixed-integer linear program

$$\begin{aligned} &\min_{x,t} \quad t, \\ &\sum_i x_i = k, \qquad x_i \in \{0,1\}, \\ \pm\sum_i (1-x_i)\widehat{d}_{ihj} + \sum_i (1-x_i)\rho_{ihj} \le t \quad &\text{for both signs and every } h<j. \end{aligned} \tag{18}$$

MLOC sets radii to zero. The objective is not monotone with cardinality, so replacing equality by inequality is a different problem. In particular, two contributions $v, -v$ give optimum ranges $0, \mathrm{osc}(v), 0$ for $k = 0, 1, 2$. A token cap is also a different feasible set and is not silently substituted.

On the simultaneous event

$$|(\ell_{ih} - \ell_{ij}) - (\widehat{\ell}_{ih} - \widehat{\ell}_{ij})| \le \rho_{ihj} \quad \forall i,\ h<j,$$

each omitted pairwise sum is at most its estimated absolute value plus the sum of its radii. Taking the maximum proves $\epsilon(S) \le C(S)$ for all subsets on the same event. The maximum in Eq. (8) is the exact robust objective for the independent pairwise box, but pairwise boxes need not describe a jointly realizable logit-error vector. It can consequently overestimate the true uncertainty set. A feasible MILP solution yields a valid value for this objective independently of whether the solver certifies optimality.

For constant radii, $\sum_{i\notin S}\rho = (n-k)\rho$; this proves Eq. (9) and equality of exact minimizer sets. Expanding $C(S) \le (1+\tau)C(S_0)$ gives

$$\widehat{\epsilon}(S) \le (1+\tau)\widehat{\epsilon}(S_0) + \tau(n-k)\rho.$$

This explains both the constant uncertainty role and the nominal slack induced by a robust-relative threshold. Time-limited MILP incumbents can still differ; the identity is a statement about exact objectives, not identical solver trajectories.

Refinement implementation. CPC fixes $C_0 = C(S_0)$, enumerates one-for-one swaps, and selects the deterministic largest strict decrease in estimated forward KL among swaps with $C \le (1+\tau)C_0$. The controlled and exhaustive branches stop when no improving feasible swap remains or after 20 passes. The neural branch has no pass cap and stops at its numerical single-exchange local optimum. Every iterate remains within the threshold tolerance by construction and its computed estimated KL never increases. The recorded implementations use a probability floor in their selection KL; the exact-arithmetic theory instead uses Eq. (3), and displayed outcomes are re-evaluated in log space. With unlimited passes and exact strict comparisons, finitely many subsets and strictly decreasing loss imply termination. A finite pass limit does not by itself certify a local optimum; that property was checked separately in the exhaustive panels. Floating-point feasibility and improvement tolerances are recorded in Appendix G.

C.2 Proof of the KL optimality certificate

Proof of Corollary 2. For every $T \in \mathcal{F}$, Theorem 1 and monotonicity imply

$$\mathcal{J}(P, r_T) \ge L_P(\mathrm{osc}(r_T)) \ge L_P(a_-).$$

Minimizing gives the lower bound on D_*. Feasibility of S gives $D_* \le d$, establishing the additive certificate and the ratio when $a_- > 0$. If S_0 attains the minimum range e_*, KL-decreasing refinement satisfies

$$\mathcal{J}(P, r_S) \le \mathcal{J}(P, r_{S_0}) \le U_P(e_*), \qquad D_* \ge L_P(e_*).$$

For $e_* > 0$, Eq. (7) bounds the ratio by $(1-\beta)/\beta$. If $e_* = 0$, r_{S_0} is constant, so initialization has zero KL. Nonnegativity and nonincrease force every subsequent iterate to have zero KL as well. □

A lower bound valid on all k-subsets remains valid on a certificate-constrained subfamily. Thus the a posteriori certificate can bound either all-subset or constrained KL regret. Under a constant-radius RLOC solve, a valid robust-objective lower bound ℓ_- gives nominal range lower bound $a_- = [\ell_- - (n-k)\rho]_+$. For nonconstant radii this subtraction is not valid without additional reasoning. A RLOC minimizer need not be a range minimizer in that case, so the $(1-\beta)/\beta$ consequence does not apply automatically. A numerical MILP bound must also be verified within an explicit numerical tolerance; it is not a symbolic proof of an exact optimum.

Certified endpoint errors, rather than presumed exact descent. Let $D_T = \mathcal{J}(P, r_T)$ use one common reference, and let $\widetilde{D}_T$ be the computed selection loss. Suppose verified bounds $\nu_T \geq 0$ on endpoint errors satisfy $|\widetilde{D}_T - D_T| \leq \nu_T$ for $T = S_0, S$, and the recorded trajectory obeys $\widetilde{D}_S \leq \widetilde{D}_{S_0}$. Writing $e_0 = \mathrm{osc}(r_{S_0})$ gives

$$D_S \leq \min\{\widetilde{D}_S + \nu_S,\ U_P(e_0) + \nu_{S_0} + \nu_S\}. \tag{19}$$

Indeed, $D_S \leq \widetilde{D}_S + \nu_S \leq \widetilde{D}_{S_0} + \nu_S \leq D_{S_0} + \nu_{S_0} + \nu_S$, and Theorem 1 bounds D_{S_0}. Combining this with any verified a_- yields an additive regret bound by subtracting $L_P(a_-)$ from the right side. At exact range initialization with $e_* > 0$, the ratio is at most $(1-\beta)/\beta + (\nu_{S_0} + \nu_S)/L_P(e_*)$. When $e_* = 0$, only the additive bound $D_S \leq \nu_{S_0} + \nu_S$ follows. Errors do not accumulate per swap because only the two endpoints are compared. A probability-floor parameter, solver gap, or improvement tolerance is *not* by itself a bound on ν_T. No endpoint error bounds or resulting ratios are claimed to have been measured for the reported neural evaluation. This is a conditional numerical qualification of the existing theorem, not a new empirical result.

True, nominal, and direct guarantees must stay distinct. Corollary 2 compares losses with a *common fixed reference* and residual family. Using $\widehat{P}, \widehat{r}$ therefore gives a nominal certificate. Estimated refinement may increase true KL if references or residuals differ. The conditional transfer interval in Eq. (14) is one way to transport separately justified nominal information, but no result substitutes estimator masses for target masses without such a bound. The completed real-model study did not evaluate L_P, U_P, or the derived optimality certificate. Their significance here is a positive characterization and the matching obstruction, not an unmeasured performance claim.

