# LOCAL SUBSTITUTIONS DO NOT COMPOSE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A model can be nearly insensitive to individual identity substitutions while still depending on the correspondence among many inputs. When those identities are forgotten together, the loss need not remain small. The relevant question is therefore not only whether one identity can replace another, but which distinctions can be erased jointly and which relations must remain. We formulate this question through *behavioral quotients*, which identify concrete states under declared transformations and measure approximation error for a fixed response and query distribution. Under uniform coalition queries, we derive sharp bounds on block quotient error from singleton substitutions and finite query certificates valid after partition selection. We further show that every singleton discrepancy can vanish as the number of bound objects grows while accurate block quotients still require exponentially many states at a fixed positive error tolerance. Coordinated renaming that preserves the bindings instead gives an exact quotient with polynomially many states. Controlled Transformer experiments exhibit the corresponding finite size separation in learned responses. Finally, we decompose shared prediction error into residual identity dependence, disagreement across aligned instances, and fitting error, with direct diagnostics for the first two terms.

## 1 INTRODUCTION

Replacing a concrete identity can leave a model's response almost unchanged even when the response depends on relations among many inputs. If many identity distinctions are then erased together, those relations may be lost. The resulting question is not only whether individual substitutions have small effects, but whether the corresponding distinctions can be forgotten jointly without losing behavior. Behavioral tests probe responses to controlled input changes (Ribeiro et al., 2020; Li et al., 2024), while studies of entity binding examine how models maintain correspondences among inputs (Feng & Steinhardt, 2024). When several positions refer to the same object, an independent exchange can break a binding that coordinated renaming would preserve. The response may depend on which inputs belong together rather than on the names assigned to them. Even when each such change has only a small effect, discarding many correspondences can produce substantial behavioral loss. The missing step is therefore to relate local exchanges to the error of a joint abstraction while making the preserved relational structure explicit.

To address this gap, we formulate identity erasure as an approximation problem and introduce *behavioral quotients*, representations that identify concrete states under a declared transformation family. For a fixed response and query distribution, quotient error is the loss of the best predictor that uses only the resulting abstract state. We relate this error to *contextual substitutions*, which exchange identities while keeping the surrounding context fixed. For independent exchanges within blocks under uniform coalition queries, aggregate discrepancies certify quotient error, including after the partition is selected from finite queries. Figure 1 previews the central phenomenon. Every isolated identity merge remains below the target error tolerance, so each distinction appears individually dispensable. Yet these local certificates do not compose: the smallest feasible block quotient still requires 243 states. A coordinated quotient that preserves the latent pairing avoids this failure and is exact with only 56 states. Reuse asks for more: aligned quotient states must still support consistent responses across instances.

Our main contributions are summarized below.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

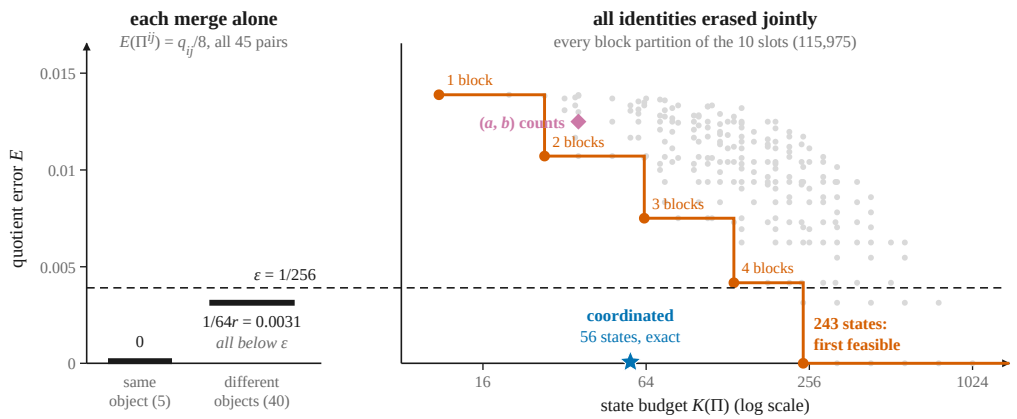



Figure 1: Individually safe identity merges need not compose. Relation-count response $v_r$ with $r = 5$ objects and tolerance $\varepsilon = 1/256$ (dashed). Left: every isolated slot merge is feasible. Its quotient error is zero for the 5 bound pairs and $1/(64r)$ for the 40 cross-object pairs. Right: quotient error versus state budget $K(\Pi)$ for all block partitions (grey), with the best achievable error at each budget shown in orange. The smallest feasible block quotient has 243 states. Aggregate $(a, b)$ counts lose the pairing at 36 states, whereas coordinated renaming preserves it and is exact with 56 states (star). Exact values for $v_r$ and learned-model results agree closely.

