RELATIONAL RESPONSE FIELDS: A GENERAL THEORY OF BLACK-BOX LLM RESPONSE CONSISTENCY AND RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Black-box language-model reliability is commonly pursued by sampling, prompting, voting, verifying, or iteratively revising individual answers. We ask a prior question: *what determines whether a collection of black-box responses is recoverable at all?* We represent responses to typed transformations of a query as a *relational response field* (RRF). Edge transports encode how valid responses must change under paraphrase, scaling, decomposition, refactoring, or other task symmetries; anchors encode independently trusted evidence such as execution or a verifier. For relation operator D, anchor operator A, and at most k corrupted response nodes, we identify $\gamma_k(D, A)$ as the intrinsic difficulty of black-box response recovery. It is positive exactly when every k-node corruption is identifiable; it gives a deterministic stability bound proportional to $1/\gamma_k$; and a matching two-point minimax lower bound shows that no estimator can improve this dependence. Thus consistency is not truth: relation-only methods are blind to null directions, including shared hallucinations. We derive sparse field-repair algorithms while separating information-theoretic identifiability from the stronger null-space conditions required by convex optimization. Controlled theorem tests and black-box mathematics/code experiments evaluate four theory-fixed consequences: consistency–truth separation, anchor phase transitions, redundancy saturation, and cross-model, cross-task prediction of repair difficulty. The results support $\gamma_k(D, A)$ as a measurable property of a response-recovery instance, rather than a score attached to one repair heuristic.

1 INTRODUCTION

A black-box large language model (LLM) can be queried but not differentiated, inspected, or retrained. Reliability methods therefore act at inference time: they alter prompts (Zhou et al., 2023b; Pryzant et al., 2023; Yang et al., 2024b), sample multiple chains and vote (Wang et al., 2023), ask the model to critique itself (Madaan et al., 2023; Shinn et al., 2023; Miao et al., 2024), search over reasoning trees (Yao et al., 2023a), or select with learned and executable verifiers (Cobbe et al., 2021; Chen et al., 2021; Li et al., 2022). Metamorphic tests instead transform an input and check a known relation among outputs (Chen et al., 2020; Segura et al., 2016). These approaches appear operationally different, yet they repeatedly create the same latent object: a family of answers connected by known relations. This paper makes that object explicit. Let $G = (V, E)$ index transformed queries. The model response at node i is mapped by a black-box parser to z_i; the collection $z = (z_i)_{i \in V}$ is a *relational response field*. An edge $e = (i, j)$ carries a typed transport T_e and asks that a valid field satisfy $z_j = T_e z_i$. An identity transport represents a paraphrase, a scalar map represents unit conversion, a structured map represents decomposition, and an execution map represents semantics-preserving code transformation. Stacking the residuals $z_j - T_e z_i$ defines a typed relation defect operator D. Trusted evidence, execution, constraints, or human labels define an anchor operator A. The important question is not whether more responses are available. It is whether the combined relations and anchors can distinguish the true field from a sparsely corrupted one.

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.

We show that this question has an exact answer:

$$\gamma_k(D, A) = \min_{\substack{h \neq 0 \\ \|h\|_{0,\mathrm{g}} \leq 2k}} \frac{\sqrt{\|Dh\|_2^2 + \|Ah\|_2^2}}{\|h\|_2}. \tag{1}$$

The group support counts response nodes, not scalar coordinates. The factor $2k$ is unavoidable: the difference between two fields, each corrupted on k nodes, can occupy $2k$ nodes. Equation (1) is a restricted minimum singular value of the *instance-specific* relation–anchor operator. Its role in RRF recovery is exact, not metaphorical.

First, $\gamma_k > 0$ is necessary and sufficient for uniform identification of every k-node corruption. Second, if observations are perturbed by norm ϵ, any feasible estimate incurs error at most $2\epsilon/\gamma_k$. Third, a two-point construction gives worst-case error at least a constant times ϵ/γ_k for every estimator. Hence the inverse margin, rather than model accuracy, query count, graph density, or the objective value of a particular algorithm, is the intrinsic condition number of black-box response recovery.

This viewpoint cleanly separates *consistency* from *truth*. If $h \in \ker D$, then z and $z + h$ have identical relation defects. A shared hallucination transported coherently across every paraphrase can therefore be perfectly consistent. Anchors matter because they remove such gauge directions. Their value depends on placement: one independent execution constraint can increase γ_k more than hundreds of duplicated paraphrases. This distinction also prevents a common theoretical mistake. Positive γ_k guarantees identifiability for the ideal sparse decoder, but a tractable group-ℓ_1 decoder requires a robust group null-space property. We state both conditions and do not attribute an algorithmic guarantee to identifiability alone.

Contributions.

- **Object and operator.** We introduce RRFs and a typed relation defect operator that place paraphrase consistency, metamorphic relations, execution, and verification in one black-box inverse problem.
- **Intrinsic difficulty.** We prove that $\gamma_k(D, A)$ is the exact uniform identifiability threshold and the minimax stability scale for k-node response recovery.
- **Impossibility and phase structure.** We characterize consistency blindness, sparse gauge ambiguities, anchor phase transitions, and why normalized duplicate relations cannot improve recoverability.
- **Repair algorithms.** We derive ideal, convex, and discrete candidate-field decoders, with separate guarantees for each and a local extension to nonlinear transports.
- **Prediction rather than only improvement.** Experiments test four consequences fixed by the theory. In particular, we evaluate whether γ_k predicts repair difficulty across models and across mathematics/code tasks after controlling for raw model quality.

Our novelty claim is deliberately narrow. Restricted singular values, group-sparse recovery, graph signal processing, sheaf Laplacians, and metamorphic testing are established subjects (Donoho, 2006; Candès et al., 2006; Eldar et al., 2010; Ricaud et al., 2013; Hansen & Ghrist, 2019). The contribution is the RRF formulation and the identification, proof, and empirical use of its relation–anchor margin as the fundamental recoverability quantity for black-box LLM responses.

2 Relational response recovery

Response fields. Let $G = (V, E)$ be a finite directed multigraph with $V = [n]$. Node i contains a transformed query x_i and a finite-dimensional real Hilbert space $\mathcal{H}_i$. A model produces text $y_i \sim P_\theta(\cdot \mid x_i)$; an external parser ϕ_i maps it to $z_i = \phi_i(y_i) \in \mathcal{H}_i$. The black-box assumption permits samples and external parsing but no logits, gradients, hidden states, or parameter updates. The product space $\mathcal{H} = \bigoplus_{i=1}^n \mathcal{H}_i$ is partitioned into node groups, and

$$z = (z_1, \ldots, z_n) \in \mathcal{H}, \quad \mathrm{supp}_{\mathrm{g}}(z) = \{i : z_i \neq 0\}, \quad \|z\|_{0,\mathrm{g}} = |\,\mathrm{supp}_{\mathrm{g}}(z)|.$$

Typed transports and defects. Each edge $e = (i, j, t)$ has a type t and a known bounded linear transport $T_e : \mathcal{H}_i \to \mathcal{H}_j$. With a relation-family weight $w_e > 0$, define

$$(Dz)_e = \sqrt{w_e}\,(z_j - T_e z_i), \qquad D : \mathcal{H} \to \mathcal{E} := \bigoplus_{e \in E} \mathcal{H}_{\mathrm{head}(e)}. \tag{2}$$

Weights are normalized within a relation family: splitting one observation into m identical copies assigns weight w/m to each. This convention makes D^*D invariant to literal duplication and prevents a query-count artifact from masquerading as information.

An anchor is an externally observable affine constraint $Az \approx b$, where $A : \mathcal{H} \to \mathcal{K}$ is linear after absorbing offsets into b. Examples are selected labels, symbolic substitution residuals, unit tests, retrieved evidence constraints, or a calibrated verifier. We stack

$$B = \begin{bmatrix} D \\ A \end{bmatrix}, \qquad \|Bh\|_2^2 = \|Dh\|_2^2 + \|Ah\|_2^2. \tag{3}$$

Observation model. Let $z^\star$ denote an unknown valid field and let $y = z^\star + e^\star$ be the parsed black-box field, where $e^\star$ is nonzero on at most k response nodes. Clean relations may have intrinsic mismatch $q = Dz^\star$, and anchors obey $b = Az^\star + \xi_A$. Since y is observed, recovering $z^\star$ is equivalent to recovering $e^\star$ from

$$s := \begin{bmatrix} Dy - q \\ Ay - b \end{bmatrix} = Be^\star + \xi, \qquad \xi = \begin{bmatrix} 0 \\ -\xi_A \end{bmatrix}, \tag{4}$$

or from its fully noisy version with $\|\xi\|_2 \le \epsilon$. In the common exact-relation case, $q = 0$. This formulation clarifies what an algorithm observes and where noise enters; it is an ordinary group-sparse inverse problem whose matrix is induced by typed response relations.

Definition 1 (Relational observability margin)**.** For corruption budget k, define

$$\gamma_k(D, A) := \min_{\substack{h \in \mathcal{H} \setminus \{0\} \\ \|h\|_{0,\mathrm{g}} \le 2k}} \frac{\|Bh\|_2}{\|h\|_2} = \min_{1 \le |S| \le 2k} \sigma_{\min}(B_S), \tag{5}$$

where B_S restricts B to node groups in S. The minimum exists because the number of supports is finite and each restricted unit sphere is compact.

The margin depends on the query transformations, response representation, relation normalization, anchors, and budget k. It does not depend on the repair algorithm. It is nonincreasing in k and nondecreasing when unnormalized informative rows are added. Equivalently, $\gamma_k > 0$ iff the group spark of B exceeds $2k$.

Examples. For equivalent prompts with scalar answers, D is a weighted graph incidence matrix. Its nullspace contains a constant shift on each connected component; an anchor fixes that gauge. For a scaling query with expected answer cz_i, the edge row is $[-c, 1]$. For code, z_i can be an execution signature and $T_e = I$ for semantics-preserving refactors. For a decomposition, z_i contains compatible blocks and T_e selects or aggregates coordinates. Nonlinear relations are treated through a residual map F and its restricted Jacobian; Section 6 states the scope, and Appendix G gives the local theorem.

3 The intrinsic difficulty of recovery

We first isolate the failure of consistency-only reasoning, then give an exact recovery characterization and matching stability laws. Complete proofs, including all edge cases, are in Appendices D–E.

Theorem 2 (Consistency blindness)**.** *Suppose an estimator receives only Dz. For every nonzero $h \in \ker D$, the fields z and $z + h$ induce exactly the same observation. Consequently, if the admissible class contains both fields, no estimator based only on relation defect can be correct on both. If $\ker D$ contains a transported global shift, a perfectly consistent shared hallucination is indistinguishable from truth.*

The statement is deterministic and does not depend on model stochasticity. Repeated sampling may reveal independent errors, but it cannot remove a systematic error direction that lies in $\ker D$. Anchors alter the relevant nullspace from $\ker D$ to $\ker D \cap \ker A = \ker B$.

Theorem 3 (Exact uniform identifiability)**.** *Fix k. The following are equivalent:*

(i) every k-node error $e^\star$ is the unique solution of $Be = Be^\star$ among errors satisfying $\|e\|_{0,\mathrm{g}} \le k$;

(ii) $\ker B$ *contains no nonzero vector supported on at most* $2k$ *nodes;*

(iii) $\gamma_k(D, A) > 0$*;*

(iv) $\mathrm{spark}_{\mathrm{g}}(B) > 2k$*.*

Thus $\gamma_k > 0$ *is necessary and sufficient, not merely sufficient, for uniform black-box recovery under the stated corruption model.*

Proof idea. If two k-sparse errors produce the same observation, their difference is a $2k$-sparse null vector. Conversely, partition the support of any $2k$-sparse null vector into two sets of size at most k; its two parts produce distinct but observationally identical errors. Finite dimensionality makes positivity of the restricted singular values equivalent to absence of such a vector.

Theorem 4 (Deterministic stability)**.** *Let* $s = Be^\star + \xi$ *with* $\|e^\star\|_{0,\mathrm{g}} \le k$ *and* $\|\xi\|_2 \le \epsilon$*. If* $\widehat{e}$ *is any* k*-node-sparse estimate with* $\|B\widehat{e} - s\|_2 \le \epsilon$*, then, whenever* $\gamma_k > 0$*,*

$$\|\widehat{e} - e^\star\|_2 \le \frac{2\epsilon}{\gamma_k(D, A)}. \tag{6}$$

For unequal feasibility radii ϵ_1, ϵ_2*, the numerator is* $\epsilon_1 + \epsilon_2$*. The same bound holds for recovered fields because* $\widehat{z} - z^\star = -(\widehat{e} - e^\star)$*.*

Stability alone does not establish that $1/\gamma_k$ is unavoidable. The next result does. Let $\mathcal{E}_k(R) = \{e : \|e\|_{0,\mathrm{g}} \le k, \|e\|_2 \le R\}$ and consider deterministic noise of radius ϵ. The minimax risk is

$$\mathfrak{R}_k(B; R, \epsilon) = \inf_{\widehat{e}} \sup_{e \in \mathcal{E}_k(R),\ \|\xi\| \le \epsilon} \|\widehat{e}(Be + \xi) - e\|_2.$$

Theorem 5 (Matching minimax obstruction)**.** *For every* B *and* k*,*

$$\frac{1}{2}\min\left\{R, \frac{2\epsilon}{\gamma_k(D, A)}\right\} \;\le\; \mathfrak{R}_k(B; R, \epsilon) \;\le\; \min\left\{R, \frac{2\epsilon}{\gamma_k(D, A)}\right\}, \tag{7}$$

with the convention $\epsilon/0 = +\infty$*; constants depend only on whether* R *bounds each candidate or their separation. In particular, if* $\gamma_k = 0$ *and amplitudes are unbounded, worst-case ambiguity is unbounded. No estimator can improve the* $1/\gamma_k$ *dependence uniformly.*

The lower bound takes a unit $2k$-sparse direction attaining γ_k, scales it until its image can be canceled by admissible noise, and splits its support into two k-sparse candidates. Their observation balls intersect, so one datum is compatible with both. Theorem 4 supplies the matching upper scale. This pair of results is the central reason to call $\gamma_k(D, A)$ the *intrinsic difficulty* of black-box LLM response recovery.

Consequences predicted before fitting an algorithm.

1. *Consistency–truth separation:* small $\|Dz\|$ cannot rule out a large component in $\ker D$.
2. *Anchor phase transition:* recovery changes qualitatively when the last $2k$-sparse null direction is removed and γ_k becomes positive.
3. *Redundancy saturation:* normalized duplicate rows leave D^*D, hence γ_k, unchanged.
4. *Difficulty prediction:* among instances at comparable noise and corruption scale, error must grow as the margin decreases, regardless of the model family or task that generated the field.

These are falsifiable structural predictions. The experiments test them separately from claims about average accuracy.

4 Sparse response-field repair

The ideal decoder directly implements Theorem 3:

$$\widehat{e}_0 \in \arg\min_{e} \|e\|_{0,\mathrm{g}} \quad \text{s.t.} \quad \|Be - s\|_2 \le \epsilon, \qquad \widehat{z} = y - \widehat{e}_0. \tag{8}$$

Algorithm 1 Sparse Relational Response Field Repair

Require: Parsed field y; typed D; anchors (A, b); relation target q; budget/noise parameters.
1: Form $B = [D; A]$ and $s = [Dy - q; Ay - b]$.
2: **if** the field and budget permit exact support search **then**
3: compute $\gamma_k = \min_{|S| \le 2k} \sigma_{\min}(B_S)$ and solve Equation (8);
4: **else if** responses have continuous editable representations **then**
5: solve Equation (9) or (10) by group proximal updates;
6: **else**
7: solve the discrete candidate-field energy over transported model candidates;
8: **end if**
9: **return** repaired field $\widehat{z} = y - \widehat{e}$ and certificate $(\gamma_k, \|B\widehat{e} - s\|)$.

It is exact under $\gamma_k > 0$ but combinatorial. For small RRFs we enumerate supports, which also computes γ_k exactly. For larger fields we use the convex relaxation

$$\widehat{e}_1 \in \arg\min_e \sum_{i \in V} \omega_i \|e_i\|_2 \quad \text{s.t.} \quad \|Be - s\|_2 \le \epsilon, \tag{9}$$

or its penalized form

$$\min_e \ \frac{1}{2}\|Be - s\|_2^2 + \tau \sum_i \omega_i \|e_i\|_2. \tag{10}$$

The equivalent field-space objective combines relation defect, anchor mismatch, and a group-sparse edit penalty. It changes parsed answers, not model parameters.

Proposition 6 (Algorithmic condition)**.** *If B satisfies the group robust null-space property of order k with constants $0 < \rho < 1$ and $\tau_B > 0$, then a solution of Equation (9) obeys*

$$\|\widehat{e}_1 - e^\star\|_2 \le C_1 \frac{\sigma_k(e^\star)_{2,1}}{\sqrt{k}} + C_2 \epsilon,$$

where C_1, C_2 depend only on (ρ, τ_B). Positive γ_k is necessary for uniform recovery but, by itself, is not a certificate for the convex relaxation.

We solve Equation (10) by proximal gradient. The group soft-thresholding map is applied nodewise, and a step size below $\|B\|_{2\to 2}^{-2}$ yields the standard $O(1/t)$ objective gap; accelerated updates give $O(1/t^2)$. Discrete outputs require a different decoder. For each node we retain a finite candidate set extracted from model samples, transport candidates across edges, and minimize relation plus anchor energy over candidate fields. This prevents a continuous average of two integers or two programs from being reported as a response.

The returned margin is a certificate about the *instance*. A low objective value is merely evidence that the chosen algorithm found a consistent candidate; it cannot replace the margin because it does not quantify nearby indistinguishable fields.

5 Experiments: testing a difficulty measure

The experiments test consequences of the theory rather than only final accuracy. We separate (i) synthetic theorem verification, where $z^\star$, B, support, and γ_k are exactly known; (ii) natural real-model fields, which include candidate absence and multi-node errors; and (iii) *real-error replay*, which inserts authentic wrong model responses into an exactly $k = 1$ field and varies only the prespecified operator design. Complete protocols, raw-log handling, uncertainty, and all cells appear in Appendices J–L.

5.1 Four theory-predicted phenomena

Synthetic fields have $d = 4$ typed fibers with transports $T_{ij} = M_j M_i^{-1}$. Margins are computed by enumerating every node support of size at most $2k$; no spectral surrogate is substituted for the defined quantity.

Table 1: Theory-predicted phenomena and synthetic verification.

Prediction	**Intervention**	**Result**
Consistency $\neq$ truth	Add a null-space shift	Defect 2.5×10^{-16} despite truth error 0.877
Anchor transition	Anchor six two-node components	$\gamma_1 : 0 \to 0.524$; exact recovery begins
Duplicate saturation	Repeat each relation $1 \to 16$ times	Normalized γ_2 stays at 0.309
Spectral difficulty	Vary graph, anchors, and k	Larger γ_k consistently yields lower error

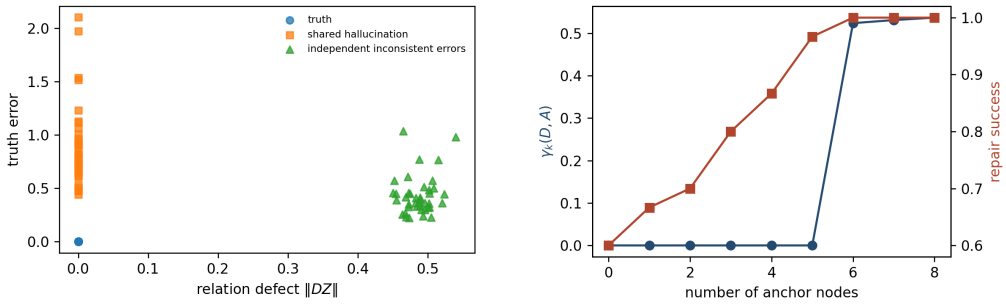

Figure 1: **Consistency blindness and anchor phase transition.** A shared transported hallucination has virtually zero defect despite large truth error (left). Recovery changes when anchors remove the final admissible sparse null direction (right).

Figure 1 gives the two qualitative discontinuities. Duplicate saturation distinguishes deterministic information from variance reduction: independent repeated calls may reduce observation noise, but literal copies do not alter the normalized Gram operator. Exact-search and convex repair are reported separately, because positive margin certifies the former while the latter additionally needs the group robust null-space property.

5.2 Real checkpoints and tasks

We query Qwen2.5-0.5B-Instruct, Phi-3-mini-4k-instruct, Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct (Yang et al., 2024a; Abdin et al., 2024). Mathematics has 128 integer-expression items with identity, scaling, and translation transformations. Code has 64 integer-function tasks with argument renaming, equivalent specifications, decomposition prompts, eight public tests, and 16 hidden tests. Each item produces four deterministic transformed responses; self-consistency and reflection logs support baselines. Recovery uses parsed outputs only, with no logits or model states.

Natural fields test external validity but often violate the $k = 1$ theorem through shared errors or missing correct candidates. Table 2 therefore reports candidate recall explicitly. We do not interpret failure outside the assumed class as evidence against or for the margin.

5.3 Cross-model, cross-task recovery difficulty

The primary predictive test replays *authentic* errors from the raw logs under an exactly controlled sparse model. For each eligible item, three nodes contain a response observed to be correct and one node contains a distinct wrong response produced by the same checkpoint. Code replays use real wrong programs whose public execution signature differs from a correct program. Ten operator designs range from disconnected zero-margin graphs to connected and anchored positive-margin graphs. Three matched-count triangle/isolate designs separate observability from relation and anchor counts. Responses are held fixed across designs.

Pooled rank association is 0.449 with cluster-bootstrap 95% interval $[0.417, 0.481]$. A grouped logistic model controls for checkpoint, task, raw error, relation count, and anchor count; adding standardized γ_1 improves held-out AUC by 0.033 and has coefficient 3.653. A 2,000-draw within-

Table 2: Natural real-model logs (percent). Math uses exact answer accuracy; code uses hidden pass@1. Methods have disclosed but unequal query budgets, so this table characterizes candidate generation rather than a compute-matched leaderboard.

Task	Model	Raw	Self-cons.	Rel.-major.	Reflect	Verifier	Cand. recall
Math.	Qwen2.5-0.5B	19.5	20.3	18.0	19.5	22.7	22.7
	Phi-3-mini	37.5	34.4	34.4	10.9	37.5	37.5
	Qwen2.5-7B	78.1	82.0	80.5	73.4	85.2	85.2
	Qwen2.5-14B	87.5	89.8	89.1	84.4	92.2	92.2
Code	Qwen2.5-0.5B	31.2	60.9	68.8	68.8	75.0	75.0
	Phi-3-mini	93.8	100.0	100.0	93.8	100.0	100.0
	Qwen2.5-7B	96.9	100.0	100.0	96.9	100.0	100.0
	Qwen2.5-14B	98.4	100.0	100.0	98.4	100.0	100.0

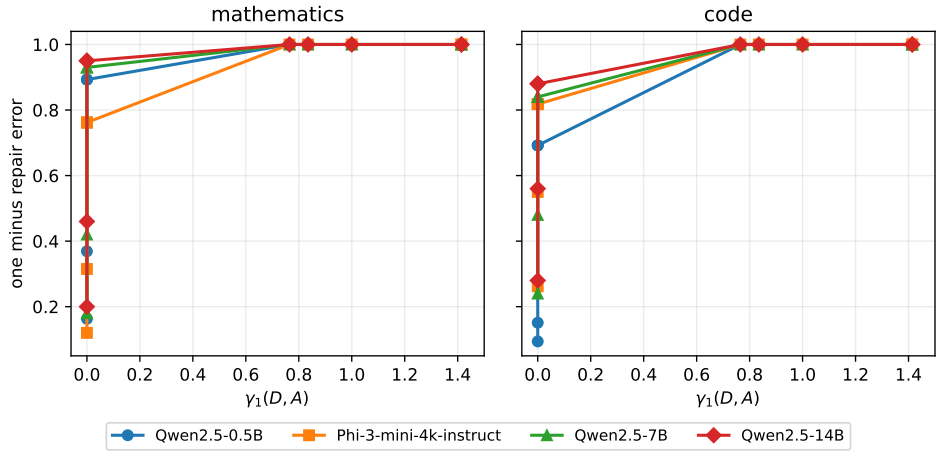

Figure 2: **One quantity across model and task strata.** Each point aggregates one prespecified relation–anchor design on real-error replays. Larger $\gamma_1(D, A)$ predicts lower full-field reconstruction error for all four checkpoints and both response spaces.

item permutation test gives $p = 0.00050$. These repeated-design analyses test whether the same pre-repair operator quantity orders field recovery after model and task change.

What this establishes. The theorem already establishes that γ_k is the worst-case condition number on the stated class. The replay experiment adds empirical content: authentic model errors in scalar mathematics and execution-signature code respond to operator interventions in the predicted order. It therefore supports the interpretation of $\gamma_k(D, A)$ as a measurable black-box recovery difficulty, not a post-hoc score of SPARSE-RRF.

What it does not establish. The controlled replay uses correctness labels to construct an evaluation stress test; it is not a deployable selector. Natural response fields can contain more than one wrong node, a coherent shared hallucination, a parser collision, or no correct candidate. Those failure categories appear separately in Appendix L. The result is cross-model and cross-task evidence across two families and two domains, not a universal scaling law for all LLMs.

6 DISCUSSION, SCOPE, AND LIMITATIONS

RRFs change the unit of analysis from an answer to a field of answers. The central theoretical object is not a consistency score but the restricted observability of the operator that makes sparse response errors visible. Three distinctions are essential. First, D tests relational compatibility while A attaches the field to truth; neither is interchangeable with the other. Second, γ_k characterizes information,

Table 3: Within-stratum rank prediction on authentic-error replays. All ten designs of an item stay in one group for inference.

