IDENTIFICATION LIMITS AND A PROTECTED CORRECTION FOR LLM-DERIVED GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Downstream statistical and machine learning methods increasingly use LLM judgments as external guidance, such as penalty weights, pairwise comparisons or prior scores, often treating the elicited signal as having a stable meaning. Yet when an LLM is shown task-matched evidence, its judgment may combine an evidence-free preference with an evidence-induced shift. We show that repeated querying can estimate this composite judgment arbitrarily precisely without identifying its decomposition. Whenever the queried evidence configurations cannot be combined to reproduce the empty, evidence-free configuration, multiple preference-shift decompositions remain observationally equivalent. This ambiguity persists under adaptive querying and cannot be resolved by increasing the query budget. We therefore develop Selective Correction, whose guarantees depend only on observable quantities. A data-only reference procedure identifies comparisons that could change the decision, an evidence gate restricts queries to those supported by task-matched evidence and the LLM supplies their direction, with its certainty allowed only to attenuate the correction. We prove that any reference ordering whose margin exceeds the sum of its local correction budgets is preserved for every possible set of LLM responses. Across seven datasets, Selective Correction improves over the data-only Reference in every setting while certifying more than 93% of reference orderings in every audited setting. Randomization controls further separate the gains due to LLM content from those due to selective targeting alone.

1 INTRODUCTION

Large language models (LLMs) increasingly provide external guidance to downstream statistical and machine learning procedures, as relevance scores, pairwise comparisons, penalty weights, prior inclusion probabilities or warm starts (Jeong et al., 2025; Capstick et al., 2025; Zhang et al., 2025; Skinner et al., 2026; Feng et al., 2025; Bryan et al., 2025; Liu et al., 2024). These procedures typically treat the elicited signal as having a stable semantic interpretation, for instance the model's prior knowledge of which variables matter for a prediction task. However, this interpretation depends on the information supplied in the prompt, since an LLM shown task-matched evidence returns a judgment that combines an evidence-free preference over the items with an evidence-induced shift. The standard remedies for unstable judgments, including repeated sampling, prompt rewording and self-consistency aggregation (Wang et al., 2023), reduce stochastic variation within a fixed information configuration but leave these two components entangled. Consequently, an elicited judgment can become arbitrarily precise while what it represents remains unidentified.

We formalize this gap by writing the judgment under information configuration d as $\mu(d) = q + Z\Delta d$, where $q = \mu(0)$ is the evidence-free preference and $Z\Delta d$ the evidence-induced shift. Whenever the configurations available to a protocol have an affine hull that excludes the empty configuration $d = 0$, distinct pairs (q, Δ) induce the same law for the entire elicitation record under unlimited repetition and adaptive querying (Proposition 1). Additional budget therefore sharpens the composite $\mu(d)$ without recovering q. Conversely, under a known link and adequate comparison coverage, a design whose hull contains the empty configuration identifies q, which makes the obstruction structural rather than a consequence of noisy responses. This structure leaves downstream methods with two gaps: (i) a procedure that reads an elicited comparison as the evidence-free prefer-

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.

ence commits to one decomposition among many that the record cannot distinguish and (ii) because elicitation is outcome-blind, even an identified q need not help the prediction task.

In this work, we ask which properties of a downstream decision can be guaranteed uniformly over this ambiguity and introduce Selective Correction, which limits the authority of LLM guidance using observable quantities alone. The data-only reference procedure determines where a correction could affect the selected set, task-matched evidence determines which comparisons are admissible and the LLM supplies their direction, while validation sets the correction strength. We prove that for any query set and any nonnegative weights, every corrected score remains within a deterministic local correction budget of its reference score, so any reference ordering whose margin exceeds the two budgets is preserved (Theorem 1). The guarantee thus holds whether or not the judgments are correct, consistent or attributable to a particular latent preference, so the certified part of the decision is immune to (i). Since the guarantee concerns stability rather than predictive improvement, validation over a grid containing zero decides whether the guidance helps, which addresses (ii). Across seven datasets spanning transcriptomic, clinical, tabular and financial prediction, Selective Correction improves on the data-only Reference in every setting, matches or exceeds a broad Global Correction in the headline metric of every dataset and certifies more than 93% of reference orderings in every audited setting. Together, the theory and experiments separate two questions that are often conflated in LLM-guided learning, namely what an elicited signal can be identified to mean and how much authority it can safely be given while that meaning remains unresolved.

1.1 Summary of contributions

Our contributions form a single chain. **(i) Structural ambiguity.** Any configuration design whose affine hull misses the empty configuration leaves an exact observational equivalence between distinct decompositions, under adaptive querying and stopping (Proposition 1), with a minimax lower bound and a failure of top-k identification at every budget (Corollary 1). **(ii) A requirement on guarantees.** Since outcome-blind records also leave the alignment between q and the downstream target open, a downstream guarantee must hold uniformly over the parameterizations that induce the same record (Section 2.3). **(iii) A correction with an observable certificate.** Any reference ordering whose margin exceeds two local correction budgets survives every admissible judgment, with an unimprovable coefficient, a top-k certificate and exact fallback to the Reference (Theorem 1). **(iv) Empirical evaluation.** A simulation confirms that the equivalence is exact and that repetition converges at the full rate to a point other than q (Section 4.1). On seven datasets, the shuffled and random controls separate the value of targeting from that of the judgments' content (Section 4.2).

1.2 Related work

LLM-derived guidance in downstream methods. Existing methods elicit an importance score (Jeong et al., 2025), a Bayesian prior (Capstick et al., 2025), penalty weights (Zhang et al., 2025), inclusion probabilities (Skinner et al., 2026), a learned degree of trust (Zhang et al., 2026; Chen & Yi, 2026), information for hypothesis testing (Bryan et al., 2025), a proposal in Bayesian sampling (Feng et al., 2025), a warm start (Liu et al., 2024) or structured investor views (Lee et al., 2025b). Such signals are usually evaluated through downstream performance (Kristiadi et al., 2024), while prompt sensitivity makes their interpretation pressing (Sclar et al., 2024; Zheng et al., 2023). We instead ask what a comparison can represent and bound how far it can move a decision.

Identification and measurement. The closest formal work (Chen et al., 2026) studies identification from LLM reports without varying the evidence configuration, whereas we characterize an obstruction that survives several configurations and the designs that escape it. Unlike corrections of machine predictions with labeled data (Angelopoulos et al., 2023; Egami et al., 2023), we work with a latent relevance signal and build on classical measurement and paired-comparison foundations (Campbell & Fiske, 1959; Ford, 1957; Bradley & Terry, 1952).

2 Identification limits of LLM elicitation

This section separates the observable composite from its decomposition, shows that the record leaves the decomposition open and derives what a downstream guarantee may depend on.

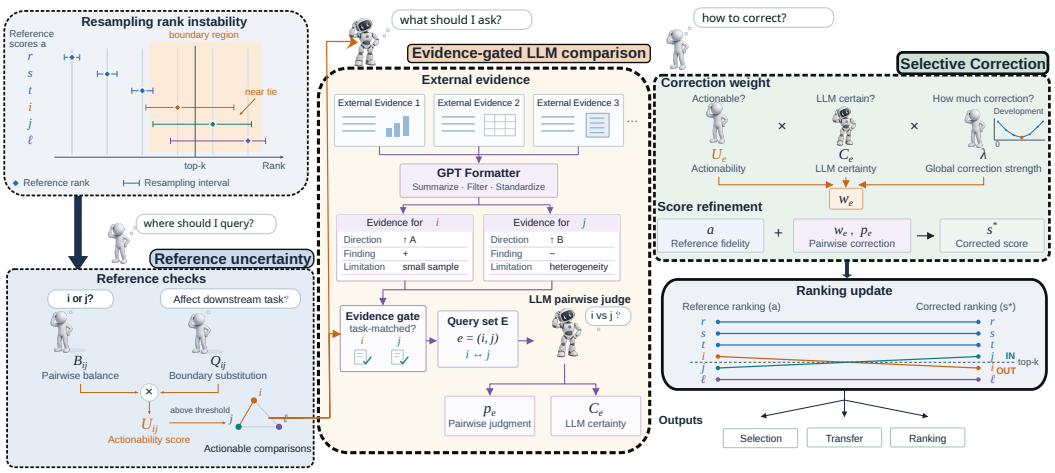

Figure 1: Overview of Selective Correction. (1) Resampling the reference selector yields the actionability score $U_{ij} = Q_{ij}B_{ij}$, which locates comparisons whose order is unstable and consequential. (2) An evidence gate admits comparisons with task-matched evidence, on which the LLM returns a judgment p_e whose certainty C_e scales the weight. (3) The weights $w_e = \lambda U_e C_e$ enter an objective anchored at the reference scores a, whose solution $s^\star$ moves each item by at most its local correction budget (Theorem 1).

2.1 Fixed-information elicitation

A downstream learner must order p candidate items (features, variables or factors) and consults an LLM by pairwise comparison (Bradley & Terry, 1952). For a pair $e = (u_e, v_e)$, the model sees the two items together with an information configuration, the evidence the prompt presents, such as variable documentation, ontology entries or retrieved abstracts. The latent comparison margin under the configuration actually used decomposes as

$$m_e \;=\; b_e \;+\; \delta_e, \tag{1}$$

where the composite judgment m_e adds the evidence-induced shift δ_e to the evidence-free preference b_e that the model expresses when no evidence is presented. The response is binary with $\mathbb{P}(Y_e = 1) = G(m_e + \kappa\, o)$, where G is known, continuous and strictly increasing and the randomized presentation order $o \in \{-1, +1\}$ has an effect κ that the design separates out (Bradley & Terry, 1952; Davidson & Beaver, 1977). Leaving the dependence across responses open, we let a protocol query in rounds. At round t it chooses, as a function of the history H_{t-1}, a batch A_t of n actions and observes the responses $Y_t \in \{0, 1\}^n$, where the ith action specifies a pair e_i, a configuration d_i and a rendering θ_i (wording, decoding and presentation order). Writing $m_e(d) = b_e + \delta_e(d)$ for the margin of pair e under configuration d, we impose one restriction on their joint law.

Assumption 1 (Batch contrast sufficiency)**.** For every round t, history h_{t-1} and batch A_t, the responses depend on the preferences and shifts only through the queried margins, in the sense that

$$\mathbb{P}\big(Y_t \in \cdot \mid H_{t-1} = h_{t-1},\; A_t\big) \;=\; K_{h_{t-1}, A_t}\big(\cdot;\; m_{e_1}(d_1), \ldots, m_{e_n}(d_n)\big) \tag{2}$$

for a probability kernel K that does not depend on the preferences and shifts.

Assumption 1 states that the answers depend on the latent preferences and shifts only through how far apart the model places each queried pair under the configuration shown. It is the batch form of the difference structure of classical paired comparison (Thurstone, 1927; Bradley & Terry, 1952; Luce, 1959), whose simplest instance takes K to be the product of the $\mathrm{Bernoulli}\big(G(m_{e_i}(d_i) + \kappa\, o_i)\big)$ laws. Moreover, score-difference models describe LLM judgments well, since comparative assessment with LLMs is competitive with dedicated evaluators once position effects are corrected (Liusie et al., 2024a) and recovers a ranking from a small fraction of the pairs (Liusie et al., 2024b). Because the kernel may depend arbitrarily on the history, the rendering and the batch, the assumption also

accommodates adaptive protocols, within-batch dependence and the position-dependent and non-transitive judgments reported for LLM judges (Xu et al., 2025). It cannot simply be dropped, since a batch law that depends on the evidence-free preferences directly separates the aliased parameter points of Proposition 1 (Proposition 3). Finally, b_e is an elicitation-channel estimand, distinct from the items' true relevance, whereas the measured quantity m_e is a property of both the items and the prompt.

2.2 Evidence confounding under replication

Since the responses depend on b_e and δ_e only through their sum, repeated queries estimate m_e ever more precisely without separating its parts. Moreover, varying the prompt is no remedy by itself, since the one-source-at-a-time ablation changes the configuration at every query and still cannot separate them. We collect the items' attributes in a centered matrix $Z \in \mathbb{R}^{p\times r}$ ($\mathbf{1}^\top Z = 0$) of full column rank, their evidence-free preferences in $q \in \mathbb{R}^p$ and the sensitivity of these preferences to S evidence sources in $\Delta \in \mathbb{R}^{r\times S}$, with $\mathbf{1}^\top q = 0$ fixing the additive constant that the record cannot see (Proposition 2). A configuration $d \in \{0,1\}^S$ induces the composite score vector $\mu(d) := q + Z\Delta d$, with $m_e = \mu_{u_e}(d) - \mu_{v_e}(d)$ and $q = \mu(0)$, so that $b_e = q_{u_e} - q_{v_e}$.

A protocol can recover q without querying the empty configuration, since whenever weights α_d summing to one place the queried configurations at the origin, $\sum_d \alpha_d d = 0$, the vector $q = \sum_d \alpha_d \mu(d)$ is an affine combination of measured composites. When no such weights exist, the empty configuration lies outside the affine hull of the set $\mathcal{D}$ of configurations a protocol may query ($0 \notin \mathrm{aff}(\mathcal{D})$), so that some $c_0 \in \mathbb{R}^S$ satisfies $c_0^\top d = 1$ throughout $\mathcal{D}$. For $\gamma \in \mathbb{R}^r$, we then define the *aliasing map*

$$T_\gamma(q,\Delta) \;:=\; \big(q + Z\gamma,\; \Delta - \gamma c_0^\top\big), \tag{3}$$

which moves $\mu(d)$ by $Z\gamma\,(1 - c_0^\top d)$ and therefore leaves it unchanged at every $d \in \mathcal{D}$, sliding the evidence-free preference by $Z\gamma$ while compensating exactly in the source sensitivities. Finally, let $\Pi(\mathcal{D})$ denote the protocols whose queried configurations stay in $\mathcal{D}$ and are otherwise unrestricted, allowing any number of repeated queries, any rewording or decoding change, any aggregation rule, any stopping rule and any adaptive choice of pair and configuration. The following proposition shows that no protocol in this class distinguishes (q,Δ) from $T_\gamma(q,\Delta)$.