C.3 Complete construction for arbitrary bounded exchanges

Proof of Theorem 3. Write $A_\beta = (1-\beta)/\beta > 1$. Choose

$$0 < \theta < \min\left\{\tau, \frac{\zeta}{2A_\beta}\right\}.$$

Then $A_\beta/(1+\theta) > A_\beta - \zeta/2$. Let $a > 0$, $b = (1+\theta)a$, $k = q+1$, and choose L so that $2L - b > (1+\tau)a$. In two coordinates set

$$r_A = (a, 0), \quad r_B = (0, b), \quad w = (L, -L).$$

Create k atoms in each of two disjoint blocks. Every A atom is r_A/k; the B atoms are $r_B/k + c_i w$, with $c_1 = \cdots = c_{k-1} = 1$ and $c_k = -(k-1)$. Their block sums are r_A, r_B. Set the base logits to $b_0 = \log(1-\beta, \beta) - r_A - r_B$; the full posterior is exactly $P = (1-\beta, \beta)$ and has a unique winner. Estimates equal true scores and all interval radii vanish.

As $n = 2k$, omitted k-sets correspond bijectively to selected k-sets. A mixed omitted set contains $t \in \{1, \ldots, k-1\}$ A atoms and a nonempty proper B subblock. The latter's coefficient sum c is a nonzero integer: it is positive without the last B atom and at most -1 with that atom. The residual contrast is

$$\frac{ta - (k-t)b}{k} + 2Lc.$$

The absolute value of the first term is at most b because $b > a$. Hence every mixed set has range at least $2L - b > (1+\tau)a$. The pure omitted blocks have ranges a and b, with $a < b < (1+\tau)a$. The unique range minimizer is therefore pure A; only pure A, B satisfy the certificate threshold. Their selected complements require $k = q+1$ replacements, so neither has a neighbor under any exchange of at most q atoms.

The two KL losses are

$$D_A(a) = \phi_{1-\beta}(a), \qquad D_B(a) = \phi_\beta((1+\theta)a).$$

Both are positive for $a > 0$, and

$$\lim_{a\to\infty} \frac{D_A(a)}{D_B(a)} = \frac{1-\beta}{\beta(1+\theta)} > A_\beta - \zeta/2.$$

Thus a finite sufficiently large a gives $D_A/D_B > A_\beta - \zeta > 1$. For that instance B is the constrained KL optimum, while every feasible bounded-exchange trajectory from A is stationary. This proves the theorem with entirely finite logits and atom vectors. For the unrestricted k-set optimum D_*, $D_* \le D_B$; combined with Corollary 2, the same example shows sharpness of the universal approximation constant. No claim about a finite-instance attainment of the limiting factor is needed. □

An additional fixed-atom family. The preceding sharp construction varies atom scale to approach the exact constant for a fixed β. For completeness, the following independent construction shows that diverging ratios as $\beta \downarrow 0$ do not require changing the atom vectors.

Construction with fixed atom vectors and three hypotheses. Fix $q \ge 1$, put $k = q + 1$ and $n = 2k$, and choose any $a > 0$, $b = (1 + \tau/2)a$ and $L > (2 + \tau)a/2$. In three coordinates define

$$r_A = (a, 0, 0), \qquad r_B = (0, 0, b), \qquad w = (L, -L, 0).$$

Create two disjoint blocks of k atoms. For block A, set every contribution to r_A/k. For block B, set

$$\ell_{B,i} = r_B/k + c_i w, \qquad c_1 = \cdots = c_{k-1} = 1, \quad c_k = -(k-1).$$

The blocks sum to r_A and r_B, respectively. We describe omitted subsets, which are in bijection with selected k-subsets because $n = 2k$. Pure omitted blocks have ranges a, b.

Every mixed omitted k-subset contains $t \in \{1, \ldots, k-1\}$ atoms from A and a nonempty proper subset of B. The sum c of its selected B coefficients is a nonzero integer: without the last B atom it is positive, and with that atom it is at most -1. The difference between its first two residual coordinates is $ta/k + 2Lc$. Hence

$$\operatorname{osc}(r_{\text{mixed}}) \ge |ta/k + 2Lc| \ge 2L - a > (1+\tau)a.$$

Therefore the unique global range minimizer is omitted block A, and the only two sets satisfying the CPC threshold $C \le (1+\tau)a$ are omitted blocks A, B. Their selected complements differ by $k = q + 1$ replacements. The graph allowing any exchange of at most q atoms has two isolated nodes. Even accepting nonimproving feasible moves cannot connect them.

For $p \in (0, 1/3)$ take

$$P_p = ((1-p)/2, (1-p)/2, p), \qquad b_0 = \log P_p - r_A - r_B.$$

The full posterior generated by these atoms and base is exactly P_p, whose minimum mass is $\beta = p$. Scores are finite for every instance and the estimates equal the true contributions. Their losses are

$$D_A(p) = \frac{1-p}{2}a + \log\left(\frac{1-p}{2}e^{-a} + \frac{1-p}{2} + p\right),$$

$$D_B(p) = pb + \log(1 - p + pe^{-b}).$$

As $p \downarrow 0$,

$$D_A(p) \longrightarrow \log\cosh(a/2) > 0, \qquad D_B(p) = p(b + e^{-b} - 1) + O(p^2).$$

The coefficient $b + e^{-b} - 1$ is strictly positive for $b > 0$. Thus $D_B(p) > 0$, block B is the constrained KL optimum for small enough p, and

$$\frac{D_A(p)}{D_B(p)} \sim \frac{\log\cosh(a/2)}{b + e^{-b} - 1}\,\frac{1}{\beta}.$$

Each of the finitely many mixed subsets has unequal first two residual coordinates. Its limiting KL under $(1/2, 1/2, 0)$ is positive. Consequently B is also the unconstrained k-subset KL optimum for all sufficiently small p. This strengthens, but is not required for, the constrained claim.

A full-posterior tie is unnecessary: replace the first two masses by $(1-p)\theta, (1-p)(1-\theta)$ for any fixed $1/2 < \theta < 1$. For small p, the winner is unique and the minimum mass is still p. The feasible graph is unchanged. Now $D_A \to \theta a + \log(\theta e^{-a} + 1 - \theta) > 0$, while D_B retains the same expansion. The ratio is again $\Theta(1/\beta)$. □

Boundaries. For $q = 1$ both constructions give four-atom single-swap obstructions. The general q statements rule out a uniform repair by any fixed larger exchange radius. At fixed a, L, only the base posterior changes along the diverging-ratio sequence; unbounded atom norms are unnecessary for that claim. Allowing $a \to \infty$ and then $p \downarrow 0$ also gives unbounded additive gaps. This does not contradict Corollary 2: precisely the mass condition $\beta > 0$ deteriorates along the family. No statement of typical or neural failure frequency follows from this construction. The existing graphs concern single exchanges only.