Model–task stratum	Replay fields	Spearman ρ
Phi-3-mini–code	220	0.397
Qwen2.5-0.5B–code	650	0.506
Qwen2.5-7B–code	180	0.442
Qwen2.5-14B–code	110	0.418
Phi-3-mini–mathematics	1260	0.460
Qwen2.5-0.5B–mathematics	560	0.385
Qwen2.5-7B–mathematics	420	0.468
Qwen2.5-14B–mathematics	260	0.451
Pooled	3660	0.451

whereas null-space conditions characterize a particular tractable decoder. Third, raw edge count is not information: relation weights must be normalized so duplicated prompts do not inflate the margin.

6.1 In what sense is $\gamma_k(D, A)$ a new difficulty object?

The word *intrinsic* is conditional, not metaphysical. Once the response representation, typed transports, anchors, corruption budget, and measurement norm are fixed, $\gamma_k(D, A)$ depends on the observation problem and not on the recovery algorithm. It has four properties expected of a genuine condition number. It is *decisive*: positivity is equivalent to uniform noiseless identifiability. It is *quantitative*: the same scalar controls the sharp noise amplification rate. It is *unavoidable*: a matching two-point lower bound applies to every estimator, including procedures unrelated to RRF optimization. It is also *design-sensitive*: adding a genuinely informative measurement can increase the margin, whereas duplicating a normalized relation family cannot manufacture information. These statements jointly distinguish γ_k from a score chosen because it happens to correlate with one repair heuristic.

Cross-model and cross-task prediction is therefore a severe empirical test, not the definition of the object. Model identity changes the distribution, magnitude, and semantics of errors; mathematics and code use different fibers, transports, parsers, and correctness tests. The theory predicts that after these nuisance factors are stratified, fields with larger normalized margins should remain easier to reconstruct. Our authentic-error replay preserves real model mistakes while changing only which observable relation–anchor design is available. Grouped inference keeps all designs from one underlying item together. Thus the measured association cannot be created by leaking nearly identical designs across train and test folds. The increase in held-out discrimination after adding γ_k, together with positive within-stratum rank correlations, is the central evidence that the same mathematical object organizes recovery difficulty beyond one model–task pair. Natural generation remains separately reported because candidate absence and dense errors test a larger pipeline than the sparse-recovery theory.

This interpretation has explicit failure criteria. The proposal would be weakened if positive-margin fields routinely admitted two indistinguishable k-sparse explanations; that would contradict the identifiability theorem or expose a violated parser/operator assumption. It would also be weakened if, after normalization and within-item controls, γ_k carried no held-out predictive information in either domain; then the worst-case object could be mathematically valid but empirically uninformative for observed LLM errors. Conversely, a successful repair at one zero-margin instance does not refute the theory, because the impossibility result is uniform: it asserts the existence of indistinguishable alternatives, not failure on every favorable instance. These distinctions keep the central claim falsifiable without confusing a deterministic limit with an average-case performance law.

6.2 Separation from adjacent uncertainty notions

Self-consistency, semantic entropy, and verifier confidence summarize the distribution of generated answers (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024), whereas γ_k characterizes the observability of sparse response perturbations after the candidate field is fixed. These quantities are

therefore complementary: a model may be confidently and relationally consistent yet wrong along a direction in $\ker D$, while diverse surface responses may still be recoverable when their parsed field is well anchored. Moreover, γ_k is not a graph spectral gap or a model-level score. It is the restricted minimum singular value of the stacked relation–anchor operator $B = [D; A]$ over supports of size at most $2k$, and therefore depends on the typed transports, anchor placement, response representation, norm, and corruption budget. Two instances with the same untyped graph may have different margins, and a globally singular operator may still have positive γ_k if all of its null vectors are too dense to be admissible. The resulting certificate is only as reliable as the parser, relation specification, and anchors; externally auditable constraints such as symbolic substitution and code execution are therefore preferable to relations scored by another LLM.

Consequences for repair-system design. The margin separates three operational regimes. When $\gamma_k = 0$, recovery has an information deficit: no optimizer, reflection prompt, or additional iteration over the same measurements can provide a uniform guarantee. The appropriate action is acquisition, such as anchoring an uncovered component or adding a relation that intersects the sparse null witness. When γ_k is positive but comparable to the operator or residual uncertainty, the problem is identifiable yet noise-limited; conservative certificates, better anchors, and calibrated weighting matter more than a more elaborate decoder. When the margin is comfortably above uncertainty, remaining failures diagnose optimization, parsing, corruption-model mismatch, or candidate absence. This trichotomy converts a post-hoc repair score into a decision rule about whether to acquire information, improve measurement quality, or improve computation. Aggregate accuracy can reward a method for receiving easier designs. Reporting performance conditional on margin, candidate recall, and parser success separates algorithmic error from information-theoretic hardness. When a valid noise radius is available, $\gamma_k \|\widehat{e} - e^\star\|_2 / \epsilon$ measures the gap to the deterministic benchmark. Gains from a better decoder and gains from acquiring information are both useful, but they are different contributions. A minimizing support and singular vector also identify the nodes and semantic direction responsible for poor observability. Query selection can target this witness, recompute the margin, and stop when improvement saturates. This is the constructive use of $\gamma_k(D, A)$ as a condition number rather than a confidence feature. Several limitations delimit the claim. The global theorems are finite-dimensional and linear; nonlinear transports receive only a local Jacobian guarantee. Text parsing can collapse distinct meanings or separate equivalent ones. Sparse node corruption is appropriate for localized response failures but not for diffuse representation bias. Computing γ_k exactly is combinatorial; relaxations and lower certificates are needed for large fields. Finally, small mathematics and code experiments test mechanism and cross-domain prediction, not frontier-model leaderboard performance. They cannot establish that every natural-language task admits useful typed transports. Within those limits, Theorems 3–5 support a precise conclusion: for a specified black-box response representation, relation system, anchor system, and corruption budget, $\gamma_k(D, A)$ is the intrinsic worst-case condition number of recovery. A repair method can approach that limit or fail to do so, but it cannot evade it.

7 Conclusion

We introduced relational response fields as a general mathematical model for black-box LLM response recovery, in which typed relations and external anchors define a structured observation operator over a field of parsed responses. The central quantity, the restricted observability margin $\gamma_k(D, A)$, exactly characterizes uniform identifiability of k-node corruptions, determines the optimal deterministic stability scale, and appears in a matching minimax lower bound; hence its inverse is the intrinsic worst-case condition number of the specified recovery problem. This characterization also makes precise why consistency alone cannot certify truth: relation-only procedures are blind to sparse or coherent directions in $\ker D$, and only informative anchors or additional relations can remove these ambiguities. By separating information-theoretic recoverability from the stronger conditions required by tractable convex decoders, the framework distinguishes limitations of the observation design from failures of a particular repair algorithm. The accompanying experiments verify the predicted consistency–truth separation, anchor transition, redundancy saturation, and cross-model, cross-task ordering of recovery difficulty. More broadly, the results suggest that black-box reliability should be analyzed not only through model accuracy or heuristic confidence, but through the observability of the response system induced by queries, relations, parsers, and trusted evidence.

Ethics Statement

This work studies the reliability and recoverability of black-box large language model responses through theoretical analysis and controlled experiments on mathematics and code tasks. The study does not involve human subjects, private or personally identifiable information, or interventions on real users. The evaluated models and tasks are used solely for research purposes, and the proposed framework is intended to characterize and improve the reliability of model outputs. While response-repair techniques could in principle be applied in downstream systems, their guarantees depend on the validity of the specified parsers, relations, anchors, and corruption assumptions; accordingly, we do not claim that the framework alone certifies the factual or safety properties of deployed language-model systems.

Reproducibility Statement

We provide detailed information to support reproduction of both the theoretical and empirical results. The assumptions, definitions, and complete proofs of the main identifiability, stability, minimax, convex-recovery, and nonlinear-extension results are given in the appendices. The appendix further specifies the synthetic constructions, model checkpoints, mathematics and code tasks, prompts, parsing rules, execution procedures, baselines, operator designs, statistical analyses, robustness checks, and deterministic rerun procedures used in the experiments.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Francis R. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

Afonso S. Bandeira, Amit Singer, and Daniel A. Spielman. A cheeger inequality for the graph connection laplacian. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1611–1630, 2013.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

T Tony Cai, Lie Wang, and Guangwu Xu. New bounds for restricted isometry constants. *IEEE Transactions on Information Theory*, 56(9):4388–4394, 2010.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Tsong Y Chen, Shing C Cheung, and Shiu Ming Yiu. Metamorphic testing: a new approach for generating next test cases. *arXiv preprint arXiv:2002.12543*, 2020.

Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, TH Tse, and Zhi Quan Zhou. Metamorphic testing: A review of challenges and opportunities. *ACM Computing Surveys (CSUR)*, 51(1):1–27, 2018.

Fan R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best k-term approximation. *Journal of the American mathematical society*, 22(1):211–231, 2009.

David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Yonina C. Eldar, Patrick Kuppinger, and Helmut Bölcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58(6):3042–3054, 2010.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In *International Conference on Machine Learning*, 2024.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013.

Jakob Hansen and Robert Ghrist. Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology*, 3:315–358, 2019.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2021.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Martin Josifoski, Carlos Guestrin, Christopher Potts, and Matei Zaharia. DSPy: Compiling declarative language model calls into self-improving pipelines. In *International Conference on Learning Representations*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Yujia Li, David Choi, Junyoung Chung, et al. Competition-level code generation with AlphaCode. *Science*, 378(6624):1092–1097, 2022.

Hunter Lightman, Vineet Kosaraju, Yura Burda, et al. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Lek-Heng Lim. Hodge laplacians on graphs. *SIAM Review*, 62(3):685–715, 2020.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Ning Miao, Yee Whye Teh, and Tom Rainforth. Selfcheck: Using LLMs to zero-shot check their own step-by-step reasoning. In *International Conference on Learning Representations*, 2024.

Marcus J. Min, Yangruibo Ding, Luca Buratti, Saurabh Pujar, Gail Kaiser, Suman Jana, and Baishakhi Ray. Beyond accuracy: Evaluating self-consistency of code large language models with Identity-Chain. In *International Conference on Learning Representations*, 2024.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. CodeGen: An open large language model for code with multi-turn program synthesis. In *International Conference on Learning Representations*, 2023.

Antonio Ortega, Pascal Frossard, Jelena Kovačević, José M. F. Moura, and Pierre Vandergheynst. Graph signal processing: Overview, challenges, and applications. *Proceedings of the IEEE*, 106 (5):808–828, 2018.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In *Conference on Empirical Methods in Natural Language Processing*, pp. 7957–7968, 2023.

Shashank Ranjan and Mathukumalli Vidyasagar. Tight performance bounds for compressed sensing with conventional and group sparsity. *IEEE Transactions on Signal Processing*, 67(11):2854–2867, 2019.

Benjamin Ricaud, David I Shuman, and Pierre Vandergheynst. On the sparsity of wavelet coefficients for signals on graphs. 8858:422–428, 2013.

Michael Robinson. Sheaves are the canonical data structure for sensor integration. *Information and Inference: A Journal of the IMA*, 6(4):423–444, 2017.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*, 2023.

Aliaksei Sandryhaila and José M. F. Moura. Discrete signal processing on graphs. *IEEE Transactions on Signal Processing*, 61(7):1644–1656, 2013.

Sergio Segura, Gordon Fraser, Ana B. Sanchez, and Antonio Ruiz-Cortes. A survey on metamorphic testing. *IEEE Transactions on Software Engineering*, 42(9):805–824, 2016.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30(1):20–36, 2011.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations of large language models on reasoning and planning tasks. In *International Conference on Learning Representations*, 2025.

Joel A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, 2024b.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023b.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

Denny Zhou, Nathanael Schärli, Le Hou, et al. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023a.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *International Conference on Learning Representations*, 2023b.

A Supplement roadmap and claim ledger

This supplement is intentionally self-contained. It serves three audiences: readers familiar with sparse inverse problems but not black-box LLM evaluation; readers familiar with LLM consistency and verification but not restricted singular-value arguments; and readers seeking enough detail to reproduce every reported number. Table 4 maps each main-paper claim to its assumptions, proof, and empirical check. No theorem relies on an experimental observation, and no experimental result is presented as proof of a theorem whose assumptions cannot be checked.

Table 4: Claim ledger.

Claim	Proof location
Consistency cannot identify truth	Appendix D
Uniform k-error identifiability iff $\gamma_k > 0$	Appendix D
Stability scales as $1/\gamma_k$	Appendix D
No estimator beats $1/\gamma_k$ uniformly	Appendix E
Group-ℓ_1 recovery	Appendix F
Nonlinear local recovery	Appendix G
Duplicate relations saturate	Appendix C
Margin predicts across model/task	Appendix K

A.1 Reading order

Appendix B develops the notation from first principles. Appendix C studies the geometry of typed defects, gauge freedom, normalization, and computation of γ_k. Appendices D and E contain complete proofs of the information-theoretic results. Appendix F covers ideal and tractable decoders; Appendix G treats nonlinear transports. Appendix H asks how to choose queries and anchors. Appendix I gives a detailed comparison with neighboring theories and methods.

The remaining appendices specify experiments and extensions. Appendix J describes the synthetic construction and prespecified validation conditions. Appendix K gives prompts, parsing, execution, baselines, and statistics for mathematics and code. Appendix L contains extended tables and robustness checks. Appendix documents the artifact, environment, and deterministic rerun procedure. Appendix M works through concrete fields by hand. Appendix N inventories failure modes. Appendices O–Q develop stochastic-noise bounds, the grouped cross-domain inference protocol, and exact margins for canonical graph designs.

A.2 Scope of the phrase "intrinsic difficulty"

The phrase always refers to a fixed tuple

$$\mathfrak{I} = (\mathcal{H}, G, (T_e, w_e)_{e\in E}, A, k, \|\cdot\|_2),$$

and to uniform recovery over errors supported on at most k node groups. Change the representation, transports, weights, anchors, corruption model, or norm and the numerical margin changes. This dependence is a feature: recovery cannot be assigned a meaningful condition number without specifying what is observed and what alternatives must be distinguished. We do *not* claim that γ_k is an intrinsic property of a pretrained model in isolation, a universal scalar ranking all tasks, or a guarantee that semantic parsers are correct.

A.3 Notation summary

Vectors are elements of finite-dimensional real Hilbert spaces and are identified with coordinates only when needed. $B = [D; A]$ is the stacked relation–anchor operator. $S \subseteq V$ denotes a set of nodes, $\mathcal{H}_S = \bigoplus_{i\in S} \mathcal{H}_i$, and B_S is the restriction of B to $\mathcal{H}_S$. Singular values use Euclidean norms induced by the chosen inner products. The group $\ell_{2,1}$ norm is $\|h\|_{2,1} = \sum_i \omega_i \|h_i\|_2$. Constants named C may change between displays; theorem-specific constants receive subscripts.

B Mathematical foundations of relational response fields

This appendix develops the framework without assuming prior exposure to graph signal processing or compressed sensing. The purpose is not to reprove every standard theorem, but to expose each modeling choice used later.

B.1 Response spaces, parsing, and typed transports

Finite-dimensional fibers and direct sums Each query node i has a response space $\mathcal{H}_i$. The spaces may have different dimensions. For example, one node can encode a scalar final answer, another a

pair of subproblem answers, and a third a length-r execution signature. To compare all responses simultaneously, form the Hilbert direct sum

$$\mathcal{H} = \bigoplus_{i=1}^{n} \mathcal{H}_i = \{(z_1, \dots, z_n) : z_i \in \mathcal{H}_i\}, \qquad \langle z, h \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \langle z_i, h_i \rangle_{\mathcal{H}_i}.$$

The induced norm satisfies $\|z\|_2^2 = \sum_i \|z_i\|_2^2$. A node is the atomic corruption unit. The group support and group cardinality are

$$\operatorname{supp}_{\mathrm{g}}(h) = \{i \in V : \|h_i\|_2 > 0\}, \qquad \|h\|_{0,\mathrm{g}} = |\operatorname{supp}_{\mathrm{g}}(h)|.$$

This is not token sparsity. A corrupted program may alter many tokens while still occupying one node group; conversely, a one-token error in each of ten answers occupies ten groups. The choice matches the claim that at most k response instances, rather than k coordinates, are unreliable.

For $S \subseteq V$, let $P_S : \mathcal{H} \to \mathcal{H}$ retain groups in S and set all others to zero. We write $h_S = P_S h$ and $\mathcal{H}_S = P_S \mathcal{H}$. If $B : \mathcal{H} \to \mathcal{Y}$ is linear, B_S denotes B restricted to $\mathcal{H}_S$. In coordinates, B_S is formed by taking all columns belonging to node groups in S.

From text samples to a mathematical field The model output y_i is a string. A parser ϕ_i maps it to z_i. Examples include:

1. the last signed integer in a response;
2. a normalized rational number or symbolic expression;
3. a vector of truth values for atomic claims;
4. a vector of program outputs on a fixed test suite;
5. an embedding, provided distances and transports are specified independently of the answer being evaluated.

The parser can return a distinguished invalid symbol. In the finite-dimensional linear analysis we encode invalidity as an additional indicator coordinate or exclude the candidate and report parse failure separately. Silently coercing an invalid program to the zero vector would confound syntax failure with a valid zero-valued function.

The response field is random before sampling because $y_i \sim P_\theta(\cdot \mid x_i)$. The recovery theorems condition on the realized parsed field y. Randomness matters for average-case empirical performance but not for deterministic identifiability: if two fields give the same observation, no amount of estimator randomization distinguishes them on that realization.

Typed transports An edge $e = (i, j, t)$ specifies a relation type t and a map $T_e : \mathcal{H}_i \to \mathcal{H}_j$. The type records semantics; the map records its representation-level action. Several canonical cases are useful.

Equivalence. Two paraphrases should have the same canonical answer, hence $T_e = I$. If their raw representations differ, each node first maps to a shared canonical coordinate system.

Equivariance. If a transformation g acts on inputs and $\rho(g)$ acts on answers, then a valid model response should obey $z_{g \cdot x} = \rho(g) z_x$. Unit conversions, coordinate permutations, variable renaming, and sign reversal fit this pattern. A group representation is not required; it is enough to know the edge map actually used.

Decomposition. Suppose a parent answer is $z_i = (u, v)$ and a child query asks only for u. Then T_e is a projection. Conversely, an edge from children to a parent can be represented by augmenting the graph with an aggregation node whose fiber contains the joint answer. Directed multiedges permit more than one relation between the same pair.

Execution. Code strings are nonlinear objects, but their execution signatures on fixed tests are vectors. A semantics-preserving refactor uses $T_e = I$ in signature space. If a transformed function changes units or argument order, T_e permutes or transforms the signature accordingly.

Logic. For Boolean vectors, implication, negation, and conjunction are nonlinear over $\mathbb{R}$ in their raw form. One may use a suitable lifted representation, a discrete candidate decoder, or the local nonlinear theory. The global linear theorem applies only when the residual is genuinely linear in the chosen coordinates.

B.2 Defects, anchors, and the sparse corruption model

Defect operator as a weighted coboundary For each edge $e = (i, j)$, the residual is $r_e(z) = z_j - T_e z_i$. Stacking weighted residuals yields

$$D = \begin{bmatrix} \sqrt{w_{e_1}}[-T_{e_1} \;\; I] \\ \vdots \\ \sqrt{w_{e_m}}[-T_{e_m} \;\; I] \end{bmatrix},$$

with zero blocks in columns unrelated to the edge. The defect energy is

$$\mathcal{E}_D(z) = \frac{1}{2}\|Dz - q\|_2^2.$$

When $q = 0$, $\ker D$ is the space of globally compatible fields. With identity transports on an ordinary graph, D is a weighted incidence matrix and D^*D is the graph Laplacian. With general linear transports it is a connection- or sheaf-like coboundary; D^*D is a positive semidefinite relation Laplacian. We use this analogy for intuition but do not require a full sheaf structure.

The target q permits valid nonzero relation residuals. For example, if two outputs differ by a known offset c, define either an affine residual $z_j - T_e z_i - c$ or augment each fiber with a constant coordinate to recover linearity. Since identifiability concerns differences of candidates, affine offsets cancel and the same operator D governs the margin.

Anchors and truth attachment Relations compare answers; anchors provide information that is not generated by transporting another model answer. Formally an anchor is $Az \approx b$. Four cases recur:

1. **direct coordinate anchor:** a human or trusted source supplies z_i;
2. **functional anchor:** substitution into an equation produces a residual linear in the canonical answer;
3. **execution anchor:** hidden tests constrain a program's execution signature;
4. **probabilistic verifier:** a calibrated score is linearized or incorporated through the nonlinear extension.

An anchor need not reveal the complete answer. A row can constrain a projection or parity. What matters is whether the combined rows remove sparse null directions. An anchor generated by the same model without independent information should not automatically be treated as truth; at best it is another noisy relation with potentially correlated errors.

Corruption, relation mismatch, and observation noise Let $z^\star$ be the target field. The parsed field is $y = z^\star + e^\star$. The ideal sparse model assumes $\|e^\star\|_{0,\mathrm{g}} \le k$. Relations satisfy $Dz^\star = q + \xi_D$ and anchors satisfy $Az^\star = b + \xi_A$. Then

$$\begin{bmatrix} Dy - q \\ Ay - b \end{bmatrix} = \begin{bmatrix} D \\ A \end{bmatrix} e^\star + \begin{bmatrix} \xi_D \\ \xi_A \end{bmatrix}. \tag{11}$$

This derivation is worth stating because it prevents a sign ambiguity. If repair is parameterized as $z = y + \Delta$, then $\Delta = -e^\star$ and the right-hand side changes sign. The implementation fixes one convention and tests it against known synthetic errors.

Relation mismatch and observation noise are mathematically indistinguishable after stacking, but scientifically different. A large ξ_D may indicate a bad transport, whereas a large ξ_A may indicate an unreliable verifier. The artifact records them separately even when the theorem uses their combined norm.

Why the support size is $2k$ Suppose two hypotheses e_1, e_2 each corrupt at most k nodes. Their difference $h = e_1 - e_2$ is supported on $\mathrm{supp}_{\mathrm{g}}(e_1) \cup \mathrm{supp}_{\mathrm{g}}(e_2)$, which can contain $2k$ nodes. Any uniform uniqueness condition must therefore exclude nonzero null vectors up to size $2k$. Replacing $2k$ by k would certify only uniqueness relative to zero, not pairwise uniqueness among all k-sparse alternatives.

The factor is tight. Let h be a null vector supported on exactly $2k$ nodes. Partition its support into S_1, S_2 of size k and set $e_1 = h_{S_1}$, $e_2 = -h_{S_2}$. Then $e_1 - e_2 = h$ and $Be_1 = Be_2$. Neither candidate violates the budget.

B.3 Sparse inverse-problem interpretation and canonical examples

Relation to classical sparse inverse problems After Equation (4), the algebra resembles compressed sensing (Donoho, 2006; Candès et al., 2006; Foucart & Rauhut, 2013). The distinctions are in how B is built and interpreted:

- columns are grouped by transformed query rather than arbitrary signal coordinates;
- rows are typed relations and truth anchors rather than generic measurements;
- $\ker D$ has semantic gauge directions, including coherent hallucinations;
- normalized duplicate queries must not create fictitious information;
- the quantity is evaluated as an instance-level difficulty predictor across black-box model outputs.

The mathematics should therefore be understood as a specialized inverse-problem theory for a new response object, not as a claim that restricted singular values themselves were unknown.

A complete two-node example Let two equivalent queries have scalar outputs. Then

$$D = [-1 \quad 1].$$

The consistent fields are (c, c), so $\ker D = \mathrm{span}\{(1, 1)\}$. If $k = 1$, the difference of two one-node errors can use both nodes; $\gamma_1(D, 0) = 0$ because $(1, 1)$ is two-node sparse. Anchor the first node:

$$A = [1 \quad 0], \qquad B^*B = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}.$$

Now every nonzero vector is observed, and

$$\gamma_1(D, A) = \sigma_{\min}(B) = \sqrt{\frac{3 - \sqrt{5}}{2}} > 0.$$

If both outputs share the same additive hallucination c, $D(c, c) = 0$ but $A(c, c) = c$. The anchor does not improve consistency; it makes the consistent error observable.

A component-size subtlety Anchors are not always necessary for sparse recovery. Consider an identity-transport connected component of size $m > 2k$. Its constant null vector occupies all m nodes and therefore is excluded from the $2k$-sparse set. The relation operator can identify every k-sparse error even though it cannot identify an arbitrary dense shift. Conversely, an unanchored component of size at most $2k$ contributes a sparse null vector and forces $\gamma_k = 0$. The experiments choose component sizes deliberately when testing anchor phase transitions.

C Geometry of typed relation and anchor operators

This appendix studies properties of D, A, $B = [D; A]$, and γ_k that are used in proofs and experiment design.

C.1 Laplacian geometry, gauge freedom, and monotonicity

Relation Laplacian and energy The operator $L_D = D^*D$ is positive semidefinite because

$$\langle z, L_D z \rangle = \|Dz\|_2^2 = \sum_{e=(i,j)} w_e \|z_j - T_e z_i\|_2^2 \geq 0.$$

Its nullspace equals $\ker D$. Adding anchors gives

$$L_B = B^*B = D^*D + A^*A, \qquad \langle z, L_B z\rangle = \|Dz\|^2 + \|Az\|^2.$$

The unrestricted smallest eigenvalue of L_B measures global observability. RRF recovery needs a *restricted* version because only differences of sparse corruptions matter. For a support S,

$$\lambda_{\min}(B_S^* B_S) = \sigma_{\min}(B_S)^2,$$

and γ_k^2 is the minimum of these eigenvalues over $|S| \le 2k$.