- We derive sharp upper and lower bounds on block quotient error from singleton contextual substitutions, together with simultaneous finite query certificates that remain valid after partition selection.
- We prove a fixed error separation for a bounded relation count family with $r$ bound objects. All singleton discrepancies are $O(1/r)$, yet accurate block quotients require $2^{\Omega(r)}$ states, while coordinated renaming gives an exact quotient with $O(r^3)$ states.
- We demonstrate the corresponding separation in five trained Transformers. At $r = 5$ and error tolerance $1/256$, exhaustive evaluation finds a minimum feasible block state count of 243, while a coordinated quotient with 56 states has near zero error.
- We separate residual identity dependence from disagreement across aligned instances and fitting error. A four query diagnostic estimates the first two terms, and a finite state learning bound quantifies estimation error on a fixed abstract interface.

## 2 Related Work

Behavioral tests measure how model responses change under controlled input variations (Ribeiro et al., 2020; Li et al., 2024). Interaction analyses describe how subsets contribute jointly to the response (Deng et al., 2022; Li & Zhang, 2023; Lu et al., 2023). Our question begins when these observations are used to justify erasing concrete identity distinctions. An input can contribute to a response without its concrete identity being indispensable. We therefore measure the optimal prediction error after a proposed identity erasure and relate it to aggregate substitution discrepancies, including when the quotient is selected from the same finite queries. Causal abstraction tests whether a higher level model reproduces intervention behavior (Geiger et al., 2025). Li et al. (2026) use pairwise interchange interventions to locate input regions where a proposed interpretation is more or less faithful. Our quotient instead groups input identities and induces orbits of concrete coalitions. We certify the response information lost within those orbits without requiring access to internal activations. Across instances, this still leaves a separate question of whether aligned quotient states support the same response.

Preserving behavior under a change of names depends on how the corresponding relations are treated. Role–filler binding has a long history in compositional representations (Martin & Doumas, 2020), and recent studies examine entity binding in language models and architectures that separate relations from object features (Feng & Steinhardt, 2024; Altabaa et al., 2024). We ask what these

relations imply for quotient size. Independent exchanges can destroy correspondence that coordinated renaming preserves. Our fixed error separation shows that the resulting state budgets can differ exponentially even when every individual substitution becomes weak, and the controlled experiment instantiates this contrast in a learned response. Our analysis also draws on symmetry and statistical estimation. Partial symmetry and symmetric influence have been studied for Boolean functions (Blais et al., 2015). Reynolds averaging supplies the invariant projection, and Johnson slice analysis supplies the exchange spectrum (Sannai et al., 2024; Filmus, 2016). In our setting, the spectrum relates local variation to quotient error, while simultaneous concentration of pairwise estimates (Hoeffding, 1963) preserves the certificate when the partition is chosen from the data.

## 3 BEHAVIORAL QUOTIENTS

We formalize the distinction between replacing one identity and erasing several distinctions jointly. Contextual substitutions measure the former, while a quotient specifies which concrete states are identified and how much response information is lost by that identification.

### 3.1 QUERY PROTOCOL AND BLOCK QUOTIENTS

Let $N = [n]$ be a set of players and let $v : 2^N \to [0, 1]$ denote the response being studied. A coalition $S$ specifies one concrete binary state. In the controlled experiment of Section 6, $u \in S$ means that field $u$ has value one and $u \notin S$ means that it has value zero; every rendered model input still contains all fields. More generally, a neural application must fix the masking, baseline, or conditional completion protocol that defines $v(S)$. To compare two identities while holding the external context fixed, let $A, B, C$ be pairwise disjoint with $|A| = |B|$ and define

$$D(A, B \mid C) = v(C \cup A) - v(C \cup B). \tag{1}$$

The main results use uniform coalition queries,

$$S \sim u = \mathrm{Unif}(2^N), \tag{2}$$

with nonuniform extensions discussed in Appendix D. For singleton substitutions, $C$ is uniform on $2^{N\setminus\{i,j\}}$ and