C.4 Why unit weights and exact cardinality matter

With $\ell_i = (a, 0)$ and $b = (-na, 0)$, $P = (1/2, 1/2)$ and every k-subset has residual $((n-k)a, 0)$, so

$$D_{\mathrm{KL}}(P \| Q_S) = \log \cosh((n-k)a/2).$$

No selector can provide universal small loss at $k < n$ without controlling contribution scale or structure. In the other direction, if omitted atoms partition into blocks of residuals v_j, the seminorm triangle inequality gives $\epsilon(S) \le \sum_j \mathrm{osc}(v_j)$. Complete cancelling blocks have zero contribution. These facts explain the generator-specific advantage of preserving adjacent pairs. They are not general guarantees about real-text sentence independence or about weighted Bayesian coresets.

D Decision validity, issuance, and usefulness

Proof of Proposition 5. On $I \cap B$, the subset omits at least one atom and every target obeys $e_t \le R_t < \gamma_t$. Hence it belongs to O. For each target in O, the range winner-preservation calculation above gives the same unique winner for full and selected contexts, proving $O \subseteq W$. Therefore $I \setminus W \subseteq B^c$. Taking probabilities proves the last statement. If $\Pr(I) = \pi_I > 0$, it only implies $\Pr(W^c \mid I) \le \min\{1, \alpha/\pi_I\}$. The validity condition must hold for the actually selected pipeline; marginal guarantees for each nonadaptively chosen subset are not enough. □

The range–margin rule is sufficient, not necessary. For $a > 0$, take full logits $z = (a, 0)$ and subset logits $z_S = (2a, 0)$. They have the same unique winner, but $r = z - z_S = (-a, 0)$ has range a, equal to the full margin $\gamma = a$. Thus the strict exact-range test fails. With exact nominal residual $\widehat{r} = r$ and aggregate uncertainty $U = \{0\}$, the signed test in Eq. (13) nevertheless succeeds: its margin is $2a - a = a > 0$. This can be realized by a literal additive subset with $n = 2, k = 1$, $b = 0$, $\ell_1 = (2a, 0)$ retained and $\ell_2 = (-a, 0)$ omitted. Consequently zero useful issuance for the evaluated range–margin construction does not imply that every valid signed construction must also issue zero certificates. This example is an exact algebraic distinction, not a new benchmark or a claim about untested neural methods.

Reading the completed counts. A row in $W \setminus O$ retains both unique full winners under the recorded deterministic subset-argmax rule although the target-wise exact-range sufficient test fails. A row in O with no issuance shows a failure to attain that oracle opportunity, not evidence that any learned bound can do so. A common valid radius is an information restriction beyond target-wise exact radii. Thus the recorded O counts are not the exact capacity of a common-radius calibrator. Subset-logit ties are not excluded from the recorded W counts: its definition requires uniqueness only of the full winner and agreement with the deterministic subset argmax. By contrast, O implies uniqueness of the subset winner. Full retention would make both residuals zero and is explicitly excluded from I, O, W. Across constructions and operating points, the repeated bases are retained for comparisons but are not pooled into independent Bernoulli trials.

Split conformal under its actual assumptions. Fix $N \ge 1$ and $\alpha \in (0, 1)$. Assume that, conditional on the fitted pipeline, the N calibration units and a new unit are exchangeable within a dataset stratum. This holds when those units are jointly exchangeable and independent of the fitting data. For a frozen selector and nonnegative nominal predictor a, put $Z_i = [\max_t e_{it} - a_i]_+$. With $j = \lceil (N+1)(1-\alpha) \rceil$, choose the jth order statistic q, or $q = \infty$ if $j > N$. Conditional exchangeability of the $N + 1$ scores and a uniform random tie order imply the test rank is uniform; without randomization, including ties is conservative. Therefore $\Pr(Z_{\mathrm{new}} \le q) \ge 1 - \alpha$, conditionally on the fitted pipeline and hence marginally. This is the standard split-conformal rank argument, not

a new calibration theorem (Lei et al., 2018, Section 2.2). Since $\max_t e_t \le a + [\max_t e_t - a]_+$, $R = a + q$ simultaneously bounds the target-family errors with this marginal probability. Proposition 5 then controls false issuance, not issuance frequency. Fitting a on the calibration outcomes or splitting correlated source units does not automatically justify this frozen-score argument; a separate conditional-exchangeability or other validity argument is needed. The detected source overlap in the completed study is why its reported coverage is observational rather than claimed as a validated guarantee. No new nonexchangeable conformal procedure is asserted (Barber et al., 2023).

E AGGREGATE-ERROR STABILITY AND CERTIFICATION LIMITS

The results below use finite real scores, natural logarithms and strictly positive distributions on the same finite support. They are deterministic implications, not coverage guarantees for fitted uncertainty sets. The reference-weight comparison is established Dragomir–Jensen theory; Pinsker and robust counterparts are also established tools. We derive their consequences for omitted-evidence residuals and give exact signed winner tests. No priority claim for the combined formulations is made.

E.1 PROOF OF THE KL TRANSFER COROLLARY

Lemma 6 (Reference-weight comparison)**.** *For a fixed finite vector u, let $\lambda = \min_h P_h/\widehat{P}_h$ and $\Lambda = \max_h P_h/\widehat{P}_h$. Then*

$$\lambda \mathcal{J}(\widehat{P}, u) \le \mathcal{J}(P, u) \le \Lambda \mathcal{J}(\widehat{P}, u). \tag{20}$$

Proof. For finite u, the formula for $\mathcal{J}$ extends continuously and nonnegatively to probability vectors with zero coordinates. The map $P \mapsto \mathcal{J}(P, u)$ is concave: $P^\top u$ is linear and $\log \sum_h P_h e^{-u_h}$ is concave. Jensen's inequality gives $\mathcal{J}(P, u) \ge 0$. Normalization implies $\lambda \le 1 \le \Lambda$. If $\lambda < 1$, the vector $R = (P - \lambda\widehat{P})/(1 - \lambda)$ is a probability vector. Concavity applied to $P = \lambda\widehat{P} + (1 - \lambda)R$ gives the lower bound. Similarly, writing $\widehat{P} = \Lambda^{-1}P + (1 - \Lambda^{-1})R'$ gives the upper bound. If either coefficient equals one, the two references coincide. This is the finite comparison of Dragomir (2006, Theorem 1) applied to the convex function $-\log$ and values e^{-u_h}. □

Lemma 7 (Square-root residual stability)**.** *For a fixed positive P and finite $r, \widehat{r}$,*

$$\left|\sqrt{\mathcal{J}(P, r)} - \sqrt{\mathcal{J}(P, \widehat{r})}\right| \le \frac{\operatorname{osc}(r - \widehat{r})}{\sqrt{8}}. \tag{21}$$

The universal constant $1/\sqrt{8}$ cannot be decreased.