Gauge directions A gauge direction is a nonzero h with $Dh = 0$. For identity transports on a connected graph, gauges are constants. For invertible transports that are path-consistent, choose a reference node r and maps M_i satisfying $T_{ij} = M_j M_i^{-1}$. Every gauge has form

$$h_i = M_i u, qquadu \in \mathcal{H}_r.$$

Thus the nullspace dimension equals the latent fiber dimension for each connected component. A cycle inconsistency can reduce the nullspace: transporting around a cycle constrains u to the fixed space of the holonomy product. The general framework does not assume path consistency, so the nullspace is computed from D itself.

An anchor removes a gauge h exactly when $Ah \neq 0$. Directly anchoring one full-rank node in a path-consistent component removes all dense gauge directions in that component. Partial anchors may remove only a subspace. Sparse recovery is less demanding: it requires removal only of gauges and cancellations whose support is at most $2k$.

Proposition 7 (Group-spark characterization)**.** *Define* $\mathrm{spark}_{\mathrm{g}}(B) = \min\{\|h\|_{0,\mathrm{g}} : h \neq 0, Bh = 0\}$, *with value* $+\infty$ *if* B *is injective. Then*

$$\gamma_k(B) > 0 \quad \Longleftrightarrow \quad \mathrm{spark}_{\mathrm{g}}(B) > 2k.$$

Proof. If a null vector has support at most $2k$, it is feasible in the definition of γ_k and gives zero. Conversely, if every such restricted kernel is trivial, each finite-dimensional B_S is injective and has positive smallest singular value. Taking the minimum over finitely many supports preserves positivity. □

Basic monotonicity properties

Proposition 8 (Budget monotonicity)**.** *If* $k_1 \le k_2$, *then* $\gamma_{k_1}(B) \ge \gamma_{k_2}(B)$.

Proof. The feasible set of perturbations for k_1 is contained in that for k_2. □

Proposition 9 (Row augmentation)**.** *Let* C *be any additional linear observation and* $\widetilde{B} = [B; C]$. *Then* $\gamma_k(\widetilde{B}) \ge \gamma_k(B)$.

Proof. For every h, $\|\widetilde{B}h\|^2 = \|Bh\|^2 + \|Ch\|^2 \ge \|Bh\|^2$. □

This proposition uses unnormalized energy. If all rows are globally renormalized after augmentation, the numerical margin can decrease. We therefore state the weighting convention whenever comparing designs.

C.2 Normalization, coordinates, and margin computation

Why duplicate normalization is necessary Suppose one relation row block is R with total reliability weight w. Duplicating it m times without reweighting changes its Gram contribution from wR^*R to mwR^*R and increases singular values by up to $\sqrt{m}$ despite observing no new fact. Under family normalization, copy r receives weight w/m, so

$$\sum_{r=1}^{m} (\sqrt{w/m}R)^* (\sqrt{w/m}R) = wR^*R.$$

Hence D^*D and every γ_k are exactly invariant. Approximate duplicates are grouped using a prespecified relation-family identifier, not post-hoc similarity of outputs. Independent transformations receive separate reliability budgets.

There are two legitimate alternatives. If duplicate model calls have independent measurement noise, averaging them can reduce noise variance; this improves the numerator ϵ in the stability ratio rather than the deterministic operator rank. If repeated calls have partially independent transports or parsers, they are not literal duplicates and their distinct rows may improve γ_k. The experiment reports both operator information and sampling variance so these effects are not conflated.

Scaling and coordinate dependence The numerical value of a singular value depends on units. If one response coordinate is measured in meters and another in millimeters, naive Euclidean norms distort the margin. We fix inner products before computing γ_k. In experiments, scalar answers are canonicalized to the original query's units, execution signatures are binary coordinates with equal weights, and relation families receive total unit weight. More generally choose positive definite metrics $W_{\mathcal{H}}$ and $W_{\mathcal{Y}}$ and define

$$\gamma_k^{(W)}(B) = \min_{\|h\|_{0,g} \le 2k} \frac{\|W_{\mathcal{Y}}^{1/2} B h\|_2}{\|W_{\mathcal{H}}^{1/2} h\|_2}.$$

Equivalently whiten domain and codomain. Cross-task comparison is meaningful only after such normalization.

Exact computation For small n and k, exact computation enumerates supports:

1. for $s = 1, \dots, \min(2k, n)$, enumerate $S \subseteq V$ with $|S| = s$;
2. form B_S by selecting all coordinates in node groups S;
3. compute $\sigma_{\min}(B_S)$ with a dense or sparse SVD;
4. return the minimum and a witness support/vector.

The number of supports is $\sum_{s=1}^{2k} \binom{n}{s}$. This is feasible for the paper's graphs ($n \le 16$, $k \le 2$) and deliberately avoids using a relaxation to test a theorem stated in terms of the exact quantity. Numerical zero is declared only when the singular value is below an absolute and relative tolerance, and the witness residual is checked directly.

Algorithm 2 Exact group-sparse margin and witness

Require: Matrix B, group column sets $\{J_i\}_{i=1}^n$, budget k.
$\gamma \leftarrow +\infty$, $(S_\star, h_\star) \leftarrow \varnothing$.
for $s = 1$ to $\min(2k, n)$ **do**
  for each $S \subseteq [n]$ with $|S| = s$ **do**
    compute a smallest right singular pair (σ, v) of B_{J_S};
    if $\sigma < \gamma$ **then**
      $(\gamma, S_\star, h_\star) \leftarrow (\sigma, S, \mathrm{embed}_S(v))$;
    end if
  end for
end for
verify $\|h_\star\|_2 = 1$, $\|Bh_\star\|_2 \approx \gamma$, and $|\operatorname{supp}_{\mathrm{g}}(h_\star)| \le 2k$.
return $(\gamma, S_\star, h_\star)$.

Certificates for larger fields Exact group spark and restricted singular values are combinatorial. Three kinds of certificates remain useful at scale.

1. **Support-conditioned certificate.** If a localization procedure proposes a support family $\mathcal{S}$, compute $\min_{S \in \mathcal{S}} \sigma_{\min}(B_S)$. This is exact only relative to $\mathcal{S}$.
2. **Coherence bound.** After group whitening, block coherence can lower-bound restricted eigenvalues through Gershgorin-type arguments. The bound is conservative but cheap.

3. **Mixed-integer search.** Encode support selection with binary variables and optimize a semidefinite or mixed-integer relaxation. A valid lower bound certifies stability even if it is not tight.

We do not use a heuristic estimate as if it were exact. Main-paper margins come from enumeration.

C.3 Operator perturbations, anchor placement, and worked spectra

Perturbing the operator Let the intended operator be B and the implemented operator be $\widetilde{B} = B + E$. For every support S, Weyl's inequality gives

$$|\sigma_{\min}(\widetilde{B}_S) - \sigma_{\min}(B_S)| \le \|E_S\|_{2\to 2} \le \|E\|_{2\to 2}.$$

Taking minima yields

$$\gamma_k(\widetilde{B}) \ge \gamma_k(B) - \|E\|_{2\to 2}. \tag{12}$$

Thus a positive empirical margin is robust only when it exceeds plausible parser and transport perturbations. We report sensitivity sweeps rather than treating a barely positive floating-point value as a scientific phase transition.

Anchor placement as restricted observability For a candidate set of anchor rows $\{a_j\}$ and selected set Q, write $B(Q) = [D; A_Q]$. The design objective

$$\max_{|Q|\le q} \gamma_k(B(Q))$$

directly targets worst-case recoverability. It differs from maximizing rank, determinant, or the unrestricted smallest eigenvalue: those criteria may improve dense directions while leaving a sparse null direction untouched. Anchor placement is generally combinatorial. Appendix H develops greedy surrogates and counterexamples.

A four-node worked spectrum Consider a chain of four scalar equivalent responses with unit edge weights and an anchor on node one. Then

$$B = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

For $k = 1$, enumerate singleton and pair supports. A singleton interior node has column norm $\sqrt{2}$; endpoint four has norm 1; adjacent pairs have correlated columns; the worst pair is obtained from the smallest singular value of

$$\begin{bmatrix} 0 & 0 \\ -1 & 1 \\ 0 & -1 \\ 0 & 0 \end{bmatrix},$$

which is positive. Removing the anchor does not create a two-sparse null vector because the chain's constant null vector occupies four nodes. Therefore $\gamma_1(D, 0) > 0$ even though D is globally singular. For $k = 2$, the four-node constant vector is feasible, so $\gamma_2(D, 0) = 0$. This example shows why corruption budget and global rank answer different questions.

D Identifiability, consistency blindness, and stability

We prove the main information-theoretic statements in full. Throughout this section, all spaces are finite-dimensional, $B = [D; A]$, and the corruption groups are the node fibers $\mathcal{H}_i$.

D.1 Consistency blindness and sparse support algebra

A decision-theoretic form of consistency blindness Theorem 2 can be strengthened from a deterministic statement about equal inputs to a lower bound for randomized procedures.

Lemma 10 (Indistinguishable fields)**.** *Let* K *be any Markov kernel that maps an observed relation defect* u *to a distribution over field estimates. If* $h \in \ker D$*, then the output laws of* K *on fields* z *and* $z + h$ *are identical.*

Proof. Linearity gives $D(z + h) = Dz + Dh = Dz$. A Markov kernel depends on the field only through its supplied observation. Equal observations induce equal output distributions. □

Proposition 11 (Two-field error under consistency-only observation)**.** *For* $h \in \ker D \setminus \{0\}$ *and any estimator* $\widehat{z} = \widehat{z}(Dz)$*,*

$$\max\left\{\mathbb{E}_z\|\widehat{z} - z\|_2, \mathbb{E}_{z+h}\|\widehat{z} - (z + h)\|_2\right\} \geq \frac{1}{2}\|h\|_2.$$

Proof. By Lemma 10, both expectations integrate over the same output law. For every realization u of the estimator, the triangle inequality gives

$$\|h\|_2 = \|(u - z) - (u - z - h)\|_2 \leq \|u - z\|_2 + \|u - z - h\|_2.$$

Taking expectation and then the larger of the two risks proves the claim. □

Proof of Theorem 2. The equality of observations follows from Lemma 10. If the admissible class contains both z and $z + h$, an estimator must receive the same input on two distinct truths. Proposition 11 quantifies the resulting error. A transported shared hallucination is a particular $h \in \ker D$: for path-consistent transports $T_{ij} = M_j M_i^{-1}$, set $h_i = M_i u$ for any nonzero latent shift u. □

The theorem does not say that consistency has no value. It says that consistency alone cannot distinguish equivalence classes $z + \ker D$. Relations can localize deviations transverse to the nullspace, and anchors can select a representative within an equivalence class.

Support algebra

Lemma 12 (Difference support)**.** *For arbitrary* $u, v \in \mathcal{H}$*,*

$$\mathrm{supp}_{\mathrm{g}}(u - v) \subseteq \mathrm{supp}_{\mathrm{g}}(u) \cup \mathrm{supp}_{\mathrm{g}}(v).$$

Consequently, if $\|u\|_{0,\mathrm{g}}, \|v\|_{0,\mathrm{g}} \leq k$*, then* $\|u - v\|_{0,\mathrm{g}} \leq 2k$*.*

Proof. If i belongs to neither support, then $u_i = v_i = 0$ and $(u - v)_i = 0$. Cardinality of the union is at most the sum of cardinalities. □

Lemma 13 (Partition of a sparse vector)**.** *If* $h \in \mathcal{H}$ *has* $\|h\|_{0,\mathrm{g}} \leq 2k$*, then there exist* $u, v \in \mathcal{H}$ *with* $\|u\|_{0,\mathrm{g}}, \|v\|_{0,\mathrm{g}} \leq k$ *and* $h = u - v$*.*

Proof. Partition $S = \mathrm{supp}_{\mathrm{g}}(h)$ into disjoint S_1, S_2 with $|S_1|, |S_2| \leq k$. Set $u = h_{S_1}$ and $v = -h_{S_2}$. Then $u - v = h_{S_1} + h_{S_2} = h$. □

The two lemmas explain both directions of the $2k$ condition: differences of feasible errors are $2k$-sparse, and every $2k$-sparse ambiguity can be represented as a difference of feasible errors.

D.2 Identifiability and exact sparse recovery

Equivalence between nullspace, spark, and positive margin For a support S, define the restricted unit sphere

$$\mathbb{S}_S = \{h \in \mathcal{H}_S : \|h\|_2 = 1\}.$$

It is compact. The continuous function $h \mapsto \|Bh\|_2$ therefore attains its minimum on $\mathbb{S}_S$, equal to $\sigma_{\min}(B_S)$.

Lemma 14 (Finite-support compactness)**.** *The infimum in Definition 1 is attained. Moreover,* $\gamma_k(B) = 0$ *iff some nonzero* $h \in \ker B$ *satisfies* $\|h\|_{0,\mathrm{g}} \leq 2k$*.*

Proof. There are finitely many supports $S \subseteq V$ with $1 \le |S| \le 2k$. On each support the minimum is attained on $\mathbb{S}_S$. The minimum over the finite family is attained. It is zero iff one restricted minimum is zero, which holds iff B_S has a nontrivial null vector. □

Proof of Theorem 3. We establish a cycle of implications.

$(i) \Rightarrow (ii)$ by contraposition. Suppose $0 \neq h \in \ker B$ and $\|h\|_{0,\mathrm{g}} \le 2k$. Lemma 13 writes $h = u - v$ with u, v each k-sparse. Since $B(u - v) = 0$, $Bu = Bv$. The two distinct feasible errors violate uniform uniqueness.

$(ii) \Rightarrow (i)$. Suppose e_1, e_2 are k-sparse and $Be_1 = Be_2$. Then $h = e_1 - e_2 \in \ker B$ and Lemma 12 gives $\|h\|_{0,\mathrm{g}} \le 2k$. Assumption (ii) forces $h = 0$, hence $e_1 = e_2$.

$(ii) \Leftrightarrow (iii)$ is Lemma 14. Finally, by definition $\mathrm{spark}_{\mathrm{g}}(B)$ is the smallest support size of a nonzero null vector, so $(ii) \Leftrightarrow (iv)$. □

Exact decoder

Corollary 15 (Correctness of the ideal decoder)**.** *Let $s = Be^\star$, $\|e^\star\|_{0,\mathrm{g}} \le k$, and $\gamma_k(B) > 0$. Every minimizer of*

$$\min_e \|e\|_{0,\mathrm{g}} \quad \textit{subject to } Be = s$$

equals $e^\star$.

Proof. The true error is feasible, so a minimizer has support at most k. Theorem 3 gives uniqueness within that class. □

If $\gamma_k = 0$, failure is worst-case, not universal: a particular $e^\star$ may still be uniquely recoverable. The theorem is a uniform statement over all supports and amplitudes.

D.3 Stability, misspecification, and theorem scope

Stability under bounded observation noise

Lemma 16 (Restricted lower inequality)**.** *For every h satisfying $\|h\|_{0,\mathrm{g}} \le 2k$,*

$$\|Bh\|_2 \ge \gamma_k(B)\|h\|_2.$$

Proof. The result is homogeneous. It is trivial for $h = 0$; otherwise normalize $h/\|h\|_2$ and apply Definition 1. □

Proof of Theorem 4. Set $h = \widehat{e} - e^\star$. Lemma 12 gives $\|h\|_{0,\mathrm{g}} \le 2k$. Since $s = Be^\star + \xi$ and $\|\xi\|_2 \le \epsilon$,

$$\|Bh\|_2 \le \|B\widehat{e} - s\|_2 + \|s - Be^\star\|_2 \le 2\epsilon.$$

Lemma 16 yields $\gamma_k\|h\|_2 \le 2\epsilon$. Dividing by positive γ_k proves Equation (6). Because $\widehat{z} = y - \widehat{e}$ and $z^\star = y - e^\star$, their difference is $-(\widehat{e} - e^\star)$ and has the same norm. □

Corollary 17 (Separate relation and anchor tolerances)**.** *Suppose the true and estimated errors have relation residual tolerances $\epsilon_D^\star, \widehat{\epsilon}_D$ and anchor tolerances $\epsilon_A^\star, \widehat{\epsilon}_A$. Then*

$$\|\widehat{e} - e^\star\|_2 \le \frac{\sqrt{(\epsilon_D^\star + \widehat{\epsilon}_D)^2 + (\epsilon_A^\star + \widehat{\epsilon}_A)^2}}{\gamma_k(B)}.$$

Proof. Bound the relation and anchor blocks separately by triangle inequalities and apply the Euclidean norm of the stacked vector. □

Approximate sparsity and model mismatch Theorem 4 assumes both truth and estimate are k-sparse. If $e^\star$ has a small diffuse tail, let $e^s tar_k$ be a best k-group approximation and write $r = e^\star - e^\star_k$. The observation becomes

$$s = Be^s tar_k + (Br + \xi).$$

Any k-sparse feasible estimate at tolerance $\epsilon + \|Br\|$ obeys

$$\|\widehat{e} - e^s tar_k\|_2 \le \frac{2(\epsilon + \|Br\|_2)}{\gamma_k(B)},$$

and therefore

$$\|\widehat{e} - e^\star\|_2 \le \frac{2(\epsilon + \|Br\|_2)}{\gamma_k(B)} + \|r\|_2.$$

This is an oracle-style bound; tractable convex decoders obtain more familiar tail terms under a robust null-space property in Appendix F.

Misspecified corruption budget If the true support size is $k_\star$ but the analysis uses $k < k_\star$, uniform identification is not guaranteed and the tail acts as mismatch. If $k \ge k_\star$, the theorem remains valid but $\gamma_k \le \gamma_{k_\star}$ can be smaller, producing a more conservative certificate. Reporting a favorable margin at a budget smaller than the plausible number of corrupted nodes is invalid. The experimental artifact computes margins for a range of budgets and marks the one used for each repair.

Random estimators and probability statements The deterministic bound holds conditionally on any event where the noise norm and support constraints are satisfied. If ξ is random and $\mathbb{P}(\|\xi\| \le \epsilon_\delta) \ge 1 - \delta$, then

$$\mathbb{P}\left(\|\widehat{e} - e^\star\| \le 2\epsilon_\delta/\gamma_k\right) \ge 1 - \delta$$

for every k-sparse feasible estimator. This conversion introduces no independence assumption between relation and anchor noise; only a valid bound on their stacked norm is needed.

What the theorem does not imply Positive γ_k does not imply that majority vote succeeds, that a local gradient method finds the ideal sparse solution, or that the parser corresponds to semantic truth. It says that the supplied measurements contain enough information to distinguish all k-sparse alternatives, with a quantified condition number. Algorithmic and representation errors remain separate terms.

E Minimax optimality and modulus of continuity

This appendix proves that the inverse dependence on γ_k cannot be improved by a different estimator. We use deterministic norm-bounded noise because it directly matches the stability guarantee and requires no distributional assumptions.

E.1 Decision setup and the restricted modulus of continuity

Parameter class and observation balls For radius $R > 0$, define

$$\mathcal{E}_k(R) = \{e \in \mathcal{H} : \|e\|_{0,\mathrm{g}} \le k,\ \|e\|_2 \le R\}.$$

Under parameter e, the set of admissible observations is the closed ball

$$\mathcal{O}_\epsilon(e) = \{Be + \xi : \|\xi\|_2 \le \epsilon\}.$$

An estimator is any function, deterministic or randomized, from the observation space to $\mathcal{H}$. Randomization cannot improve the two-point lower bound under norm loss, so we state the proof for deterministic outputs and condition on internal randomness if present.

Two parameter values are observationally confusable exactly when their balls intersect. In a Hilbert space,

$$\mathcal{O}_\epsilon(e_1) \cap \mathcal{O}_\epsilon(e_2) \ne \varnothing \quad \Longleftrightarrow \quad \|B(e_1 - e_2)\|_2 \le 2\epsilon. \tag{13}$$

The forward direction follows by the triangle inequality. For the reverse direction, the midpoint $u = (Be_1 + Be_2)/2$ lies within half the separation of both centers.

Restricted modulus Define the ambiguity modulus

$$\omega_k(B; R, \epsilon) := \sup \{\|e_1 - e_2\|_2 : e_1, e_2 \in \mathcal{E}_k(R),\ \|B(e_1 - e_2)\|_2 \le 2\epsilon\}.$$

This quantity is an exact geometric measure of the largest pair that noise can make indistinguishable.

Proposition 18 (Modulus upper bound)**.** *If $\gamma_k(B) > 0$, then*

$$\omega_k(B; R, \epsilon) \le \min\left\{2R, \frac{2\epsilon}{\gamma_k(B)}\right\}.$$

Proof. Any difference $h = e_1 - e_2$ has group support at most $2k$, so $\|Bh\| \ge \gamma_k\|h\|$. The confusability constraint gives $\|h\| \le 2\epsilon/\gamma_k$. The radius constraints independently give $\|h\| \le \|e_1\| + \|e_2\| \le 2R$. □

Proposition 19 (Constructive modulus lower bound)**.** *For every B, k, R, ϵ,*

$$\omega_k(B; R, \epsilon) \ge \min\left\{R, \frac{2\epsilon}{\gamma_k(B)}\right\},$$

where $2\epsilon/0 = +\infty$.

Proof. By Lemma 14, choose a unit vector $h_\star$ with group support at most $2k$ and $\|Bh_\star\| = \gamma_k$. Lemma 13 gives $h_\star = u - v$ with u, v each k-sparse and $\|u\|, \|v\| \le 1$ because they occupy disjoint subsets of the coordinates of a unit vector.

Let $a = \min\{R, 2\epsilon/\gamma_k\}$, with $a = R$ if $\gamma_k = 0$. Set $e_1 = au$ and $e_2 = av$. Both belong to $\mathcal{E}_k(R)$. Their separation is $\|e_1 - e_2\| = a\|h_\star\| = a$, while

$$\|B(e_1 - e_2)\| = a\gamma_k \le 2\epsilon.$$

They are therefore feasible in the supremum and establish the bound. □

The factor-of-two gap in the radius term arises because a general pair in the radius-R ball can be $2R$ apart, whereas splitting an arbitrary $2k$ witness only guarantees that each part has norm at most the witness norm. The ϵ/γ_k scaling, which is the scientific claim, is exact up to universal constants.

E.2 Two-point lower bounds and the main minimax theorem

Two-point minimax lemma

Lemma 20 (Intersecting observations imply error)**.** *If $e_1, e_2 \in \mathcal{E}_k(R)$ have intersecting observation balls, then every estimator $\widehat{e}$ satisfies*

$$\sup_{e \in \{e_1, e_2\}} \sup_{u \in \mathcal{O}_\epsilon(e)} \|\widehat{e}(u) - e\|_2 \ge \frac{1}{2}\|e_1 - e_2\|_2.$$

Proof. Choose a common observation $u \in \mathcal{O}_\epsilon(e_1) \cap \mathcal{O}_\epsilon(e_2)$. For $a = \widehat{e}(u)$,

$$\|e_1 - e_2\| \le \|a - e_1\| + \|a - e_2\|.$$

At least one term is at least half the left-hand side. The same argument holds after conditioning on estimator randomization; averaging cannot reduce both expected distances below half their separation. □

Proof of the main minimax theorem

Proof of Theorem 5. For the lower bound, Proposition 19 constructs e_1, e_2 separated by

$$a = \min\{R, 2\epsilon/\gamma_k\}$$

with intersecting observation balls. Lemma 20 gives minimax risk at least $a/2$.

For the upper bound, the zero estimator has worst-case risk at most R. Independently, choose a k-sparse residual-minimizing estimator

$$\widehat{e} \in \arg\min_{\|e\|_{0,\mathrm{g}} \le k} \|Be - s\|_2.$$

The true $e^\star$ attains residual at most ϵ, so the minimizer also has residual at most ϵ. Theorem 4 gives risk at most $2\epsilon/\gamma_k$. Selecting the better of the two estimators proves the upper bound $\min\{R, 2\epsilon/\gamma_k\}$. □

E.3 Degenerate regimes, stochastic noise, and interpretation

Zero margin and unbounded ambiguity When $\gamma_k = 0$, Lemma 14 supplies a nonzero $2k$-sparse $h \in \ker B$. Scaling h does not change its observation. If no amplitude radius is imposed, split th into two k-sparse candidates for arbitrary t; their separation grows without bound while observations remain equal. Thus worst-case risk is infinite even with $\epsilon = 0$. With radius R, Theorem 5 yields at least $R/2$.

This is stronger than saying that a particular algorithm fails. It says the data do not contain enough information for any procedure to guarantee recovery on the class.

Gaussian-noise corollary The deterministic result has a standard stochastic analogue. Suppose $s = Be + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2 I)$. Take two candidates whose mean separation in observation space has Kullback–Leibler divergence

$$\mathrm{KL}(P_{e_1} \| P_{e_2}) = \frac{\|B(e_1 - e_2)\|_2^2}{2\sigma^2}$$

bounded by a constant. Scaling the minimizing direction to $\|B(e_1 - e_2)\| \asymp \sigma$ and applying Le Cam's method gives expected error at least $c \min\{R, \sigma/\gamma_k\}$ for a universal $c > 0$. A full constant optimization is unnecessary for the paper's deterministic experiments; the purpose of the corollary is to show that the same condition number appears under familiar stochastic noise.

Why average-case model accuracy cannot replace the margin Model accuracy describes a distribution over errors. The minimax margin describes the geometry of what the observation operator can distinguish. A highly accurate model can occasionally fail exactly along a nearly invisible direction and be unrecoverable; a weaker model with independent localized errors can be easy to repair if γ_k is large. Conversely, a small γ_k does not predict that typical errors must be large; it predicts that some admissible errors are hard. Empirical prediction is therefore assessed after controlling for raw error scale and model/task strata, and is presented as evidence about how often natural errors align with difficult directions, not as a logical consequence of the minimax theorem.

F Algorithmic theory and optimization details

Information-theoretic identifiability does not specify how to compute the recovered field. This appendix separates three decoders: exact support search, convex group-sparse recovery, and discrete candidate-field optimization.