Proposition 1 (Observational equivalence of aliased parameters)**.** *Under the comparison model of Section 2.1, including Assumption 1, suppose that $0 \notin \mathrm{aff}(\mathcal{D})$, with c_0 as above. Then for every $\gamma \in \mathbb{R}^r$ and every protocol $\pi \in \Pi(\mathcal{D})$, the elicitation record has the same law under both parameter points,*

$$\mathbb{P}_{q,\Delta} \;=\; \mathbb{P}_{T_\gamma(q,\Delta)},$$

as laws of the record stopped at the protocol's own stopping time. In particular, the total variation distance between the two record laws is zero at every budget and under every stopping rule.

Proposition 1 establishes an observational equivalence in the sense of Rothenberg (1971). When $0 \notin \mathrm{aff}(\mathcal{D})$, every point of the orbit $\{q + Z\gamma : \gamma \in \mathbb{R}^r\}$ lies in the identified set for q (Manski, 2003), because the transformed parameters induce the same record law. The ambiguity is thus structural, since sampling noise and completion instability average away with the budget, whereas rewording, decoding and calibration change how margins map to responses but leave the margins themselves fixed. Self-consistency aggregation, temperature sweeps, paraphrase and adaptive querying within $\mathcal{D}$ therefore buy precision about the queried composites while the orbit survives every budget. The following corollary quantifies what the ambiguity costs, where $\tilde{q} := q + Z\gamma$ denotes the aliased evidence-free preference and $\mathbb{P}^\star$ the law the two parameter points share.

Corollary 1 (Cost of observational equivalence)**.** *Under the conditions of Proposition 1, fix γ with $Z\gamma \neq 0$.*

- *(i) For every estimator $\hat{q}$ measurable in the record, $\max\big\{\mathbb{E}_{\mathbb{P}^\star}\|\hat{q}-q\|,\, \mathbb{E}_{\mathbb{P}^\star}\|\hat{q}-\tilde{q}\|\big\} \;\geq\; \frac{1}{2}\|Z\gamma\|$, a bound that no budget and no computation reduces.*
- *(ii) Let $\psi\colon \mathbb{R}^p \to \mathcal{Y}$ be measurable with $\mathcal{Y}$ standard Borel and let $\hat{\psi}$ be any record-measurable estimate of $\psi(q)$. If $\psi(q) \neq \psi(\tilde{q})$ then $\min\big\{\mathbb{P}^\star(\hat{\psi} = \psi(q)),\, \mathbb{P}^\star(\hat{\psi} = \psi(\tilde{q}))\big\} \leq \frac{1}{2}$. Taking $\psi = S_k$, the top-k set of the evidence-free preference, which is the target of top-k rank aggregation (Chen & Suh, 2015), is not identified whenever $S_k(q) \neq S_k(\tilde{q})$.*

Part (i) is a two-point bound in the sense of Le Cam (Tsybakov, 2009). For $k < p$, some γ moves the top-k set, so part (ii) applies (Section A.7). Section A proves both parts and also records the Fisher-null direction of the record likelihood.

2.3 Downstream guarantees under an unidentified decomposition

Proposition 1 is the negative half of an exact criterion. With a known rendering-invariant link, both presentation orders and connected comparisons at configurations spanning $\mathrm{aff}(\mathcal{D})$, the evidence-free preference q is identified exactly when $0 \in \mathrm{aff}(\mathcal{D})$ and the full decomposition exactly when this hull has dimension S (Section A.6), so the empty configuration plus one configuration per source is minimal. Precision about the composite therefore does not imply identification of q, which the geometry of the design decides. Identification in turn does not imply alignment with the downstream task. Under outcome-blind elicitation, where the model never observes the downstream outcomes, the record law is a function of (q, Δ) and the protocol alone, so settings in which q is aligned or anti-aligned with the downstream target induce the same record law. These are the two gaps of Section 1, of which even a fully identifying design closes only the first.

We therefore ask which properties of a downstream decision can be guaranteed uniformly over the ambiguity that the record leaves unresolved, a question that two design principles answer. First, identification-robust influence requires a downstream guarantee to be stated in quantities that the record fixes, so that it holds for every decomposition and alignment consistent with the record. A safeguard that rests on the faithfulness of p_e to b_e fails this test, since parameter points that the record cannot separate disagree about b_e. The bound of Theorem 1, in contrast, holds for every admissible answer at once, with an envelope that depends on the query set and the actionability scores alone. Second, decision relevance requires external guidance to be spent where the reference procedure leaves the decision unsettled, since these are the comparisons on which an external judgment can change the decision. The first principle thus fixes what the guarantee may depend on and the second fixes where the guidance is worth buying, which together determine the architecture of Section 3.

3 Selective Correction and its protected region

Throughout, the Reference denotes the condition in which no external guidance is used. A reference selector, such as the ℓ_1 path, elastic net, sure independence screening or a gradient-boosted permutation ranking, is trained on the available data and produces the reference score vector $a \in \mathbb{R}^p$, whose ordering is the reference ranking and whose top-k set $S_k(a)$ is the reference top-k set. Since a is computed from the data alone, it is unrelated to the evidence-free preference q of the model.

Selective Correction implements the two principles of Section 2.3, with query targeting realizing decision relevance and the protected-region guarantee realizing identification-robust influence. Instead of resolving the ambiguity of Proposition 1, it makes its guarantee invariant to that ambiguity by dividing authority among four sources, each acting through observable quantities. First, the data decide where external guidance could matter, by resampling the reference selector. Second, the evidence gate decides which of these comparisons are admissible. Third, the model supplies a direction on each admitted comparison, weighted by a certainty that can only shrink its influence. Finally, validation sets the correction strength λ, while Theorem 1 certifies what none of these steps can overturn. Figure 1 summarizes the pipeline that Sections 3.1 to 3.4 develop.

3.1 Query targeting by reference uncertainty

Following stability selection (Meinshausen & Bühlmann, 2010; Shah & Samworth, 2013), we refit the reference selector on R resamples of the training data and define, for each pair (i, j) of items,

$$Q_{ij} := \mathbb{P}\big(\text{exactly one of } i, j \in S_k\big), \qquad B_{ij} := 1 - 2\big|\mathbb{P}(a_i > a_j) - \tfrac{1}{2}\big|, \qquad U_{ij} := Q_{ij}\, B_{ij}, \tag{4}$$

where the probabilities are taken over resamples. The substitution probability Q_{ij} measures selection-boundary instability, the pairwise balance B_{ij} measures whether the reference procedure orders the pair consistently and their product, the *actionability score* $U_{ij} \in [0, 1]$, is large only when reordering i and j is both plausible and consequential. A pair firmly inside or outside the selected set has $Q_{ij} \approx 0$ however unstable its order, whereas a boundary pair the reference orders decisively has $B_{ij} \approx 0$.

We call a pair an actionable comparison when U_{ij} exceeds a threshold fixed before any query (Section D.7). The query set E consists of the actionable comparisons that also pass the evidence gate, a binary criterion specified in advance that requires both items to carry retrievable task-matched evidence. If either leaves E empty, the procedure returns the Reference unchanged. As a principled criterion of decision relevance rather than an optimal allocation of queries, the actionability score depends on the reference procedure alone, since targeting asks where the data leave a decision unresolved, a property that no report of the model's certainty can measure.

3.2 THE CORRECTION OBJECTIVE

Each $e \in E$ returns the model's pairwise judgment $p_e = \mathbb{P}(u_e \succ v_e) \in [0,1]$, the reported order-symmetrized preference probability, which serves as a normalized report and requires no calibration. The corrected score vector $s^\star$ minimizes

$$F(s) \;:=\; \tfrac{1}{2}\|s-a\|_2^2 \;+\; \sum_{e\in E} w_e\, \mathrm{CE}\big(p_e,\, \sigma(s_{u_e}-s_{v_e})\big), \qquad w_e \;:=\; \lambda\, U_e\, C_e, \tag{5}$$

where $U_e := U_{u_e v_e}$ (and likewise Q_e and B_e), $\lambda \geq 0$ is the correction strength, $w_e \geq 0$ the correction weight of comparison e, $C_e := 1 - \mathrm{Ent}(p_e)/\log 2$ the LLM certainty, σ the logistic function, CE the binary cross-entropy and Ent the binary entropy. The quadratic term anchors the corrected scores at the Reference, while the pairwise term permits local changes only where a queried judgment supports them. The certainty measures the decisiveness of a report rather than its accuracy and only modulates a comparison already admitted to E, leaving the protection below intact because $0 \leq w_e \leq \lambda U_e$ whatever the model reports.

The objective in equation 5 has three properties that suit the design. First, the objective reconciles mutually inconsistent comparisons in a single strongly convex problem with a unique solution (Boyd & Vandenberghe, 2004). Second, differentiating gives $s^\star - a = -N^\top f$ with $|f_e| \leq w_e$, where N is the signed incidence matrix of the query graph, so the correction is a sum of bounded contributions along that graph and needs no truncation step. Third, the objective returns a at $\lambda = 0$, so recovering the Reference is a selectable outcome of validation over a grid containing zero. We call the method Selective Correction and refer to the same objective on a broad query set fixed independently of reference uncertainty as Global Correction, the natural ablation, which the guarantee below also covers (Section C.1).

3.3 THE PROTECTED-REGION GUARANTEE

We now state the guarantee, which holds for any query set and any nonnegative weights.

Theorem 1 (Protected-region guarantee)**.** *Let E be any query set, $\{w_e\}$ any finite nonnegative correction weights and $\{p_e\} \subset [0,1]$ any pairwise judgments. Then F in equation 5 has a unique global minimizer $s^\star$, and, writing $\rho_j := \sum_{e\ni j} w_e$ for item j's* local correction budget*,*

$$\big|s_j^\star - a_j\big| \;\leq\; \rho_j \quad \textit{for every } j, \qquad \textit{hence} \qquad a_i - a_j > \rho_i + \rho_j \implies s_i^\star > s_j^\star$$

for every pair i, j, queried or not. The bound holds uniformly over all admissible judgments $p_e \in [0,1]$.

Theorem 1 establishes that the corrected score of item j never leaves its correction interval $I_j := [a_j - \rho_j,\, a_j + \rho_j]$, whatever the model reports. The bound is deterministic and uniform over $p_e \in [0,1]$, so it covers any LLM, adversarial or mutually inconsistent judgments and uncalibrated reports. Since $C_e \leq 1$, the answer-free envelopes $\bar\rho_j := \lambda \sum_{e\ni j} U_e \geq \rho_j$ certify a region fixed before any query (Corollary 3). Consequently, the reference ordering survives wherever the reference margin $a_i - a_j$ exceeds the two local correction budgets, for queried and unqueried pairs alike. An item never queried has $\rho_j = 0$ and keeps its reference score exactly. We call the set of orderings with this property the *protected region*. This permits us to certify the deployed decision itself, since $\min_{i\in S_k(a)}(a_i - \rho_i) > \max_{j\notin S_k(a)}(a_j + \rho_j)$ implies $S_k(s^\star) = S_k(a)$. Computed from a, the query set and the weights alone, the certificate meets the requirement of Section 2.3.

Moreover, neither threshold can be lowered. With two items, one comparison of weight w, a report $p = 0$ and a reference gap M, the displacement of the first item tends to its budget w as $M \to \infty$, so an unqueried item just below the first overtakes it although their margin approaches $\rho_i + \rho_j$. For a

Algorithm 1 Selective Correction. Procedure and reported certificate.

Require: data, reference selector, k, R, threshold, grid of λ containing 0
Ensure: $s^\star$, $S_k(s^\star)$, budgets $\{\rho_j\}$, intervals $\{I_j\}$, top-k certificate
Step 1: Reference. Fit the reference selector and record a and $S_k(a)$.
Step 2: Locate. Refit on R resamples and keep the comparisons whose U_{ij} in equation 4 exceeds the threshold.
Step 3: Gate. Keep those whose two items both carry task-matched evidence, forming E.
Step 4: Query. Elicit p_e for each $e \in E$ in both presentation orders and set $w_e = \lambda U_e C_e$.
Step 5: Correct. Minimize equation 5 over the grid and choose λ on validation data only.
return $S_k(s^\star)$ with $\{\rho_j\}$, $\{I_j\}$ and the top-k certificate.

queried pair the threshold sharpens to $\rho_i + \rho_j - w_{ij}$ (Section C.6). The proof is elementary by design, since the quadratic anchor makes F strongly convex while stationarity writes each displacement as a sum of pairwise terms bounded by their weights. Without the anchor, the pairwise objective need not have a unique finite minimizer. Section C gives the proof and a version with additive slack for approximate solvers.

However, because the guarantee holds for any query set and any nonnegative weighting, it applies verbatim to Global Correction and provides no evidence that the targeting rule or the evidence gate is well chosen (Section C.1), which we assess empirically in Section 4.2. Moreover, a certified ordering survives even when the guidance would have corrected it. Validation over a grid containing $\lambda = 0$ therefore settles the resulting trade between foregone improvement and avoided damage.

3.4 The procedure

In the last step of Algorithm 1 the guarantee becomes part of the output, since the budgets $\{\rho_j\}$, the correction intervals and the top-k certificate are known once λ is fixed and are reported with the selected set. Every threshold and grid is fixed before the held-out evaluation, at which point the analysis is frozen.

4 Experiments

We evaluate identification on simulated records with planted parameters (Section 4.1) and Selective Correction on real data. The real-data experiments and the ablation use seven datasets on two tasks. Prediction covers top-k selection on GSE272769, CREDIT-G and Darmanis, cross-cohort transfer on Renal TCMR and Breast pCR and transfer across two temporal batches on Osteoporosis, while full ranking uses JKP153. Each dataset follows its prespecified evaluation design, described with the data and their processing in Section D.1. We name each setting by its reference selector and selected-set size k. Thus EN-50, GBM-10, SIS-20 and LASSO-10 denote the elastic net, the gradient-boosted permutation ranking, sure independence screening (Fan & Lv, 2008) and the ℓ_1 path with $k = 50$, 10, 20 and 10, while Top-k denotes the single prespecified selector of a dataset.