Proof. Put $v = r - \widehat{r}$, $r_t = \widehat{r} + tv$, $Q_t = \mathcal{T}_{r_t}(P)$ and $g(t) = \mathcal{J}(P, r_t)$. Differentiating the finite sum gives

$$g'(t) = (P - Q_t)^\top v.$$

For probability vectors A, B, subtracting $\min_h v_h$ shows that $|(A - B)^\top v| \le \operatorname{osc}(v)\mathrm{TV}(A, B)$: the positive and negative parts of $A - B$ each have mass $\mathrm{TV}(A, B)$. Pinsker's inequality, with half-ℓ^1 total variation, gives

$$|g'(t)| \le \operatorname{osc}(v)\sqrt{g(t)/2}.$$

For every $\varepsilon > 0$, even when $g(t) = 0$,

$$\left|\frac{d}{dt}\sqrt{g(t) + \varepsilon}\right| \le \frac{\operatorname{osc}(v)}{\sqrt{8}}\sqrt{\frac{g(t)}{g(t) + \varepsilon}} \le \frac{\operatorname{osc}(v)}{\sqrt{8}}.$$

Integration on $[0, 1]$ and the limit $\varepsilon \downarrow 0$ prove the bound. To see sharpness, take $P = (1/2, 1/2)$, $\widehat{r} = 0$ and $r = (t, 0)$. Then $\mathcal{J}(P, r) = \log\cosh(t/2) = t^2/8 + O(t^4)$, so the ratio of the left side to $\operatorname{osc}(r)$ tends to $1/\sqrt{8}$ as $t \to 0$. □

The Pinsker step. For completeness, take $H = \{h : A_h > B_h\}$. If $A = B$ the inequality is immediate. Otherwise positive distributions give $x = A(H), y = B(H) \in (0, 1)$. The log-sum inequality, obtained from Jensen's inequality for $t \log t$ on H and its complement, gives $D_{\mathrm{KL}}(A\|B) \geq d(x\|y)$, where d is binary KL. For fixed y,

$$\frac{d^2}{dx^2}\{d(x\|y) - 2(x-y)^2\} = \frac{1}{x(1-x)} - 4 \geq 0.$$

The expression and its derivative vanish at $x = y$; hence $d(x\|y) \geq 2(x-y)^2$. Since $x - y = \mathrm{TV}(A, B)$, the claimed Pinsker bound follows. This proof is included for completeness, not as a contribution; see Agrawal & Horel (2021) for broader divergence comparisons.

Classical self-bounding interpretation. For the same $g(t)$, defined on all real t, $0 \leq g''(t) = \mathrm{Var}_{Q_t}(v) \leq L := \mathrm{osc}(v)^2/4$. If $L > 0$, the smoothness descent inequality and nonnegativity give

$$0 \leq g(t - g'(t)/L) \leq g(t) - g'(t)^2/(2L).$$

Consequently $|g'(t)| \leq \sqrt{2Lg(t)}$, yielding the same square-root constant; $L = 0$ is immediate. This is exactly the principle of Glaeser (1963, Lemma I), attributed there to Malgrange, applied to $g + \varepsilon$ and then $\varepsilon \downarrow 0$. It is not a new inequality. For a learning-theoretic formulation see Srebro et al. (2010, Lemmas 2.1 and 3.1), which state a looser constant.

Proof of Corollary 4. Apply Lemma 7 under P, then Lemma 6, to obtain

$$\sqrt{D} \leq \sqrt{\mathcal{J}(P, \widehat{r})} + \delta/\sqrt{8} \leq \sqrt{\Lambda \widehat{D}} + \delta/\sqrt{8}.$$

For the lower bound, compare references at r first, then perturb under $\widehat{P}$:

$$\sqrt{D} \geq \sqrt{\lambda}\sqrt{\mathcal{J}(\widehat{P}, r)} \geq \sqrt{\lambda}\,[\sqrt{\widehat{D}} - \delta/\sqrt{8}]_+.$$

Squaring gives the more explicit interval

$$\lambda[\sqrt{\widehat{D}} - \delta/\sqrt{8}]_+^2 \leq D \leq (\sqrt{\Lambda \widehat{D}} + \delta/\sqrt{8})^2. \tag{22}$$

The log-ratio condition implies $\Lambda/\lambda \leq e^{\eta}$. Together with $\lambda \leq 1 \leq \Lambda$ this gives $\lambda \geq e^{-\eta}$ and $\Lambda \leq e^{\eta}$, proving Eq. (14). □

Sharpness of reference factors and boundary cases. Fix $t \neq 0$, $\widehat{P}_p = (p, 1-p)$, and

$$P_p = \frac{(e^{\eta}p, 1-p)}{1 - p + e^{\eta}p}, \qquad r = \widehat{r} = (t, 0).$$

The log-ratio oscillation is η. For a binary mass x on the first coordinate, $\mathcal{J}((x, 1-x), (t, 0)) = x(t + e^{-t} - 1) + O(x^2)$. The leading coefficient is positive for $t \neq 0$, so the loss ratio tends to e^{η} as $p \downarrow 0$. Exchanging references gives the lower-factor sharpness. Every distribution in the sequence is positive. This establishes sharpness of the reference factors when $\delta = 0$, not joint optimality of the combined interval for positive δ, η. If $\delta = \eta = 0$, the losses coincide. If $\widehat{D} = 0$, positivity forces $\widehat{r}$ to be constant, and the upper bound reduces to $\delta^2/8$; the known $F(\delta)$ can sharpen it. If no atom is omitted and both residuals use the full-minus-full definition, they are zero. For $m = 1$ every loss and oscillation is zero. For reverse KL, the tilted references must be exchanged: their log-ratio oscillation is at most $\eta + \delta$, not automatically η.

Ordering and access to scores. Let U_S, L_S denote the upper and lower bounds of Eq. (14) for subset S. Then $U_S < L_T$ implies $D_S < D_T$. There is no assertion that these intervals are informative on the current neural panel. A specialization avoids a full-target score at certificate evaluation: given subset logits $s = f_T(S)$ and $\widehat{r}$, set $P_0 = \mathrm{softmax}(s + \widehat{r})$ and $D_0 = D_{\mathrm{KL}}(P_0\|\,\mathrm{softmax}(s))$. A valid aggregate error bound $\mathrm{osc}(r - \widehat{r}) \leq \delta$ implies $\mathrm{osc}(\log P - \log P_0) \leq \delta$. Thus Corollary 4 applies with reference P_0, nominal loss D_0 and $\eta = \delta$. This is a conditional construction, not the existing CPC objective or evidence of a computational advantage; obtaining a valid δ still requires justification.