$$q_{ij} = \mathbb{E}_C D(i, j \mid C)^2. \tag{3}$$

Thus $q_{ij}$ is the average squared cost of replacing $i$ by $j$ under this declared query protocol. We first study identity erasure through independent exchanges within blocks. Let $\Pi = \{B_1, \ldots, B_r\}$ be a partition of $N$. Its block quotient retains only the number of selected identities in each block,

$$t_\Pi(S) = \big(|S \cap B_1|, \ldots, |S \cap B_r|\big), \tag{4}$$

and is invariant under

$$G_\Pi = \mathrm{Sym}(B_1) \times \cdots \times \mathrm{Sym}(B_r). \tag{5}$$

An identity that should remain visible can stay in a singleton block. Among all predictors that use only the quotient state, the optimal one is

$$P_\Pi v(S) = \mathbb{E}[v(T) \mid t_\Pi(T) = t_\Pi(S)], \qquad E(\Pi) = \|v - P_\Pi v\|_2^2. \tag{6}$$

Because $P_\Pi v$ is already the best predictor available after the identification, $E(\Pi)$ is the irreducible behavioral loss caused by this quotient rather than fitting error of a particular predictor. Its structural state budget is

$$K(\Pi) = \prod_{B \in \Pi} (|B| + 1), \qquad c(\Pi) = \log_2 K(\Pi). \tag{7}$$

Here $K(\Pi)$ counts states within the declared quotient family and does not characterize the size of a particular model architecture.

### 3.2 EXACT EQUIVALENCE AND OPTIMAL PROJECTION

Before allowing approximation, exact contextual interchangeability, permutation invariance, representation by block counts, and zero quotient error coincide.

**Theorem 1** (Exact equivalence for block quotients)**.** *For a fixed partition $\Pi$, the following are equivalent: (i) $D(i, j \mid C) = 0$ for every pair $i, j$ in the same block and every external context $C$; (ii) $v(gS) = v(S)$ for all $g \in G_\Pi$ and $S \subseteq N$; (iii) $v(S) = F(t_\Pi(S))$ for some $F$; and (iv) $E(\Pi) = 0$. Moreover, if $i \sim j$ means $i = j$ or the pair is interchangeable in every external context, then $\sim$ is an equivalence relation and its classes form the unique coarsest exact block partition.*

Transpositions within each block generate $G_\Pi$, and equal block counts characterize its orbits, which gives the equivalence. The exact case is canonical; the remainder of the paper studies approximate quotients, where several distinct partitions can be feasible.

**Theorem 2** (Optimal projection and orbit substitution identity)**.** *For any $F$ that depends only on $t_\Pi$,*

$$\|v - F(t_\Pi)\|_2^2 = E(\Pi) + \|P_\Pi v - F(t_\Pi)\|_2^2. \tag{8}$$

*If we first draw an orbit according to its probability under $u$ and then draw $S, T$ independently and uniformly within that orbit, then*

$$\boxed{E(\Pi) = \frac{1}{2}\mathbb{E}[v(S) - v(T)]^2 = \frac{1}{2}\mathbb{E}D(S \setminus T, T \setminus S \mid S \cap T)^2.} \tag{9}$$

Equation equation 9 gives quotient error a direct behavioral meaning: it is half the expected disagreement between two concrete states that the quotient treats as the same. We next ask whether one common set of local substitution measurements can certify the error of many proposed identity erasures.

## 4 Local substitutions certify block quotients

A proposed block quotient identifies many concrete states at once, while a singleton substitution probes only one local exchange. Within each quotient orbit, these exchanges form the edges of an exchange graph. We ask when variation along those edges controls the total variation that remains inside the orbit. For a partition $\Pi$, aggregate the discrepancies of the pairs it intends to merge,

$$J(\Pi) = \sum_{B \in \Pi} \frac{1}{4|B|} \sum_{\substack{i<j \\ i,j \in B}} q_{ij}. \tag{10}$$

Let

$$\kappa_\Pi = \sum_{B \in \Pi} \frac{d_B(|B| - d_B + 1)}{|B|}, \qquad d_B = \lfloor |B|/2 \rfloor, \tag{11}$$

where singleton blocks contribute zero.