4.1 Identification

We evaluate the observational equivalence of Proposition 1 on simulated elicitation records across six query designs and increasing budgets. The records follow the logistic comparison model of Section 2.1 with $p = 12$ items, $r = 3$ attributes, $S = 2$ evidence sources and order effect $\kappa = 0.15$. Each configuration queries a fixed graph of 18 comparisons at budgets from 576 to 9216 queries, with 500 replicates per design and budget (Section B). Two designs satisfy $0 \notin \mathrm{aff}(\mathcal{D})$ and lose q, namely SAT with both sources at every query and the one-source-at-a-time ablation OAT, whereas the other four, MIN, CROSS, BASEFREE and DEF, identify it. We evaluate each design by (1) the Fisher information along its aliasing direction, (2) the convergence rate of the error component along the orbit $\{q + Z\gamma\}$ and (3) the error against each equivalent truth $q + cZ\gamma^\dagger$, where $q + Z\gamma^\dagger$ is the limit of the SAT estimator.

As shown in Figure 2, the identifying designs converge at the parametric rate as expected, with orbit-component slopes between -0.503 and -0.516. However, a blind design carries essentially no

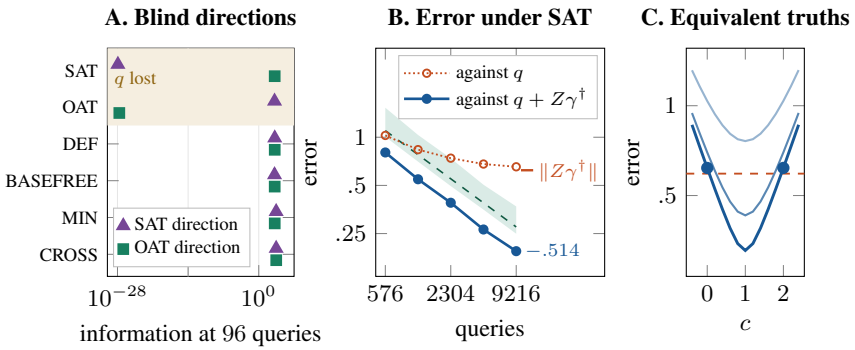

Figure 2: Simulation of Proposition 1 (Section B). (A) Fisher information along the aliasing directions of SAT and OAT at 96 queries per comparison and configuration. (B) SAT error against q and against the equivalent truth $q + Z\gamma^{\dagger}$, with the identifying designs as a band and a dashed guide of slope $-\frac{1}{2}$. (C) Error against each truth $q + cZ\gamma^{\dagger}$ at 576, 2304 and 9216 queries (light to dark), with the bound of Corollary 1 dashed.

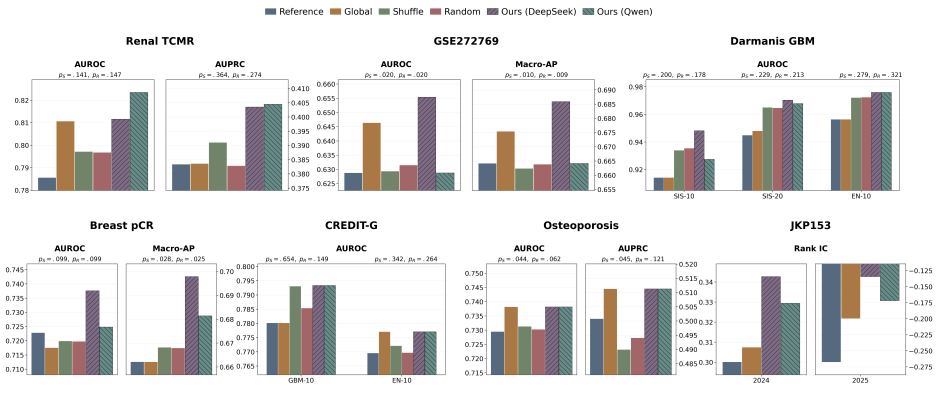

Figure 3: Downstream results on the seven datasets. Hatched bars are Selective Correction with DeepSeek and with Qwen. The values p_{S} and p_{R} are tail probabilities under the shuffled and random controls. The JKP panel adds the held-out 2025 period. Correction strengths are chosen on validation data and all results are reported on held-out or external test data.

information along its aliasing direction (Figure 2A). In particular, under SAT the orbit component stays flat, with slope -0.0003 over a $16\times$ budget increase. Scored against the equivalent truth $q + Z\gamma^{\dagger}$, the same SAT error falls at slope -0.514 while the error against q plateaus at 0.654 (Figure 2B). The pair of truths at $c = 0$ and $c = 2$ attains errors 0.6536 and 0.6540 against the bound 0.6219 of Corollary 1 (Figure 2C). These results demonstrate that repetition sharpens precision about the composite while leaving its decomposition unresolved, exactly as Proposition 1 predicts.

4.2 Targeting and Semantic Value

We compare three methods and two controls on each dataset, leaving additional real-data benchmarks to Sections D.2 and D.5. (i) The Reference uses the reference selector without external guidance. (ii) Global Correction applies equation 5 on a broad query set fixed independently of reference uncertainty, with at least the LLM query budget of Selective Correction. (iii) Selective Correction queries only the actionable, evidence-eligible comparisons of Section 3.1. (iv) Two controls keep the selective targeting but replace the judgments, either by reassigning the measurements across the targeted pairs (shuffled control, tail probability p_{S}) or by removing the pair-specific signal (random control, p_{R}). Both corrected methods choose λ on validation data over a grid containing zero. Judgments come from the DeepSeek V4.1 Flash API and are replicated by a locally deployed Qwen model, Qwen3-32B on the six prediction datasets and Qwen2.5-14B-Instruct on JKP153.

Table 1: Downstream results. Δ_{G} and Δ_{Ref} are Ours minus Global and minus Reference. Bold marks the best method. Protected is the percentage of reference orderings certified by Theorem 1, with no bound violation in any audit. Darmanis gives ranges over three settings and the protection of SIS-10†. GSE272769 protection is fold-weighted.

Experiment	Metric	Ref.	Global	Ours	Δ_{G}	Δ_{Ref}	Protected (%)
Renal TCMR, Top-50	AUPRC	.3834	.3836	**.4035**	+.0199	+.0201	95.84
GSE272769, EN-50	AUROC	.6287	.6463	**.6554**	+.0091	+.0267	93.48
Breast pCR, SIS-20	AUROC	.7228	.7175	**.7376**	+.0201	+.0148	96.73
CREDIT-G, GBM-10	AUROC	.7801	.7801	**.7933**	+.0132	+.0132	–
Osteoporosis, Top-10	AUROC	.7295	**.7382**	**.7382**	.0000	+.0087	93.76
Darmanis	AUROC	three settings			+.0195 to +.0341		98.80†
JKP153, 2024	Rank IC	.3001	.3074	**.3427**	+.0353	+.0426	–

As shown in Figure 3 and Table 1, Selective Correction improves on the Reference on all seven datasets. It also outperforms Global Correction in the headline metric of every dataset except Osteoporosis, where the two tie. Secondary metrics narrow the margin, with Renal TCMR and CREDIT-G EN-10 tying in AUROC (Figure 3). Since Global Correction queries at least as much, these gains come from where the guidance is spent. The ordering persists in the held-out 2025 JKP period, where Selective Correction reaches a Rank IC of -0.134 against -0.200 for Global Correction. Moreover, the controls separate two sources of gain. On GSE272769 both AUROC and macro-AP exceed both controls, Breast SIS-20 separates for macro-AP and Osteoporosis separates from the shuffled control, so in these settings the gain derives from the pair-specific content of the judgments. On CREDIT-G GBM-10 and Darmanis, in contrast, the controls rise with Selective Correction, so there the gain derives from correcting the targeted comparisons. These results demonstrate that selective targeting spends a limited budget where guidance changes the decision.

4.3 Protection and robustness

Auditing every setting that meets the conditions of Theorem 1, we find no displacement beyond its local correction budget and certified percentages of reference orderings above 93% throughout (Table 1). Moreover, the benefit carries over to Qwen, which improves on the Reference in eight of the nine reported prediction settings and returns it exactly in the ninth, where validation selects $\lambda = 0$, with zero bound violations in all nine audits (Section D.3). These results demonstrate that the guarantee certifies most of the reference ordering while leaving the ambiguous comparisons free to move.

Finally, the ablation on Breast SIS-20 supports the full weighting, since the $Q + B + C$ construction performs best and $B + C$ retains part of the gain, whereas $Q + B$ and $Q + C$ fall below the Reference (Section D.4).

5 Discussion

LLM guidance remains valuable within these limits, since the composite score vector $\mu(d)$ is estimable to any precision and its content adds value on three datasets. What the record may leave open is its interpretation (Proposition 1). A downstream guarantee should therefore depend only on what the record certifies. Selective Correction is one architecture that meets this requirement, with a protected-region guarantee that bounds how far the guidance can move the reference scores for every admissible judgment (Theorem 1).

Lastly, the guarantee certifies stability rather than correctness and preserves an erroneous reference ordering as faithfully as a correct one, which resampling-based targeting deprioritizes because the data order it consistently. The limitation thus lies with the targeting rule, since the theorem holds for any query set and would certify a rule that also reaches stable errors with a weighting that does not vanish on them, a natural direction for future work. More broadly, the precision of an elicited composite, the identification of its interpretation and the validity of the downstream decision are distinct properties of LLM-derived guidance, none of which delivers the others.

Ethics statement. The authors have read and adhere to the ICLR Code of Ethics. Six of the seven datasets are publicly available through GEO, OpenML and the JKP factor database and are used under their terms of access. The osteoporosis cohort is a proprietary hospital dataset that is not publicly available because of institutional data governance restrictions. The method bounds how far external LLM guidance can move a data-driven decision. It does not remove biases that LLM judgments may carry from their training data, which is one reason to report the protected region with every output.

Reproducibility statement. Section 3 and Algorithm 1 fully specify Selective Correction. Sections A and C provide complete proofs of Proposition 1, Corollary 1 and Theorem 1, together with all conditions used. Section B details the simulation designs, budgets, numbers of replicates and the frozen specification underlying Figure 2. Section D documents the source of each dataset, preprocessing and evidence construction, the elicitation models, the resampling behind query targeting (Section D.7) and the protocol for selecting the correction strength (Tables 8 and 9). Code is available at `https://anonymous.4open.science/r/Identification-Limits-and-Protected-Corrections-for-LLM-Judgments/`.

The use of large language models. LLMs are the object of study in this paper. DeepSeek V4.1 Flash and Qwen models produce the pairwise judgments, while a GPT-assisted workflow organizes the external evidence into packets (Section D.1). In preparing the paper, LLM-based assistants were also used for (i) editing and polishing author-written text, (ii) searching the literature, with every reference verified by the authors, (iii) polishing and checking proofs of the theoretical results, which the authors verified, and (iv) writing and reviewing simulation and real-data code. The authors take full responsibility for all content of the paper.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Jordan G. Bryan, Hongqian Niu, and Didong Li. Incorporating LLM-derived information into hypothesis testing for genomics applications. bioRxiv preprint, 2025.

Donald T. Campbell and Donald W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959.

Alexander Capstick, Rahul Krishnan, and Payam Barnaghi. AutoElicit: Using large language models for expert prior elicitation in predictive modelling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 6746–6777. PMLR, 2025.

Jiangyu Chen and Ban Yi. Evidence-gated LLM priors for multi-objective Bayesian optimization. *arXiv preprint arXiv:2606.01730*, 2026.

Xiaohong Chen, Ashesh Rambachan, and Elie Tamer. Partial identification from LLM prompts. *arXiv preprint arXiv:2606.15031*, 2026.

Yuxin Chen and Changho Suh. Spectral MLE: Top-K rank aggregation from pairwise comparisons. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 371–380, 2015.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024.

Spyros Darmanis, Steven A. Sloan, Derek Croote, Marco Mignardi, Sophia Chernikova, Peyman Samghababi, Ye Zhang, Norma Neff, Mark Kowarsky, Christine Caneda, Gordon Li, Steven D. Chang, Ian David Connolly, Yingmei Li, Ben A. Barres, Melanie Hayden Gephart, and Stephen R. Quake. Single-cell RNA-Seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Reports*, 21(5):1399–1410, 2017.

Roger R. Davidson and Robert J. Beaver. On extending the Bradley-Terry model to incorporate within-pair order effects. *Biometrics*, 33(4):693–702, 1977.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68589–68601, 2023. doi: 10.52202/075280-3000. arXiv:2306.04746.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Jean Feng, Avni Kothari, Luke Zier, Chandan Singh, and Yan Shuo Tan. Bayesian concept bottleneck models with LLM priors. In *Advances in Neural Information Processing Systems*, volume 38, pp. 104188–104219, 2025. doi: 10.52202/085713-3141. arXiv:2410.15555.

Lester R. Ford, Jr. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8P2):28–33, 1957. doi: 10.1080/00029890.1957.11989117.

Daniel P. Jeong, Zachary C. Lipton, and Pradeep Ravikumar. LLM-Select: Feature selection with large language models. *Transactions on Machine Learning Research*, 2025. arXiv:2407.02694.

Agustinus Kristiadi, Felix Strieth-Kalthoff, Marta Skreta, Pascal Poupart, Alán Aspuru-Guzik, and Geoff Pleiss. A sober look at LLMs for material discovery: Are they actually good for Bayesian optimization over molecules? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25603–25622. PMLR, 2024. arXiv:2402.05015.

Joseph Lee, Shu Yang, Jae Young Baik, Xiaoxi Liu, Zhen Tan, Dawei Li, Zixuan Wen, Bojian Hou, Duy Duong-Tran, Tianlong Chen, and Li Shen. Knowledge-driven feature selection and engineering for genotype data with large language models. In *AMIA Informatics Summit*, 2025a.