E.2 WHY SEPARATE COMPONENT MAGNITUDES DO NOT IDENTIFY DIRECT ERROR

Proposition 8 (Cancellation in the aggregate residual error)**.** *Use Eq.* (12)*, let* $q = n-k$*,* $\widehat{r} = r_E(S)$*,* $r = f_T(V) - f_T(S)$ *and* $N = \Delta_T(V) - \Delta_T(S)$*. If estimator increments are* $\ell_{E,i} = f_E(\{i\}) - f_E(\varnothing)$*, then*

$$E_S := r - \widehat{r} = M + N = f_T(V) - f_T(S) - \sum_{i \notin S} f_E(\{i\}) + q f_E(\varnothing), \tag{23}$$

$$|\operatorname{osc}(r) - \operatorname{osc}(\widehat{r})| \le \operatorname{osc}(E_S) \le \operatorname{osc}(r) + \operatorname{osc}(\widehat{r}). \tag{24}$$

Arbitrarily large separate ranges of M, N *are compatible with zero direct error. In general, their marginal distributions do not identify aggregate-error quantiles.*

Proof. Expand the two components:

$$M = \sum_{i \notin S} [f_T(\{i\}) - f_E(\{i\})] - q[f_T(\varnothing) - f_E(\varnothing)],$$

$$N = f_T(V) - f_T(S) - \sum_{i \notin S} f_T(\{i\}) + q f_T(\varnothing).$$

Adding cancels all target singleton and target empty-context terms, proving Eq. (23). Triangle and reverse-triangle inequalities for the oscillation seminorm prove Eq. (24). For the counterexample choose $n = 3$, $S = \{1, 2\}$ and a nonconstant vector v. Set estimator empty and singleton scores, and target singleton scores, to zero. Set $f_T(\varnothing) = tv$ and $f_T(S) = f_T(V) = s$. Then $r = \widehat{r} = 0$, $M = -tv$ and $N = tv$. Both component ranges diverge as $|t| \to \infty$, although the full and subset posteriors coincide. Choosing s with a unique maximum also preserves a unique winner. For the marginal claim, let Z be a symmetric random sign, $A > 0$, $q > 0$ and $\operatorname{osc}(v) = 1$. The pairs $(M, N) = (qAZv, -qAZv)$ and $(qAZv, qAZv)$ have identical marginal laws for each component, but aggregate ranges zero and $2qA$, respectively, almost surely. These decompositions can be realized by singleton-derived scores: take zero estimator and target singleton scores, set $f_T(\varnothing) = -M/q$, and choose $f_T(V) - f_T(S) = M + N$ with V, S distinct from the empty and singleton contexts. Thus the ambiguity is compatible with the score decomposition itself. □

The estimator empty-context term remains in Eq. (23); not all reference errors disappear. What fails is the inference that the isolated term $-q[f_T(\varnothing) - f_E(\varnothing)]$ creates an unavoidable direct-error floor. Marginal laws do not identify the exact aggregate quantile, but they can give conservative upper bounds: if $\Pr(\operatorname{osc}(M) > a) \le \alpha_1$ and $\Pr(\operatorname{osc}(N) > b) \le \alpha_2$, the triangle inequality and union bound imply $\Pr(\operatorname{osc}(M + N) > a + b) \le \alpha_1 + \alpha_2$ without independence. Equation (24) also bounds means with identical weights; a mean alone is not a calibrated high-probability radius. The decomposition in Figure 4 remains descriptive, not a causal account of a proposed repair.

E.3 SIGNED ROBUST WINNER CERTIFICATES

Let $s = f_T(S)$ be observed subset logits, and suppose the aggregate error $E_S = r - \widehat{r}$ lies in a supplied nonempty compact set $U \subset \mathbb{R}^m$. Write $h_U(v) = \max_{e \in U} v^\top e$, and let $v_{hj} = \mathbf{e}_h - \mathbf{e}_j$ for coordinate basis vectors $\mathbf{e}_h$.

Proposition 9 (Exact signed robust counterpart)**.** *Define* $\mu_{aj} = s_a - s_j + \widehat{r}_a - \widehat{r}_j$*. Candidate* a *is the unique full winner for every* $e \in U$ *if and only if*

$$\mu_{aj} > h_U(-v_{aj}) \quad \textit{for all } j \ne a. \tag{25}$$

If a *is also the unique subset winner, this is an exact robust preservation test. For* $m \ge 2$*, the smallest uniform direct-range bound is*

$$R_U = \max_{h \ne j} \{\widehat{r}_h - \widehat{r}_j + h_U(v_{hj})\}. \tag{26}$$

For $m = 1$*, the test is vacuous and* $R_U = 0$*.*

Proof. The full a-versus-j margin equals $\mu_{aj} + v_{aj}^\top e$. Its minimum on U is $\mu_{aj} - h_U(-v_{aj})$, attained by compactness. Hence strict positivity of every minimum is necessary and sufficient. Equality allows failure of unique victory. For the scalar bound, the supremum over e commutes with the finite maximum over pairs in $\operatorname{osc}(\widehat{r} + e)$, yielding Eq. (26). □

Dependence inside a contrast matters. This is a support-function specialization of established robust counterparts, not a new general robust-optimization method (Ben-Tal & Nemirovski, 1999). Suppose $M = A\xi$, $N = B\xi$ for a common $\|\xi\|_2 \le 1$. The aggregate support in direction v is $\|(A+B)^\top v\|_2$, by Cauchy–Schwarz with equality in the maximizing direction. Allowing separate perturbations instead gives $\|A^\top v\|_2 + \|B^\top v\|_2$, which can be arbitrarily larger when $B = -A$. These are uncertainty-set statements, not probabilistic independence assumptions. Once all aggregate competitor supports are known, further dependence between competitors does not change the exact deterministic test: robust victory requires every contrast to be positive for every admissible error.

An oscillation ball. For $U = \{e : \mathbf{1}^\top e = 0,\ \mathrm{osc}(e) \le \delta\}$ and $m \ge 2$, each contrast support is δ: the range bounds it above, and centered coordinates $e_h = \delta/2$, $e_j = -\delta/2$, with all others zero, attain it. Equation (25) becomes $s_a - s_j + \widehat{r}_a - \widehat{r}_j > \delta$. Keeping favorable signed nominal residuals can be less conservative than requiring the subset margin to exceed $\mathrm{osc}(\widehat{r}) + \delta$.