F.1 Ideal and convex sparse decoders

The ideal residual decoder Given $s = Be^\star + \xi$ and noise radius ϵ, define

$$\widehat{e}_0 \in \arg\min_e \|e\|_{0,\mathrm{g}} \quad \text{subject to} \quad \|Be - s\|_2 \le \epsilon. \tag{14}$$

If the true error is k-sparse, it is feasible. Any solution with support at most k satisfies Theorem 4. Exact support search solves

$$\min_{S \subseteq V,\ |S| \le k} \ \min_{e \in \mathcal{H}_S} \|Be - s\|_2.$$

For fixed S, the inner minimizer is a least-squares solution $e_S = B_S^\dagger s$ and the residual is $\|(I - B_S B_S^\dagger)s\|$. Enumerating all supports costs $O(\binom{n}{k})$ least-squares problems. The paper uses this decoder only where $n \le 16$ and $k \le 2$.

Exact search has two roles. It is an estimator whose guarantee matches the information-theoretic theorem, and it is a diagnostic separating operator difficulty from optimization failure. If exact search fails when γ_k is large, the cause is noise, model mismatch, candidate parsing, or an incorrect implementation; it cannot be blamed on a convex relaxation.

Group norms Let positive weights ω_i correct for group dimension or prior reliability. Define

$$\|e\|_{2,1,\omega} = \sum_{i=1}^{n} \omega_i \|e_i\|_2, \qquad \sigma_k(e)_{2,1,\omega} = \inf_{\|v\|_{0,\mathrm{g}} \le k} \|e - v\|_{2,1,\omega}.$$

With equal-dimensional groups and no prior preference, $\omega_i = 1$. If group dimensions differ, $\omega_i = \sqrt{\dim \mathcal{H}_i}$ prevents large groups from being selected merely because they contain more coordinates. We never tune group weights on test labels.

The constrained convex decoder is

$$\widehat{e}_1 \in \arg\min_{e} \|e\|_{2,1,\omega} \quad \text{subject to} \quad \|Be - s\|_2 \le \epsilon. \tag{15}$$

Its penalized counterpart is

$$\widehat{e}_\lambda \in \arg\min_{e} \frac{1}{2}\|Be - s\|_2^2 + \lambda \|e\|_{2,1,\omega}. \tag{16}$$

These are group basis-pursuit denoising and group Lasso formulations (Yuan & Lin, 2006; Bach, 2008; Eldar et al., 2010).

Group robust null-space property

Definition 21 (Group robust NSP)**.** The operator B satisfies the ℓ_2 group robust null-space property of order k with constants $0 < \rho < 1$ and $\tau > 0$ if, for every $h \in \mathcal{H}$ and every $S \subseteq V$ with $|S| \le k$,

$$\|h_S\|_2 \le \frac{\rho}{\sqrt{k}} \|h_{S^c}\|_{2,1} + \tau \|Bh\|_2. \tag{17}$$

For unequal group weights, the cardinality factors are replaced by the corresponding weighted compatibility constants.

The null-space property controls dense kernel vectors whose mass is concentrated on a small set. Such vectors do not violate group spark, yet they can cause a group-ℓ_1 decoder to prefer the wrong representation.

Example 22 (Positive margin does not certify group ℓ_1)**.** Let $k = 1$ and choose a 2×3 matrix whose one-dimensional kernel is spanned by $h = (1, 0.4, 0.4)$. Every two columns are linearly independent, so no nonzero two-sparse null vector exists and $\gamma_1 > 0$. However, for $S = \{1\}$, $\|h_S\|_1 = 1 > 0.8 = \|h_{S^c}\|_1$; the exact null-space property fails. Therefore some one-sparse signal is not recovered by ℓ_1 minimization. Grouping scalar coordinates separately gives the RRF analogue.

Theorem 23 (Stable convex recovery)**.** *Assume Definition 21. Let $s = Be^\star + \xi$ with $\|\xi\|_2 \le \epsilon$, and let $\widehat{e}_1$ solve Equation (15). Then*

$$\|\widehat{e}_1 - e^\star\|_2 \le C_1(\rho) \frac{\sigma_k(e^\star)_{2,1}}{\sqrt{k}} + C_2(\rho, \tau)\epsilon. \tag{18}$$

One admissible choice under the unweighted convention is

$$C_1 = \frac{2(1+\rho)^2}{1-\rho}, \qquad C_2 = \frac{2\tau(3+\rho)}{1-\rho},$$

and sharper constants follow from the standard block-sparse proof.

Proof outline with the key inequalities. Set $h = \widehat{e}_1 - e^\star$ and let S index the k groups of $e^\star$ with largest norms. Feasibility gives $\|Bh\|_2 \le 2\epsilon$. Optimality of $\widehat{e}_1$ and decomposability of the group norm give the cone inequality

$$\|h_{S^c}\|_{2,1} \le \|h_S\|_{2,1} + 2\|e^\star_{S^c}\|_{2,1}. \tag{19}$$

Apply the group robust NSP to S, use $\|h_S\|_{2,1} \le \sqrt{k}\|h_S\|_2$, and solve the resulting inequality for $\|h_S\|_2$ and $\|h_{S^c}\|_{2,1}$. Partition S^c into blocks of k groups in decreasing norm and use the block Stechkin inequality to control $\|h_{S^c}\|_2$ by $\|h_{S^c}\|_{2,1}/\sqrt{k}$ plus the first block. Combining terms yields Equation (18). The argument is the group analogue of robust sparse recovery proofs in Foucart & Rauhut (2013); Ranjan & Vidyasagar (2019). □

This theorem proves Proposition 6. In exact k-sparsity, the approximation term vanishes. The stability constant is governed by the algorithmic NSP constants, not solely by γ_k.

F.2 Field-space objectives, optimization, and tuning

Penalized objective in field coordinates Write $z = y - e$. Expanding Equation (16) gives

$$\min_z \frac{1}{2} \left\| \begin{bmatrix} Dz - q \\ Az - b \end{bmatrix} \right\|_2^2 + \lambda \sum_i \omega_i \|z_i - y_i\|_2.$$

This is Sparse RRF Repair: it seeks a field that fits typed relations and anchors while changing few response nodes. The edit penalty is groupwise because changing one answer representation is one event. A coordinatewise ℓ_1 penalty is available when within-node sparsity is scientifically meaningful, but its recovery budget differs from the paper's k.

Proximal-gradient solver Let $f(e) = \frac{1}{2}\|Be - s\|^2$ and $g(e) = \lambda \sum_i \omega_i \|e_i\|_2$. The gradient $\nabla f(e) = B^*(Be - s)$ is L-Lipschitz with $L = \|B\|_{2\to 2}^2$. For step $0 < \eta \le 1/L$,

$$e^{t+1} = \mathrm{prox}_{\eta g}\left(e^t - \eta B^*(Be^t - s)\right).$$

The proximal map is group soft thresholding:

$$[\mathrm{prox}_{\eta g}(u)]_i = \left(1 - \frac{\eta\lambda\omega_i}{\|u_i\|_2}\right)_+ u_i,$$

with output zero when $u_i = 0$. Standard convex analysis yields

$$F(e^t) - F(e_\lambda^\star) \le \frac{\|e^0 - e_\lambda^\star\|_2^2}{2\eta t}.$$

FISTA acceleration improves the objective gap to $O(t^{-2})$. The implementation uses a power iteration estimate of L, multiplies by a safety factor, checks monotonic objective decrease, and stops only when both relative objective change and the proximal-gradient mapping are small.

Regularization and validation The penalty λ trades measurement fit against edit sparsity. In synthetic experiments with known noise, we set λ from a calibration grid and freeze it before test trials. In black-box tasks, selection uses only training/calibration fields and observable residuals; correctness labels are reserved for evaluation. We report sensitivity across a logarithmic grid. Choosing λ separately for every test answer by looking at truth would turn the repair method into an oracle.

The constrained decoder has the conceptual advantage of an explicit noise radius. The penalized decoder is easier to optimize. A Pareto curve maps one parameterization to the other under convexity, but the correspondence can be nonunique. The artifact records the actual parameter rather than implying equivalence at an unspecified value.

F.3 Discrete repair, diagnostics, complexity, and black-box scope

Discrete candidate-field decoder Numerical and code outputs often live in a discrete semantic space. Let $\mathcal{C}_i = \{c_{i1}, \ldots, c_{im_i}\} \subset \mathcal{H}_i$ be parsed candidates from model samples and transported candidates. Define

$$\widehat{z} \in \operatorname*{arg\,min}_{z_i \in \mathcal{C}_i} \frac{1}{2}\|Dz - q\|_2^2 + \frac{\beta}{2}\|Az - b\|_2^2 + \mu \sum_i \ell_i(z_i; y_i), \tag{20}$$

where ℓ_i penalizes deviation from the raw response without consulting ground truth. Exact enumeration is used for tiny candidate products. For pairwise edge energies, min-sum message passing, integer programming, or branch-and-bound is available.

The candidate decoder has a basic limitation: it cannot output truth if neither truth nor a transport-equivalent candidate is in the candidate closure. We therefore report *candidate recall*, the fraction of instances whose candidate closure contains a correct field. Repair accuracy is bounded above by this quantity; omitting it can make selection performance look worse than the candidate generator or, conversely, hide an oracle candidate injection.

Code repair in execution-signature space For code, the optimizer selects among complete generated programs rather than editing arbitrary signature coordinates. Each program is executed on public relation tests and, for evaluation only, hidden tests. The public signature supplies D and designated anchor constraints; hidden tests compute pass rate. Two programs with the same public signature are observationally equivalent even if their source differs. The source-level modified-token count is measured after selection and does not enter γ_k.

Execution failure, timeout, and unsafe behavior are distinct categorical outcomes. The runner executes generated code in a subprocess with a wall-clock limit, restricted builtins, no network, and a temporary directory. Invalid programs receive an invalid candidate marker; they are not assigned the all-zero functional signature.

Optimization diagnostics Every run records:

1. primal objective and residual norm;
2. number of active node groups;
3. proximal-gradient mapping norm or exact support optimality gap;
4. computed γ_k and witness support;
5. candidate recall for discrete tasks;
6. wall-clock time and random seed.

These diagnostics distinguish a low-margin instance from a solver that stopped early. The acceptance report fails if an algorithmic comparison mixes unconverged and converged runs.

Computational complexity One proximal step costs one multiplication by B and B^*, linear in the number of nonzero relation blocks. Memory is likewise sparse. Exact γ_k computation dominates small experiments at $O(\sum_{s\le 2k}\binom{n}{s})$ restricted SVDs. Candidate-field enumeration costs $\prod_i m_i$ in the worst case; pairwise structure permits dynamic programming only on low-treewidth graphs. None of these procedures requires model gradients or weights after responses have been logged.

Black-box boundary The model is used only to produce strings at requested queries. Parsing, transport, anchor evaluation, margin computation, and repair operate externally. The artifact includes model identifiers and generation code but excludes model weights. A hosted API can replace a local model without changing the inverse problem, provided the response log and decoding metadata are retained.

G Local theory for nonlinear transports and parsers

Many useful response relations are nonlinear. This appendix gives a local result based on a restricted Jacobian margin. It does not claim global recovery for arbitrary semantic maps.

G.1 Nonlinear measurements and local regularity assumptions

Nonlinear measurement map Let $\mathcal{M}:\mathcal{H}\to\mathcal{Y}$ stack relation and anchor residuals:

$$\mathcal{M}(z)=\begin{bmatrix}(\sqrt{w_e}\,r_e(z))_{e\in E}\\ a(z)\end{bmatrix}.$$

For linear transports, $\mathcal{M}(z)=Bz-c$. For nonlinear transports, $r_e(z)=z_j-T_e(z_i)$ and a may contain nonlinear verification residuals. Suppose $\mathcal{M}$ is Fréchet differentiable near the target $z^\star$ with Jacobian $J_\star=D\mathcal{M}(z^\star)$.

Define the local restricted Jacobian margin

$$\gamma_k(J_\star)=\min_{0<\|h\|_{0,\mathrm{g}}\le 2k}\frac{\|J_\star h\|_2}{\|h\|_2}.$$

This is the linear margin of the tangent inverse problem. A positive value is not enough by itself: nonlinear remainder terms may cancel the linear signal outside a sufficiently small neighborhood.

Quadratic remainder assumption

Assumption 24 (Restricted second-order remainder)**.** There exist $L \geq 0$ and $r > 0$ such that for every h with $\|h\|_{0,\mathrm{g}} \leq 2k$ and $\|h\|_2 \leq r$,

$$\|\mathcal{M}(z^\star + h) - \mathcal{M}(z^\star) - J_\star h\|_2 \leq \frac{L}{2}\|h\|_2^2. \tag{21}$$

The condition follows from a Lipschitz Jacobian on the relevant sparse line segments. It is stated only on sparse directions because dense curvature is irrelevant to the local k-error class.

Theorem 25 (Local restricted injectivity)**.** *Under Assumption 24, if $\gamma_k(J_\star) > Lr/2$, then for every nonzero $2k$-sparse h with $\|h\|_2 \leq r$,*

$$\|\mathcal{M}(z^\star + h) - \mathcal{M}(z^\star)\|_2 \geq \left(\gamma_k(J_\star) - \frac{Lr}{2}\right)\|h\|_2 > 0. \tag{22}$$

Hence $\mathcal{M}$ is injective from $z^\star$ along the specified sparse neighborhood.

Proof. The reverse triangle inequality, Definition of $\gamma_k(J_\star)$, and Equation (21) give

$$\|\mathcal{M}(z^\star + h) - \mathcal{M}(z^\star)\| \geq \|J_\star h\| - \frac{L}{2}\|h\|^2 \geq \left(\gamma_k(J_\star) - \frac{L}{2}\|h\|\right)\|h\|.$$

Use $\|h\| \leq r$. □

G.2 Local identifiability, stability, and certificates

Local stability

Corollary 26 (Nonlinear local stability)**.** *Suppose $\widehat{z} - z^\star$ is $2k$-sparse, has norm at most r, and both fields fit the same nonlinear observation to tolerance ϵ. If $\gamma_k(J_\star) > Lr/2$, then*

$$\|\widehat{z} - z^\star\|_2 \leq \frac{2\epsilon}{\gamma_k(J_\star) - Lr/2}.$$

Proof. Their measurement difference is at most 2ϵ. Apply Equation (22) to $h = \widehat{z} - z^\star$ and divide. □

This bound has the expected structure: the effective margin is the tangent margin minus a curvature penalty. If r approaches $2\gamma_k/L$, the certificate vanishes. The result is local and centered at $z^\star$; an implementable certificate may use a candidate point and an operator perturbation bound.

Candidate-centered certificate Suppose the Jacobian is evaluated at $\widetilde{z}$ rather than the unknown truth and

$$\|D\mathcal{M}(\widetilde{z}) - J_\star\|_{2\to 2} \leq \delta_J.$$

Weyl's inequality implies

$$\gamma_k(J_\star) \geq \gamma_k(D\mathcal{M}(\widetilde{z})) - \delta_J.$$

A valid local certificate therefore requires

$$\gamma_k(D\mathcal{M}(\widetilde{z})) > \delta_J + Lr/2.$$

The terms δ_J and L must come from analytic bounds or held-out perturbation tests; setting them to zero because the Jacobian is convenient would turn a local approximation into an unsupported global claim.

G.3 Examples, sparse Gauss–Newton repair, and failure regimes

Nonlinear examples

Multiplicative relations. If responses are positive scalars and a relation requires $z_j = z_i^c$, work in log coordinates to obtain the linear relation $\log z_j = c \log z_i$. This is preferable to local linearization when positivity and numerical stability are controlled.

Boolean constraints. A clause residual such as $r(z) = z_1 z_2 - z_3$ is polynomial. Around a binary candidate, its Jacobian may have a positive restricted margin, but alternative binary assignments can exist outside the local ball. A discrete candidate decoder gives a global finite search and should be preferred.

Program behavior. Source-to-signature execution is discontinuous, so a Jacobian theorem is inappropriate. Map each complete program to a discrete signature and optimize over candidates. The local theory applies only to a differentiable surrogate representation, not to execution itself.

Semantic embeddings. If T_e is a learned semantic transport, its Jacobian and curvature inherit model uncertainty. A large numerical margin can be meaningless if the embedding collapses truth distinctions. External task anchors and parser audits are therefore required.

Implicit-function perspective Classical local identifiability often follows from an injective Jacobian. RRF recovery needs only restricted injectivity: $J_\star$ may have dense null directions while remaining injective on all $2k$-node subspaces. Theorem 25 is a quantitative sparse inverse-function statement with an explicit neighborhood. It does not require $J_\star$ to be square or globally full rank.

Gauss–Newton sparse repair For nonlinear residuals, one practical iteration linearizes at z^t:

$$\Delta^t \in \arg\min_{\Delta} \frac{1}{2}\|\mathcal{M}(z^t) + J_t\Delta\|_2^2 + \lambda\|\Delta\|_{2,1}, \qquad z^{t+1} = z^t + \alpha_t\Delta^t.$$

A line search enforces decrease in the true nonlinear objective. Local convergence requires regularity beyond the main theorem: stable support identification, bounded Jacobian variation, and suitable step acceptance. The paper uses this method only in an auxiliary curvature experiment and does not cite the linear γ_k theorem as a global convergence guarantee.

Failure outside the local region Consider $\mathcal{M}(z) = \sin z$ in one dimension at $z^\star = 0$. The Jacobian margin is one, yet $\mathcal{M}(0) = \mathcal{M}(\pi) = 0$. No positive tangent margin rules out the distant alternative. Similarly, semantic transformations can have multiple globally compatible answer fields. Anchors or a restricted candidate class are required to exclude them. This elementary example is why the paper repeatedly uses "local" for nonlinear statements.

H Designing queries, relations, and anchors

The margin is not merely diagnostic. It supplies an objective for choosing which transformed queries and external checks to acquire under a budget.

H.1 Design objectives and structural limits

Design problem Let D_0 contain mandatory relations, and let $\{C_j\}_{j\in\mathcal{J}}$ be candidate relation or anchor row blocks with costs $c_j > 0$. For $Q \subseteq \mathcal{J}$, define

$$B(Q) = \begin{bmatrix} D_0 \\ (C_j)_{j\in Q} \end{bmatrix}.$$

The robust design problem is

$$\max_{Q\subseteq\mathcal{J}} \gamma_k(B(Q)) \quad \text{subject to} \quad \sum_{j\in Q} c_j \le C. \tag{23}$$

This criterion targets the worst sparse ambiguity. A design may have high total rank, trace, or determinant while leaving one $2k$-node support nearly invisible; Equation (23) directly penalizes that failure.

Monotonicity but not generic submodularity With fixed row weights, adding a block cannot decrease γ_k. However, the set function $Q \mapsto \gamma_k(B(Q))$ need not be submodular. Two rows can be individually useless on different null directions but jointly remove the final ambiguity, creating complementarity rather than diminishing returns. Consequently the classical $(1 - 1/e)$ greedy guarantee does not follow without additional structure.

Example 27 (Complementary anchors)**.** Let a two-dimensional latent gauge $u = (u_1, u_2)$ be invisible to D. Candidate anchor C_1 observes only u_1 and C_2 only u_2. The unrestricted minimum singular value remains zero after either one alone and becomes positive after both. The marginal gain of the second anchor is larger when the first has already been selected, violating diminishing returns.

H.2 Witness-guided coverage and informative measurements

Worst-witness greedy design Exact enumeration returns not only γ_k but a minimizing support and unit witness h_t. A natural adaptive rule chooses the row block that sees this witness most strongly:

$$j_t \in \arg\max_{j \notin Q_t} \frac{\|C_j h_t\|_2^2}{c_j}. \tag{24}$$

Then set $Q_{t+1} = Q_t \cup \{j_t\}$ and recompute the worst witness. This procedure resembles cutting-plane methods: each acquisition attacks the currently least observable direction. It is not globally optimal in general, but every selected row has an interpretable purpose and monotonic improvement can be checked exactly on small graphs.

Algorithm 3 Witness-guided relation–anchor acquisition

Require: Base operator D_0, candidate blocks C_j, costs c_j, budget C, sparsity k.
1: $Q \leftarrow \varnothing$.
2: **while** a feasible candidate remains **do**
3: compute (γ, h) attaining or approximating $\gamma_k(B(Q))$;
4: choose feasible j maximizing $\|C_j h\|^2 / c_j$;
5: $Q \leftarrow Q \cup \{j\}$;
6: **end while**
7: **return** Q, the margin trajectory, and all witness supports.

Component coverage theorem for identity transports Consider scalar identity transports on an undirected graph. Each connected component C contributes a constant null vector $\mathbf{1}_C$. Direct full-coordinate anchors remove that vector if at least one node in C is anchored.

Proposition 28 (Small-component anchor necessity)**.** *If an unanchored connected component C has $|C| \le 2k$, then $\gamma_k(D, A) = 0$. If every unanchored component has size greater than $2k$ and D has no other null directions, then no* null *vector is $2k$-sparse; hence $\gamma_k > 0$.*

Proof. For the first claim, $\mathbf{1}_C$ is a nonzero null vector of D and A supported on at most $2k$ nodes. For the second, every nonzero null vector is constant and nonzero on at least one entire unanchored component, hence has support greater than $2k$. Apply the group-spark characterization. □

The proposition explains the synthetic phase transition: six disjoint two-node components with $k = 1$ require one anchor per component. The sixth anchor removes the final two-node gauge and changes γ_1 from zero to positive.

Independent relations versus paraphrase volume Suppose a set of paraphrases all induces the same canonical equality row after parsing. Under family normalization, their deterministic information is unchanged. An independent relation changes the row space: scaling, inverse checking, decomposition, or an execution constraint may observe a direction that equality misses. The design objective therefore favors diversity in operator action, not surface-form diversity alone.

This does not imply that paraphrase sampling is useless. It can estimate stochastic model variability, increase candidate recall, and reduce variance when calls are independent. Those benefits enter the error distribution and noise radius. They are distinct from increasing the deterministic restricted singular value.

H.3 Reliability-aware deployment and design diagnostics

Reliability-aware design Candidate rows differ in noise. If row block C_j has covariance Σ_j, whiten it as $\Sigma_j^{-1/2} C_j$ when the covariance is estimable and nonsingular. Then γ_k measures signal

relative to noise. With uncertain covariance, use conservative weights and propagate an operator perturbation interval. An extremely strong but biased verifier should not receive infinite weight; it can create a large computed margin around the wrong anchor target.

One robust formulation maximizes a worst-case margin over an uncertainty set $\mathcal{U}_j$:

$$\max_{Q} \min_{\widetilde{C}_j \in \mathcal{U}_j,\, j \in Q} \gamma_k \left([D_0; (\widetilde{C}_j)_{j \in Q}] \right).$$

Exact solution is difficult, but Equation (12) supplies a conservative lower certificate by subtracting operator-norm uncertainty.

Model-adaptive versus model-agnostic design A model-agnostic design chooses transformations and anchors before seeing outputs. It tests whether the operator geometry predicts outcomes across model families. A model-adaptive design can use observed residuals to decide which relation to query next. The latter may be more efficient but introduces selection dependence. Our primary cross-model experiment uses fixed design schedules; witness-guided acquisition is evaluated separately so that predictive claims are not driven by peeking at correctness.

Stopping rules A principled acquisition loop can stop when one of three conditions holds:

1. a certified lower bound on γ_k exceeds the target $2\epsilon/\delta$, guaranteeing error at most δ for the ideal decoder;
2. candidate recall or parser validity, rather than observability, becomes the bottleneck;
3. the marginal gain per cost falls below a prespecified threshold.

Stopping because the current candidate has low relation defect is unsafe: a nullspace hallucination can have zero defect before the field is identifiable.

A counterexample to degree-based anchoring High-degree nodes are tempting anchor targets. Consider two components: a dense clique of ten nodes and a two-node pair, with $k = 1$. The highest-degree node lies in the clique, but the clique's constant gauge is ten-node dense and does not threaten γ_1. The unanchored pair contributes a two-node null vector and forces $\gamma_1 = 0$. Anchoring either node of the pair is optimal; anchoring another clique node leaves the margin zero. Degree and centrality do not replace sparse observability.

Design as an experimental variable The paper varies anchor coverage, relation independence, duplicate count, and corruption budget to manipulate γ_k without changing the underlying ground-truth field. This intervention is stronger evidence than a passive correlation: the theory predicts in advance which design changes should alter the margin and which should not. Cross-model and cross-task prediction then asks whether the same geometric quantity orders empirical repair difficulty after these interventions.

I Detailed relation to prior theories and black-box methods

This appendix positions RRFs carefully. The paper does not claim the first use of consistency, graph constraints, metamorphic transformations, sparse recovery, or restricted singular values. Its claim is that black-box LLM response families form a useful recovery object and that the relation–anchor restricted margin is the exact condition number for sparse recovery of that object.

I.1 Mathematical foundations and neighboring inverse-problem theories

Compressed sensing and sparse inverse problems Compressed sensing studies recovery of sparse vectors from underdetermined linear measurements (Donoho, 2006; Candès et al., 2006). Uniform uniqueness is governed by spark or absence of sparse null vectors; stable and tractable recovery uses restricted isometry, restricted eigenvalue, coherence, or null-space properties (Tropp, 2006; Candes, 2008; Bickel et al., 2009; Foucart & Rauhut, 2013). Group sparsity replaces coordinate support with predefined blocks (Yuan & Lin, 2006; Bach, 2008; Eldar et al., 2010). Our proof architecture deliberately follows this hierarchy:

1. group spark and $\gamma_k > 0$ characterize information-theoretic uniqueness;
2. the restricted lower singular value gives deterministic stability for sparse feasible estimates;
3. a group robust null-space property gives a convex-decoder theorem;
4. minimax two-point arguments show the inverse-margin dependence is unavoidable.