Youngbin Lee, Yejin Kim, Juhyeong Kim, Suin Kim, and Yongjae Lee. LLM-enhanced Black–Litterman portfolio optimization. *arXiv preprint arXiv:2504.14345*, 2025b. CIKM 2025 Workshop on Financial AI.

Dawei Li, Zhen Tan, and Huan Liu. Exploring large language models for feature selection: A data-centric perspective. *ACM SIGKDD Explorations Newsletter*, 26(2):44–53, 2025. doi: 10.1145/3715073.3715077.

Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance Bayesian optimization. In *International Conference on Learning Representations*, 2024. arXiv:2402.03921.

Adian Liusie, Potsawee Manakul, and Mark Gales. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 139–151, 2024a.

Adian Liusie, Vatsal Raina, Yassir Fathullah, and Mark Gales. Efficient LLM comparative assessment: A product of experts framework for pairwise comparisons. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6835–6855, 2024b.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, 2003.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Thomas J. Rothenberg. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design, or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. arXiv:2310.11324.

Rajen D. Shah and Richard J. Samworth. Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75 (1):55–80, 2013.

Caleb Skinner, Yihan Guo, and Meng Li. LLM sparsity prior for robust feature selection. *arXiv preprint arXiv:2605.23102*, 2026.

A. Springall. Response surface fitting using a generalization of the Bradley-Terry paired comparison model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 22(1):59–68, 1973.

L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. arXiv:2203.11171.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. Investigating non-transitivity in LLM-as-a-judge. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69583–69612, 2025.

Tom Tianze Yang, Tim Tianyi Yang, Shaoshan Liu, Fuyuan Lyu, and Xue Liu. Ice-search: A language model-driven feature selection approach. *arXiv preprint arXiv:2402.18609*, 2024.

Erica Zhang, Ryunosuke Goto, Naomi Sagan, Jurik Mutter, Nick Phillips, Ash Alizadeh, Kangwook Lee, Jose Blanchet, Mert Pilanci, and Robert Tibshirani. LLM-Lasso: A robust framework for domain-informed feature selection and regularization. *arXiv preprint arXiv:2502.10648*, 2025.

Erica Zhang, Naomi Sagan, Danny Tse, Fangzhao Zhang, Mert Pilanci, and Jose Blanchet. Learning when to trust LLM priors: A validated framework for semantic prior integration. *arXiv preprint arXiv:2601.21410*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. Datasets and Benchmarks Track; arXiv:2306.05685.

Appendix

This appendix collects the proofs of the theoretical results, the full identification simulation and the additional downstream experiments. Its content is organized as follows.

Appendix		Content
A	A.1	The multi-source comparison model and the evidence-free preference.
	A.2	Conditions invoked by the results.
	A.3	Trajectory factorization for adaptive protocols (Lemma 1).
	A.4	Cancellation of configuration-level intercepts (Proposition 2).
	A.5	Formal version and proof of Proposition 1 (Proposition 3).
	A.6	Dimension of the unresolved set and the identification criterion.
	A.7	Proof of Corollary 1.
B		Designs and full results of the identification simulation.
C	C.1	Proof of Theorem 1 and its validity for any query set.
	C.2	Top-k certificate (Corollary 2).
	C.3	Envelope registrable before querying (Corollary 3).
	C.4	Certificate for approximate solvers (Proposition 4).
	C.5	The role of nonnegative weights.
	C.6	Tightness of the protected region and sharpness of both thresholds.
D	D.1	Datasets, preprocessing and evidence construction.
	D.2	Comparison with LLM-based feature-selection methods.
	D.3	Cross-model replication with Qwen.
	D.4	Component ablation of the correction weight.
	D.5	Direct LLM-only selection.
	D.6	Correction-strength selection and training, validation and test data.
	D.7	Resampling for query targeting and the actionability threshold.

A Proof of Proposition 1

This appendix states the model behind Proposition 1 and proves it together with Corollary 1. Section A.1 states the model and Section A.2 collects the conditions the results invoke. Section A.3 establishes the trajectory factorization on which both propositions rest. Section A.4 shows that configuration-level intercepts cancel, Section A.5 states and proves the nonidentification result and Section A.6 counts what a design leaves unresolved. Finally, Section A.7 proves Corollary 1, together with the lemma that makes its top-k clause nonvacuous.

A.1 The general multi-source model

We consider p items with r observed attributes collected in a centered matrix $Z \in \mathbb{R}^{p\times r}$ of full column rank and S evidence sources. A prompting regime is summarized by its information configuration $d \in \{0,1\}^S$, the indicator of which sources are presented. Under configuration d, item j carries the score $\mu_j(d)$, where

$$\mu(d) \;:=\; q + Z\, g(d), \qquad g(0) = 0, \tag{6}$$

is the composite score vector of configuration d. Elicitation is by pairwise comparison. When the pair (j,l) is queried under configuration d with rendering θ, which comprises the wording, the decoding parameters and a randomized presentation order $o(\theta) \in \{-1,+1\}$, the response is binary with

$$\mathbb{P}\big(Y = 1 \mid j,l,d,\theta\big) \;=\; G\big(\mu_j(d) - \mu_l(d) + \kappa\, o(\theta)\big), \tag{7}$$

where G is known, continuous and strictly increasing.

Each ingredient of equation 7 is standard. Paired comparison through a latent score is the model of Bradley & Terry (1952), the additive within-pair order term is its order-effect extension (Davidson & Beaver, 1977) and item covariates entering the score linearly give its covariate form (Springall, 1973). The same model is fitted at scale to pairwise model comparisons in LLM evaluation (Chiang et al., 2024). The new element here is the configuration index d. At the level of a batch of queries,

Assumption 1 allows responses within a batch to depend on one another arbitrarily and asks only that the batch law depend on the parameters through the score contrasts, the margins of Section 2.1, at the queried actions. Conditionally independent responses are therefore a special case and are not assumed. The logistic form of G is likewise a working choice that serves estimation only, since every identification statement below is link-robust.

The normalization $g(0) = 0$ fixes the target

$$q \;=\; \mu(0), \tag{8}$$

which we call the evidence-free preference, namely the composite the model returns under the empty configuration with no evidence presented. It is the vector version of b_e in equation 1. Importantly, q is the estimand of the comparison channel at the empty configuration and is distinct from an item's true quality, true relevance or source-free ground truth. Identifying q thus determines what the model measures when nothing is shown and leaves open whether that measurement is correct for a task. Finally, the structure $g(d) = \Delta d$, whose columns are the source sensitivities, is the form on which the aliasing map below acts.

A.2 Conditions invoked

The modeling commitment. The results invoke batch contrast sufficiency (Assumption 1) together with the monotone comparison model, under which the batch response law depends on the parameters only through the score contrasts at the queried actions and each marginal is equation 7 with G known, continuous and strictly increasing. The nonidentification result is link-robust and uses only the first half.

Design conditions. The attributes Z are centered and of full column rank, a property of the constructed item set that is checked before querying and places no restriction on the LLM.

Dependence. The conditionally independent model, whose batch law is the product of its marginals, is a special case of Assumption 1. It is used only to state Lemma 1 in its simplest form, while the general batch form of the assumption replaces it throughout, so no result below needs conditional independence.

Conventions. We adopt three normalizations. First, $\mathbf{1}^\top q = 0$, which fixes the free additive constant that Proposition 2 shows the record cannot see. Second, Z is centered, which costs nothing, since the $\mathbf{1}$-component of any attribute contributes only item-constant intercepts and these cancel by the same proposition. Third, win probabilities are interior, which is automatic at any fixed parameter.

A.3 Trajectory factorization

Both propositions reduce a statement about the law of an adaptively gathered record to a statement about per-query response laws, a reduction that this subsection supplies. All action, response and record spaces are standard Borel. An action is $A_t = (j_t, l_t, \theta_t, d_t)$, whose configuration component takes values in $\mathcal{D}$ in Proposition 3. For batched elicitation, an action specifies the whole selected batch and a response its joint response vector, which keeps the spaces standard Borel, while a choice made between responses is represented as a further action step.

Fix a parameter point P and a protocol $\pi \in \Pi$ given by the kernels $\pi_t(dA_t \mid h_{t-1})$. Let $K_P(dy \mid A)$ denote the response kernel, which under Assumption 1 depends on P only through the queried contrasts and, in the conditionally independent case, is history-free. We write $h_t = (h_{t-1}, A_t, y_t)$ for the history and ν_P for the record law on $\Omega = \prod_{t\ge1}(\mathbb{A} \times \mathbb{Y})$ given by the Ionescu–Tulcea extension of $\pi_1, K_P, \pi_2, K_P, \ldots$.

Lemma 1 (Trajectory factorization, multi-configuration form)**.** *In the setting above, the following hold.*

(i) For every T and every bounded measurable f,

$$\begin{aligned}\mathbb{E}_{\nu_P} f(H_T) = \int \pi_1(dA_1 \mid h_0) \int K_P(dy_1 \mid A_1) \cdots \\ \int \pi_T(dA_T \mid h_{T-1}) \int K_P(dy_T \mid A_T)\, f(h_T).\end{aligned} \tag{9}$$

(ii) If two parameter points P and P' induce the same response kernel at every action in the protocol's action space, which under Assumption 1 holds whenever all queried contrasts agree, then $\nu_P = \nu_{P'}$.

(iii) Under the hypothesis of (ii), the stopped-record laws under P and P' coincide for every adapted stopping time τ with values in $\{1, 2, \dots\} \cup \{\infty\}$.

Proof of Lemma 1. Part (i) is the Ionescu–Tulcea construction written out for the first T coordinates. By (i), the finite-dimensional laws of the record are determined by the parameter-free protocol kernels and the collection $\{K_P(\cdot \mid A)\}_A$. Under the hypothesis of (ii), equation 9 therefore coincides for all T and f. Since the finite-coordinate measurable cylinders form a π-system generating the product σ-algebra, it follows that $\nu_P = \nu_{P'}$, which proves (ii).

For (iii), we define the stopped-record space and the stopping map

$$\mathbb{S} := \bigsqcup_{t<\infty} (\mathbb{A} \times \mathbb{Y})^t \sqcup \Omega, \qquad \Phi(\omega) := \begin{cases} \omega_{1:\tau(\omega)} & \text{on } \{\tau < \infty\}, \\ \omega & \text{on } \{\tau = \infty\}. \end{cases}$$

The map Φ is measurable and parameter-free, so the stopped law is $\Phi_{\#}\nu_P$. Equality of trajectory laws therefore transfers to stopped laws. Randomized protocols are already covered, since their randomization lives inside the kernels π_t. This completes the proof. □

The factorization extends to dependent responses, since under the general batch form of Assumption 1 the kernel $K_P(dy_t \mid A_t)$ in equation 9 is replaced by the history-dependent kernel $K_P(dy_t \mid A_t, h_{t-1})$. For every realized history h_{t-1} and every batch action the protocol can select after it, the assumption ensures that two parameter points inducing identical realized contrasts at the actions that batch queries also induce the same conditional response-batch law. This is exactly the pathwise equality the argument requires. Since the Ionescu–Tulcea extension applies verbatim to history-dependent kernels, Lemma 1 holds in this general form, so within-batch dependence costs nothing in Propositions 2 and 3.

Lemma 1 delivers law equality once the per-action response laws agree at every action the protocol can take, so with varying configurations, contrast equality is needed at every queried configuration. The two propositions below secure it in different ways, since the contrast equation 10 of Proposition 2 is free of the intercept at every configuration, whereas Lemma 2 gives the contrast equality that Proposition 3 needs at every configuration in $\mathcal{D}$.

A.4 Cancellation of configuration-level intercepts

We first show that an intercept shared by all items within a configuration is invisible to every protocol.

Proposition 2 (Cancellation of configuration-level intercepts)**.** *Suppose each item's score carries an arbitrary configuration-level intercept, $\eta_j(d) = \beta(d) + \mu_j(d)$. Then, for every configuration set, every protocol $\pi \in \Pi$ and every parameter point, the following hold.*

(i) The record law does not depend on $\beta(\cdot)$.

(ii) The identified set for $\beta(\cdot)$ is all of $\mathbb{R}^{|\mathcal{D}|}$.

(iii) The identified functional of the scores is at most the contrast vector.

The argument uses only Assumption 1, allows arbitrary across-query dependence and requires no link form or monotonicity.

Proof of Proposition 2. The proof proceeds in three parts, one for each claim.

Part 1: Invariance of the record law. Fix a pair (j, l), a configuration d and a rendering θ. Under the extended score,

$$\eta_j(d) - \eta_l(d) = \mu_j(d) - \mu_l(d), \tag{10}$$

and the right-hand side does not involve $\beta(\cdot)$. By Assumption 1, the response law of the query is a function of this contrast and θ, so by equation 10 it is the same for every choice of $\beta(\cdot)$. Since

equation 10 is free of β at every configuration, the response kernels of any two choices of $\beta(\cdot)$ agree at every action the protocol can take. Lemma 1, in either form, then shows that the law of the entire record under any $\pi \in \Pi$ and any configuration set is a functional of the per-query response laws and hence is constant in $\beta(\cdot)$.

Part 2: Identified set of the intercepts. Fix configuration laws generated by (β_0, q, Δ). For any candidate $\beta(\cdot)$, the parameter point (β, q, Δ) reproduces the identical record law by Part 1, so every $\beta(\cdot)$ lies in the identified set.

Part 3: Identified functional. By Parts 1 and 2, the record leaves the configuration-level intercepts free, so the identified functional of the scores is at most the contrast vector. This completes the proof. □

A.5 Nonidentification under a restricted configuration design

The following proposition is the formal version of Proposition 1. Fix a nonempty configuration set $\mathcal{D} \subseteq \{0,1\}^S$ with $0 \notin \mathrm{aff}(\mathcal{D})$ and let $c_0 \in \mathbb{R}^S$ satisfy $c_0^\top d = 1$ at every $d \in \mathcal{D}$. Part 1 of the proof shows that such a c_0 exists exactly when $0 \notin \mathrm{aff}(\mathcal{D})$. The aliasing map

$$T_\gamma \colon (q, \Delta) \;\longmapsto\; (q + Z\gamma,\; \Delta - \gamma c_0^\top), \qquad \gamma \in \mathbb{R}^r,$$

moves the evidence-free preference. We write $\Pi(\mathcal{D})$ for the protocols of Π whose queried configurations lie in $\mathcal{D}$ at every history. These protocols may select pairs, configurations and renderings adaptively, aggregate arbitrarily and stop at any adapted stopping time.