A conditional omission barrier. If the *aggregate* set U contains $\{q\kappa tv : |t| \le 1\}$ with $\kappa \ge 0$, a necessary condition for robust victory is

$$\mu_{aj} > q\kappa |v_a - v_j| \quad (j \ne a), \qquad R_U \ge q\kappa\, \mathrm{osc}(v).$$

The first statement follows by choosing the adverse endpoint in each contrast. For the second let $r_\pm = \widehat{r} \pm q\kappa v$. Then $2q\kappa\,\mathrm{osc}(v) = \mathrm{osc}(r_+ - r_-) \le \mathrm{osc}(r_+) + \mathrm{osc}(r_-) \le 2R_U$. If U is exactly the segment plus a centered oscillation ball of radius β, the exact threshold in Eq. (25) is $q\kappa|v_a - v_j| + \beta$ because support functions add under Minkowski sums. Common-shift modes have no contrast effect. A large separate mismatch term does not establish the assumed aggregate-set inclusion, as Proposition 8 shows. Also μ_{aj} depends on S, so this is not an unconditional monotonicity theorem in the budget.

Computation and probability are separate. Testing a fixed candidate uses $m-1$ support evaluations, and a scalar range bound uses $m(m-1)$. For a supplied $m \times d$ ellipsoidal factor these cost $O(md)$ and $O(m^2 d)$, respectively. These counts exclude uncertainty fitting and subset selection. Exactness is relative to U: if it is an outer approximation containing unattainable errors, failure of the test does not prove that the actual model can change winners. If an error set has marginal coverage $1-\alpha$ for the actually selected subset under a frozen pipeline, a false issued certificate implies set failure, giving $\Pr(\text{false and issued}) \le \alpha$. For issuance probability $p > 0$, only $\Pr(\text{false} \mid \text{issued}) \le \min(1, \alpha/p)$ follows. Coverage for each fixed subset is not enough for arbitrary adaptive selection without a simultaneous event or a validity argument for the selected pipeline. None of these deterministic statements repairs the source-dependence limitation of the current data.

F SUPPLEMENTARY GRAPH AND TRANSFER DIAGNOSTICS

The graph audit was specified after the original exact outcomes were known. Nodes are all fixed-cardinality subsets satisfying the saved incumbent-relative threshold plus 10^{-10}; edges exchange one atom. For terminal node S, feasible component G_S and feasible set G, estimated-KL regret decomposes exactly as

$$\widehat{D}_S - \min_{T \in G} \widehat{D}_T = \left(\widehat{D}_S - \min_{T \in G_S} \widehat{D}_T\right) + \left(\min_{T \in G_S} \widehat{D}_T - \min_{T \in G} \widehat{D}_T\right).$$

The two terms are the within-component and component-separation gaps. Nonoptimality uses 10^{-9} and improving edges use 10^{-12}. Among nonoptimal cells, summed gaps are $0.230012 \simeq 0.017221 + 0.212790$ (individually rounded). Forty-eight of all 96 graphs are disconnected, including graphs where CPC nevertheless attains an optimum. Component membership was independently recomputed using set-based breadth-first search after the sparse-graph implementation. This is a computational cross-check, not independent human review.

The structural panel saved its plan, 20 latent tensors and subset universe before evaluating objectives. For regime index $r = 0, 1, 2, 3$ and replicate $j = 0, \ldots, 4$, seeds are $2026090900 + 100r + j$. Noise and heterogeneous radii use seed plus 10000, with radii $0.6925(0.5 + U[0,1])$ after Gaussian perturbations; constant radii are 0.6925. It uses the archived generator, identity order, lexicographic

Setting	Target rows	Median η	Median δ	Median $\log_{10}(U/D)$	$U < F(\widehat{\epsilon}+\delta)$
Atomic 0.20	886	11.06	10.50	4.70	16/886
Atomic 0.35	886	11.06	7.75	4.72	21/886
Paired 0.20	850	8.37	6.25	3.87	12/850
Paired 0.35	850	8.37	5.50	3.75	14/850

Table 5: CPC oracle transfer diagnostics among actual omissions. Each base has two target rows. Tightness ratios exclude $D \leq 10^{-12}$ only; all other counts include them. U is Eq. (14)'s upper bound, $\widehat{\epsilon} = \mathrm{osc}(\widehat{r})$, and the final column counts strict improvements over the known range-only upper bound using the same aggregate allowance. No uncertainty was learned or refitted.

ties and exact robust initialization. The prediction was at least one nonoptimal terminal component without a global constrained optimum, not a minimum failure rate. All forty cells were retained, and a second graph implementation verified component sizes and gaps. The chosen dimensions and question were motivated by earlier results; this is a prospectively fixed structural check, not preregistered confirmation of the entire paper.

The supplied saved-score audit reports evaluation of all 52,000 neural evaluation method-target rows, and all enumerated subsets in the original exact panel. The actual nominal full reference reconstructs estimator empty plus all singleton increments; it is not replaced by the direct estimator full posterior. True KL is recomputed in float64 log space, and both sides of the transfer interval and known $F(\epsilon)$ and $F(\widehat{\epsilon}+\delta)$ bounds pass numerical checks. Table 5 reports all four operating points. The positive-loss denominators for its tightness ratios are 883, 877, 841 and 841, respectively. The transfer upper bound improves on the stronger oracle $F(\epsilon)$ in only 7, 7, 0 and 2 target rows, respectively. Zero interval-separated orderings occur for each CPC-versus-KL-greedy and CPC-versus-LLR comparison, each using 1,000 paired target rows. These eight comparisons reuse bases and targets. In the exact panel, CPC's upper bound improves on $F(\epsilon)$ in 24/96 cells and on $F(\widehat{\epsilon}+\delta)$ in 30/96; no CPC lower bound is positive above 10^{-12}. The diagnostic does not establish a useful learned certificate.

Among actual omissions, the atomic settings have 443 bases each and the paired settings have 425. Joint agreement counts with the unique full winner are 172, 216, 223 and 232 in operating-point order. The oracle full-range test succeeds on 92, 126, 142 and 146; the oracle signed oscillation-ball test succeeds on 59, 88, 86 and 89. These tests use different uncertainty descriptions and are not a dominance comparison. This count requires a unique full-target winner and equality with the deterministic subset argmax, but does not exclude subset ties. It therefore differs from both a two-sided unique-winner test and the earlier plain argmax-equality diagnostic, which permits full-context ties. Full retention accounts separately for 57 atomic and 75 paired bases at each fraction. These oracle successes do not change the observed zero useful issuance of the calibrated pipeline.