**Theorem A (Local certification and forbidden merges).** *For every real function $v$ and every partition $\Pi$,*

$$\boxed{E(\Pi) \le J(\Pi) \le \kappa_\Pi E(\Pi).} \tag{12}$$

*If $i, j$ lie in the same block, then*

$$\boxed{E(\Pi) \ge q_{ij}/8.} \tag{13}$$

*Both constants in equation 12, and the factor $1/8$ in equation 13, are attainable over the unrestricted function class.*

Conditioned on the block counts, the quotient predictor is constant on one exchange graph. The quantity $J(\Pi)$ is its Dirichlet energy: it measures variation across neighboring states obtained by within-block swaps. The quotient error $E(\Pi)$ measures the total variation around the best constant on each orbit. The exchange spectrum has smallest positive eigenvalue one and largest eigenvalue $\kappa_\Pi$, which gives Equation equation 12; Appendix A derives the spectrum and sharpness results. The two directions support different conclusions. If $J(\Pi)$ is small, the complete proposed quotient is certified to have small global error. Conversely, a pair with $q_{ij} > 8\varepsilon$ cannot appear in any block quotient with error at most $\varepsilon$. For an isolated merge these statements meet exactly. Let $\Pi^{ij}$ have $\{i, j\}$ as its only nontrivial block. Then $J(\Pi^{ij}) = q_{ij}/8$ and $\kappa_{\Pi^{ij}} = 1$, so Theorem A gives

$$\boxed{E(\Pi^{ij}) = q_{ij}/8.} \tag{14}$$

Thus a small $q_{ij}$ can make that one merge genuinely safe. The next question is whether many individually safe merges remain safe when their identity distinctions are erased together.

### 4.1 CERTIFICATION FROM FINITE QUERIES AFTER SELECTION

The population discrepancies are unknown in practice, and the same estimates may be used both to choose a partition and to certify it. We therefore require one event that controls all pairwise estimates simultaneously. For each $i < j$, draw $M$ independent uniform contexts and define

$$\widehat{q}_{ij} = \frac{1}{M} \sum_{m=1}^{M} D(i, j \mid C_m)^2, \qquad h = \sqrt{\frac{\log(n(n-1)/\delta)}{2M}}. \tag{15}$$

**Theorem B (Finite sample certificate after selection).** *With probability at least $1 - \delta$, all pairs satisfy*

$$|\widehat{q}_{ij} - q_{ij}| \le h. \tag{16}$$

*On the same event, every partition simultaneously satisfies*

$$E(\Pi) \le \widehat{J}(\Pi) + w_\Pi h, \qquad w_\Pi = \frac{n - |\Pi|}{8}. \tag{17}$$

*Therefore a partition chosen after observing the same $\widehat{q}$ remains certified whenever its bound is at most the target error $\varepsilon$.*

The proof first controls all $\binom{n}{2}$ pairwise estimates by Hoeffding's inequality. Every partition is a fixed weighted sum of those pairs, so the same event controls all partitions without a separate union bound over the exponentially many partitions. This is what permits selection and certification from the same substitution estimates. The same event can sometimes certify minimality. Define $\widehat{q_{ij}} = \max(0, \widehat{q}_{ij} - h)$. If a selected partition $\Pi$ has upper error certificate at most $\varepsilon$ and

$$\widehat{q_{ij}} > 8\varepsilon \quad \text{for every pair } i, j \text{ from different blocks}, \tag{18}$$

then every feasible block partition refines $\Pi$, so $\Pi$ is the unique feasible block quotient with the smallest state count. This condition is sufficient rather than necessary; Appendix E gives a nonunique approximate example, while Appendix D records thresholded pairwise variants and other finite-query qualifications. The certificates therefore answer a positive question: a proposed joint erasure can be certified from shared local measurements. They do not imply that individually safe erasures compose. We now show that this distinction is fundamental when the response depends on a relation that must move coherently.

## 5 RELATIONAL BINDING AND COLLECTIVE DEPENDENCE

The preceding section certifies a complete proposed quotient and characterizes the error of one isolated merge. We now construct a regime in which every isolated merge is feasible at the target tolerance, yet erasing many identity distinctions jointly still requires an exponentially large block quotient. The obstruction is a correspondence among inputs rather than any individually decisive identity.