Proposition 3 (Nonidentification under a restricted configuration design)**.** *Assume Assumption 1 and let Z be centered with full column rank. Then the following hold uniformly over every protocol $\pi \in \Pi(\mathcal{D})$, budget and stopping rule.*

(i) (Observational equivalence.) The stopped-record laws under (q, Δ) and $T_\gamma(q, \Delta)$ are identical, $\mathrm{TV} = 0$, while q moves by $\|Z\gamma\|$ and the normalization $\mathbf{1}^\top q = 0$ is preserved.

(ii) (Fisher-null direction.) Along $\epsilon \mapsto T_{\epsilon\gamma}(q, \Delta)$ the record likelihood ratio is a.s. constant, so $(Z\gamma, -\gamma c_0^\top)$ lies in the null space of any record Fisher information, wherever that information is defined.

(iii) (No consistent estimator.) For any estimator $\hat{q}$ measurable in the record and any budget,

$$\max\Big\{\mathbb{E}_{(q,\Delta)}\|\hat{q} - q\|,\; \mathbb{E}_{T_\gamma(q,\Delta)}\|\hat{q} - (q + Z\gamma)\|\Big\} \;\geq\; \tfrac{1}{2}\|Z\gamma\|, \tag{11}$$

and the minimax risk over any class Θ_B containing both (q, Δ) and $T_\gamma(q, \Delta)$ is at least $\sup\{\|Z\gamma\|/2 : T_\gamma(q, \Delta) \in \Theta_B\}$.

No link form, monotonicity or graph condition is required. Assumption 1 cannot simply be dropped, since a batch law that depended on q directly, for instance through $\mu(0)$, would separate the two parameter points.

The proof rests on the following identity, which shows that the aliasing map leaves every queried contrast unchanged.

Lemma 2 (Invariance of contrasts under aliasing)**.** *For every pair (j, l) and every $d \in \mathcal{D}$,*

$$\big[(q_j + z_j^\top\gamma) - (q_l + z_l^\top\gamma)\big] + (z_j - z_l)^\top(\Delta - \gamma c_0^\top)d = (q_j - q_l) + (z_j - z_l)^\top\Delta d. \tag{12}$$

Moreover, $\mathbf{1}^\top q = 0$ implies $\mathbf{1}^\top(q + Z\gamma) = 0$, while $Z\gamma \neq 0$ for every $\gamma \neq 0$.

Proof of Lemma 2. Expanding the left-hand side of equation 12 gives

$$\begin{aligned}
&\big[(q_j + z_j^\top\gamma) - (q_l + z_l^\top\gamma)\big] + (z_j - z_l)^\top(\Delta - \gamma c_0^\top)d \\
&\quad = (q_j - q_l) + (z_j - z_l)^\top\Delta d + (z_j - z_l)^\top\gamma\,\big(1 - c_0^\top d\big) \\
&\quad = (q_j - q_l) + (z_j - z_l)^\top\Delta d,
\end{aligned}$$

where the last step uses $c_0^\top d = 1$. The identity therefore holds at every configuration in $\mathcal{D}$. Since Z is centered, $\mathbf{1}^\top Z = 0$, so $\mathbf{1}^\top(q + Z\gamma) = \mathbf{1}^\top q = 0$. Finally, Z has full column rank and is therefore injective, so $Z\gamma \neq 0$ for $\gamma \neq 0$. □

Proof of Proposition 3. The proof proceeds in four parts. First, we characterize the design condition $0 \notin \mathrm{aff}(\mathcal{D})$ through the functional c_0. Second, we prove the observational equivalence (i). Third, we derive the Fisher-null direction (ii). Fourth, we derive the two-point bound (iii).

Part 1: Characterization of the design condition. Fix $d^1 \in \mathcal{D}$ and write $\mathrm{aff}(\mathcal{D}) = d^1 + V$ with $V = \mathrm{span}\{d - d^1 : d \in \mathcal{D}\}$. The origin lies outside the affine hull exactly when $d^1 \notin V$. In that case a linear functional c_0 can annihilate V and take the value 1 at d^1, so that $c_0^\top d = 1$ for every $d \in \mathcal{D}$. Conversely, such a functional equals 1 on the whole affine hull, which therefore excludes 0.

Part 2: Observational equivalence. By Lemma 2, the aliasing map leaves every queried contrast unmoved. By Lemma 1 in its general form, the stopped record law is a functional of those contrasts for every $\pi \in \Pi(\mathcal{D})$. The two record laws therefore coincide exactly. The same lemma shows that q moves by $Z\gamma$, of norm $\|Z\gamma\|$, and that the normalization $\mathbf{1}^\top q = 0$ is preserved.

Part 3: Fisher-null direction. By Part 2 applied to $\epsilon\gamma$, the record law under $T_{\epsilon\gamma}(q, \Delta)$ equals that under (q, Δ) for every ϵ. On any dominating measure, the likelihood ratio is therefore a.s. constant in ϵ. The score in the direction $(Z\gamma, -\gamma c_0^\top)$ thus vanishes a.s. and any record Fisher information annihilates it.

Part 4: Two-point bound. Write $\mathbb{P}^\star$ for the common record law from Part 2 and $\tilde{q} := q + Z\gamma$. For any estimator, the triangle inequality inside the expectation, under the single measure $\mathbb{P}^\star$, gives

$$\mathbb{E}_{\mathbb{P}^\star}\|\hat{q} - q\| + \mathbb{E}_{\mathbb{P}^\star}\|\hat{q} - \tilde{q}\| \;\geq\; \|q - \tilde{q}\| \;=\; \|Z\gamma\|. \tag{13}$$

The larger of the two terms is therefore at least $\|Z\gamma\|/2$, which is equation 11. The minimax statement follows by taking the supremum over admissible γ. This completes the proof. □

The argument requires little of the comparison model. For a fixed parameter-independent protocol, the record law is determined by the conditional response kernels at the actions and histories that the protocol can select. Equality of those kernels at two parameter points therefore gives equality of the stopped-record laws, so the identified set of any parameter contains every parameter that induces the same kernels at the selectable actions. Under conditionally independent, history-free sampling, these kernels reduce to the configuration kernels $\{K_d\}_{d\in\mathcal{D}}$. This reduction uses the independence as well as the marginals, since two configurations can carry the same marginal response law while a jointly queried pair is comonotone at one parameter point and antitone at the other. The factorization also runs in one direction, from equal kernels to observational equivalence, so a protocol that never selects an action where the kernels of two parameter points differ cannot distinguish them.

At the other extreme, the design $\mathcal{D} = \{0\}$ consists of the empty configuration alone and therefore falls outside the hypothesis of Proposition 3. Under the full model of Section A.1, its composite is $\mu(0) = q$, so the known strictly monotone link recovers the contrasts of q, which do not involve Δ. This design therefore identifies q as its composite and leaves Δ unrestricted.

A.6 Dimension of the unresolved set

Under stronger conditions, the parameter changes a design leaves unresolved can be counted exactly. Suppose in addition that every queried composite and the shared order effect are identified, which requires the link and sufficient comparison coverage beyond the hypotheses of Proposition 3. Fix $d^1 \in \mathcal{D}$ and let $V = \mathrm{span}\{d - d^1 : d \in \mathcal{D}\}$. The parameter changes such a design cannot see are then exactly

$$\big\{(-Z\Gamma d^1,\ \Gamma)\ :\ \Gamma \in \mathbb{R}^{r\times S},\ \Gamma|_V = 0\big\}, \qquad \text{of dimension} \qquad r\,(S - \dim V). \tag{14}$$

Indeed, a change (u, Γ) leaves every composite fixed exactly when

$$u + Z\Gamma d = 0 \qquad \text{for every } d \in \mathcal{D}.$$

Differencing two configurations and using the injectivity of Z gives $\Gamma|_V = 0$. The remaining equation then fixes $u = -Z\Gamma d^1$. Such a change moves the evidence-free preference by $-Z\Gamma d^1$, which vanishes for every admissible Γ precisely when $d^1 \in V$, that is, when $0 \in \mathrm{aff}(\mathcal{D})$. Under these added hypotheses, the count thus supplies the converse of Proposition 3. The evidence-free preference q is identified if and only if the empty configuration lies within affine reach of the queried configurations, while the whole of Δ is identified if and only if $\dim V = S$.

A.7 Proof of Corollary 1

Proof of Corollary 1. We treat the two parts in turn, writing $\mathbb{P}^\star$ for the common record law of Proposition 1 and $\tilde{q} := q + Z\gamma$.

Part 1: Minimax bound. Part (i) restates Proposition 3(iii) at the level of the main text. For $\hat{q}$ measurable in the record, both expectations are taken under $\mathbb{P}^\star$, so equation 13 gives

$$\mathbb{E}_{\mathbb{P}^\star}\|\hat{q} - q\| + \mathbb{E}_{\mathbb{P}^\star}\|\hat{q} - \tilde{q}\| \;\geq\; \|Z\gamma\|,$$

and the larger term is at least $\|Z\gamma\|/2$.

Part 2: Nonidentification of decisions. Let $\hat{\psi}$ be measurable in the record. By Proposition 1, its law is the same under the two parameter points, so both probabilities are evaluated under the single measure $\mathbb{P}^\star$. When $\psi(q) \neq \psi(\tilde{q})$, the events $\{\hat{\psi} = \psi(q)\}$ and $\{\hat{\psi} = \psi(\tilde{q})\}$ are disjoint, so their probabilities sum to at most 1 and the smaller is at most $1/2$. This completes the proof. □

Two remarks make precise what Corollary 1 asserts. First, the bound in (i) concerns record-measurable estimators only. Without this restriction the display fails, since an estimator allowed to read the parameter attains zero risk at each parameter point separately. Second, (ii) bounds the probability of naming the right value, as distinct from a useful one, so it shows that the record cannot distinguish the two parameter points and makes no claim that either value is preferable.

The hypothesis of part (ii) is satisfiable whenever $k < p$, as the following lemma shows.

Lemma 3 (Satisfiability of the top-k hypothesis)**.** *Let Z be centered with full column rank and let $1 \leq k < p$. For every $q \in \mathbb{R}^p$ there exists $\gamma \in \mathbb{R}^r$ with $S_k(q) \neq S_k(q + Z\gamma)$.*

Proof of Lemma 3. Fix any $\gamma_0 \neq 0$ and put $v := Z\gamma_0$. Injectivity of Z gives $v \neq 0$ and centering gives $\mathbf{1}^\top v = 0$. A nonzero vector summing to zero is nonconstant. Let A^+ and A^- be the top-k sets under the lexicographic orders of (v_i, q_i) and of $(-v_i, q_i)$, with the tie rule of S_k breaking any remaining ties. They differ, since equality would force every entry of v in the common set to be both at least and at most every entry outside it, which makes v constant.

Consider $q + tv$ for $t \in \mathbb{R}$. For $|t|$ large, the ordering of $q + tv$ is governed by v. In particular, if $v_i > v_j$, then

$$(q + tv)_i - (q + tv)_j \;=\; (q_i - q_j) + t\,(v_i - v_j)$$

is positive for all sufficiently large t and negative for all sufficiently negative t, whereas for $v_i = v_j$ it equals $q_i - q_j$. It follows that $S_k(q + tv) = A^+$ for all large enough t and $S_k(q + tv) = A^-$ for all sufficiently negative t. Since $A^+ \neq A^-$, at least one of them differs from $S_k(q)$, so the corresponding $\gamma = t\gamma_0$ proves the claim. □

B The identification simulation in full

Design. This appendix details the simulation summarized in Section 4.1 and Figure 2. We plant one (q, Δ, Z) and simulate $44{,}800$ elicitation records under six query designs, with 500 replicates per design and budget. Every prediction was registered in a frozen specification beforehand and a separate reader bootstrapped every interval over the frozen records. Two designs lose q, namely SAT, which shows every source at every query, and OAT, the one-source-at-a-time ablation, whose two configurations satisfy $c_0^\top d = 1$ with $c_0 = (1, 1)$. Since Proposition 1 applies to OAT directly, varying the configuration is not by itself a remedy. Among the four designs that identify q, MIN is the empty configuration plus one configuration per source, CROSS adds the configuration with both sources, BASEFREE identifies q without the empty configuration and DEF is deficient in one source direction.

Exactness of the equivalence. Along the predicted aliasing directions, the Fisher information per unit-length direction is at most 3.5×10^{-29}, the single-query total variation is at most 3.4×10^{-17} with no response cell differing, while the log-likelihood deviation of 4.5×10^{-13} is one unit in the last place of the log-likelihood itself. Under DEF all three quantities are exactly 0, since its null direction shifts a source the design never queries. The informative comparison is therefore between designs on the same directions, where a design that identifies q carries information of order 10^3 and

one that does not carries 10^{-28} (Figure 2A, at 96 queries per comparison and configuration, where information is linear in the budget). Each blind design is blind only along its own aliasing direction and fully informed along the other's, so blindness is a joint property of a design and a direction. For SAT the null space is six-dimensional and Proposition 1 names three of those dimensions, while the other three move the sources against each other and leave q unchanged.

Precision without identification. Splitting the error along the orbit separates precision from identification cleanly. The component along the orbit has fitted slope -0.0003 under SAT (interval $[-0.0008, 0.0002]$) and stays flat over a $16\times$ budget increase, while its slope under OAT is -0.0072. Under the four designs that identify q, the same component falls at slopes between -0.503 and -0.516, every interval inside the registered band $[-0.65, -0.35]$, so a single quantity has slope zero under the blind designs and slope $-\frac{1}{2}$ under the identifying ones. The total errors of the two blind designs inherit this contrast more weakly, with slopes -0.160 and -0.184 against -0.502 to -0.518, because a total mixes a component that converges with one that does not. Decomposing OAT's error the other way gives -0.518 orthogonal to the aliasing direction, -0.524 on the centered aliasing component and -0.510 on the observed quantities, all inside the band.