G Experimental protocol and computational scope

Controlled generators. All generators draw $b \sim N(0, 0.25^2 I)$ and center each contribution across hypotheses. Aligned contributions use a centered unit Gaussian direction, lognormal strengths with log-mean -0.3 and log-standard-deviation 0.6, and coordinate noise of standard deviation 0.20. Cancellation alternates $v, -v+\eta$ with centered Gaussian directions and noise standard deviation 0.08. Near-tie contributions begin with coordinate standard deviation 0.35 and adjust the final atom so the centered total is near a fresh standard-deviation-0.02 vector. Pathways use five centered Gaussian centers of coordinate standard deviation 0.7, alternating scales 1 and -0.8 across blocks, with standard-deviation-0.15 noise. Estimator perturbations are independent Gaussian coordinates at the condition's σ.

Seeds and selection. With condition index $c \in \{0, \ldots, 35\}$ and replicate s, the evaluation seed is $20260821 + 1000c + s$ and the development seed adds 10^7. Evaluation perturbations add 999999 to the latent seed; development perturbations use $20260821 + 2 \times 10^7 + 1000c + s$. Conditions follow the generator, noise and fraction order stated in the main text. Each condition produces 15×5 CPC development records and 50×9 evaluation records. All methods share the same instance and exact budget. Position takes the prefix; lexical ranks term frequency–inverse document frequency

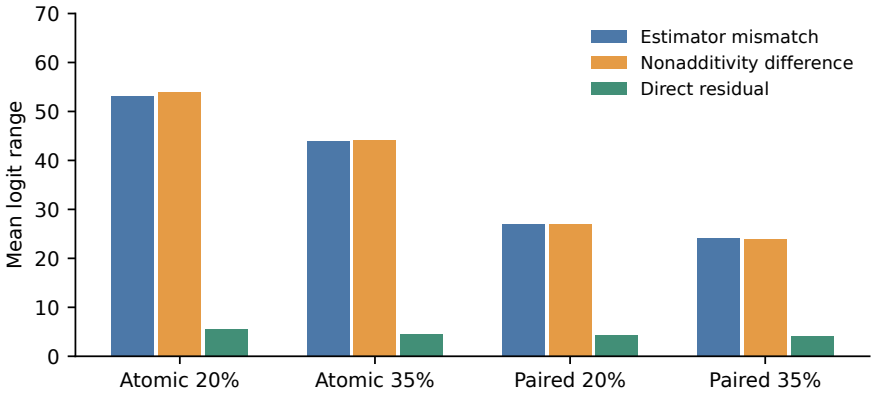

Figure 4: Completed neural mismatch/nonadditivity diagnostic. Bars show the mean ranges of M, N and the direct full-minus-subset residual on 1,000 target rows (500 bases) per setting. They are not scalar additive shares: Eq. (12) is a vector identity and includes the estimator residual. Dependence and cancellation, rather than independent accumulation, explain why large separate ranges need not imply a large direct residual. No causal intervention or new run is represented.

(TF-IDF) similarity to the query using unigrams and bigrams; coverage greedily combines weighted uncovered terms, query matches and a redundancy penalty; LLR norm ranks centered contribution norms; KL-greedy sequentially minimizes estimated forward KL; MLOC, RLOC and CPC use the objectives in Section 3. No method row is discarded on the basis of its result.

Exact diagnostic. The three seeds per regime are $2026090700 + 100r + s$ for $r = 0, 1, 2, 3$ and $s = 0, 1, 2$. Noise is 0.25, cardinalities are 7 and 8, and CPC allows 20 passes at $\tau = 0.05$. The constant pair radius is 0.6925; the heterogeneous construction multiplies it by $0.5 + U_i$ with the prospectively saved uniform atom draws. These are deterministic uncertainty sets, not confidence statements. The original cell artifacts record all $\binom{16}{7} = 11{,}440$ or $\binom{16}{8} = 12{,}870$ subset objectives. The comparator's objective-call allowance matches refinement calls, excluding initialization; both pipelines' initialization and stopping rules are preserved in the script. Neither these counts nor the controlled solver time limit establish equal total computational cost.

Neural scoring. The estimator and two target checkpoints are specified in Section 5. Exact model revisions and acquisition records accompany the supplementary implementation. The context cap is 8,192 tokens, with a single prompt per scoring microbatch. Answer options for that prompt are scored together; the microbatch setting does not mean that options are sampled. No model training or stochastic answer generation is performed. Scores sum the causal token log probabilities of the answer-label suffix without an end-of-sequence (EOS) score. The scorer separately tokenizes the prompt (including its special tokens) and a space-prefixed answer label (without special tokens), then concatenates the token IDs. This fixed token-sequence convention is not necessarily equivalent to retokenizing the combined string. Full, empty and selected contexts use the recorded prompt construction. The reported evaluation used CUDA scoring. Capped item budgets and both atomizations use Eq. (16), with Python rounding (ties to even). The supplementary package documents the scoring configuration, checkpoint identities, and data-preparation workflow.

Neural baselines and calibration. The retained ledger includes CPC, RLOC, MLOC, additive KL-greedy, LLR norm, lexical, maximal marginal relevance (MMR) lexical (Carbonell & Goldstein, 1998), lead and five random selections with seeds 17, 29, 43, 71 and 113. Lexical and MMR use the question followed by all answer-option texts as the query, without access to the gold option. TF-IDF is fitted to that query and its candidate atoms using unigrams and bigrams with ℓ^2 normalization. MMR greedily maximizes 0.7 times query similarity minus 0.3 times maximum similarity to an already selected atom; smaller atom indices break exact ties. The main comparison reports CPC versus LLR norm; it is not a claim of superiority over the entire ledger. Neural solver defaults differ from the controlled campaign: 60-second limits and requested relative gap 10^{-3} in the saved implementation. Neural CPC has no swap-pass cap, unlike the 20-pass controlled and exact

routines. It requires a computed improvement larger than 10^{-12} and allows certificate tolerance 10^{-10}. Selection objectives floor each probability at 10^{-15} and then renormalize each distribution; the stable-metric audit recomputes displayed neural KL from saved logits via log-sum-exp without changing selections or development choices. Thus the exact-KL refinement theorem is not an unconditional theorem about the recorded floating-point trajectory. The residual calibration has 120 RACE, 100 DREAM and 80 SciFact bases. Settings and the nonnegative development predictor are frozen before these residuals are evaluated.

Development-only selection and predictor fitting. The included development scripts evaluate six interval candidates per atomization (constant or adaptive radii at interval-fitting levels $0.05, 0.10, 0.20$), crossed with $\tau \in \{0, 0.02, 0.05, 0.10, 0.20\}$ and the two retained fractions. Each setting uses all 300 development bases and both targets. The completed-bank validators require 36,000 selections and 72,000 target-method rows. For each atomization and fraction, candidates with a Wilson lower bound of at least 0.90 on joint direct-error coverage are ranked by median robust radius, then mean target direct KL, then a stable key. If none qualifies, the fallback maximizes empirical coverage, with mean KL and the stable key breaking ties. After that choice, nonnegative least squares fits the positive part of joint direct error minus nominal range, with features $(1, \widehat{\epsilon}, C, k)$. The selector and this predictor are frozen before residual calibration and evaluation. These are the recorded rules, not a new search or a claim that source-level independence has been established.