### 5.1 A DIFFUSE RELATION-COUNT FAMILY

Consider $r$ objects, where object $i$ has two bound slots $f_i$ and $r_i$. For a coalition $S$, define

$$z_u(S) = 2\mathbf{1}\{u \in S\} - 1 \in \{-1, +1\}. \tag{19}$$

The true correspondence is the matching $M_r = \{\{f_i, r_i\} : i \in [r]\}$. We measure agreement along these bound pairs through

$$Q_r(S) = \frac{1}{\sqrt{r}} \sum_{i=1}^{r} z_{f_i}(S) z_{r_i}(S). \tag{20}$$

Under uniform coalitions the $r$ summands are independent Rademacher variables, so $\|Q_r\|_2^2 = 1$ for every $r$. The total relational signal therefore remains normalized even though one substitution will affect only a vanishing fraction of it. To obtain a bounded response, let

$$\psi_B(t) = \max(-B, \min(B, t)), \qquad v_{r,B}(S) = \frac{1}{2} + \frac{\psi_B(Q_r(S))}{2B}. \tag{21}$$

We use $B = 4$ and write $v_r = v_{r,4}$. Exchanging the two slots of one true pair leaves their product unchanged. Exchanging slots from different objects changes only the terms involving their partners, so the corresponding singleton discrepancy decreases with $r$. The question is whether these weak local changes can be ignored jointly.

### 5.2 How block quotients retain relational signal

Let $\Pi = \{B_1, \dots, B_p\}$ be an arbitrary partition of the $2r$ slots and let $b_a = |B_a|$. Write $m_{aa}$ for the number of true matching edges contained in $B_a$ and $m_{ab}$ for the number joining $B_a$ and $B_b$.

**Proposition 1** (Relational signal retained by a block quotient)**.**

$$\boxed{\|P_\Pi Q_r\|_2^2 = \frac{1}{r}\left[\sum_{a:b_a\geq 2}\frac{m_{aa}^2}{\binom{b_a}{2}} + \sum_{a<b}\frac{m_{ab}^2}{b_a b_b}\right].} \tag{22}$$

*Consequently,*

$$\boxed{\|Q_r - P_\Pi Q_r\|_2^2 \geq \left(1-\frac{p}{r}\right)_+.} \tag{23}$$

The formula makes the loss mechanism explicit. Under uniform coalitions, the degree-two monomials $z_u z_v$ are orthonormal. Matching edges whose endpoints fall in the same pair of blocks are averaged into the same orbit term, producing the squared multiplicities $m_{ab}^2$; different block pairs remain orthogonal. A block pair can therefore preserve only the relational signal carried by the combinations of endpoints that it still distinguishes. Summing these contributions gives the exact expression above, and bounding each block contribution shows that a quotient with $p$ blocks preserves at most a $p/r$ fraction of the normalized matched signal. Appendix B gives the complete derivation. Clipping does not remove this obstruction. Since $Q_r$ is sub-Gaussian, $\|Q_r - \psi_B(Q_r)\|_2 \leq (2/B)e^{-B^2/4}$. At $B = 4$ this leaves a fixed-error separation.

**Theorem C (Diffuse relational separation).** *Let $v_r = v_{r,4}$ and $\varepsilon_0 = 1/256$. Then:*

1. *For each true pair, $q_{f_i,r_i} = 0$, and every pair of slots belonging to different objects satisfies*

$$q_{ij} \leq \frac{1}{8r}. \tag{24}$$

2. *For every block partition $\Pi$ with $p$ blocks,*

$$\boxed{E_{v_r}(\Pi) \geq \frac{1}{64}\left[\sqrt{\left(1-\frac{p}{r}\right)_+} - \frac{e^{-4}}{2}\right]_+^2.} \tag{25}$$

3. *If $E_{v_r}(\Pi) \leq \varepsilon_0$, then*

$$\frac{p}{r} \geq c_0 := 1 - \left(\frac{1}{2} + \frac{e^{-4}}{2}\right)^2, \qquad c_0 \approx 0.7408, \tag{26}$$

*and therefore*

$$\boxed{K(\Pi) \geq 2^{c_0 r}.} \tag{27}$$

4. *Under coordinated renaming*

$$G_{\mathrm{bind}} = \{f_i \mapsto f_{\sigma(i)},\ r_i \mapsto r_{\sigma(i)} : \sigma \in \mathrm{Sym}(r)\}, \tag{28}$$

*the response is invariant and*

$$\boxed{E_{G_{\mathrm{bind}}} = 0, \qquad K(G_{\mathrm{bind}}) = \binom{r+3}{3}.} \tag{29}$$

The same behavior therefore admits two qualitatively different abstractions. Independent block exchange needs exponentially many states at the fixed tolerance, while coordinated renaming removes the object labels and preserves only the histogram of the four bound-pair states, giving a polynomial exact quotient.