Convergence to an equivalent truth. The flat component reflects convergence to a different target. Every point of the orbit induces the same record, so the same responses give the same estimate and only the target moves. Re-scoring the frozen records against $q + Z\gamma^\dagger$ is therefore exact arithmetic and requires no second simulation. Scored this way, the SAT error falls from 0.803 to 0.194 at slope -0.514 (interval $[-0.526, -0.502]$), while the error against q plateaus at 0.654 (Figure 2B), with the corresponding slope -0.518 for OAT. The estimator thus converges at the full rate to a target at distance 0.622 from the one that generated the data, which accounts for the plateau.

Since the record cannot distinguish the points of the orbit, the error against $q + cZ\gamma^\dagger$ is minimized at $c = 1$ at every budget and the curve is symmetric about this point (Figure 2C). The pair $c \in \{0, 2\}$ is the two-point problem of Corollary 1, with errors 0.6536 and 0.6540 against a bound of 0.6219, so the corollary is close to tight here. Finally, a calibration regression over six planted offsets from 0 to 2.25 gives slope 0.9985 (standard error 0.00057), while the observed offset is 0.0005 where the theory puts it at zero. The simulation thus finds the mechanism of Proposition 1 exactly where the theory places it. The proposition itself covers every admissible configuration set, every γ and every protocol in $\Pi(\mathcal{D})$, which no single planted parameter point can exhaust, while Sections 4.2 and 4.3 address whether the composite is useful downstream.

C Proof of Theorem 1

This appendix proves Theorem 1 and develops the certificates that the main text reports. Section C.1 proves the theorem and Sections C.2 to C.4 derive the top-k certificate, the envelope that can be registered before querying and the version for approximate solvers. Section C.5 records what the argument loses under signed weights and Section C.6 evaluates its tightness numerically.

Throughout, E is a finite set of oriented pairs with $u_e \neq v_e$ and at most one edge per unordered pair, whose signed incidence matrix $N \in \{-1, 0, 1\}^{|E|\times p}$ has $N_{e,u_e} = +1$ and $N_{e,v_e} = -1$. We write $w \in \mathbb{R}^{|E|}_{\geq 0}$ for the correction weights, $\mathbf{p} \in [0,1]^{|E|}$ for the judgments and $\rho_j := \sum_{e\ni j} w_e$, where the sum runs over the edges incident to j at either endpoint. Cross-entropy is $\mathrm{CE}(y, \hat{y}) := -y\log\hat{y} - (1-y)\log(1-\hat{y})$, with the target first. Since $\log\sigma(t) = t - \log(1+e^t)$ and $\log(1-\sigma(t)) = -\log(1+e^t)$, the loss on a single edge and its derivatives are

$$\begin{aligned} \ell_y(t) \;&:=\; \mathrm{CE}\big(y, \sigma(t)\big) \;=\; \log(1+e^t) - yt, \\ \ell_y'(t) \;&=\; \sigma(t) - y, \qquad \ell_y''(t) \;=\; \sigma(t)\big(1-\sigma(t)\big) \;>\; 0. \end{aligned} \tag{15}$$

C.1 Existence, uniqueness and the coordinatewise bound

Proof of Theorem 1. The proof proceeds in three parts. First, we show that F has a unique global minimizer. Second, we bound each coordinate of its displacement from a. Third, we deduce the order statement.

Part 1: Existence and uniqueness. Each summand $w_e\ell_{p_e}(\cdot)$ of equation 5 is convex, because $w_e \geq 0$ and ℓ_{p_e} is convex by equation 15. The objective F is therefore the sum of the 1-strongly

convex quadratic $\frac{1}{2}\|s-a\|^2$ and a convex function, so it is itself 1-strongly convex. Consequently, F is coercive and attains a unique global minimum $s^\star$.

Part 2: Coordinatewise bound. By equation 15, the stationarity condition at $s^\star$ reads

$$s^\star - a \;=\; -N^\top \big[w \odot \big(\sigma(Ns^\star) - \mathbf{p}\big)\big], \tag{16}$$

whose jth coordinate is a signed sum, over the edges incident to j, of the terms $w_e(\sigma(\cdot)-p_e)$. Since $p_e \in [0,1]$ and σ takes values in $(0,1)$, each factor satisfies $|\sigma(\cdot)-p_e|<1$, so that

$$\begin{aligned} \big|s_j^\star - a_j\big| \;&\le\; \sum_{e \ni j} w_e \,\big|\sigma(\cdot) - p_e\big| \\ &\le\; \sum_{e \ni j} w_e \;=\; \rho_j. \end{aligned} \tag{17}$$

Part 3: Order statement. If $a_i - a_j > \rho_i + \rho_j$, then equation 17 gives

$$s_i^\star - s_j^\star \;\ge\; (a_i - \rho_i) - (a_j + \rho_j) \;>\; 0.$$

This completes the proof. □

Since $|\sigma - p_e| < 1$ strictly, the bound equation 17 is attained only in the limit, so it is conservative and never vacuous for finite weights. Three features of the argument are worth isolating. First, it uses no property of p_e beyond membership in $[0,1]$, so the bound is uniform over all possible answers, including adversarial ones. Second, it uses no property of how E or w were chosen, so it holds without any targeting rule, entropy weight or evidence gate. The theorem thus certifies the correction architecture itself, including corrections that query globally or weight uniformly. For the same reason, it provides no evidence for the targeting rule of Section 3.1 or for the evidence gate, whose value Section 4.2 assesses empirically. Third, an item incident to no queried edge has $\rho_j = 0$, so equation 17 gives $s_j^\star = a_j$ exactly, which makes exact fallback a special case of the theorem.

C.2 Top-k preservation

The coordinatewise bound yields a certificate for the selected set itself.

Corollary 2 (Top-k preservation)**.** *Let $S_k(a)$ be the reference top-k set. If*

$$\min_{i \in S_k(a)} \big(a_i - \rho_i\big) \;>\; \max_{j \notin S_k(a)} \big(a_j + \rho_j\big), \tag{18}$$

then $S_k(a)$ and $S_k(s^\star)$ are uniquely defined and $S_k(s^\star) = S_k(a)$.

Proof of Corollary 2. Since $\rho \ge 0$, condition equation 18 forces $a_i > a_j$ strictly for every $i \in S_k(a)$ and every $j \notin S_k(a)$. The vector a therefore carries no tie at the $k/(k{+}1)$ boundary, so $S_k(a)$ is the unique such set. For any $i \in S_k(a)$ and $j \notin S_k(a)$, equation 17 gives

$$s_i^\star \;\ge\; a_i - \rho_i \;>\; a_j + \rho_j \;\ge\; s_j^\star,$$

so every member of $S_k(a)$ outranks every nonmember under $s^\star$. The strict separation also rules out a tie at the boundary of $s^\star$. Since $|S_k(a)| = k$, the two top-k sets coincide. □

Checkable once the weights are fixed, condition equation 18 is the top-k certificate that Algorithm 1 returns.

C.3 Registrability of the region before querying

The realized budget ρ_j is known only once the judgments are in, since C_e is a function of the answer, whereas a weaker region is known before any judgment arrives.

Corollary 3 (Pre-registrable envelope)**.** *Under the deployed weighting $w_e = \lambda U_e C_e$, for every j,*

$$\big|s_j^\star - a_j\big| \;\le\; \rho_j \;\le\; \bar\rho_j \;:=\; \lambda \sum_{e \ni j} U_e, \tag{19}$$

and every statement of Sections C.1 and C.2 holds with $\bar\rho_j$ in place of ρ_j.

Proof of Corollary 3. Certainty is a decisiveness measure bounded by one, so $0 \le w_e \le \lambda U_e$ for every e. Summing over the edges incident to j gives $\rho_j \le \bar{\rho}_j$. Theorem 1 gives the first inequality. Each statement of Sections C.1 and C.2 is monotone in the budgets, so it continues to hold with the larger budgets $\bar{\rho}_j$. □

The actionability scores U_e come from resamples of the training data, so $\bar{\rho}$ depends only on the query set, the targeting rule and the correction strength. The region it defines can therefore be registered before a single query is issued, for a λ fixed in advance or for the largest value on the grid. This makes λ a design lever, since the certainty factor can only shrink a budget fixed beforehand. Because each ρ_j is weakly increasing in λ and under the addition of an incident edge, the protected region $\{(i,j) : a_i - a_j > \rho_i + \rho_j\}$ weakly shrinks as the correction strength grows or the query set densifies and can even be empty. Its realized size is a property of the deployed instance and is reported with the output. No cap on any further factor is needed, because Sections 3.1 and 3.2 admit evidence eligibility through membership in E rather than as a multiplier.

C.4 Approximate optimizers

A numerical solver returns an approximate stationary point in place of $s^\star$, to which the certificate extends with an additive slack.

Proposition 4 (Certificate under approximate stationarity)**.** *Let $\tilde{s}$ satisfy $|\partial_j F(\tilde{s})| \le \varepsilon_j$ for every j. Then, for every j,*

$$\left|\tilde{s}_j - a_j\right| \;\le\; \rho_j + \varepsilon_j, \tag{20}$$

and the order and top-k statements hold with ρ_j replaced by $\rho_j + \varepsilon_j$ throughout.

Proof of Proposition 4. By equation 15, the partial derivatives of F are

$$\partial_j F(s) \;=\; (s_j - a_j) + \big[N^\top\big(w \odot (\sigma(Ns) - \mathbf{p})\big)\big]_j.$$

Rearranging at $s = \tilde{s}$ and applying the term-by-term bound of Part 2 in the proof of Theorem 1 gives

$$\left|\tilde{s}_j - a_j\right| \;\le\; \left|\partial_j F(\tilde{s})\right| + \sum_{e \ni j} w_e \;\le\; \varepsilon_j + \rho_j.$$

The order and top-k statements then follow as in Part 3 of that proof and in Corollary 2. □

The solver reports the ε_j. When it reports only a norm, $\varepsilon_j = \|\nabla F(\tilde{s})\|_\infty$ serves for every j, so the protected region degrades gracefully and measurably.

C.5 The role of nonnegative weights

Under the weighting $w_e = \lambda U_e C_e$, with $\lambda \ge 0$, $U_e = Q_e B_e \ge 0$ a product of $[0,1]$-valued factors and $C_e \in [0,1]$, nonnegativity holds by construction. This subsection records what the argument of Section C.1 would lose under a signed weighting, which is the natural way to encode a judgment one wishes to oppose.

The Hessian of F is

$$\nabla^2 F(s) \;=\; I + N^\top \operatorname{diag}\big(w_e \sigma_e (1 - \sigma_e)\big) N, \qquad \sigma_e := \sigma\big((Ns)_e\big),$$

so along a direction x,

$$x^\top \nabla^2 F(s)\, x \;=\; \|x\|^2 \;+\; \sum_{e \in E} w_e\, \sigma_e (1 - \sigma_e)\, (Nx)_e^2. \tag{21}$$

Two distinct consequences follow. First, the modulus of 1-strong convexity is lost as soon as any $w_{e_0} < 0$. To see this, pick x with $(Nx)_{e_0} \neq 0$ and evaluate equation 21 at points s whose margins satisfy $(Ns)_{e_0} = 0$ while $|(Ns)_e| \to \infty$ on every other edge. Each other term then carries the factor $\sigma_e(1 - \sigma_e) \to 0$, so the quadratic form tends to

$$\|x\|^2 + \tfrac{1}{4} w_{e_0} (Nx)_{e_0}^2 \;<\; \|x\|^2,$$

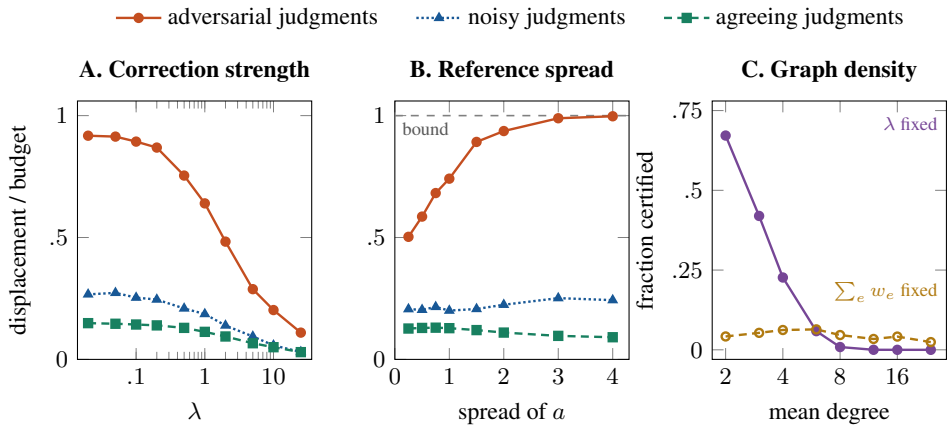

Figure 4: Theorem 1 evaluated on 60 items with 40 replicates per cell (Section C.6). The adversarial setting sends every queried comparison against the reference. (A, B) Largest displacement relative to its budget, $\max_j |s^\star_j - a_j|/\rho_j$, against the correction strength and against the spread of the reference scores. (C) Fraction of top-k pairs certified against graph density at fixed λ or fixed $\sum_e w_e$, which depends on the budgets alone.

so that the smallest eigenvalue of $\nabla^2 F(s)$ falls below 1 for some s. This step uses the restriction to at most one edge per unordered pair, since two opposite orientations of the same pair cannot have their margins driven apart and a large positive weight on the reversed copy could hold the modulus at 1. Second, convexity itself, and with it uniqueness of the minimizer, can fail once the contribution of the negative edges along some x exceeds $\|x\|^2$. It need not fail, however, since a large positive weight elsewhere contributes positive curvature and may keep F convex. Neither failure is governed by $\sum_e |w_e|$ alone, because equation 21 weights each edge by $(Nx)_e^2$ and gives positively weighted edges the opposite sign.