H Statistical analysis and additional certificate results

Controlled bootstrap. The fixed bootstrap seed is 20260906. Within each condition, multinomial weights resample its 50 evaluation instances; the same paired weights apply across methods and endpoints. The 11 scopes are pooled, four generator margins, three noise margins and three budget margins. Seven contrasts per scope comprise five comparisons $D_{\mathrm{CPC}} - 0.8D_B$ for random, position, lexical, coverage and LLR norm, plus the range and KL-cost comparisons defined in the main text. Both directional tails enter one Holm step-down family (Holm, 1979). We use an add-one tail estimate, store raw exceedances and ranks, and report pointwise percentile intervals. Clopper-Pearson intervals concern Monte Carlo tail probability, not population-effect uncertainty. Bonferroni upper limits across 154 tails followed by Holm adjustment provide the conservative Monte Carlo check.

Contrast	Estimate	95% interval	Holm p	MC upper
random	-0.5171	[-0.5572, -0.4777]	0.000154	0.001237
position	+0.1094	[+0.0915, +0.1276]	1.000000	1.000000
lexical	+0.1094	[+0.0915, +0.1276]	1.000000	1.000000
coverage	-0.1327	[-0.1559, -0.1096]	0.000154	0.001237
llr norm	-0.2018	[-0.2284, -0.1754]	0.000154	0.001237
Range: 1%	-0.6840	[-0.7565, -0.6120]	0.000154	0.001237
KL cost: 10%	-0.0393	[-0.0582, -0.0205]	0.002112	0.003994

Table 6: Pooled lower-tail threshold tests. Negative contrasts favor meeting the stated threshold. "MC upper" applies Holm to simultaneous Monte Carlo tail upper bounds. Other scopes remain in the saved 154-test family. Intervals are pointwise. The label "KL cost" permits a 10% cost and must not be interpreted as a superiority test.

These are approximate centered-bootstrap tests, not exact finite-sample population tests. Monte Carlo-tail adjustment controls simulation uncertainty within the specified procedure; it does not make its bootstrap population approximation exact. The repeated-resampling analysis does not retroactively make hypotheses prospective. The previous 198-comparison metric family is separate; no 5% family-wise guarantee is claimed over its union with these tests. Neural intervals use 5,000 source-cluster draws within dataset, retain dataset weights, and average target outcomes within base. They are descriptive, pointwise comparisons for the fixed evaluated pipeline. They do not establish calibration coverage under arbitrary within-source dependence.

Across all roles, the 62 overlapping source components comprise 56 interval-fit/residual-calibration, three development/residual-calibration, one development/evaluation, one interval-fit/evaluation and one development/interval-fit component. These counts locate dependence; they do not prove that

the final residual quantiles leaked into development selection. These facts limit the current coverage claim. We make no prospective population guarantee for this calibration design and do not selectively remove overlaps. All four calibrated constructions and all operating points are reported in Table 2 of the main text.

I FROZEN NEURAL SETTINGS

Table 7 records the selected parameters. The nominal predictor is $\widehat{\epsilon} + \beta_0 + \beta_a\widehat{\epsilon} + \beta_C C + \beta_k k$, with $\beta_a = \beta_C = 0$ in all selected settings. The paired interval candidate is labeled adaptive in the implementation, but its fitted slope is zero and floor is one, so the selected coordinate radii are constant. Pairwise radii are the sums of the two coordinate radii. The interval-fitting level is 0.20, distinct from the final residual-calibration level 0.05 reported here. For N residual-calibration scores, the quantile uses sorted rank $\lceil (N+1)(1-\alpha) \rceil$, or infinity if that rank exceeds N.

Atoms	f	τ	Coord. radius	β_0	β_k
Atomic	0.20	0.02	10.937500	3.009167	0.000000
Atomic	0.35	0.00	10.937500	3.053750	0.000000
Paired	0.20	0.00	10.437500	2.828675	0.116121
Paired	0.35	0.00	10.437500	3.115417	0.000000

Table 7: Retained neural parameters, chosen before residual calibration and evaluation and tabulated from the saved pipeline. Values are displayed to six decimal places; the same machine-readable table is included in the supplementary evidence directory.

Table 3 in the main text retains every selector in the evaluated ledger, including all five random seeds. The entries are descriptive direct-KL means; the pointwise source-cluster intervals in the main text apply only to CPC versus LLR norm. A smallest displayed mean is not an adjusted superiority result, and none of these averages validates certificate coverage.

J EVIDENCE PROVENANCE AND REPRODUCTION PACKAGE

Empirical evidence. The empirical evidence comprises the completed controlled, exact, and neural analyses, with aggregate tables, figures, frozen settings, and audit definitions. The supplementary implementation and machine-readable aggregate assets document the methods and reported outcomes. An implemented baseline is not treated as a completed evaluation. This paper makes no completed token-budget, neural second-order repair, new model-scale, or compute-matched superiority claim.

Oracle count interpretation. The exact-range oracle in Figure 2 is defined target by target before requiring both to pass. It is not a common-radius oracle. Counts are descriptive, not significance tests.

Reproduction package. The supplementary implementation contains the controlled generators, exhaustive subset enumeration and graph checks, neural scoring and selection methods, analysis scripts, configurations, and tests. Model and dataset revisions, file-level instructions, and the claim-to-evidence index are supplied with the code rather than repeated here. The exact synthetic panels can be regenerated from fixed seeds; timed mixed-integer solves may yield different incumbents across systems.

The released 1,500-example regeneration manifest and general launch workflow use a different role allocation from the reported 1,250-example neural study. They are not an exact historical split manifest. The supplementary documentation distinguishes these workflows and the additional run-specific inputs needed by the historical analysis scripts. The reported neural aggregates are retained unchanged; no new neural run or independent raw-score reaggregation is claimed.

Numerical checks. The accompanying verification script evaluates randomly chosen finite probability vectors and residuals against the extremal formulas, independently solves endpoint-

constrained minimizations, enumerates the bounded-exchange constructions for several exchange budgets, and checks the deterministic nesting of decision events. These are numerical falsification checks of transcription and formulas; they are neither formal proof verification nor empirical validation on language models. Complete analytic proofs establish the mathematical statements. The script generating Figure 2 reads explicit count records with source locations and performs no model calls, resampling, exclusion, or imputation.