**Individually safe merges need not compose.** The local weakness is operational rather than merely a failure of a sufficient test. Let $\Pi^{ij}$ merge only one pair of slots. By Equation equation 14,

$$E(\Pi^{ij}) = q_{ij}/8 \le \frac{1}{64r}. \tag{30}$$

Hence for every $r \ge 4$, every isolated merge is feasible at $\varepsilon_0 = 1/256$. Yet Theorem C requires every jointly feasible block quotient to satisfy $K(\Pi) \ge 2^{c_0 r}$. The obstruction is therefore collective: many individually acceptable identity erasures cannot in general be composed without losing the correspondence among them. For $r \le 16$, clipping is inactive. Appendix B further shows that retaining only the aggregate fact and rule counts gives error $(1 - 1/r)/64$. The coordinated quotient succeeds not because it keeps only marginal counts, but because it preserves which fields belong together. Appendix C gives a complementary exact-state construction in which some cross-object substitutions are individually decisive.

### 5.3 Stability for learned responses

The construction above specifies an exact response, whereas a trained model will generally approximate it. We therefore ask which parts of the separation survive response error. Let $w$ be any learned response with

$$\eta = \|w - v_r\|_2. \tag{31}$$

Distance to a fixed quotient subspace changes by at most $\eta$, so the block lower bound degrades continuously. In particular, $E_w(\Pi) \le \varepsilon$ implies

$$\frac{p}{r} \ge 1 - \left[8(\sqrt{\varepsilon} + \eta) + \frac{e^{-4}}{2}\right]^2, \tag{32}$$

whenever the right-hand side is positive. In the other direction, exact invariance of $v_r$ under $G_{\mathrm{bind}}$ gives

$$E_{G_{\mathrm{bind}}}(w) \le \eta^2. \tag{33}$$

Thus one approximation quantity controls both sides of the comparison: a learned response close to $v_r$ remains difficult to compress with independent blocks and remains close to the coordinated quotient. Appendix B gives the full inequalities. We next test whether a trained Transformer enters this stable regime.

## 6 Learned relational quotients in a controlled Transformer

The experiment tests three consequences of the theory in one learned response: whether the learned response depends on object pairing, whether individually feasible merges compose into a compact block quotient, and whether preserving the pairing yields a smaller faithful abstraction. All comparisons use the same model response and the same uniform distribution over complete binary states.

### 6.1 The learned response depends on object pairing

Each of $r = 5$ objects has a binary fact field and a binary rule field. Every model input explicitly contains all ten fields, and the target is the number of objects whose bound fields disagree,

$$d(S) = \sum_{i=1}^{r} \mathbf{1}\{x_{f_i} \ne x_{r_i}\}. \tag{34}$$

We train a two-layer Transformer with learned object, field, value, and positional embeddings. Clause order is randomized during training. For each of five seeds, the 1024 complete states are split into 716 training, 153 validation, and 155 untouched test states; validation loss alone selects the epoch. Architecture and optimization details are in Appendix F. For quotient analysis, we define a scalar response from the model's predictive distribution rather than from its predicted class,

$$v_\theta(S) = \sum_{d=0}^{r} p_\theta(d \mid S) \left[\frac{1}{2} + \frac{r - 2d}{8\sqrt{r}}\right]. \tag{35}$$

Table 1: Quotient state budgets and learned response errors at $r = 5$. Errors are mean $\pm$ standard deviation across five trained models.

| Representation | States | Quotient MSE |
|---|---|---|
| Object-pair block counts | 243 | $(1.65 \pm 1.68) \times 10^{-8}$ |
| Binding-preserving orbit | 56 | $(3.13 \pm 3.24) \times 10^{-8}$ |
| Aggregate $(a, b)$ counts | 36 | $0.0124641 \pm 8.59 \times 10^{-6}$ |

At $r = 5$, clipping is inactive, so this is the model expectation of the analytic relation response. All five models reach 100% accuracy on the untouched test states, and

$$\|v_\theta - v_r\|_2 = (3.34 \pm 1.03) \times 10^{-4}. \tag{36}$$

For comparison, the Bayes-optimal classifier restricted to the two aggregate counts $(a, b)$ has only 66.8% accuracy under the uniform distribution over all 1024 states. Thus separate fact and rule counts are insufficient; solving the task requires information about which fields are paired.