Statistical variants include the Dantzig selector, best k-term approximation, sharp restricted-isometry analysis, and decomposable regularizers (Candes & Tao, 2007; Cohen et al., 2009; Cai et al., 2010; Negahban et al., 2012). We cite these to locate the mathematical ingredients; none studies transformed black-box LLM response fields.

The difference lies in the construction and interpretation of the signal and operator. The unknown is not a parameter vector internal to the LLM. It is a corruption field over externally queried responses. Measurements are not arbitrary Gaussian projections; they are typed transformations and truth anchors whose nullspace has semantic meaning. Duplicate query normalization, shared hallucination gauges, and anchor placement are therefore central rather than incidental.

Calling γ_k a new theoretical object requires precision. As a formula it belongs to the family of restricted minimum singular values. As an ***RRF recoverability margin*** it is a new task-conditioned object: its arguments are the typed defect and anchor operators of a black-box response instance, and the paper proves and tests its role as a model- and task-transferable predictor. We do not rename classical linear algebra and claim its invention.

Inverse-problem conditioning Classical inverse problems distinguish identifiability, stability, and regularization (Engl et al., 1996). The smallest singular value is the condition number of a full linear inverse. RRFs require restricted conditioning because global gauge directions can be dense and irrelevant under a sparse corruption model. The key example is a connected relation graph with a dense constant null vector: B is singular globally but may be injective on every $2k$-node subspace. Conversely, a full-rank operator can have a tiny restricted singular value and be practically unstable.

The minimax result is a modulus-of-continuity argument for the restricted parameter class. Its value is conceptual: it rules out the possibility that a more sophisticated LLM workflow can overcome a missing relation-anchor direction without adding information or narrowing the error class.

Graph signal processing Graph signal processing places values on graph vertices and studies smoothness, spectra, filtering, and sampling through graph Laplacians (Ricaud et al., 2013; Sandryhaila & Moura, 2013; Ortega et al., 2018). With identity transports, an RRF is a graph signal and D^*D is a graph Laplacian. Typed transports generalize equality across nodes. RRF recovery differs from standard smooth graph-signal denoising in four ways:

- the clean field satisfies typed, often exact equivariances rather than generic low-frequency smoothness;
- errors are grouped by response node and may have arbitrary amplitude;
- anchors attach relational equivalence classes to externally checkable truth;
- the primary quantity is the worst restricted observability of $[D; A]$, not the unrestricted Laplacian spectral gap.

The ordinary spectral gap can still be informative for dense perturbations, but it need not detect a worst sparse support. Classical spectral graph theory and Hodge-Laplacian analysis provide additional context for unrestricted Laplacian geometry (Chung, 1997; Lim, 2020).

Connection Laplacians, synchronization, and sheaves Synchronization estimates latent group elements from noisy pairwise ratios; connection Laplacians transport vectors between local coordinate systems (Singer, 2011; Bandeira et al., 2013). Cellular sheaves formalize data assigned to cells with restriction maps and define sheaf Laplacians whose kernel contains globally consistent sections (Hansen & Ghrist, 2019; Robinson, 2017). Typed RRF defects have the same algebraic flavor: responses live in node fibers and edge maps compare them.

We avoid claiming that D is always a cellular-sheaf coboundary. RRF edges can be directed, fibers can differ, transports can be noninvertible, and nonlinear or affine relations may be handled by lifting rather than by a sheaf. Sheaf language is exact only under compatible restriction-map definitions. The

RRF framework is intentionally operational: any externally auditable residual map can be included, and the linear theory applies to its stacked operator.

The novelty relative to synchronization is also in the target. We do not estimate one latent group action or a globally dense assignment from pairwise ratios. We repair a sparse subset of black-box model responses and ask which relation/anchor designs make all such corruptions distinguishable.

Error-correcting interpretation The map $e \mapsto Be$ acts like a syndrome map. A $2k$-sparse null vector is analogous to a codeword of too-small distance: two k-error patterns produce the same syndrome. Group spark plays the role of minimum distance, and γ_k is a robust Euclidean distance. This analogy explains why the difference of two k-error patterns appears and why exact consistency is insufficient when a low-weight gauge exists.

RRFs differ from classical fixed codes because the parity checks are designed from semantic transformations and can be acquired adaptively. Anchors are systematic truth checks rather than additional model-generated parity alone. The analogy is used for intuition; the paper does not claim coding-theoretic optimality of the query graphs.

I.2 Black-box LLM consistency, verification, and search methods

Metamorphic testing Metamorphic testing addresses the oracle problem by applying transformations for which relations among outputs are known (Chen et al., 2020; Segura et al., 2016; Chen et al., 2018). It is a direct precursor to typed query edges. Traditional metamorphic testing usually reports whether individual source/follow-up pairs violate a relation. RRFs aggregate many typed relations into a field-level inverse problem, add anchors, characterize indistinguishable corruption patterns, and optimize a sparse repair rather than only detecting violations.

The relation is complementary. Metamorphic-testing expertise is required to specify valid T_e; RRF theory says what the resulting collection can identify. A false metamorphic relation is operator misspecification, not evidence of a model error. The artifact therefore records relation generation separately from model querying.

Self-consistency and sampling-based aggregation Self-consistency samples multiple reasoning paths and selects a majority answer (Wang et al., 2023). It is effective when errors are sufficiently independent and the correct answer has the largest mass. RRFs differ in both data and claim. Nodes are typed transformations, not only i.i.d. samples of one prompt, and transports can imply non-identity output changes. More importantly, Theorem 2 identifies a failure mode that majority cannot resolve: a systematic wrong answer transported coherently across queries.

Semantic uncertainty clusters answers by meaning to avoid treating paraphrases as distinct outcomes (Kuhn et al., 2023; Farquhar et al., 2024). SelfCheckGPT uses disagreement among black-box samples to detect factual hallucination (Manakul et al., 2023). These methods quantify uncertainty or inconsistency. RRF recoverability additionally asks whether the available relations and anchors make the correct field identifiable and stable. A field can have low semantic entropy and zero relational defect while being wrong.

Verification, reflection, and self-correction Outcome and process verifiers rerank generated solutions (Cobbe et al., 2021; Lightman et al., 2023). Self-refinement and Reflexion ask a model to critique and revise its output (Madaan et al., 2023; Shinn et al., 2023); SelfCheck constructs explicit checking prompts (Miao et al., 2024). Empirical studies show that self-verification can be limited or correlated with generation errors (Huang et al., 2024; Stechly et al., 2025).

In RRF terminology, a sound external verifier supplies anchor rows or an anchor loss. A self-critique generated by the same model is not automatically an anchor; absent independent evidence it is another response node or noisy relation. This distinction formalizes why reflection can repeat a shared misconception. It also allows hybrid systems: a verifier score can be an anchor, transformed responses supply D, and sparse repair combines them.

Prompt and workflow optimization Automatic prompt engineering and prompt optimization search for instructions that improve expected task performance (Zhou et al., 2023b; Pryzant et al.,

2023; Yang et al., 2024b; Fernando et al., 2024). Declarative pipeline optimizers such as DSPy tune prompts and demonstrations for composed systems (Khattab et al., 2024). Their optimization variables are prompts, demonstrations, or workflow modules.

RRF repair holds the black-box model and query family fixed and optimizes the realized response field. The two approaches are compatible: prompt optimization can design node queries, while RRF theory evaluates whether their relation-anchor operator has useful sparse observability. A prompt set with high validation accuracy can still have a low margin on a particular instance, and a margin-improving query may not increase average model accuracy. They answer different questions.

Search over reasoning structures Chain-of-thought, least-to-most prompting, Tree of Thoughts, and ReAct create structured trajectories (Wei et al., 2022; Zhou et al., 2023a; Yao et al., 2023a;b). A reasoning tree can be converted into an RRF if edges carry externally specified relations between parsed node outputs. Merely linking chronological steps is not enough: D requires a testable transport. RRFs therefore provide a possible analysis layer for search methods but do not replace their proposal mechanism.

Code generation and execution Code generation has unusually strong anchors because programs can be executed (Chen et al., 2021; Li et al., 2022; Nijkamp et al., 2023). Prior work evaluates pass@k, filters with tests, or measures consistency under code transformations (Min et al., 2024). In our code RRF, each generated program is a node, semantics-preserving prompt transformations define equality edges in execution-signature space, and public tests define anchors. Hidden tests remain evaluation-only. The margin concerns recoverability of response signatures; source-level correctness still depends on candidate recall and test coverage. Open code-model families such as Code Llama and StarCoder2 further illustrate why a recovery layer should remain checkpoint-agnostic (Roziere et al., 2023; Lozhkov et al., 2024).

Hallucination and truthfulness evaluation Hallucination surveys distinguish factuality errors, faithfulness errors, and unverifiable generation (Ji et al., 2023). TruthfulQA demonstrates that models can imitate common falsehoods (Lin et al., 2022). RRF consistency blindness is not a new taxonomy of hallucination. It is an operator-level explanation for why coherent shared errors evade relation-only detection. Anchors such as evidence, tools, or human labels are needed to distinguish truth within a relational equivalence class.

I.3 Scope of the unification and novelty boundary

What is and is not unified RRFs unify methods only at the level of observable response relations. They do not assert that all black-box optimization algorithms reduce to one numerical solver. A method belongs in the framework when its outputs can be parsed into a field and its checks can be represented as relation or anchor residuals. Methods that change model weights, rely on inaccessible hidden states, or operate without specified output relations fall outside the strict black-box RRF setting.

Novelty boundary The strongest defensible novelty statement is:

> We introduce relational response fields as a black-box LLM recovery object, instantiate typed relation and anchor observations as a group-sparse inverse problem, prove that their restricted observability margin is the exact uniform and minimax recovery condition number, and test whether that same quantity predicts repair difficulty across model families and task domains.

We do not claim the first consistency method, the first graph-based LLM evaluator, the first metamorphic test, or the invention of restricted singular values.

J Synthetic theorem-verification protocol

Synthetic experiments test mathematical consequences under conditions where the true field, corruption support, operators, and exact margin are known. They are not presented as LLM benchmarks.

J.1 Synthetic field construction and exact margin computation

Path-consistent typed fields Each instance has $n \in \{12, 16\}$ nodes and fiber dimension $d = 4$. Nodes are partitioned into connected components. For node i, sample an invertible map

$$M_i = Q_i \operatorname{diag}(\exp(s_i)),$$

where Q_i is obtained from the QR factorization of a standard Gaussian matrix and $(s_i)_\ell \sim \mathcal{N}(0, 0.12^2)$. Each component C receives latent vector $u_C \sim \mathcal{N}(0, I_d)$, and the clean response is

$$z_i^\star = M_i u_C, \qquad i \in C.$$

For edge (i, j) within a component, set $T_{ij} = M_j M_i^{-1}$. Then $z_j^\star = T_{ij} z_i^\star$ exactly, and the transported gauge $h_i = M_i v_C$ lies in $\ker D$.

This construction gives nontrivial typed maps while retaining analytic control of the nullspace. Condition numbers of M_i are monitored; seeds producing values above the prespecified threshold are rejected before corruption is sampled.

Sparse corruptions Choose $k \in \{1, 2\}$ nodes uniformly without replacement. At each selected node sample a Gaussian direction, normalize it, and multiply by amplitude $a \in [0.9, 2.1]$. The observed field is $y = z^\star + e^\star$. Localization accuracy is the fraction of corrupted nodes among the k groups with largest estimated edit norm. Relative field error is

$$\frac{\|\widehat{z} - z^\star\|_2}{\max(\|z^\star\|_2, 10^{-12})}.$$

Recovery success thresholds are fixed before results are inspected and are accompanied by continuous error.

Exact margin computation For every instance, the experiment enumerates all supports of size at most $2k$ and computes the smallest singular value of B_S. The minimum support and right singular vector are stored. A value below 10^{-10}, together with witness residual below 10^{-9}, is classified as zero. Results are repeated at tighter and looser tolerances to ensure that phase transitions are not numerical artifacts.

J.2 Four theorem-verification experiments

Experiment S1: consistency–truth separation For each of 40 seeds, build a single complete component with no anchors. Compare three fields:

1. the clean truth $z^\star$;
2. a shared hallucination $z_i = z_i^\star + M_i v$ with fixed latent shift v;
3. independent errors injected at three random nodes.

The shared shift is constructed algebraically in $\ker D$, so its defect should be at floating-point zero while truth error is positive. The independent errors should have positive defect. Acceptance requires both inequalities on every seed, not only in expectation.

Experiment S2: anchor phase transition Use six disjoint two-node components and $k = 1$. Add direct full-fiber anchors according to a fixed schedule that covers one new component before adding a second anchor to any component. By Proposition 28, $\gamma_1 = 0$ until all six components are anchored. At the sixth anchor it becomes positive. For each anchor count $0, \ldots, 8$ and 30 seeds, inject one corruption and run exact support repair and the proximal method.

Primary outputs are γ_1, exact-repair error, convex-repair error, support localization, and success. The phase-transition claim is accepted only if the exact margin changes at the predicted coverage count and exact repair changes in the same direction. Convex repair is diagnostic because its success additionally depends on algorithmic conditions.

Experiment S3: duplicate saturation Start from a fixed set of relation families. For duplication factors $m \in \{1, 2, 4, 8, 16\}$, replace every row block R in one family by m copies $R/\sqrt{m}$. The Gram contribution and exact margin must remain equal up to numerical tolerance. Compare against adding genuinely independent typed chords with unit family weights. Independent rows may increase γ_k until restricted rank saturates.

The test distinguishes three quantities:

1. raw edge count;
2. rank of the unweighted row space;
3. normalized γ_k.

Reporting only raw duplicate singular values would incorrectly reward repeated evidence.

Experiment S4: spectral difficulty prediction Generate random component counts, anchor counts, independent chord counts, corruption budgets, and amplitudes. For each instance compute exact γ_k before repair. To isolate condition-number scaling, report raw error amplitude and stacked observation noise. The primary response is normalized recovery error

$$E_{\mathrm{norm}} = \frac{\gamma_k \|\widehat{z} - z^\star\|_2}{\max(\epsilon, 10^{-12})},$$

alongside the unnormalized error. Theorem 4 predicts a bounded normalized error for feasible exact support estimates; the empirical algorithm may add optimization error.

Difficulty prediction is summarized by Spearman and Pearson correlations between γ_k and negative error, logistic association with success, and localization. Confidence intervals resample complete instances. A scatterplot uses all points and does not suppress zero-margin failures.

J.3 Robustness sweeps, decoder comparisons, and validation criteria

Noise and operator perturbation sweeps For additive noise levels $\epsilon \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, sample a random observation-space direction and scale it to ϵ. For operator perturbations, add a Gaussian matrix E normalized to a prescribed spectral norm δ. Equation (12) predicts that the certified margin decreases by at most δ. The experiment checks this bound and the resulting stability ratio.

Convex versus exact decoder Every small instance is repaired with exact support search. The group-proximal decoder is run with a calibration-selected regularization parameter and a convergence tolerance. Three outcomes are distinguished:

1. both decoders succeed: information and optimization are adequate;
2. exact succeeds but convex fails: the instance is identifiable but the relaxation or tuning is inadequate;
3. exact fails or is unstable: the operator/noise condition is inadequate or the corruption model is violated.

This comparison prevents algorithmic behavior from being used as the definition of theoretical difficulty.

Seeds and replication Master seed 0 controls the published run. Each experiment derives nonoverlapping seed ranges for graph construction, corruption, and noise. The artifact records every seed in result CSV files. Repeating master seeds 1–4 is an extended robustness check; no seed is discarded based on recovery performance.

Prespecified validation conditions The synthetic suite evaluates the following prespecified conditions:

1. shared hallucinations have defect below 10^{-10} and nonzero truth error;
2. the anchor transition occurs at the analytically predicted coverage count;

3. normalized literal duplication changes γ_k by less than 10^{-10};
4. independent rows weakly increase the unnormalized margin;
5. margin/error association has the predicted sign and its bootstrap interval is reported;
6. all exact-margin witnesses pass a residual check.

All conditions are reported, including any violations.

K REAL-MODEL MATHEMATICS AND CODE PROTOCOL

This appendix specifies the cross-model, cross-task experiment used to test whether $\gamma_k(D, A)$ predicts empirical repair difficulty. The reported results use responses produced by the named open checkpoints rather than simulated error profiles. Raw text logs are retained; model weights are not included in the artifact.

K.1 ESTIMAND, MODELS, AND TASK DATASETS

Research question and estimand The primary question is not "does one repair algorithm improve average accuracy?" It is:

> Holding the response log fixed, do interventions on the relation–anchor design that increase $\gamma_k(D, A)$ predict lower repair error across two model families and two task domains?

Each base item is evaluated under ten prespecified operator designs. This within-item variation changes observability while preserving the model's generated candidate responses. Model and task are treated as strata. Raw error and candidate recall are recorded because a margin cannot manufacture a correct candidate or compensate for arbitrary violation of the sparse-error model.

Models We evaluate four open-weight instruction-tuned checkpoints:

1. `Qwen/Qwen2.5-0.5B-Instruct`, a 0.5B-parameter model from the Qwen2.5 family (Yang et al., 2024a);
2. `microsoft/Phi-3-mini-4k-instruct`, a 3.8B-parameter model from the Phi-3 family (Abdin et al., 2024);
3. `Qwen/Qwen2.5-7B-Instruct`, a 7B-parameter model from the Qwen2.5 family (Yang et al., 2024a);
4. `Qwen/Qwen2.5-14B-Instruct`, a 14B-parameter model from the Qwen2.5 family (Yang et al., 2024a).

The four checkpoints span two model families and multiple parameter scales. The checkpoint-specific precision and memory configuration are recorded in the released artifact, while the generation and evaluation protocols are held fixed across checkpoints. Generation uses each checkpoint's native chat template and tokenizer. Greedy decoding is used for transformed-query fields and reflection. Self-consistency uses four samples generated with temperature 0.7, top-$p = 0.9$, and fixed random seeds. The maximum numbers of newly generated tokens are 24 for integer answers and 192 for code. The recovery procedure treats every checkpoint as a black box. After the response texts have been generated, recovery uses only the parsed outputs and does not access logits, hidden states, gradients, or model parameters. Local model weights are required only to generate the response logs; the same artifact interface could instead be populated by responses obtained from a hosted API.

Mathematics dataset The mathematics set contains 128 deterministically generated integer-expression tasks, 16 from each of eight families:

1. five-digit addition;
2. large positive subtraction;
3. two- or three-digit multiplication;
4. affine products $(a + b)7 - c$;

5. nested differences $a(b-c)+d$;
6. exact quotients $(a+c)/d$ constructed to be integral;
7. sums of squares minus an offset;
8. differences of products plus an offset.

Operands are sampled from ranges fixed in code. Answers are computed with Python integer arithmetic before prompts are created. The generated set is chosen for exact transformation control rather than benchmark breadth; established reasoning benchmarks such as GSM8K and MATH motivate future validation with curated transformations (Cobbe et al., 2021; Hendrycks et al., 2021).

Each item has four transformed queries:

1. the original expression;
2. a lexical paraphrase with the same answer;
3. a scaling query asking for s times the expression value, with $s \in \{2, 3, 4, 5\}$;
4. a translation query asking for the expression value plus shift $t \in \{-37, -23, 19, 31, 47\}$.

Parsed outputs from variants three and four are transported back to canonical coordinates by division by s and subtraction of t. In canonical space every valid field is constant, so edge transports are identity. The raw typed transformation metadata is retained to audit the canonicalization.

The system instruction is:

> You are a precise arithmetic solver. Return exactly one signed base-10 integer and no words, commas, equations, or explanation.

The strict output format reduces parser ambiguity without revealing the answer. The parser first accepts an exact signed integer; if additional text is present, it records the last standalone integer and marks that the strict format was violated. Parse failures are stored rather than silently dropped.

Code dataset The code set contains 64 tasks, four independent test instantiations for each of 16 integer-valued function families:

1. absolute difference, greatest common divisor, least common multiple;
2. digit sum, digital root, divisor count, primality indicator;
3. Fibonacci and triangular numbers, Collatz stopping time;
4. count-even, maximum adjacent gap, second-largest distinct value;
5. weighted sum, alternating sum, and count-above-threshold.

Every item contains a reference implementation, four semantically equivalent natural-language specifications, eight public tests, and 16 hidden tests. Tests are generated from fixed item-specific seeds. All outputs are bounded integers, permitting an execution-signature vector in $\mathbb{R}^8$.

The four prompt variants use the original specification, argument renaming, equivalent wording, and a decomposition hint. All require a Python function named `solve` and forbid imports. The model sees neither reference code nor test outputs.

Generated code is extracted from the first fenced Python block or from a top-level `def solve`. An AST validator rejects imports, global/nonlocal statements, classes, dunder access, and dynamic execution primitives. Code runs with a restricted builtin set, no network, and a per-call wall-clock timer. Parse failure, compile failure, timeout, noninteger output, and wrong output are separate statuses.

Public execution signatures are used by relation and anchor methods. Hidden tests are evaluation-only. A candidate passes the code task only if all 16 hidden outputs match the reference. We also report the fraction of hidden tests passed to obtain a continuous error measure.

K.2 Response-field construction, repair procedures, and outcomes

Response fields and designs Every item yields four deterministic transformed responses, hence a four-node field. The corruption budget is $k = 1$ for the primary exact-repair analysis. Ten designs are fixed independently of model outputs:

Design	Relation edges	Full-node anchor
one relation	$(0, 1)$	none
paired relations	$(0, 1), (2, 3)$	none
three-node chain	$(0, 1), (1, 2)$	none
spanning tree	$(0, 1), (0, 2), (0, 3)$	none
triangle plus isolate	$(0, 1), (0, 2), (1, 2)$	none
typed complete	all six pairs	none
tree plus anchor	spanning tree	node 0
triangle, component anchor	triangle plus isolate	node 0
triangle, isolate anchor	triangle plus isolate	node 3
complete plus anchor	all six pairs	node 0

The first three designs and the unanchored triangle contain a null direction supported on at most two nodes, so $\gamma_1 = 0$. Anchoring the triangle component still leaves the isolated node invisible. In contrast, anchoring the isolate makes the remaining null direction occupy all three triangle nodes, outside the $2k$ class, so its restricted margin is positive despite global singularity. Connected four-node designs are also positive for $k = 1$. These matched relation/anchor counts prevent a regression from identifying observability by row count alone. All margins are computed by exact support enumeration separately for the scalar mathematics fiber and eight-dimensional code signature fiber.

The designs are interventions on B, not selected after observing answers. Their purpose is to create a range of theoretically meaningful recovery difficulty on identical response logs.

Mathematics repair Let z_i be the canonicalized numeric answer. For a chosen design, form

$$s = \begin{bmatrix} Dz \\ Az - b \end{bmatrix}.$$

When an anchor is present, b is the exact canonical value supplied by the deterministic arithmetic evaluator at node zero. Enumerate all supports of size at most one, solve restricted least squares for the error, and choose the support with minimum residual. The repaired answer is the rounded median of the corrected node values.

This procedure is deliberately simple. Its role is to test the operator condition, not to conceal performance inside a learned reranker. If more than one node is wrong, the $k = 1$ model is violated; this is recorded through the true error count and residual. Robustness results for $k = 2$ appear in Appendix L.

Code repair and candidate selection Each deterministic program produces an eight-dimensional public execution signature. Coordinates are scaled by a robust statistic of observed candidate magnitudes. The exact sparse decoder repairs the signature field under the chosen design. Because an arbitrary corrected signature is not source code, the method returns the generated program whose observed signature is closest to the repaired coordinatewise median. This is a discrete candidate-field projection.

Candidate recall is one if at least one of the four transformed programs passes all hidden tests. No selector can exceed candidate recall. Anchor designs use the full public reference signature at node zero; hidden signatures never enter A, hyperparameter selection, or candidate choice.

Baselines The same raw logs support five baselines:

1. **raw:** greedy response to the original prompt;
2. **self-consistency:** four stochastic samples of the original prompt, aggregated by integer mode for mathematics and execution-signature medoid for code;

3. **relation majority:** four deterministic transformed responses, canonicalized and aggregated by mode/medoid;
4. **reflection:** one additional greedy call shown the original answer and asked to check or correct it;
5. **verifier reranking:** choose among transformed candidates using exact arithmetic error for mathematics or number of public code tests passed.

The verifier is intentionally strong and external. Query budgets are reported: raw uses one generation, reflection two sequential generations, and field-based methods four transformed generations. The paper does not claim a compute-matched leaderboard advantage.

Primary outcomes For mathematics, correctness is exact integer equality and continuous error is

$$\min\left\{2, \frac{|\widehat{a} - a^\star|}{\max(|a^\star|, 1)}\right\}.$$

For code, correctness is hidden pass@1 and continuous error is one minus the hidden-test pass fraction. Additional outcomes are parse rate, candidate recall, observable residual, edit norm, and estimated support.

An *opportunity case* contains at least one correct and at least one incorrect deterministic transformed candidate. Such cases isolate selection/repair from candidate-generation ceilings. All-item and opportunity-only analyses are both reported; the primary pooled regression uses candidate-recall cases and controls raw error.

K.3 Statistical validation, data integrity, and compute environment

Statistical analysis The analysis has four layers.

Within-stratum association. For each model×task stratum, compute Spearman and Pearson correlations between γ_k and negative repair error. Opportunity cases are used when at least 20 are available; otherwise all items are reported with a flag.

Pooled association. Compute pooled correlation over all item-design rows. This descriptive number is not treated as independent-sample inference because each item appears under ten designs.

Grouped predictive model. Fit logistic regression for repair correctness with task, model, and standardized raw error. Compare against the same model plus standardized γ_k. Five-fold cross-validation groups all designs of one model-task-item together, preventing variants of the same response log from leaking across folds. Report baseline AUC, full AUC, ΔAUC, and the standardized margin coefficient.