The bound itself survives a negative weight. Each loss term is 1-Lipschitz, so F remains coercive and has minimizers. The stationarity identity equation 16, which holds at every critical point whatever the sign of w, still gives $|s_j - a_j| \le \sum_{e \ni j} |w_e|$ there. What is lost is uniqueness, so the certificate no longer refers to a single corrected score vector. The deployed weighting avoids this by supplying nonnegativity by construction.

C.6 Tightness of the protected region

Since Theorem 1 is deterministic, the bound can be read off once the objective is minimized, with no estimation required. The informative quantities are how much slack the bound carries and what moves that slack, because a bound that never comes close to binding certifies less than it appears to. The study therefore measures a property of the objective, independent of any model or task, namely how much of the reference ordering the guarantee certifies at a given correction strength. We minimize F by damped Newton on 60 items with 40 replicates per cell, sweeping the correction strength, the density of the query graph and the separation of the reference scores. The query graphs are random with a planted cycle for connectivity, the actionability scores U_e are drawn instead of computed from resamples and the reference scores are Gaussian. The first two enter only through ρ_j, on which the theorem is uniform, whereas the third is swept. The anchor makes F strongly convex, so a unique minimizer exists however extreme the reports. Every replicate converged and cells are reported only if they did. Across every cell of every sweep, no displacement exceeded its budget and no certified pair reversed. The sweeps use $w_e = \lambda U_e$, an admissible weighting under the theorem, in place of the deployed $w_e = \lambda U_e C_e$. Since $C_e \le 1$, the deployed budgets are no larger.

Source of the slack. Stationarity of F gives, for every item,

$$s^\star_j - a_j \;=\; -\sum_{e \ni j} w_e \, N_{ej} \big(\sigma(s^\star_{u_e} - s^\star_{v_e}) - p_e\big), \tag{22}$$

where $N_{ej} \in \{-1, +1\}$ is the entry of the signed incidence matrix of Section C. The theorem amounts to $|\sigma - p| \leq 1$ applied term by term, so slack arises in exactly two ways, from queried comparisons whose implied probability sits near the reported one and from incident comparisons that pull against each other. Both are visible in the sweeps.

Attainment of the bound. Because the guarantee is uniform over $p_e \in [0, 1]$, the adversarial reports probe the worst case. As the reference scores separate, every queried $\sigma(a_{u_e} - a_{v_e})$ approaches 0 or 1 while the adversarial report sits at the opposite extreme, so by equation 22 every term saturates. The measured ratio $\max_j |s^\star_j - a_j|/\rho_j$ accordingly rises to 0.997 (Figure 4B). Two items suffice to show that the coefficient of the displacement bound cannot be lowered. With one edge of weight w, a report $p = 0$ and a reference gap M, stationarity gives

$$D \;=\; M - 2w\,\sigma(D), \qquad \frac{|s^\star_1 - a_1|}{w} \;=\; \sigma(D) \;\to\; 1 \quad \text{as } M \to \infty,$$

where $D := s^\star_1 - s^\star_2$ is the corrected margin. Consequently, no constant below one holds uniformly over reference scores, weights and judgments. The ordering threshold is sharp in the same sense. An isolated third item with reference score $a_1 - M'$ for some $M' < w$ overtakes the first once M is large, since $s^\star_1 \to a_1 - w$, although their margin M' approaches the threshold $\rho_1 + \rho_3 = w$ as $M' \uparrow w$. For a pair (i, j) that is itself queried with weight w_{ij}, the threshold sharpens. If $D_{ij} := s^\star_i - s^\star_j \leq 0$, then $\sigma(D_{ij}) \leq \frac{1}{2}$, so the shared comparison moves the margin by at most w_{ij} while the remaining comparisons move it by at most $(\rho_i - w_{ij}) + (\rho_j - w_{ij})$, which gives $D_{ij} \geq (a_i - a_j) - (\rho_i + \rho_j - w_{ij})$. The ordering of a queried pair therefore survives whenever $a_i - a_j > \rho_i + \rho_j - w_{ij}$. Instance-dependent refinements remain possible and one is immediate. Since each corrected margin lies within $\rho_{u_e} + \rho_{v_e}$ of its reference margin,

$$|s^\star_j - a_j| \;\leq\; \sum_{e \ni j} w_e\,\sigma\big(|a_{u_e} - a_{v_e}| + \rho_{u_e} + \rho_{v_e}\big) \;\leq\; \rho_j,$$

with strict inequality whenever $\rho_j > 0$. Against judgments that disagree only noisily, the realized displacement stays four to five times inside the budget, so the certificate is conservative in practice and unimprovable as a universal constant.

A single effective lever. The budget grows linearly in λ while the displacement it permits does not keep pace, so the slack widens and the certificate weakens as the correction strength grows. The certified fraction of top-k pairs falls from 0.99 to 0 as λ runs from 0.02 to 2 (Figure 4A), which leaves the guarantee informative in the small-λ regime the method uses. It also falls from 0.67 to 0 as the mean degree runs from 2 to 16. This second decline suggests that a sparse query set buys protection, which would make the targeting rule of Section 3.1 load-bearing for Theorem 1. Most of the decline, however, reflects the total correction. Holding the expected total correction $\sum_e w_e$ fixed and varying only how it is spread, the certified fraction stays between 0.064 at mean degree 6 and 0.024 at degree 24 instead of falling to zero (Figure 4C). The advantage of a sparse graph therefore lies largely in correcting less. The remaining effect is real but small. It must exist because certification compares each reference margin against $\rho_i + \rho_j$, so the placement of the weight matters as well as its total. For example, three items on a path with total weight 2 and $a = (3, 0, -10)$ certify $1 \succ 2$ under $(w_{12}, w_{23}) = (0.1, 1.9)$ and fail to do so under $(1.9, 0.1)$. This sweep leaves the targeting rule itself untested, since it varies density at random instead of placing weight by equation 4, although targeting could move the protected region by placing weight away from items whose margins are narrow.

D Additional downstream experiments

D.1 Dataset-specific experimental setup and evidence construction

To evaluate LLM-derived guidance across heterogeneous domains, we consider seven datasets spanning transcriptomics, clinical prediction, tabular risk prediction and financial factor ranking (Table 2). For each task, the candidate variable universe is constructed by data-only preprocessing, while external evidence is collected independently and organized into structured evidence packets. A GPT-assisted evidence construction workflow organizes the retrieved information, assesses evidence sufficiency and task alignment and converts heterogeneous sources into structured records.

Table 2: Summary of the downstream datasets and their external evidence. Variables gives the dimension before candidate-universe construction, which each dataset description reports together with the resulting candidate universe. Tokens is the mean number of input tokens per LLM judgment.

Dataset	Task	Samples	Variables	External evidence	Tokens
Renal TCMR	Rejection prediction	403/300 (train/test)	11,929 genes	Rejection cohorts, literature	2,475
GSE272769	Sepsis outcome prediction	161	23,623 genes	Independent sepsis cohorts	575
Breast pCR	Treatment response prediction	306/182 (train/test)	12,034 genes	Treatment-response cohorts	605
CREDIT-G	Credit risk prediction	1,000	20 variables	Credit-risk sources	425
Osteoporosis	Risk prediction	892/727 (train/test)	Clinical variables	Clinical literature	1,379
Darmanis	Tumor-region classification	632 cells	23,257 genes	Literature, spatial atlases	442
JKP153	Factor ranking	1,200 months	153 factors	Asset-pricing literature	930

Instead of raw retrieved documents, the judgment LLM receives frozen evidence packets containing source information, task context, quantitative evidence when available and known limitations, from which every measurement is generated. The evidence itself is domain dependent, since transcriptomic tasks use independent expression cohorts and biomedical evidence, clinical tasks use feature-specific medical evidence, credit prediction uses domain-specific feature information and financial experiments use factor definitions and asset-pricing literature.

D.1.1 Renal TCMR transcriptomics

The Renal TCMR experiment predicts T-cell-mediated rejection (TCMR) status from renal allograft biopsy gene expression, using GSE36059 as the training cohort and GSE48581 as an external test cohort. The GEO-provided GPL570 expression matrices are processed through probe annotation and gene-level aggregation, which collapses multiple probes mapped to the same gene by their median into 11,929 genes, after which highly variable gene (HVG) selection constructs a candidate universe of 2,000 genes.

External evidence is collected from renal-transplant biomedical resources, including independent rejection-related transcriptomic studies and biomedical literature. Each packet preserves the gene identity, source identity, cohort and study context, target comparison, biological interpretation, reported direction, quantitative evidence when available and known limitations. A pairwise query presents the packets of two candidate genes and contains about 2,475 input tokens on average.

D.1.2 Sepsis mortality transcriptomics (GSE272769)

The GSE272769 experiment predicts 30-day outcome in adult sepsis from the whole-blood transcriptomes of 161 samples, publicly available from NCBI GEO under accession GSE272769. The GEO-provided transcript-cluster expression matrix is mapped to 23,623 gene-level features, from which HVG selection constructs a 1,500-gene candidate universe. External evidence is collected from independent sepsis transcriptomic studies, including GAinS/E-MTAB-4451, VANISH/E-MTAB-7581 and GSE95233. Its packets summarize study design, outcome definition, gene-level association, effect direction, quantitative statistics when available, sample information and study limitations. The pairwise prompts present these source-specific packets in place of aggregated conclusions and contain about 575 input tokens per judgment on average.

D.1.3 Breast cancer pathological complete response

The Breast pCR experiment predicts pathological complete response from pretreatment breast tumor gene-expression profiles, using GSE25055 for training and GSE25065 for external testing. The GEO-deposited processed Affymetrix matrices pass through probe annotation, probe-to-gene mapping, protein-coding gene filtering and duplicate probe aggregation, which reduce 22,283 probes to 12,034 gene-level features before HVG selection yields a 2,000-gene candidate universe. External evidence is constructed from independent treatment-response transcriptomic cohorts, including GSE41998 and GSE32646. For each gene, the packet summarizes response-related association, effect direction, discrimination statistics, significance information, sample counts and limitations. A pairwise query contains about 605 input tokens on average.

D.1.4 GERMAN CREDIT RISK PREDICTION

The CREDIT-G experiment evaluates credit-risk prediction on the German Credit dataset from OpenML dataset 31, which contains 1,000 samples and 20 semantic variables. No variable filtering is applied, because the original variables are interpretable credit-application attributes. Continuous variables are standardized and categorical variables encoded for the downstream predictors, while evidence construction remains at the level of the original variables. External evidence is collected from credit-risk sources and covers feature interpretation, domain meaning and empirical relationships with credit outcomes. Its packets contain feature descriptions, source information, empirical findings and applicability limitations. A feature-level query contains about 425 input tokens on average.

D.1.5 OSTEOPOROSIS CLINICAL PREDICTION

The osteoporosis experiment uses a proprietary hospital clinical cohort for risk prediction, which is not publicly available because of institutional data governance restrictions. The cohort consists of two temporal batches collected at different times. The first batch (892 records) is used for training and for choosing the correction strength, while the second batch (727 records), collected later, is held out as the temporal test cohort, so every reported Osteoporosis result is a temporal external validation on later patients. After clinical quality control and variable eligibility filtering, the candidate universe contains 37 clinical variables. External evidence is constructed through feature-specific medical literature retrieval, whose results the GPT-assisted workflow organizes and assesses for relevance to the target task, including population alignment, measurement definition, outcome relevance and study design. A clinical feature query contains about 1,379 input tokens on average.

D.1.6 DARMANIS GLIOBLASTOMA TRANSCRIPTOMICS

The Darmanis experiment classifies glioblastoma tumor-core against tumor-periphery cells in the publicly available single-cell dataset of Darmanis et al. (2017) (GEO accession GSE84465). It uses a processed gene-expression matrix of 632 cells and 23,257 genes in place of raw sequencing files, from which expression transformation and HVG selection construct a candidate universe of 2,000 genes. External evidence combines biomedical literature retrieval with independent spatial and anatomical transcriptomic resources, including the Ivy Glioblastoma Atlas Project/GSE107559, GSE116520 and GSE13276. A gene-level query contains about 442 input tokens on average.

D.1.7 JKP153 EQUITY FACTOR RANKING

The JKP153 experiment evaluates LLM-derived guidance for ranking 153 equity factors from the Jensen–Kelly–Pedersen (JKP) universe, whose data are obtained from the public JKP factor database. No filtering is required, since the factors are defined by the original JKP metadata and return series. External evidence is constructed from factor metadata, original asset-pricing papers, working papers, replication studies and robustness literature. Each factor packet, from which the LLM measurements are generated, contains the factor identity, economic interpretation, construction information, canonical source, empirical findings and replication or robustness information when available. A factor-level query contains about 930 input tokens on average. The 2025 period is held out and evaluated after the analysis was frozen, separately from the 2024 validation period.

D.2 COMPARISON WITH OTHER FEATURE-SELECTION METHODS

We compare Selective Correction with recent LLM-based feature-selection methods under a common downstream protocol, in which every method uses the same candidate universe, selected-set size, data split and downstream evaluator within each experiment. The comparison includes the score-based variant of LLM-Select (Jeong et al., 2025), Data-centric LLM feature selection (Li et al., 2025), FREEFORM (Lee et al., 2025a), ICE-SEARCH (Yang et al., 2024) and LLM-Lasso (Zhang et al., 2025).

As shown in Table 3, Selective Correction attains the highest AUROC in seven of the nine settings, namely CREDIT-G, Osteoporosis and five of the six Darmanis settings, whereas LLM-Select Score is highest on Breast pCR and FREEFORM on Darmanis SIS-20. These results show that Selec-

Table 3: Comparison with recent LLM-based feature-selection methods under the same downstream protocol. All entries are AUROC. Bold marks the best value in each row.