### 6.2 Individually feasible merges do not yield a compact block quotient

At the theorem tolerance $\varepsilon_0 = 1/256$, the largest learned singleton discrepancy satisfies

$$q_{ij}/8 < 0.00312 < \varepsilon_0. \tag{37}$$

By Equation equation 14, every isolated merge is therefore feasible for every trained response. We then enumerate all $115{,}975$ set partitions of the ten slots and evaluate each quotient exactly on all 1024 states. By number of blocks, the best four-block quotient has error

$$0.0041526 \pm 4.11 \times 10^{-6} > \varepsilon_0, \tag{38}$$

whereas the best five-block error is $(1.65 \pm 1.68) \times 10^{-8}$. Separately minimizing the structural state count $K(\Pi)$ over all feasible partitions gives

$$\boxed{K^{\min}_{\text{block}} = 243} \tag{39}$$

for the analytic target and for every learned response. The conclusion therefore does not follow merely from needing five blocks: it uses the exhaustive search over the state budgets of all partitions. At the stricter tolerance $10^{-3}$, the true object pairing is also the unique feasible perfect pairing among all 945 pairings; Appendix F reports this control and the full frontier.

### 6.3 A compact quotient preserves the binding

The same learned responses admit a much smaller quotient when object names move coherently with their bound fields. Table 1 compares three representations. The 56-state value is the orbit count of the declared coordinated transformation group, not a claim of global minimality over arbitrary representations. Its near-zero error shows that the learned response can discard concrete object names while preserving the correspondence. The smaller aggregate-count representation fails because it discards that correspondence as well. A positional-order control for the seed-0 model and all seed-level results are reported in Appendix F.

## 7 Reuse across instances

The controlled experiment establishes a faithful relational quotient for one behavioral interface. Reuse asks for a stronger statement. Different instances may each admit a small quotient while assigning different responses to the same aligned abstract state. A reusable abstraction therefore requires both within-instance identity compression and agreement across instances.

### 7.1 The alignment interface

Let $A$ denote an instance, let $Y = v_A(S)$ be its response, and let $T = t_A(S)$ map concrete inputs into a common state space. The alignment must be fixed before the test, or learned on independent

discovery data; encoding the instance identity in $T$ would make agreement vacuous. Define

$$F_A(T) = \mathbb{E}[Y \mid A, T], \qquad F_*(T) = \mathbb{E}[Y \mid T]. \tag{40}$$

Here $F_A$ is the best response rule for one instance after compression, while $F_*$ is the best rule shared across the aligned population.

**Theorem D (Decomposition for aligned representations).** *For every shared predictor $F(T)$,*

$$\begin{aligned}\mathbb{E}[(Y - F(T))^2] = {} & \underbrace{\mathbb{E}[(Y - F_A(T))^2]}_{E_{\mathrm{id}}} \\ & + \underbrace{\mathbb{E}[(F_A(T) - F_*(T))^2]}_{E_{\mathrm{temp}}} \\ & + \underbrace{\mathbb{E}[(F_*(T) - F(T))^2]}_{E_{\mathrm{fit}}}.\end{aligned} \tag{41}$$

The decomposition follows by writing $Y - F = (Y - F_A) + (F_A - F_*) + (F_* - F)$. The three terms are orthogonal because the first has zero conditional mean given $(A, T)$ and the second has zero conditional mean given $T$. Thus small $E_{\mathrm{id}}$ establishes faithful compression within instances, while reusable computation additionally requires small $E_{\mathrm{temp}}$. Appendix E gives a counterexample with zero within-instance error and nonzero cross-instance disagreement.

### 7.2 Direct diagnostics and learning on a fixed interface

The two representation errors can be separated without first fitting a shared predictor. Draw a state $T = t$ from its population marginal, then draw $A, A'$ independently from $\mathcal{L}(A \mid T = t)$. Within $(A, t)$ draw independent responses $Y_1, Y_2$, and within $(A', t)$ draw $Z_1, Z_2$. Define

$$W_{\mathrm{temp}} = \tfrac{1}{2}(Y_1 - Z_1)(Y_2 - Z_2), \qquad W_{\mathrm{id}} = \tfrac{1}{4}[(Y_1 - Y_2)^2 + (Z_1 - Z_2)^2]. \tag{42}$$