Within-item permutation test. Under the null that design margins do not order outcomes, permute the ten margin values within each model-task-item block. Recompute pooled Spearman correlation for 2,000 permutations. The one-sided p-value is $(1 + \#\{\rho_{\text{null}} \geq \rho_{\text{obs}}\})/(2001)$. A 1,000-repetition cluster bootstrap resamples complete item blocks and reports a 95% interval.

These analyses test transfer of one geometric quantity across strata while respecting repeated designs. They do not prove a universal causal law for arbitrary tasks.

Prespecified empirical validation conditions The real-model suite evaluates whether:

1. raw logs exist for both named checkpoints;
2. both task domains are present;
3. every design margin is finite and auditable;
4. pooled association has the predicted sign;
5. the adjusted standardized margin coefficient is positive;
6. the within-item permutation test has $p < 0.05$.

All validation outcomes are retained in the reported results, and the main-paper claims are calibrated to the observed evidence.

Data integrity and leakage prevention Ground-truth expressions and reference programs generate labels and hidden tests. They are never inserted into model prompts. Public code tests are used only by methods labeled as anchored/verifier methods. Hidden tests are loaded by the evaluation stage after candidate selection. Dataset seeds, prompt strings, raw outputs, parsed values, design matrices, and selected indices are stored so that every result can be reconstructed without querying the model again.

Hardware and software Generation runs on the user-supplied AutoDL server with many NVIDIA GeForce RTX 4090 (48 GB). The environment uses Python 3.12, PyTorch 2.5.1 with CUDA 12.4, Transformers 5.14.1, SciPy, scikit-learn, pandas, and NumPy. LaTeX compilation uses TeX Live and the supplied ICLR 2027 style. Exact package versions are written to the artifact manifest after the final run.

L Extended empirical results and audits

This appendix contains the complete stratified results used by the main paper. Tables are generated from machine-readable CSV/JSON outputs by `scripts/export_latex_tables.py`; numerical values are not transcribed manually.

L.1 Core recovery results and predictive statistics

Natural response-field baselines Table 5 reports raw, self-consistency, transformed-relation majority, reflection, verifier reranking, and candidate recall for the unmodified real-model logs. These results characterize the candidate generator and should not be confused with the controlled $k = 1$ theorem test.

Table 5: Natural real-model response logs. Values are percent exact mathematics accuracy or code hidden pass@1. Candidate recall is the fraction with at least one correct transformed candidate.

Task	Model	Raw	Self-cons.	Rel.-major.	Reflect	Verifier	Cand. recall
mathematics	Phi-3-mini	37.5	34.4	34.4	10.9	37.5	37.5
mathematics	Qwen-0.5B	19.5	20.3	18.0	19.5	22.7	22.7
code	Phi-3-mini	93.8	100.0	100.0	93.8	100.0	100.0
code	Qwen-0.5B	31.2	60.9	68.8	68.8	75.0	75.0

Candidate recall is an upper bound for methods restricted to the deterministic transformed candidate set. Reflection can generate a new candidate and is therefore not bounded by that deterministic recall. The table is descriptive: methods use different sequential/query budgets, and no claim of compute-matched superiority is made.

Real-error replay The primary margin test replays authentic model errors under an exactly controlled one-node corruption model. A replay instance exists only when the log contains both a correct response and a distinct incorrect response. Three node values are populated by an observed correct response; one node receives the observed wrong response. For code, the wrong candidate must differ on the public execution signature so that the representation can, in principle, observe the error. Hidden-only bugs are analyzed separately as representation failures.

Table 6 gives every task–model–design cell. Zero-margin designs are expected to fail on some corrupt-node placements because a one- or two-node component carries an admissible null direction. Every connected four-node design has positive γ_1 and should recover the one-node error exactly in the noiseless signature space. Full-node anchors increase the margin but are not necessary when the connected component has size four and $2k = 2$.

Cross-model, cross-task prediction statistics Table 7 reports within-stratum rank association, grouped predictive performance, within-item randomization inference, and a cluster bootstrap. All ten designs of a replay instance remain in the same cross-validation fold. The permutation shuffles margin assignments within an item, preserving the model output, task, and error amplitude.

Table 6: Real-error replay by task, model, and operator design. Error is normalized full-field reconstruction error.

Task	Model	Design	n	γ_1	Exact (%)	Error
code	Phi-3-mini	one_relation	22	0.000	31.8	0.738
		paired_relations	22	0.000	68.2	0.450
		three_node_chain	22	0.000	81.8	0.182
		spanning_tree	22	0.765	100.0	0.000
		triangle_plus_isolate	22	0.000	81.8	0.182
		typed_complete	22	1.414	100.0	0.000
		tree_plus_anchor	22	0.835	100.0	0.000
		triangle_anchor_component	22	0.000	81.8	0.182
		triangle_anchor_isolate	22	1.000	100.0	0.000
		complete_plus_anchor	22	1.414	100.0	0.000
	Qwen-0.5B	one_relation	65	0.000	21.5	0.906
		paired_relations	65	0.000	40.0	0.849
		three_node_chain	65	0.000	69.2	0.308
		spanning_tree	65	0.765	100.0	0.000
		triangle_plus_isolate	65	0.000	69.2	0.308
		typed_complete	65	1.414	100.0	0.000
		tree_plus_anchor	65	0.835	100.0	0.000
		triangle_anchor_component	65	0.000	69.2	0.308
		triangle_anchor_isolate	65	1.000	100.0	0.000
		complete_plus_anchor	65	1.414	100.0	0.000
mathematics	Phi-3-mini	one_relation	126	0.000	22.2	0.880
		paired_relations	126	0.000	51.6	0.685
		three_node_chain	126	0.000	76.2	0.238
		spanning_tree	126	0.765	100.0	0.000
		triangle_plus_isolate	126	0.000	76.2	0.238
		typed_complete	126	1.414	100.0	0.000
		tree_plus_anchor	126	0.835	100.0	0.000
		triangle_anchor_component	126	0.000	76.2	0.238
		triangle_anchor_isolate	126	1.000	100.0	0.000
		complete_plus_anchor	126	1.414	100.0	0.000
	Qwen-0.5B	one_relation	56	0.000	30.4	0.837
		paired_relations	56	0.000	55.4	0.631
		three_node_chain	56	0.000	89.3	0.107
		spanning_tree	56	0.765	100.0	0.000
		triangle_plus_isolate	56	0.000	89.3	0.107
		typed_complete	56	1.414	100.0	0.000
		tree_plus_anchor	56	0.835	100.0	0.000
		triangle_anchor_component	56	0.000	89.3	0.107
		triangle_anchor_isolate	56	1.000	100.0	0.000
		complete_plus_anchor	56	1.414	100.0	0.000

The most direct interpretation is conditional: among replay fields that satisfy the paper's one-node corruption model, operator designs with larger margin make the authentic error easier to recover. This result transfers across four checkpoints and both scalar-answer and execution-signature spaces. It does not imply that γ_1 alone explains natural-field accuracy when several responses are jointly wrong or no correct candidate exists.

L.2 Task-family analyses and robustness audits

Per-family mathematics results The family breakdown in Table 8 reveals whether a pooled average is dominated by one operation. Large differences between raw accuracy and candidate recall indicate

Table 7: Predictive statistics for real-error replay. The grouped classifier controls for model, task, and raw error before adding γ_1.

Statistic	Value
Spearman γ_1 vs. $-$error (code—Phi-3-mini-4k-instruct)	0.397
Spearman γ_1 vs. $-$error (code—Qwen2.5-0.5B-Instruct)	0.506
Spearman γ_1 vs. $-$error (mathematics—Phi-3-mini-4k-instruct)	0.460
Spearman γ_1 vs. $-$error (mathematics—Qwen2.5-0.5B-Instruct)	0.385
Pooled Spearman γ_1 vs. $-$error	0.449
Grouped AUC, controls only	0.860
Grouped AUC, controls + γ_1	0.893
Grouped ΔAUC	0.033
Standardized γ_1 coefficient	3.653
Within-item permutation p	0.00050
Cluster-bootstrap 95% CI	[0.417, 0.481]

that transformed prompts sometimes expose complementary model competence; a small gap indicates a candidate-generation ceiling.

Table 8: Per-family mathematics results (percent).

Model	Family	n	Raw	Self-cons.	Rel.-major.	Verifier
Phi-3-mini	affine_product	16	0.0	0.0	0.0	0.0
	exact_quotient	16	12.5	6.2	6.2	12.5
	large_add	16	100.0	100.0	100.0	100.0
	large_subtract	16	81.2	62.5	68.8	81.2
	mixed_products	16	0.0	0.0	0.0	0.0
	multiply	16	100.0	100.0	93.8	100.0
	nested_difference	16	6.2	6.2	6.2	6.2
	squares	16	0.0	0.0	0.0	0.0
Qwen-0.5B	affine_product	16	0.0	0.0	0.0	0.0
	exact_quotient	16	0.0	0.0	0.0	6.2
	large_add	16	93.8	100.0	87.5	100.0
	large_subtract	16	43.8	43.8	43.8	56.2
	mixed_products	16	0.0	0.0	0.0	0.0
	multiply	16	18.8	18.8	12.5	18.8
	nested_difference	16	0.0	0.0	0.0	0.0
	squares	16	0.0	0.0	0.0	0.0

Per-family code results Table 9 separates arithmetic utilities, loops, and list-processing functions. Public execution medoids can improve selection when wrong programs have distinct signatures. They cannot detect a program that passes all public tests but fails a hidden corner case.

Consistency–truth audit on real logs For every natural field, we record normalized relation defect and correctness. Four cases are possible:

1. low defect, correct field;
2. high defect, mixed correctness;
3. low defect, shared wrong field;
4. parse-invalid field.

The third case is the empirical counterpart of Theorem 2. It is reported as a count and with representative raw outputs. A low-defect threshold is fixed on calibration data. We do not choose it after seeing which fields are wrong.

Table 9: Per-family code results (percent).

Model	Family	n	Raw	Self-cons.	Rel.-major.	Verifier
Phi-3-mini	absolute_difference	4	100.0	100.0	100.0	100.0
	alternating_sum	4	0.0	100.0	100.0	100.0
	collatz_steps	4	100.0	100.0	100.0	100.0
	count_above_threshold	4	100.0	100.0	100.0	100.0
	count_even	4	100.0	100.0	100.0	100.0
	digit_sum	4	100.0	100.0	100.0	100.0
	digital_root	4	100.0	100.0	100.0	100.0
	divisor_count	4	100.0	100.0	100.0	100.0
	fibonacci	4	100.0	100.0	100.0	100.0
	greatest_common_divisor	4	100.0	100.0	100.0	100.0
	least_common_multiple	4	100.0	100.0	100.0	100.0
	maximum_adjacent_gap	4	100.0	100.0	100.0	100.0
	prime_indicator	4	100.0	100.0	100.0	100.0
	second_largest_distinct	4	100.0	100.0	100.0	100.0
	triangular_number	4	100.0	100.0	100.0	100.0
	weighted_sum	4	100.0	100.0	100.0	100.0
Qwen-0.5B	absolute_difference	4	100.0	100.0	100.0	100.0
	alternating_sum	4	0.0	0.0	0.0	0.0
	collatz_steps	4	100.0	100.0	100.0	100.0
	count_above_threshold	4	100.0	100.0	100.0	100.0
	count_even	4	100.0	100.0	100.0	100.0
	digit_sum	4	0.0	50.0	100.0	100.0
	digital_root	4	0.0	25.0	100.0	100.0
	divisor_count	4	0.0	25.0	100.0	100.0
	fibonacci	4	0.0	75.0	100.0	100.0
	greatest_common_divisor	4	0.0	0.0	0.0	0.0
	least_common_multiple	4	0.0	0.0	0.0	0.0
	maximum_adjacent_gap	4	0.0	75.0	0.0	0.0
	prime_indicator	4	0.0	100.0	100.0	100.0
	second_largest_distinct	4	0.0	50.0	0.0	100.0
	triangular_number	4	0.0	100.0	100.0	100.0
	weighted_sum	4	100.0	75.0	100.0	100.0

Hidden-only code errors A generated program can agree with the reference on all eight public tests and fail a hidden test. In execution-signature space restricted to public tests, this wrong program is indistinguishable from a correct program. The issue is not a small margin of the supplied B; the representation has quotiented out the relevant semantic distinction. Adding more relation edges among identical public signatures cannot help. Additional tests expand the anchor representation and may remove the ambiguity.

We report three counts per model: parse/compile failures, public-signature-visible wrong programs, and hidden-only wrong programs. Real-error replay uses the second category for the theorem-aligned test; the third category is retained as a limitation rather than silently removed from natural-field results.

Corruption-budget robustness The primary replay has $k = 1$ by construction. Natural fields can have several wrong nodes. We recompute exact margins and repairs for $k = 1$ and $k = 2$. On four-node fields, relation-only designs necessarily have $\gamma_2 = 0$ because the dense constant gauge is now within the $2k = 4$ support class. Anchor designs can remain positive. This shift is predicted by the theory and illustrates why a margin must always be reported with its corruption budget.

For natural fields whose observed wrong-node count exceeds the assumed budget, errors are reported separately. A method is not credited with violating an impossibility theorem when its input violates the theorem's model.

Parser sensitivity Mathematics is reparsed under three rules: strict entire-string integer, last standalone integer, and first standalone integer. Strict parsing measures instruction following; last-integer parsing tolerates explanations. The main parser is declared before evaluation. Agreement among rules is reported. For code, extraction from fenced blocks is compared with extraction from a top-level function start. AST validation remains identical.

The margin depends on the parsed representation. Parser sensitivity therefore changes both outcome and operator realization; it is not merely a cosmetic preprocessing choice.

Relation-weight sensitivity The ten primary designs assign unit weight to each prespecified typed relation. Two robustness schemes are evaluated:

1. fixed total relation energy, dividing each edge weight by the number of edges;
2. calibration reliability, weighting relation types by inverse observed residual variance on a held-out split.

The first isolates row-space coverage from energy; the second expresses signal-to-noise geometry. Literal duplicate experiments always split one fixed family weight across copies.

Exact-margin tolerance Margins are recomputed with zero thresholds 10^{-8}, 10^{-10}, and 10^{-12}. Each zero-margin design has an explicit witness with residual close to machine precision; each positive four-node design stays separated from zero by orders of magnitude. Statistical conclusions are unchanged because the design ordering is not driven by borderline singular values.

L.3 Budget accounting, failure taxonomy, and complete catalogs

Query-budget accounting For each model and task:

- raw uses one greedy generation;
- reflection uses the raw generation plus one checking generation;
- self-consistency uses four stochastic generations of one prompt;
- relation majority, verifier reranking, and RRF use four deterministic transformed generations.

The real-error replay reuses already generated logs and makes no extra model calls. Operator computation and repair are negligible compared with generation at this scale, although exact γ_k enumeration grows combinatorially with n and k.

Failure-case taxonomy Every failed natural repair is assigned one nonexclusive label:

1. **candidate absence:** no correct transformed response exists;
2. **shared error:** all valid responses agree on a wrong canonical answer;
3. **budget violation:** more than k nodes are wrong;
4. **representation collision:** public code signatures agree but hidden semantics differ;
5. **zero/low margin:** the operator admits an ambiguous sparse direction;
6. **optimization/selection error:** a correct candidate exists and information is adequate, but the algorithm selects incorrectly;
7. **parse failure:** the semantic response is not represented reliably.

This taxonomy is more informative than one aggregate accuracy number. Only the fifth category is directly captured by γ_k; the theory predicts neither candidate recall nor parser correctness.

Complete dataset catalogs Tables 10 and 11 list every item. Prompt strings, tests, and reference programs are available in JSONL and are too verbose to duplicate in full in the PDF.

Table 10: Representative examples from the mathematics benchmark. The complete 128-instance catalog is released with the artifact.

Index	Family	Expression	Answer	Scale	Shift
0	large_add	909 + 770	1679	3	-23
1	large_subtract	9897 - 7758	2139	5	-37
2	multiply	59 * 56	3304	2	-23
3	affine_product	(80 + 90) * 7 - 90	1100	2	47
4	nested_difference	17 * (91 - 4) + 75	1554	5	31
5	exact_quotient	(32 + 678) / 5	142	5	47
6	squares	(27 * 27) + (24 * 24) - 179	1126	5	19
7	mixed_products	7 * 27 - 39 * 19 + 62	-490	4	31
8	large_add	928 + 934	1862	4	19
9	large_subtract	3746 - 1052	2694	3	47
10	multiply	11 * 71	781	3	-23
11	affine_product	(97 + 79) * 7 - 59	1173	2	-37
		... omitted examples ...			
58	multiply	79 * 95	7505	3	47
59	affine_product	(21 + 42) * 7 - 143	298	3	-37
60	nested_difference	13 * (92 - 82) + 108	238	2	-23
61	exact_quotient	(13 + 1752) / 5	353	5	47
62	squares	(19 * 19) + (23 * 23) - 179	711	4	47
63	mixed_products	30 * 29 - 18 * 10 + 88	778	2	-37
64	large_add	878 + 198	1076	3	19
65	large_subtract	1090 - 559	531	2	-37
66	multiply	45 * 68	3060	2	-37
67	affine_product	(74 + 92) * 7 - 140	1022	4	-37
68	nested_difference	3 * (93 - 20) + 133	352	3	19
69	exact_quotient	(61 + 1487) / 6	258	5	31
		... omitted examples ...			
116	nested_difference	5 * (98 - 5) + 158	623	2	19
117	exact_quotient	(92 + 2248) / 5	468	3	19
118	squares	(24 * 24) + (13 * 13) - 72	673	3	19
119	mixed_products	28 * 21 - 31 * 29 + 105	-206	5	31
120	large_add	795 + 618	1413	4	19
121	large_subtract	2735 - 2574	161	5	-23
122	multiply	90 * 72	6480	5	31
123	affine_product	(84 + 22) * 7 - 154	588	5	19
124	nested_difference	11 * (42 - 11) + 109	450	5	-37
125	exact_quotient	(93 + 1629) / 6	287	2	31
126	squares	(19 * 19) + (7 * 7) - 175	235	2	-37
127	mixed_products	34 * 6 - 12 * 13 + 56	104	4	-37

M WORKED EXAMPLES AND BOUNDARY CASES

This appendix gives concrete calculations for readers new to sparse inverse problems. Each example identifies the field, operator, nullspace, margin consequence, and repair interpretation.

M.1 CONSISTENCY, TRANSFORMATIONS, AND GRAPH CONNECTIVITY

Three paraphrases and a shared wrong answer Let three scalar nodes be equivalent paraphrases with edges $(1, 2)$ and $(2, 3)$:

$$D = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}.$$

Table 11: Representative examples from the code benchmark. The 64-instance catalog is included in the artifact.

Index	Family	Signature	Public tests	Hidden tests
0	absolute_difference	solve(x, y)	8	16
1	greatest_common_divisor	solve(x, y)	8	16
2	least_common_multiple	solve(x, y)	8	16
3	digit_sum	solve(n)	8	16
4	digital_root	solve(n)	8	16
5	divisor_count	solve(n)	8	16
6	prime_indicator	solve(n)	8	16
7	fibonacci	solve(n)	8	16
	... omitted examples ...			
28	second_largest_distinct	solve(xs)	8	16
29	weighted_sum	solve(xs)	8	16
30	alternating_sum	solve(xs)	8	16
31	count_above_threshold	solve(xs, t)	8	16
32	absolute_difference	solve(x, y)	8	16
33	greatest_common_divisor	solve(x, y)	8	16
34	least_common_multiple	solve(x, y)	8	16
35	digit_sum	solve(n)	8	16
	... omitted examples ...			
56	triangular_number	solve(n)	8	16
57	collatz_steps	solve(n)	8	16
58	count_even	solve(xs)	8	16
59	maximum_adjacent_gap	solve(xs)	8	16
60	second_largest_distinct	solve(xs)	8	16
61	weighted_sum	solve(xs)	8	16
62	alternating_sum	solve(xs)	8	16
63	count_above_threshold	solve(xs, t)	8	16

The true field is $(17, 17, 17)$, but the model outputs $(19, 19, 19)$. Both have zero defect. Their difference $(2, 2, 2)$ lies in $\ker D$. Majority vote, pairwise agreement, cycle consistency, and minimizing $\|Dz\|$ all prefer no change. An anchor $A = [1, 0, 0]$ with target 17 exposes the shift.

For $k = 1$, the shared error is not in the assumed class because it occupies three nodes. Interestingly, D can still identify a single-node error: its only null direction has support three, larger than $2k = 2$, so $\gamma_1(D, 0) > 0$. The same relation graph is adequate for localized errors and blind to the dense shared hallucination. This distinction is lost if one says only "the graph is singular."

One outlier among four transformed answers Suppose canonicalized answers are $z = (42, 42, 47, 42)$ on a connected four-node star. The true field is $z^\star = (42, 42, 42, 42)$ and $e^star = (0, 0, 5, 0)$. With edges $(1, 2), (1, 3), (1, 4)$,

$$Dz = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}.$$

Searching one-node supports asks which single coordinate edit can make the residual zero. Editing node three by five succeeds; editing any other one leaves a nonzero edge residual. Because every two-column restriction of D is injective for this four-node connected graph, $\gamma_1 > 0$ and the solution is unique.

If node three were isolated, its error would produce no residual. The corresponding column of D would be zero, $\gamma_1 = 0$, and no relation-only algorithm could detect it.

Scaling relation without canonicalization Let node one ask for $a+b$ and node two ask for $3a+3b$. Their scalar transport is $T(z) = 3z$, so

$$D = [-3 \quad 1].$$

A valid field has form $(u, 3u)$. A shared canonical error c appears as $(c, 3c)$ and is in the nullspace. Canonicalizing node two by division by three changes coordinates to $(z_1, z_2/3)$ and the defect row becomes $[-1, 1]$. The two formulations are equivalent only if the domain norm is transformed consistently. Computing singular values after coordinate rescaling without updating the metric can change the numerical margin.

Affine shift through homogeneous coordinates Suppose node two asks for the original expression plus t. The relation $z_2 = z_1 + t$ is affine. Lift each scalar to $\widetilde{z} = (z, 1)$ and define

$$T = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}.$$

Then $\widetilde{z}_2 = T\widetilde{z}_1$. Perturbations of valid lifted responses have zero in the homogeneous coordinate, so the sparse margin should be restricted to that tangent subspace. The experiment instead canonicalizes by subtracting t, avoiding an artificial error coordinate.

Two disconnected pairs at $k = 1$ Let four scalar nodes form components $\{1, 2\}$ and $\{3, 4\}$ with identity edges. Then

$$D = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

The vector $(1, 1, 0, 0)$ is a two-node null vector. Since $2k = 2$, $\gamma_1 = 0$. Partition it as

$$(1, 0, 0, 0) - (0, -1, 0, 0).$$

These are two distinct one-node error patterns with equal syndromes. An anchor on node one removes the first component's gauge but leaves $(0, 0, 1, 1)$; both components require coverage for uniform recovery.

A large unanchored component can be sparse-observable Take a ten-node connected identity graph with no anchor and $k = 2$. The constant gauge uses ten nodes, while differences of two feasible errors use at most four. No gauge lies in the restricted class. For a connected graph, every nonzero vector supported on fewer than ten nodes creates a boundary edge and therefore nonzero defect. Thus $\gamma_2 > 0$ even though D is rank-deficient.

This does not identify the absolute dense truth; it identifies every two-node corruption relative to the observed field. If the model makes a coherent ten-node error, the sparse model is violated and consistency remains blind.

M.2 Anchors, duplicate measurements, optimization, and noise

A partial anchor Each node response is a two-vector (u, v). Relations transport both coordinates identically. An anchor observes only u at one node. The dense gauge $(0, c)$ remains unanchored. Whether γ_k is zero depends on component size and k: if the component is small enough for the v-gauge to be $2k$-sparse, the margin is zero; otherwise localized errors can still be observable. Counting anchored nodes without considering anchor rank is insufficient.

Literal duplicate rows Let $R = [-1, 1]$. One unit-weight relation has Gram matrix

$$R^*R = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Four unnormalized copies give $4R^*R$ and double every nonzero singular value. Four normalized copies $R/2$ give

$$\sum_{j=1}^{4} (R/2)^*(R/2) = R^*R.$$

The latter correctly represents four textual copies of one deterministic fact. If the four calls provide independent noisy measurements, averaging decreases noise variance, which should be reflected in ϵ, not deterministic rank.

Positive margin but convex failure Let scalar groups and choose B with nullspace spanned by $h = (1, 0.4, 0.4)$. No two-sparse null vector exists, so $\gamma_1 > 0$ and every one-sparse vector is uniquely identified by B. Yet for support $S = \{1\}$,

$$\|h_S\|_1 = 1 > 0.8 = \|h_{S^c}\|_1.$$

The null-space property fails. If e^star is supported on coordinate one, moving along the null direction can reduce the ℓ_1 norm and cause basis pursuit to select a different dense vector. Exact search succeeds while the convex relaxation fails. This is why the main paper gives separate theorems.

Noisy stability calculation Assume $\gamma_1 = 0.5$, true observation noise norm at most 0.02, and an exact-support estimate with residual at most 0.02. Theorem 4 gives

$$\|\widehat{e} - e^\star\|_2 \le \frac{0.04}{0.5} = 0.08.$$

If an alternative design doubles the margin to one without changing noise, the bound halves. If repeated independent queries halve the noise radius without changing the normalized operator, the bound also halves. Margin improvement and noise reduction are distinct routes to easier recovery.

Near-null direction Exact nonidentifiability is not the only problem. Suppose a unit two-node perturbation h satisfies $\|Bh\| = 10^{-3}$. Then $\gamma_1 \le 10^{-3}$. Observation noise of norm 10^{-2} can conceal a perturbation roughly ten times larger than the unit normalization scale. Numerically, the field is identifiable in exact arithmetic, but statistically ill-conditioned. Reporting only that B_S has full rank misses this difficulty; the magnitude of γ_k matters.