Dataset / setting	Reference	Ours	LLM-Select Score	Data-centric	FREEFORM	ICE-SEARCH	LLM-Lasso
Breast pCR, SIS-20	0.7228	0.7376	**0.7650**	0.7322	0.7322	0.7322	0.7322
CREDIT-G, GBM-10	0.7801	**0.7933**	0.7772	0.7565	0.7736	0.7679	0.7736
Osteoporosis, Top-10	0.7295	**0.7382**	0.7187	0.7203	0.7261	0.7259	0.7244
Darmanis, SIS-10	0.9141	**0.9482**	0.7870	0.9113	0.9165	0.7144	0.8991
Darmanis, LASSO-10	0.9566	**0.9680**	0.7870	0.9113	0.9165	0.7144	0.8991
Darmanis, EN-10	0.9564	**0.9759**	0.7870	0.9113	0.9165	0.7144	0.8991
Darmanis, SIS-20	0.9448	0.9703	0.9058	0.9574	**0.9801**	0.7849	0.8914
Darmanis, LASSO-20	0.9785	**0.9915**	0.9058	0.9574	0.9801	0.7849	0.8914
Darmanis, EN-20	0.9818	**0.9879**	0.9058	0.9574	0.9801	0.7849	0.8914

Table 4: Cross-model replication in AUROC. Δ_{Qwen} is Ours (Qwen) minus the Reference and Viol. counts bound violations under the Qwen correction. For GSE272769, validation selected $\lambda = 0$ for Qwen. Bold marks the best value in each row.

Experiment / setting	Reference	Ours (DeepSeek)	Ours (Qwen)	Δ_{Qwen}	Viol.
Renal TCMR, Top-50	0.7856	0.8116	**0.8235**	+0.0379	0
GSE272769, EN-50	0.6287	**0.6554**	0.6287	0.0000	0
Breast pCR, SIS-20	0.7228	**0.7376**	0.7248	+0.0020	0
CREDIT-G, GBM-10	0.7801	**0.7933**	**0.7933**	+0.0132	0
CREDIT-G, EN-10	0.7695	**0.7770**	**0.7770**	+0.0075	0
Osteoporosis, Top-10	0.7295	**0.7382**	**0.7382**	+0.0087	0
Darmanis, SIS-10	0.9141	**0.9482**	0.9275	+0.0134	0
Darmanis, SIS-20	0.9448	**0.9703**	0.9678	+0.0230	0
Darmanis, EN-10	0.9564	**0.9759**	**0.9759**	+0.0195	0

tive Correction is competitive with recent LLM-based feature-selection procedures under the same downstream protocol.

D.3 Cross-model replication

We test whether the downstream behavior of Selective Correction depends on a particular elicitation model by repeating the nine reported prediction settings with a locally deployed Qwen3-32B model in place of the DeepSeek V4.1 Flash API. The data splits, reference selectors, targeting rules, evidence inputs and validation procedures stay unchanged. The correction strength is reselected on validation data over the same grid without access to the test data.

As shown in Table 4, Qwen improves over the data-only Reference in eight of the nine settings. The remaining setting, GSE272769, selects $\lambda = 0$ in all five outer folds and therefore reproduces the Reference exactly, so no Qwen setting falls below the Reference. All nine Qwen settings also pass the realized displacement audit with zero bound violations. Although the size of the gain varies across elicitation models, Selective Correction thus remains useful when the LLM measurements change.

The fallback on GSE272769 shows the role of validation, since DeepSeek selects nonzero correction in three of the five outer folds whereas Qwen selects $\lambda = 0$ in all five, although the two models agree closely on the elicited pairwise measurements. Agreement between two LLMs thus leaves the decision whether to correct with downstream validation, which retains authority over whether the measurements are used.

The JKP replication is reported separately because it uses a different local checkpoint, Qwen2.5-14B-Instruct, and a full-ranking task in place of top-k prediction, with the targeting and evaluation protocol otherwise following the registered JKP experiment. As shown in Table 5, selective targeting remains above the data-only ranking with both DeepSeek and Qwen in the 2024 validation period. The same ordering relative to the Reference holds in the held-out 2025 period, although Rank IC is negative for all three values in that period. Taken together, the replications show that the help-and-

Table 5: Cross-model replication for JKP153 in Rank IC. The Qwen column uses Qwen2.5-14B-Instruct in place of the Qwen3-32B model of Table 4. The year 2025 is the held-out period. Bold marks the best value in each row.

Evaluation period	Reference	Ours (DeepSeek)	Ours (Qwen)
2024 validation	0.3001	**0.3427**	0.3294
2025 held-out	-0.2674	**-0.1343**	-0.1718

Table 6: Component ablation on Breast pCR (SIS-20), in AUROC on the GSE25065 external test cohort. Δ is relative to the Reference. Bold marks the best construction.

Components retained	AUROC	Δ vs. Reference
Reference	0.7228	0.0000
$Q + B + C$	**0.7376**	**+0.0148**
$B + C$	0.7315	+0.0087
$Q + B$	0.7053	-0.0175
$Q + C$	0.6983	-0.0245

fallback behavior of Selective Correction carries over across elicitation models, while the selected strength and the resulting gain remain model- and task-dependent.

D.4 Breast SIS-20 component ablation

We ablate the three factors of the Selective Correction weight on the Breast pCR SIS-20 experiment. The full correction uses

$$w_e = \lambda Q_e B_e C_e,$$

where Q_e measures substitution around the selection boundary, B_e measures pairwise balance across reference resamples and C_e measures the certainty of the elicited LLM comparison. Each variant retains a subset of the factors, which enter the weight multiplicatively, so $Q + B + C$ denotes the full construction while $B + C$, $Q + B$ and $Q + C$ remove Q, C and B, respectively. All variants use the same Breast SIS-20 training and external-test protocol.

As shown in Table 6, the full $Q+B+C$ construction attains the highest AUROC. Removing Q while retaining B and C preserves a smaller positive gain over the Reference, whereas removing either B or the certainty factor C places the variant below the data-only Reference. The Breast improvement is therefore strongest when boundary substitution, pairwise balance and LLM certainty act together, which supports the combined weighting.

D.5 Direct LLM-only selection

We separately evaluate direct LLM-only feature selection, in which the LLM receives the task description and candidate feature identities and the selected features are evaluated with the same downstream protocol as the Reference and Selective Correction. Pointwise semantic feature scoring is repeated independently three times. We report the mean and standard deviation of the resulting downstream AUROC together with a score ensemble that averages the three LLM feature scores before top-k selection.

As shown in Table 7, Selective Correction exceeds both the three-run LLM-only mean and the score ensemble on Breast pCR, CREDIT-G, Osteoporosis and all six Darmanis settings, whereas the LLM-only ensemble attains the highest AUROC on GSE272769 and Renal TCMR. Direct semantic selection can therefore be strong on some datasets, while Selective Correction governs how external judgments modify a data-only decision and preserves the protected-region guarantee of Theorem 1.

D.6 Correction-strength selection and evaluation protocol

For each downstream dataset, the correction strength λ is selected on validation data and frozen before testing, so the data used to choose it stay disjoint from the data used to report it (Table 8). For a fixed dataset, selector and support size k, we evaluate the prespecified grid of λ values of

Table 7: Direct LLM-only feature selection compared with the data-only Reference and Selective Correction. All entries are AUROC. Mean ± SD is computed over three independent scoring runs, while "Ensemble" averages their feature scores before top-k selection. Bold marks the best value in each row.

Dataset / setting	LLM-only mean ± SD	Ensemble	Reference	Ours
Breast pCR, SIS-20	0.7032 ± 0.0352	0.6956	0.7228	**0.7376**
CREDIT-G, GBM-10	0.7586 ± 0.0041	0.7692	0.7801	**0.7933**
GSE272769, EN-50	0.5911 ± 0.0150	**0.6578**	0.6287	0.6554
Osteoporosis, Top-10	0.7253 ± 0.0031	0.7289	0.7295	**0.7382**
Renal TCMR, Top-50	0.8309 ± 0.0322	**0.8498**	0.7856	0.8116
Darmanis, SIS-10	0.8866 ± 0.0208	0.8993	0.9141	**0.9482**
Darmanis, LASSO-10	0.8866 ± 0.0208	0.8993	0.9566	**0.9680**
Darmanis, EN-10	0.8866 ± 0.0208	0.8993	0.9564	**0.9759**
Darmanis, SIS-20	0.8838 ± 0.0381	0.8634	0.9448	**0.9703**
Darmanis, LASSO-20	0.8838 ± 0.0381	0.8634	0.9785	**0.9915**
Darmanis, EN-20	0.8838 ± 0.0381	0.8634	0.9818	**0.9879**

Table 8: Training, validation and test data. Training counts are in the units of the analysis, namely sample rows for the transcriptomic experiments, sites and patients for the clinical cohort, cells and plates for the single-cell cohort and months for the factor panel. Validation units are the units on which the correction strength is chosen. Test data are never read before the strength is frozen.

Dataset	Training	Validation units	Test data
Renal TCMR	GSE36059, 403 samples	5-fold CV within training (≈322 fit / 81 validation per fold)	GSE48581, 300 samples
GSE272769	161 samples	inner 4-fold CV inside each outer training fold	5 outer folds, 32–33 samples each
Breast pCR	GSE25055, 306 samples	5-fold CV within training (≈245 fit / 61 validation per fold)	GSE25065, 182 samples
CREDIT-G	800 samples (80% of the cohort)	300 stratified $80/20$ resamples (640 fit / 160 validation)	200 samples, locked held-out split
Osteoporosis	Batch 1, 892 rows / 664 patients	25 evaluations within Batch 1 (5 folds × 5 seeds)	Batch 2, 727 rows / 538 patients
Darmanis	632 cells / 24 plates	plate-grouped 5-fold CV, 506–514 training cells per fold	held-out folds, 118–136 cells
JKP153	rolling 120-month estimation window	2024 calendar year	2025 calendar year

Table 9 on the validation units and select the λ with the largest mean validation AUROC, resolving ties in favor of the smaller λ. Where an external cohort is available, the correction strength is chosen inside the training cohort and the metric is reported on the external cohort. Where none is available, Table 8 records the fold structure that plays the validation and test roles, with the reported metric averaged over the held-out folds. Validation never reads a test outcome, since the strength is frozen before any test performance is computed. The final feature set is computed from the full training data with the frozen selector, evidence packets and LLM measurements. The downstream predictor is fitted on the training data restricted to the selected variables and evaluated on the test data.

D.7 Resampling for query targeting

The targeting quantities Q_{ij}, B_{ij} and U_{ij} of Section 3.1 come from R refits of the reference selector, each on a subsample of the training data drawn without replacement at a fixed fraction and stratified by the unit the design treats as independent. Selector hyperparameters are held at their reference values inside every resample, so a resample never retunes the selector. Every resample reads training data only, so none touches a held-out outcome or an elicited judgment. For resample b we keep the refitted score vector and its ranking $r^{(b)}$, with rank 1 for the largest score and average ranks for ties,

Table 9: Grid of correction strengths λ searched on the validation units of Table 8. Every grid contains 0.

Dataset	λ grid
Renal TCMR	$\{0, 10^{-4}, 3{\times}10^{-4}, 10^{-3}, 3{\times}10^{-3}, 10^{-2}, 0.03, 0.1, 0.3, 1\}$
GSE272769	$\{0, 0.03, 0.1, 0.3, 1, 3, 10\}$
Breast pCR	$\{0, 0.03, 0.1, 0.3, 1, 3, 10\}$
CREDIT-G	$\{0, 0.025, 0.05, 0.075, 0.10, 0.15, 0.20, 0.30\}$
Osteoporosis	$\{0, 0.03, 0.1, 0.3, 1, 3, 10\}$
Darmanis	$\{0, 0.03, 0.1, 0.3, 1, 3, 10\}$
JKP153	$\{0, 0.05, 0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2\}$

Table 10: Resampling used to compute the substitution probability Q, the pairwise balance B and the actionability score U. All resamples are drawn without replacement unless stated otherwise and none uses a held-out outcome or an elicited judgment.

Dataset	R	Fraction	Unit and stratification
Renal TCMR	200	0.80	Samples, stratified by phenotype
GSE272769	200	0.80	Samples, within each outer training fold
Breast pCR	200	0.80	Samples, within each outer training fold
CREDIT-G	300	0.80/0.20	Stratified train/test resample
Osteoporosis	200	0.80	Patients (cluster), stratified by label
Darmanis	200	14/17, 6/7 plates	Plates (cluster), stratified by label
JKP153	1000	12-month blocks	Moving-block bootstrap over 120 months

and set

$$Q_{ij} = \frac{1}{R}\sum_{b=1}^{R} \mathbf{1}\Big\{ r_i^{(b)} \le k < r_j^{(b)} \ \text{ or } \ r_j^{(b)} \le k < r_i^{(b)} \Big\},$$

$$\hat{p}_{ij} = \frac{\#\{r_i^{(b)} < r_j^{(b)}\} + \frac{1}{2}\,\#\{r_i^{(b)} = r_j^{(b)}\}}{R},$$

so that Q_{ij} is the empirical frequency with which exactly one member of the pair enters the selected set, while $\hat{p}_{ij}$ gives a tie half the weight of a strict ordering. Then $B_{ij} = 1 - 2|\hat{p}_{ij} - \frac{1}{2}|$ and $U_{ij} = Q_{ij}B_{ij}$. Candidate pairs are restricted to the selection boundary, so both items must have a median rank of at most $k + 5$ and at least one of them must have a median rank within 5 positions of k. A pair is actionable when $U_{ij} \ge 0.25$. For the full-ranking experiment, which orders all 153 factors instead of selecting a top-k set, the same reference-uncertainty construction replaces the boundary window by a rank-gap criterion, under which a pair is admissible when the two factors lie within three rank positions of each other in the median bootstrap ranking and the entropy of its directional frequency is at least 0.90, with 0.85 and 0.95 reported as sensitivity variants. Table 10 lists R, the fraction and the unit of each experiment.