M.3 Semantic collisions, nonlinear ambiguity, bias, and budget effects

Code signature collision Two programs agree on eight public tests. One implements the intended function; the other has a branch that fails only on a hidden negative input. Their public signatures are identical, so every relation and anchor defined on that signature space assigns zero difference. The corresponding semantic error is outside the representation, not merely in a small singular direction. Expanding the test set can separate the programs; adding more paraphrase edges cannot.

Nonlinear global ambiguity Let a scalar relation residual be $r(z) = \sin z$ with target zero. At $z = 0$, the derivative is one and the local margin is positive. Yet $z = \pi$ gives the same residual. A local Jacobian certificate cannot rule out the distant candidate. Discrete domain restrictions or anchors are needed for global recovery.

Model accuracy and recovery difficulty are different Model A is correct at 90% of nodes but its rare errors are coherent gauges. Model B is correct at 60% but errors are isolated and the operator has high restricted margin. A may have higher raw accuracy and lower repairability on its failures; B may be easier to repair when a candidate field is available. Cross-model experiments control raw error and evaluate whether γ_k explains additional variation. The margin is not a replacement for accuracy; it conditions recovery given an error class.

Anchor bias Suppose all relation responses correctly indicate 42 but an anchor incorrectly states 43 with very large weight. The combined operator can have a large margin while the affine target b is biased, causing stable recovery of the wrong field. The margin measures sensitivity and identifiability relative to the supplied observations, not truthfulness of b. Anchor validity is a separate assumption and must be externally audited.

Budget dependence On a four-node connected equality graph, $\gamma_1 > 0$ because the constant gauge occupies four nodes and $2k = 2$. At $k = 2$, that same gauge is admissible and $\gamma_2 = 0$. There is no contradiction: the operator can distinguish every one-node error but not every two-node error. A paper that reports γ without k omits essential information.

N Limitations, failure modes, and nonclaims

This appendix expands the concise limitations in the main paper and states how each one affects interpretation.

N.1 Modeling and measurement assumptions

Representation validity The theory operates on parsed responses $z_i = \phi_i(y_i)$. If ϕ_i discards a truth-relevant distinction, no margin computed after parsing can recover it. Public execution signatures illustrate this sharply: programs with identical public behavior but different hidden behavior collide. Semantic embeddings can create subtler collisions. The margin is intrinsic to the specified representation, not to unparsed natural-language meaning.

Mitigation requires parser audits, richer task-specific representations, and external anchors. It cannot be solved by adding a theorem assumption that "the embedding is semantic."

Transport validity A typed transformation must preserve or predictably change the correct answer. Ambiguous paraphrases, overflow-sensitive code refactors, or transformations that alter pragmatic context violate this requirement. Such errors enter as operator or target misspecification and can make a correct model appear inconsistent.

The artifact uses algebraically generated math transformations and execution-tested code specifications. Natural-language applications would need human or formal validation of relations. Equation (12) handles small operator error, not arbitrary semantic invalidity.

Anchor validity and dependence Theorems assume that anchor noise is bounded relative to a target field. A biased or adversarial anchor can stably identify the wrong field. A self-verifier generated by the same LLM may share errors with the responses and should not be modeled as independent trusted truth without evidence. The experiments use deterministic arithmetic and reference execution outputs, unusually strong anchors.

Real applications may have soft, correlated, or strategic verifiers. Extending the theory to jointly uncertain anchors and fields is important future work.

Sparse node-corruption model The group-sparse model is suitable when a small number of transformed responses fail while most are compatible with one truth field. It is unsuitable for diffuse calibration bias, widespread prompt misunderstanding, or shared hallucinations affecting all nodes. The impossibility theorem explains the last case rather than solving it.

Approximate sparsity gives an oracle tail term, but a dense error aligned with $\ker D$ can still dominate. One should report empirical wrong-node counts and budget sensitivity rather than assume $k = 1$ universally.

Worst-case versus typical-case difficulty γ_k is a worst-case restricted margin. A small value means some admissible direction is difficult; it does not imply that the model often makes that error. A large value gives uniform stability under the model assumptions; it does not guarantee that parsing, candidate generation, or anchors are correct. Cross-model prediction is therefore an empirical question about alignment between natural errors and difficult directions.

The real-error replay deliberately enforces the sparse model and uses authentic error values, providing a theorem-aligned stress test. Natural-field results are weaker evidence because they include candidate absence and budget violations.

Finite-dimensional linearity Global theorems assume finite-dimensional linear operators. Discrete candidate spaces are finite but not linear; we use the linear signature space for margins and a candidate projection for outputs. Nonlinear residuals receive only local guarantees with explicit curvature. General text-to-text semantic transport is neither proven linear nor globally identifiable.

Infinite-dimensional extensions could use restricted lower bounds on unions of subspaces, but compactness and attainment arguments require care. They are outside this paper.

Exact margin computation Computing γ_k by support enumeration is exponential in $k \log n$. The paper's exact claims are experimentally feasible because fields are small. Large agent graphs require lower certificates, coherence bounds, branch-and-bound, or task-specific structure. A heuristic estimate should be labeled as such.

This computational limitation does not invalidate the information-theoretic object, but it limits routine deployment and adaptive design at scale.

Normalization dependence Singular values depend on units and weights. Cross-task comparison is meaningful only after specifying domain/codomain metrics and relation-family normalization. The paper canonicalizes math units, robustly scales code signature coordinates, and separates duplicate family weight. Different defensible metrics can produce different numerical margins.

The invariant claim is conditional: under a fixed, declared norm and weighting, γ_k is the exact restricted condition number. There is no unit-free scalar without additional structure.

N.2 Empirical and operational limitations

Candidate recall Discrete repair cannot output a program or answer absent from its candidate closure unless it has a generative edit step. High margin cannot overcome zero candidate recall. The paper reports candidate recall and evaluates field reconstruction separately from candidate selection.

A future generative RRF algorithm could ask the LLM to instantiate a repaired signature, but then generation introduces a new stochastic channel and may fail even when the target field is identified.

Public versus hidden tests Code anchors are only as complete as public tests. Hidden tests evaluate generalization beyond the observed signature. Passing public tests does not prove program correctness. The margin certifies recovery in public signature space, not semantic equivalence over all inputs.

This limitation is fundamental to testing, not unique to RRF. Formal verification or exhaustive finite-domain execution would provide stronger anchors when available.

Small models and small tasks The experiments use four open checkpoints and 192 generated tasks. They are designed for mechanistic clarity and reproducibility on many 4090, not benchmark leadership. Results may differ for frontier APIs, long-form reasoning, multilingual tasks, or domains without exact anchors.

Cross-model transfer across two families is evidence, not universality. The paper avoids claiming that one coefficient applies to every model or task.

Compute and comparison fairness Field methods use four transformed queries, whereas raw uses one. Reflection and self-consistency have different sequential and sampling budgets. The paper reports these budgets and treats baseline accuracy as context. Its primary claim concerns difficulty prediction under fixed logs, where all operator designs reuse the same generations.

No compute-optimality claim is made.

Dataset generation Synthetic arithmetic and code tasks permit exact labels and transformations but may be easier, more regular, or less linguistically diverse than public benchmarks. They reduce contamination and relation ambiguity at the cost of ecological breadth. Repeating the study on GSM8K, MATH, HumanEval, and MBPP would require carefully validated transformations and licensing-aware artifact packaging.

N.3 Interpretive boundaries, security, nonclaims, and future work

Statistical dependence Ten operator designs reuse each response log, so rows are not independent. Grouped cross-validation, within-item permutation, and cluster bootstrap address this design. They do not correct for every analysis choice. The protocol and outcomes are specified in code and supplement, but the study is not externally preregistered.

Causal interpretation Operator design is manipulated within an item, which supports a causal interpretation for those designs under deterministic replay. However, γ_k is a function of the design, and designs can affect algorithms through features not summarized by the scalar margin. The

minimax theorem establishes a fundamental bound; empirical coefficient estimates do not prove that all performance differences are mediated only by γ_k.

Security Executing generated code is risky. The AST restrictions and timers reduce risk for tiny integer functions but are not a hardened sandbox. Production reproduction should use containers, syscall restrictions, no network, read-only filesystems, and resource quotas.

Bibliographic scope The supplement surveys the closest mathematical and LLM-method literature but cannot cover every graph-consistency, testing, uncertainty, or inverse-problem paper. The novelty claim is framed to avoid priority claims about individual ingredients.

Nonclaims For clarity, the paper does not claim:

- that consistency implies truth;
- that γ_k is an intrinsic scalar of a model independent of task and representation;
- that positive γ_k alone guarantees group-Lasso recovery;
- that all natural-language transformations are valid or linear;
- that a large margin corrects biased anchors or missing candidates;
- that the small experiments establish state-of-the-art benchmark accuracy;
- that restricted singular values, graph Laplacians, sheaves, or metamorphic testing are newly invented.

Future theoretical directions Important extensions include probabilistic error priors coupled to worst-case geometry, robust design with uncertain transports, nonasymptotic estimation of reliability weights, approximate margin certificates for large graphs, adversarial anchors, nonlinear global identifiability, and sequential query policies with stopping guarantees. Another open question is whether natural LLM error distributions concentrate near a low-dimensional subset of the worst restricted directions; such a result could connect minimax and typical-case repair.

O PROBABILISTIC NOISE, APPROXIMATE SPARSITY, AND WEIGHTED MARGINS

The main results are deterministic because bounded-noise statements expose the information geometry without committing to a data-generating distribution. This appendix connects those statements to probabilistic observation models. It also records the modifications required by heteroscedastic measurements, approximately sparse response failures, and operator error. These extensions are useful for readers accustomed to statistical inverse problems, but none is needed for the exact identifiability theorem.

O.1 RANDOM-NOISE RADII AND WEIGHTED OBSERVABILITY GEOMETRY

From a random noise law to a deterministic radius Let $B = [D; A] : \mathcal{H} \to \mathbb{R}^m$ and observe

$$s = Be^\star + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I_m). \tag{25}$$

The following statement is a direct probabilistic corollary of Theorem 4; its proof is included to make the confidence parameter and dimension dependence explicit.

Proposition 29 (Gaussian high-probability recovery)**.** *Fix $\delta \in (0,1)$ and define*

$$\epsilon_\delta = \sigma\big(\sqrt{m} + \sqrt{2\log(1/\delta)}\big).$$

Suppose $\|e^\star\|_{0,\mathrm{g}} \le k$, $\gamma_k(D,A) > 0$, and $\widehat{e}$ is any k-group-sparse vector satisfying $\|B\widehat{e} - s\|_2 \le \epsilon_\delta$. Then, with probability at least $1-\delta$,

$$\|\widehat{e} - e^\star\|_2 \le \frac{2\sigma}{\gamma_k(D,A)}\big(\sqrt{m} + \sqrt{2\log(1/\delta)}\big).$$

Proof. Write $\xi = \sigma g$ with $g \sim \mathcal{N}(0, I_m)$. The map $g \mapsto \|g\|_2$ is one-Lipschitz, and $\mathbb{E}\|g\|_2 \le \sqrt{\mathbb{E}\|g\|_2^2} = \sqrt{m}$. Gaussian concentration therefore gives

$$\mathbb{P}\{\|g\|_2 > \sqrt{m} + t\} \le e^{-t^2/2}.$$

Set $t = \sqrt{2\log(1/\delta)}$. On the resulting event, both $e^\star$ and $\widehat{e}$ are feasible for the radius-ϵ_δ constraint. Their difference is supported on at most $2k$ groups, and

$$\gamma_k(D, A)\|\widehat{e} - e^\star\|_2 \le \|B(\widehat{e} - e^\star)\|_2 \le \|B\widehat{e} - s\|_2 + \|s - Be^\star\|_2 \le 2\epsilon_\delta.$$

Dividing by the positive margin proves the result. □

The factor $\sqrt{m}$ should not be read as a penalty for adding arbitrary normalized relations. If an experiment adds rows while fixing a total measurement-energy budget, the coordinate variance or row weights must be rescaled as part of the observation model. Otherwise both signal energy and noise energy change. The correct comparison uses the whitened operator below.

For independent centered sub-Gaussian coordinates with proxy variance σ^2, a norm concentration inequality gives the same form up to a universal constant. For heavy-tailed coordinates, an ℓ_2 radius may be inappropriate: robust residual aggregation or coordinate clipping is required before the deterministic theorem can be invoked. The restricted margin itself does not turn a heavy-tailed noise law into a bounded one.

Heteroscedastic and correlated measurements Relation residuals and anchors rarely have equal reliability. Suppose

$$\xi \sim \mathcal{N}(0, \Sigma), \qquad \Sigma \succ 0.$$

Whitening gives $\widetilde{s} = \Sigma^{-1/2}s$, $\widetilde{B} = \Sigma^{-1/2}B$, and standard Gaussian noise. This motivates the covariance-aware margin

$$\gamma_{k,\Sigma}(B) := \min_{0<\|h\|_{0,\mathrm{g}}\le 2k} \frac{\|\Sigma^{-1/2}Bh\|_2}{\|h\|_2} = \min_{|S|\le 2k} \sigma_{\min}(\Sigma^{-1/2}B_S). \tag{26}$$

All deterministic identifiability and stability arguments apply verbatim to $\widetilde{B}$. In particular, correlation between two anchor measurements prevents them from being counted as two independent units of information: near-collinearity in Σ is removed by whitening. If Σ is singular because some residuals are exact linear copies, one first restricts to the support of Σ and separately retains deterministic noiseless constraints.

The covariance must be estimated without using test correctness. A defensible protocol estimates it from repeated calibration queries, freezes a regularized estimate

$$\widehat{\Sigma}_\tau = (1-\tau)\widehat{\Sigma} + \tau\frac{\mathrm{tr}(\widehat{\Sigma})}{m}I,$$

and reports sensitivity to τ. Treating a noisy LLM judge as an exact anchor corresponds to assigning zero variance and can make the weighted margin arbitrarily optimistic. A floor on all estimated variances prevents that pathology and states the assumed reliability in observable units.

O.2 Gaussian minimax bounds and approximate sparsity

A Gaussian two-point lower bound The bounded-noise minimax theorem shows a worst-case $1/\gamma_k$ law. A related dependence persists under Gaussian noise. To avoid hiding constants, we give a self-contained two-point statement. Let

$$\Theta_k(R) = \{e \in \mathcal{H} : \|e\|_{0,\mathrm{g}} \le k,\ \|e\|_2 \le R\}$$

and let $\mathfrak{R}_G(B, k, R, \sigma)$ be the minimax expected ℓ_2 error under $s \sim \mathcal{N}(Be, \sigma^2 I)$.

Proposition 30 (Gaussian local minimax obstruction)**.** *Assume the minimum defining $\gamma_k(B)$ is attained, as it is in finite dimensions. Then*

$$\mathfrak{R}_G(B, k, R, \sigma) \ge \frac{1}{8}\min\left\{R, \frac{\sigma}{\gamma_k(B)}\right\}.$$

When $\gamma_k(B) = 0$, the convention $\sigma/0 = +\infty$ yields the lower bound $R/8$.

Proof. First suppose $\gamma_k(B) > 0$. Choose a unit vector h supported on at most $2k$ groups with $\|Bh\|_2 = \gamma_k(B)$. Partition its support into disjoint sets S_1, S_2, each of size at most k, and write $h = h_{S_1} - (-h_{S_2})$. Let

$$t = \min\{R, \sigma/\gamma_k(B)\}, \qquad e_1 = th_{S_1}, \qquad e_2 = -th_{S_2}.$$

Both parameters belong to $\Theta_k(R)$ because restriction cannot increase the norm, and $\|e_1 - e_2\|_2 = t$. For $P_j = \mathcal{N}(Be_j, \sigma^2 I)$,

$$\mathrm{KL}(P_1 \| P_2) = \frac{\|B(e_1 - e_2)\|_2^2}{2\sigma^2} = \frac{t^2\gamma_k(B)^2}{2\sigma^2} \le \frac{1}{2}.$$

Pinsker's inequality gives $\mathrm{TV}(P_1, P_2) \le 1/2$. The standard two-point reduction follows directly from the triangle inequality: for every estimator, at least one of the two parameter risks is at least

$$\frac{\|e_1 - e_2\|_2}{4}\big(1 - \mathrm{TV}(P_1, P_2)\big) \ge \frac{t}{8}.$$

If $\gamma_k(B) = 0$, choose a nonzero $2k$-group-sparse $h \in \ker B$, normalize it, and use $t = R$. The two distributions are then identical and the same argument applies with total variation zero. □

The proposition is deliberately local and constant-level. Sharper Gaussian minimax rates can depend on the number of admissible supports and on the full restricted spectrum, not only its smallest value. The point relevant here is narrower: even under smooth stochastic noise, no estimator removes the inverse dependence on the least observable sparse direction.

Approximately sparse response failures A natural field can contain one dominant failure and many small discrepancies. Let $e^\star = e_k + r$, where e_k is k-group-sparse and r is a residual tail. The observation is $s = Be^\star + \xi$. A k-sparse decoder should be compared with e_k, because exact recovery of a dense $e^\star$ is outside its model class.

Proposition 31 (Oracle approximation inequality)**.** *Suppose $\|\xi\|_2 \le \epsilon$, $\gamma_k(D, A) > 0$, and $\widehat{e}$ is k-group-sparse with*

$$\|B\widehat{e} - s\|_2 \le \epsilon + \|Br\|_2.$$

Then

$$\|\widehat{e} - e^\star\|_2 \le \|r\|_2 + \frac{2(\epsilon + \|Br\|_2)}{\gamma_k(D, A)}.$$

Proof. The sparse approximation e_k is feasible because

$$\|Be_k - s\|_2 = \|Br + \xi\|_2 \le \|Br\|_2 + \epsilon.$$

Both e_k and $\widehat{e}$ are k-group-sparse, so the deterministic stability argument gives

$$\|\widehat{e} - e_k\|_2 \le 2(\epsilon + \|Br\|_2)/\gamma_k(D, A).$$

Add $\|e_k - e^\star\|_2 = \|r\|_2$ by the triangle inequality. □

The bound separates two effects. The unavoidable approximation error is $\|r\|_2$; the observable tail Br also behaves as additional measurement noise. A tail lying mostly in $\ker B$ can have small $\|Br\|_2$ yet large semantic norm, reminding us that the chosen representation and norm determine what counts as a small error.

For the group-ℓ_1 decoder, standard robust-null-space arguments replace $\|r\|_2$ by a best-k group approximation term such as $\sigma_k(e^\star)_{2,1}/\sqrt{k}$. That stronger conclusion requires the group robust null-space property and does not follow from $\gamma_k > 0$ alone.

O.3 Operator uncertainty, calibration, and scope

Operator error and calibration uncertainty Let B be the operator used by the decoder but suppose the actual measurement map is $B + \Delta$:

$$s = (B + \Delta)e^\star + \xi.$$

If $\|\Delta\|_{2\to 2} \le \eta$ and $\|e^\star\|_2 \le R$, then the nominal model sees effective noise $\widetilde{\xi} = \Delta e^\star + \xi$ with

$$\|\widetilde{\xi}\|_2 \le \epsilon + \eta R.$$

Consequently every deterministic stability statement remains valid after replacing ϵ by $\epsilon + \eta R$. This elementary reduction is important in black-box evaluation: a parser or transport estimated from finite calibration data is part of the observation channel, not free side information.

There is a second effect. The true margin can differ from the reported nominal margin. For any support S with $|S| \le 2k$, Weyl's singular-value perturbation inequality gives

$$\left|\sigma_{\min}((B + \Delta)_S) - \sigma_{\min}(B_S)\right| \le \|\Delta_S\|_{2\to 2} \le \eta.$$

Taking minima over supports yields

$$\gamma_k(B + \Delta) \ge \max\{\gamma_k(B) - \eta, 0\}. \tag{27}$$

Thus a nominal certificate smaller than the operator uncertainty is not robustly positive. A conservative artifact should report both $\widehat{\gamma}_k$ and an uncertainty radius, or report the lower certificate $(\widehat{\gamma}_k - \widehat{\eta})_+$.

What these extensions do and do not establish The deterministic margin is the common geometric quantity in all results above. Probability enters only through a noise radius, covariance metric, or distributional indistinguishability argument. This separation is useful: the same B can be studied under adversarial, Gaussian, or empirical noise without redefining its restricted observability.

The extensions do not imply that observed LLM errors are Gaussian, independent, or exactly sparse. Those assumptions are diagnostics to be checked. The real-model experiment therefore reports parser failures, candidate recall, hidden-test-only code bugs, and replay results separately. A high margin cannot repair a missing candidate, validate an incorrect transport, or convert a dense semantic failure into a sparse one. It quantifies the difficulty of the recovery problem that has actually been specified.

P Statistical validation of a cross-model, cross-task difficulty measure

The mathematical results establish what $\gamma_k(D, A)$ means for a specified observation problem. They do not by themselves establish that the object explains variation in naturally occurring LLM failures. That empirical claim requires a design that separates operator geometry from model competence, item difficulty, candidate availability, and repeated measurements of the same item. This appendix gives the complete inferential protocol used for the authentic-error replay and states which conclusions each statistic can support.

P.1 Estimands and complementary evaluation regimes

Units, indices, and estimands Let m index a model, t a task domain, i an underlying dataset item, and d a relation–anchor design. One replay field is formed by taking correct responses for an item and replacing exactly one node by an authentic incorrect response produced by the same model on the same task family. The replacement is not synthesized by a numerical noise model. It retains the model's actual arithmetic answer or executable program, subject to the replay eligibility rules in Appendix K.

For each tuple (m, t, i, d) we record

$$G_{mtid} = \gamma_1(D_{td}, A_{td}), \qquad E_{mtid} = \frac{\|\widehat{e}_{mtid} - e^\star_{mti}\|_2}{\max\{\|e^\star_{mti}\|_2, 10^{-12}\}},$$

and the exact-recovery indicator

$$Y_{mtid} = \mathbb{I}\{\widehat{z}_{mtid} = z^{\star}_{mti}\}.$$

The exact equality is equality in the task's audited semantic representation: integer equality for mathematics and equality of the complete hidden execution signature for code. Source-string equality is not required.

The primary empirical estimand is conditional association: after holding a model–task stratum and an underlying item fixed, do designs with larger G make the same authentic corruption easier to recover? This is intentionally narrower than claiming that G alone predicts raw accuracy across arbitrary benchmarks. Raw model competence affects whether a usable correct candidate exists and how often the sparse-error model applies.

Why natural evaluation and replay are both necessary The natural-generation table asks a pipeline question. It includes generation, parsing, candidate construction, relation checking, selection, and final task correctness. A failure at any stage lowers performance. That is the operational quantity a practitioner cares about, but it is a noisy test of the information-theoretic theorem.

The replay table asks a mechanism question under theorem-aligned conditions. It guarantees exactly one corrupted node, constructs the corruption from a real model error, and varies the observable design. Correct candidates are present by construction, so the experiment isolates whether the relation–anchor operator makes the error distinguishable. Reporting only replay would conceal candidate-generation limitations; reporting only natural accuracy would make a failed parser look like a counterexample to sparse identifiability. The two tables answer different questions and are not merged into one success rate.

For code, public and hidden execution signatures add another separation. The public signature defines observable relations and anchors. Hidden tests define evaluation correctness. A program that agrees on every public test but fails a hidden test is an observationally invisible bug under that design. Such cases diagnose anchor insufficiency rather than decoder failure. We retain their counts, but replay eligibility requires a wrong public signature when the purpose is to test sparse localization.

P.2 Rank, permutation, predictive, and hierarchical analyses

Within-stratum rank association Absolute margins are comparable only after fixing fiber scaling and row normalization. We impose a common normalization within each task construction and first compute Spearman association separately for every model–task stratum:

$$\widehat{\rho}_{mt} = \operatorname{corr}\big(\operatorname{rank}(G_{mtid}), \operatorname{rank}(-E_{mtid})\big). \tag{28}$$

Average ranks are used for ties. A positive value means that better conditioned designs tend to have lower error; it does not assume a linear dose–response curve. The pooled descriptive correlation is also reported, but the stratum values carry the cross-domain claim because pooling can create Simpson reversals through different base error rates.

Repeated designs from one item are dependent. Treating all (i, d) rows as independent would understate uncertainty. Confidence intervals therefore use a cluster bootstrap: within every model–task stratum, sample items with replacement, carry all ten designs of each selected item together, recompute the statistic, and take percentile endpoints over bootstrap replicates. The resampling seed and number of replicates are stored in the result JSON.

A within-item permutation test The null tested by the permutation procedure is that, conditional on an item and the multiset of observed design outcomes, assignment of margin labels to design outcomes contains no association. For each permutation b, independently shuffle the ten G values among the designs of every item, preserving the model, task, corruption, and outcome values. Let T_0 be the observed pooled rank statistic and T_b its permuted value. The one-sided Monte Carlo p-value is

$$\widehat{p} = \frac{1 + \sum_{b=1}^{B_{\mathrm{perm}}} \mathbb{I}\{T_b \geq T_0\}}{B_{\mathrm{perm}} + 1}. \tag{29}$$

The add-one correction prevents a zero p-value and is valid for Monte Carlo randomization. Because labels move only within item, the test cannot be driven by one model receiving easier items or one task having larger numerical error scales.

The permutation test has a limitation: design outcomes are deterministic functions of a generated field, not randomized interventions collected prospectively. Exchangeability is therefore a reference null for association, not a causal identification theorem. The strongest causal statement in the paper remains mathematical monotonicity under adding rows to a fixed operator, where the intervention is defined algebraically.

Grouped cross-validation and incremental prediction To test whether G adds predictive information beyond coarse nuisance variables, we compare two pre-specified logistic models for exact recovery. The control model uses model identity, task identity, design row count, anchor count, and an observable raw-disagreement measure. The augmented model adds standardized G. Standardization parameters are learned on each training fold and applied unchanged to its held-out fold.

All rows belonging to one item are assigned to the same fold. Let $\mathcal{I}_f$ be held-out items in fold f. Parameters are estimated on items outside $\mathcal{I}_f$, predictions are emitted only for $\mathcal{I}_f$, and the out-of-fold predictions are concatenated before computing area under the receiver-operating-characteristic curve. No reported AUC is a training AUC. The incremental statistic is

$$\Delta\mathrm{AUC} = \mathrm{AUC}_{\mathrm{controls}+\gamma} - \mathrm{AUC}_{\mathrm{controls}}.$$

Keeping the item grouped is essential: random row splits would let the model infer an item's outcome from nine nearly identical siblings and predict the tenth.

The fitted coefficient is not interpreted per raw unit because margins depend on normalization. We standardize the transformed margin using training-fold mean and variance, then apply those training statistics to the held-out fold. A positive out-of-fold coefficient and positive $\Delta\mathrm{AUC}$ support incremental prediction. They do not show that the logistic link is the true recovery law.

Hierarchical sensitivity model As a sensitivity analysis, one may fit a mixed-effects specification

$$\operatorname{logit}\mathbb{P}(Y_{mtid}=1) = \alpha + u_m + v_t + w_i + \beta\widetilde{G}_{td} + \theta^\top X_{mtid}, \tag{30}$$

where u_m, v_t, and w_i are model, task, and item intercepts and X contains observable controls. With only four models and two task domains, random-effect asymptotics for u_m and v_t are weak; fixed stratum effects plus item-clustered uncertainty are safer in the present experiment. Equation (30) is included to specify how the test should scale when more open-weight models and domains become available.

A continuous-error sensitivity model regresses $\log(E+\epsilon_E)$ on $\log(G+\epsilon_G)$ with item fixed effects. Because zero-margin designs can produce exact recovery on favorable instances, a hurdle model separating $G=0$ from $G>0$ is preferable to forcing one linear slope through the discontinuity. Rank statistics remain the primary analysis because they require fewer functional assumptions.

P.3 Confounds, falsification, uncertainty, and interpretation

Confounds and the controls that address them

More rows. A larger margin may accompany more relation rows. Row count is included as a control, and normalized duplicate families provide a negative control: they increase count without changing $B^\top B$ or γ_k.

More anchors. Anchors can simultaneously raise the margin and directly reveal truth. This is not a spurious relationship; anchors are part of the observation problem. Nevertheless, anchor count is controlled so the statistic tests more than the binary presence of an anchor. Designs with equal counts but different placement test coverage geometry.

Item difficulty. All design comparisons reuse the same generated field. Item grouping in permutation, bootstrap, and cross-validation prevents between-item competence from being mistaken for a spectral effect.

Model competence. Associations are reported within model–task strata before pooling. A stronger model may produce fewer replay-eligible errors, which changes precision and selection, but it cannot by itself create a within-item design ordering.

Parser success. Natural accuracy uses all generated rows and reports parse rate. Replay uses only fields for which the typed semantic representation is defined. The eligibility count is therefore part of the result, not an invisible preprocessing loss.

Candidate recall. A selector cannot return an absent truth. Natural results report candidate recall, and replay fixes recall by construction. Improvements in replay recovery cannot be attributed to generating more candidates for high-margin designs because the candidate set is held fixed across designs.

Negative controls and falsification checks Four negative controls are built into the artifact.

1. **Consistency-only fields.** A transported common-mode error has zero relation defect. Any method using only D should fail to distinguish it from truth.
2. **Uncovered components.** A design with an unanchored sparse component has $\gamma_k = 0$. Recovery may succeed on some nodes, but uniform exact recovery should fail for corruptions inside the invisible component.
3. **Normalized duplicates.** Repeating an identical relation with weights divided by the square root of the multiplicity preserves $B^\top B$ and the margin. An implementation reporting spectral improvement fails this audit.
4. **Shuffled margins.** Within-item permutation destroys the alignment between design geometry and outcomes while preserving all marginal distributions. Predictive improvement should disappear under this shuffle.

The theory would face a substantive empirical challenge if the observed statistic were indistinguishable from its shuffled reference in every model–task stratum, if its sign reversed systematically across domains after normalization, or if row-count controls explained all held-out gain. A single favorable correlation is not enough; direction, grouping, negative controls, and out-of-fold performance must agree.

Multiplicity, uncertainty, and reporting discipline The primary outcomes and analyses are fixed as follows: exact replay recovery, normalized replay error, stratum Spearman correlations, one within-item permutation test, one item-cluster bootstrap interval, and one grouped cross-validation AUC comparison. Family-level tables, alternative error normalizations, and hidden-only code analyses are secondary. This ordering prevents selecting the most favorable statistic after seeing results.

The artifact stores full-precision numbers; the paper rounds only for readability. A p-value is reported with its Monte Carlo resolution $1/(B_{\mathrm{perm}}+1)$. Confidence intervals describe sampling variation over the finite item construction, not uncertainty over all possible prompts, models, or future benchmarks. Model decoding is seeded, but deterministic GPU kernels are not assumed unless the environment provides them; raw generations are therefore included so every downstream statistic can be reproduced without regenerating text.

Interpretation ladder The empirical evidence supports claims at three distinct levels.

1. **Mechanism verification:** anchor coverage, duplicate normalization, and error-versus-margin trends behave as the linear theory predicts on controlled instances.
2. **Cross-stratum prediction:** the same normalized object orders replay difficulty within multiple model–task strata and improves grouped held-out discrimination.
3. **External scope:** the result suggests, but does not prove, usefulness for other models, tasks, semantic representations, or natural dense failures.

Only the first level follows tightly from the theorem assumptions. The second is the paper's decisive evidence for treating $\gamma_k(D, A)$ as more than a method-specific score. The third remains a program for subsequent work. Maintaining this ladder prevents a small but controlled experiment from being narrated as a universal benchmark law.

Q EXACT MARGINS AND ANCHOR GEOMETRY FOR CANONICAL GRAPH DESIGNS

This appendix derives γ_k for elementary response fields without relying on numerical singular-value routines. The calculations explain the anchor phase transition and redundancy saturation experiments and clarify a subtle point: global injectivity, sparse injectivity, and numerical conditioning are different properties.

Q.1 LAPLACIAN FORMULATION AND CLOSED-FORM CANONICAL EXAMPLES

Identity transports and anchored graph Laplacians Let every node fiber be $\mathbb{R}^d$, every transport be the identity, and every relation edge $e = (i, j)$ have weight $w_e > 0$. Choose an arbitrary orientation and let $C \in \mathbb{R}^{|E|\times n}$ be the weighted incidence matrix, whose row for e is $\sqrt{w_e}(e_j - e_i)^\top$. Then

$$D = C \otimes I_d, \qquad D^\top D = L_w \otimes I_d,$$

where $L_w = C^\top C$ is the weighted graph Laplacian. Suppose a scalar anchor of strength $a_i \geq 0$ acts isotropically on node i. With $Q = \mathrm{diag}(a_1^2, \ldots, a_n^2)$,

$$B^\top B = (L_w + Q) \otimes I_d. \tag{31}$$

For a node set S, restriction of columns gives

$$B_S^\top B_S = ((L_w + Q)_{SS}) \otimes I_d.$$

Consequently

$$\gamma_k(D, A)^2 = \min_{\varnothing \neq S \subseteq V,\ |S|\leq 2k} \lambda_{\min}((L_w + Q)_{SS}). \tag{32}$$

This is an exact principal-submatrix formula, not a graph-level heuristic. It also shows why the smallest eigenvalue of the full anchored Laplacian can be overly pessimistic when $k \ll n$: γ_k minimizes only over small principal supports.

Proposition 32 (Null space and sparse anchor coverage)**.** *Let the graph have connected components $V_1, \ldots, V_c$. A component is anchored if $a_i > 0$ for at least one i in that component. Then:*

1. *B is globally injective if and only if every component is anchored;*
2. *$\gamma_k(D, A) = 0$ if and only if some unanchored component has at most $2k$ nodes.*

The statements hold for every fiber dimension $d \geq 1$.

Proof. Because $\|Bx\|_2^2 = \sum_{(i,j)\in E} w_{ij}\|x_i - x_j\|_2^2 + \sum_i a_i^2\|x_i\|_2^2$, a vector lies in $\ker B$ exactly when it is constant on each connected component and its constant is zero on every anchored component. Thus the null space contains one copy of $\mathbb{R}^d$ for every unanchored component, proving the first claim.

Every nonzero null vector is constant and nonzero on at least one complete unanchored component. Its group support therefore contains all nodes of that component. The smallest possible nonzero null support is the size of the smallest unanchored component. By the exact identifiability equivalence, $\gamma_k = 0$ exactly when that support size is at most $2k$. □

This proposition refines the informal slogan "one anchor per component." One anchor per component is necessary and sufficient for recovery without a sparsity restriction. For k-sparse recovery, a large unanchored component can have no $2k$-sparse null vector, so γ_k can be positive even though B is globally singular. The field is then identifiable only over the declared sparse class.

A two-node field in closed form Consider two scalar nodes joined by one identity relation and anchor node 1 with strength $a \geq 0$:

$$B = \begin{bmatrix} -1 & 1 \\ a & 0 \end{bmatrix}, \qquad B^\top B = \begin{bmatrix} 1 + a^2 & -1 \\ -1 & 1 \end{bmatrix}.$$

For $k = 1$, supports of size at most 2 include the full field, so

$$\gamma_1(D, A)^2 = \frac{2 + a^2 - \sqrt{a^4 + 4}}{2}. \tag{33}$$

At $a = 0$ the margin is zero: adding the same scalar to both responses changes neither relation residual. For every $a > 0$ the margin is positive. As $a \downarrow 0$, a Taylor expansion gives $\gamma_1^2 = a^2/2 + O(a^4)$; hence recovery is identifiable but badly conditioned under a weak anchor. As $a \to \infty$, $\gamma_1^2 \to 1$, so relation strength becomes the bottleneck. Merely making an already strong anchor stronger cannot push the margin beyond the information carried through the edge.

The d-dimensional isotropic case repeats each eigenvalue d times and has the same margin. If the transport is an orthogonal matrix T, change variables at node 2 by $x_2 \mapsto T^{-1}x_2$; singular values are unchanged. A nonorthogonal transport changes the fiber metric and must be normalized before margins across relation types are compared.

Q.2 Redundancy, path geometry, and measurement design

Duplicate relations under fixed information budget Suppose one relation row block is R. Repeating it r times without normalization creates

$$D_{\text{raw}} = \begin{bmatrix} R \\ \vdots \\ R \end{bmatrix}, \qquad D_{\text{raw}}^\top D_{\text{raw}} = rR^\top R.$$

This increases singular values by $\sqrt{r}$, but it also assumes r independent measurements each with the original precision. Prompt duplication does not justify that assumption when copies are deterministic or perfectly correlated.

Under a fixed total family weight, each copy receives coefficient $r^{-1/2}$:

$$D_{\text{norm}} = \frac{1}{\sqrt{r}} \begin{bmatrix} R \\ \vdots \\ R \end{bmatrix}.$$

Then

$$D_{\text{norm}}^\top D_{\text{norm}} = R^\top R. \tag{34}$$

For any fixed anchor A, the stacked Gram matrix and every restricted principal Gram matrix are identical to those with one copy. Therefore

$$\gamma_k(D_{\text{norm}}, A) = \gamma_k(R, A)$$

for every k. The synthetic artifact constructs the repeated matrix, rescales its rows, and recomputes the margin; it does not insert the one-copy value into the result table.

If copies have independent random noise, whitening may legitimately yield more information. If their noise correlation is ρ, the gain lies between the independent and identical-copy extremes and is determined by the inverse covariance, not the raw number of prompts. This is why Appendix O defines $\gamma_{k,\Sigma}$.

A path graph with one anchor Let $V = \{1, \dots, n\}$, unit-weight edges connect consecutive nodes, and node 1 has anchor strength $a > 0$. For a scalar vector x,

$$\|Bx\|_2^2 = a^2 x_1^2 + \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2.$$

Write $\Delta_i = x_{i+1} - x_i$. Since

$$x_j = x_1 + \sum_{i=1}^{j-1} \Delta_i,$$

Cauchy–Schwarz gives

$$|x_j|^2 \le 2|x_1|^2 + 2(j-1)\sum_{i=1}^{j-1}|\Delta_i|^2 \le 2|x_1|^2 + 2n\sum_{i=1}^{n-1}|\Delta_i|^2.$$

Summing over j yields

$$\|x\|_2^2 \le 2n|x_1|^2 + 2n^2\sum_{i=1}^{n-1}|\Delta_i|^2 \le C_{n,a}\|Bx\|_2^2, \quad C_{n,a} = \max\{2n/a^2, 2n^2\}. \tag{35}$$

Hence the global smallest singular value satisfies

$$\sigma_{\min}(B) \ge C_{n,a}^{-1/2}.$$

The n^{-1} scaling of this elementary bound captures the fact that an error far from the anchor is communicated through a long chain. Stronger discrete Poincare inequalities improve constants but not the qualitative lesson: connectivity establishes visibility, while path length controls conditioning.

For $k \ll n$, formula (32) can give a larger restricted margin because no admissible witness occupies the entire path. This illustrates why k belongs in the definition of the difficulty object. A relation design may be poor for dense field reconstruction yet adequate for one localized response failure.

Adding measurements and comparing anchor placements Let B be an existing operator and let C be any additional normalized measurement block. For every support S,

$$[B;C]_S^\top [B;C]_S = B_S^\top B_S + C_S^\top C_S \succeq B_S^\top B_S.$$

Therefore

$$\gamma_k([B;C]) \ge \gamma_k(B). \tag{36}$$

Equality is possible when C is blind to a minimizing witness of B. The theorem justifies greedy query acquisition but also shows why counting new rows is insufficient.

Suppose $h^\star$ is a unit $2k$-sparse witness attaining $\|Bh^\star\|_2 = \gamma_k(B)$. For a candidate scalar measurement $c^\top$, the energy of this witness becomes

$$\|Bh^\star\|_2^2 + |c^\top h^\star|^2.$$

A candidate with large $|c^\top h^\star|$ attacks the current weakest direction. Recomputing the exact margin after each candidate remains necessary because the minimizing support or vector can switch. The witness score is a principled screening rule, not an exact submodular guarantee.

Anchor placement has the same geometry. In an identity-transport component, anchoring a node on which $h^\star$ has negligible mass barely changes the current bottleneck; anchoring a high-mass node can increase it sharply. Once that direction is lifted, another component or support can become limiting, producing the piecewise-smooth phase curves observed in exact enumeration.

Q.3 Typed transports and audit procedures

Typed transports and gauge transformations The Laplacian formulas above use identity transports only for transparency. Let every edge transport T_{ij} be invertible and suppose there exist node maps U_i such that

$$T_{ij} = U_j U_i^{-1}.$$

This is a flat transport system. Under the change of variables $\widetilde{x}_i = U_i^{-1}x_i$, relation consistency becomes $\widetilde{x}_j - \widetilde{x}_i = 0$. If every U_i is orthogonal in the declared fiber metric, the change is an isometry and the restricted margin is exactly that of an identity-transport graph with transformed anchors.

If the U_i are merely invertible, let

$$\kappa_- = \min_i \sigma_{\min}(U_i), \qquad \kappa_+ = \max_i \sigma_{\max}(U_i).$$

The block-diagonal change of coordinates bounds norms by $\kappa_-\|\widetilde{x}\| \le \|x\| \le \kappa_+\|\widetilde{x}\|$. Margins in the two coordinate systems can therefore differ by condition-number factors. Comparing relation types without declaring fiber metrics can mistake a change of units for easier recovery.

Non-flat transports have nontrivial cycle holonomy: transporting around a cycle need not return the original vector. Then a globally consistent nonzero field may not exist, and the kernel of D is governed jointly by graph topology and transport products. The general restricted singular-value definition remains valid; the simple component criterion does not. This is precisely why the framework begins with typed linear operators rather than assuming every relation is equality.

Audit recipe for a new design For a finite candidate design, an exact audit proceeds as follows.

1. Fix fiber coordinates, norms, transport maps, row weights, anchor maps, and k before observing repair outcomes.
2. Assemble $B = [D; A]$ using the same parser and normalization used at inference.
3. Enumerate every nonempty node support S with $|S| \leq 2k$ and compute $\sigma_{\min}(B_S)$ with a documented numerical tolerance.
4. Store the minimizing support and right singular vector as a witness. Verify directly that its group support and Rayleigh quotient match the report.
5. For zero margins, construct the indistinguishable pair from a sparse null vector. For positive margins, perturb along the witness to check the predicted $1/\gamma_k$ sensitivity.
6. If rows are estimated or stochastic, repeat the computation after whitening and report an operator-uncertainty lower certificate.

This recipe turns γ_k from an abstract symbol into a falsifiable certificate. Every value in the experiment artifact is accompanied by enough operator information to rerun these checks without model weights.

Table 12: Direct RRF repair on unmodified natural response fields. RRF-D performs relation-only repair, while RRF-D+A augments the same decoder with arithmetic or public-execution anchors. Results are exact-answer accuracy for mathematics and hidden-test pass@1 for code, reported as percentages.

Task	Model	RRF-D	RRF-D+A
Math	Qwen2.5-0.5B	20.3	22.7
	Phi-3-mini	35.9	37.5
	Qwen2.5-7B	82.8	85.2
	Qwen2.5-14B	90.6	92.2
Code	Qwen2.5-0.5B	71.9	75.0
	Phi-3-mini	100.0	100.0
	Qwen2.5-7B	100.0	100.0
	Qwen2.5-14B	100.0	100.0

Q.4 Direct RRF Repair on Natural Response Fields

The controlled replay experiments isolate the effect of the relation–anchor operator under an exactly one-node corruption model. We additionally evaluate whether the proposed decoder can be applied directly to the original, unmodified response fields generated by the four checkpoints. Each field contains the parsed responses to the four prespecified transformations of one mathematics or code item. Unlike the replay setting, these natural fields may contain multiple incorrect nodes, coherent errors shared across transformations, parsing failures, or no correct candidate. For each node, we construct a finite candidate set from its original parsed response and the valid responses transported from the other nodes. We then apply the discrete candidate-field decoder described in Section **??**. The decoder searches over candidate fields that modify at most one response node and selects the field with the smallest normalized relation–anchor energy. Ties are resolved deterministically by preferring fewer edits and then retaining the original response. No ground-truth correctness labels or hidden code tests are used during candidate construction or repair. We report two variants. **RRF-D** uses only the normalized typed relation operator D, and therefore selects fields according to cross-transformation compatibility. **RRF-D+A** uses the same candidate sets and decoder but additionally incorporates the available anchor operator A. Mathematics anchors are derived from externally checkable arithmetic constraints, while code anchors use execution signatures on the public tests. Hidden tests are reserved exclusively for final code evaluation. Table 12 reports exact-answer accuracy for mathematics and

hidden-test pass@1 for code. Because the two variants differ only in whether independent anchors are included, their comparison isolates the empirical value of attaching relational consistency to external evidence. An improvement from RRF-D to RRF-D+A indicates that relations alone leave ambiguity that the anchors help resolve.

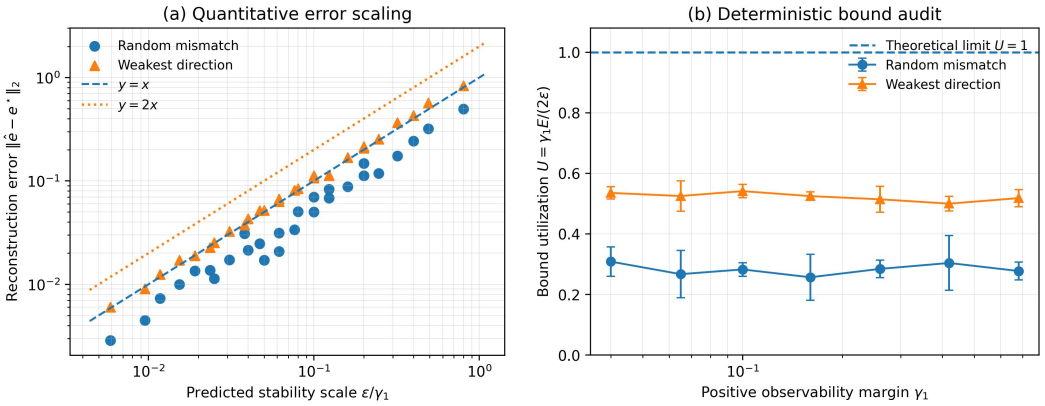

Figure 3: Quantitative stability under positive-margin relation mismatch. All operator designs satisfy $\gamma_1(D, A) > 0$. Left: reconstruction error $E = \|\widehat{e} - e^\star\|_2$ as a function of the predicted stability scale ϵ/γ_1. Circles denote random mismatch directions and triangles denote weakest-direction perturbations; the lines $y = x$ and $y = 2x$ indicate the minimax scale and the deterministic upper-bound scale, respectively. Right: bound utilization $U = \gamma_1 E/(2\epsilon)$ as a function of γ_1. The horizontal line $U = 1$ is the deterministic stability limit from Theorem 4.

Q.5 Quantitative Stability at Positive Margin

The preceding experiments test the qualitative transition from non-identifiability to identifiability and whether larger margins predict easier recovery across model–task strata. We next test the sharper quantitative prediction of Theorems 4 and 5: when the corruption budget and mismatch radius are fixed, reconstruction error should be organized by the scale $\epsilon/\gamma_k(D, A)$, rather than merely by whether the margin is zero.

Positive-margin operator family. We construct a normalized family of four-node relation operators with identity transports and edge set

$$(1, 2), \qquad (2, 3), \qquad (3, 4).$$

For bridge parameter $\delta > 0$, the relation-family weights are

$$w_{12} = w_{34} = \frac{1}{2+\delta}, \qquad w_{23} = \frac{\delta}{2+\delta},$$

so the total relation weight remains one for every design. The resulting operator is

$$D_\delta z = \begin{bmatrix} \sqrt{w_{12}}(z_2 - z_1) \\ \sqrt{w_{23}}(z_3 - z_2) \\ \sqrt{w_{34}}(z_4 - z_3) \end{bmatrix}.$$

We use

$$\delta \in \{10^{-4}, 3{\times}10^{-4}, 10^{-3}, 3{\times}10^{-3}, 10^{-2}, 3{\times}10^{-2}, 10^{-1}, 0.3, 1\}.$$

Every $\delta > 0$ produces a connected relation graph and satisfies $\gamma_1(D_\delta, 0) > 0$. As the bridge becomes weaker, however, a field that is approximately constant on nodes $\{1, 2\}$ and zero elsewhere becomes increasingly difficult to observe, causing γ_1 to approach zero continuously. We compute each margin exactly by enumerating all node supports of size at most two.

Relation mismatch and exact decoding. For a one-node error $e^\star$, the observed residual is

$$s = D_\delta e^\star + \xi_D, \qquad \|\xi_D\|_2 = \epsilon.$$

Recovery uses the exact residual-minimizing decoder

$$\widehat{e} \in \arg\min_{\|e\|_{0,g}\le 1} \|D_\delta e - s\|_2.$$

Because the true error has residual at most ϵ, the minimizing estimate also has residual at most ϵ. Theorem 4 therefore implies

$$E := \|\widehat{e} - e^\star\|_2 \le \frac{2\epsilon}{\gamma_1(D_\delta, 0)}.$$

We evaluate a Cartesian grid of positive mismatch radii and retain only the unsaturated regime in which the amplitude constraint does not truncate the theoretical scale.

We use two mismatch constructions. For *random mismatch*, ξ_D is sampled uniformly from the observation-space sphere of radius ϵ, with fixed seeds and repeated trials. This evaluates typical robustness but need not attain the worst restricted direction. For the *weakest-direction* stress test, let

$$h_\delta^\star \in \arg\min_{\substack{h\neq 0 \\ \|h\|_{0,g}\le 2}} \frac{\|D_\delta h\|_2}{\|h\|_2}, \qquad \|h_\delta^\star\|_2 = 1.$$

We split its support into two one-node errors, $h_\delta^\star = e_\delta^{(0)} - e_\delta^{(1)}$, scale their separation to $2\epsilon/\gamma_1$, and place the observation at the midpoint of their two residuals. The observation is then within distance ϵ of both candidates, so every decoder must incur error at least ϵ/γ_1 on one member of the pair. This construction directly targets the minimax obstruction rather than relying on randomly sampled noise to discover it.

Quantitative audit. Figure 3 plots reconstruction error against the predicted scale

$$x = \frac{\epsilon}{\gamma_1}, \qquad y = E.$$

The lines $y = x$ and $y = 2x$ show, respectively, the unsaturated minimax lower scale and the deterministic upper-bound scale. The right panel reports the bound-utilization ratio

$$U = \frac{\gamma_1 E}{2\epsilon}.$$

Under the assumptions of Theorem 4, exact feasible decoding requires $U \le 1$, up to numerical tolerance, while a weakest-direction construction attaining the minimax scale gives $U \approx 1/2$.

Across all positive-margin designs, reconstruction error has pooled rank association $\rho = 0.959$ with ϵ/γ_1. A log–log regression gives slope $\widehat{\beta} = 1.021$ with 95% confidence interval $[0.942, 1.099]$. The weakest-direction trials concentrate near the $y = x$ reference scale, whereas random mismatch generally produces smaller error because it is not systematically aligned with the least observable restricted direction. The maximum observed utilization is $\max U = 0.578$, and the fraction exceeding one beyond numerical tolerance is 0.00%.

These results test a stronger claim than a zero-to-positive phase transition. Every evaluated design is identifiable, yet its sensitivity to relation mismatch changes quantitatively with the inverse observability margin. The weakest-direction construction further shows that the $1/\gamma_1$ dependence is not only an upper-bound artifact: perturbations aligned with the least observable admissible direction approach the minimax scale predicted by Theorem 5.