Then

$$\boxed{\mathbb{E}W_{\mathrm{temp}} = E_{\mathrm{temp}}, \qquad \mathbb{E}W_{\mathrm{id}} = E_{\mathrm{id}}.} \tag{43}$$

Shared support is essential: a state observed in only one instance provides no testable evidence of agreement across instances. Appendix D gives the conditional-expectation proof and further sampling details. Once the common interface is fixed, its state count controls only the remaining finite-data estimation problem. For a $K$-state interface, the lookup regressor described in Appendix D, trained from $m$ independent teacher queries, has expected error at most $E_{\mathrm{id}} + E_{\mathrm{temp}} + \min\{1/4, K/[2(m + 1)]\}$. The three stages are therefore distinct: remove dispensable identity within each instance, test agreement on the aligned states, and only then estimate the shared response from finite data.

## 8 Discussion and conclusion

The central implication is that local substitutability is not the same as joint identity erasure. Local measurements remain useful: aggregate substitution discrepancies can certify the error of a complete proposed quotient, and an isolated merge has exactly the error predicted by its pairwise discrepancy. But individually acceptable merges need not compose. Relations can be distributed across many weak local effects, so deciding what can be forgotten requires evaluating the joint abstraction rather than only the strongest individual substitution. This changes how intervention results should be interpreted. A large response change after one independent substitution can reflect a broken binding rather than an indispensable name. Conversely, small changes under every isolated substitution do not establish that all corresponding identities are jointly irrelevant. The relevant object is a declared quotient: which concrete states it identifies, which relations that identification preserves, and how much response information is lost. In our construction, independent block exchange loses a diffuse correspondence and requires exponentially many states at fixed error, while coordinated renaming removes object names but preserves the binding. The controlled Transformer shows that the same distinction can arise after learning rather than only in a prescribed analytic response. A faithful quotient for one instance is still not a reusable computation. Reuse additionally requires aligned states to support the same response, measured by $E_{\mathrm{temp}}$. Our experiments establish the within-interface phenomenon in a controlled learned system; extending it to pretrained language models requires valid interventions and a justified transformation family under the target prompt distribution.

### AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript, to check the correctness and internal consistency of mathematical derivations and proofs, and to assist with experimental plotting.

### REPRODUCIBILITY STATEMENT

All assumptions and main mathematical statements are specified in the paper, with complete derivations and counterexamples in the appendices. Appendix F reports the controlled Transformer architecture, optimization settings, train, validation, and test splits, seed level results, quotient enumeration protocol, order control, and numerical checks. Appendix G summarizes the proof and numerical verification status of the main claims.

## REFERENCES

Awni Altabaa, Taylor Webb, Jonathan D. Cohen, and John Lafferty. Abstractors and relational cross-attention: An inductive bias for explicit relational reasoning in transformers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XNa6r6ZjoB.

Eric Blais, Amit Weinstein, and Yuichi Yoshida. Partially symmetric functions are efficiently isomorphism testable. *SIAM Journal on Computing*, 44(2):411–432, 2015. doi: 10.1137/140971877. URL https://doi.org/10.1137/140971877.

Huiqi Deng, Qihan Ren, Hao Zhang, and Quanshi Zhang. Discovering and explaining the representation bottleneck of DNNs. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2111.06236.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zb3b6oKO77.

Yuval Filmus. An orthogonal basis for functions over a slice of the boolean hypercube. *The Electronic Journal of Combinatorics*, 23(1):P1.23, 2016. doi: 10.37236/4567. URL https://arxiv.org/abs/1406.0142. Revised arXiv version 1406.0142v2, 2022.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://jmlr.org/papers/v26/23-0058.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://www.jstor.org/stable/2282952.

Mingjie Li and Quanshi Zhang. Does a neural network really encode symbolic concepts? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20452–20469. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23at.html.

Puyin Li, Jiyuan Tan, Ahmad Jabbar, Thomas Icard, and Atticus Geiger. Bucketing the good apples: A method for diagnosing and improving causal abstraction. *arXiv preprint arXiv:2605.02234*, 2026. URL https://arxiv.org/abs/2605.02234.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2961–2984. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.163. URL https://aclanthology.org/2024.acl-long.163/.

Xiaolei Lu, Jianghong Ma, and Haode Zhang. Asymmetric feature interaction for interpreting model predictions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4666–4678, 2023. doi: 10.18653/v1/2023.findings-acl.286. URL https://aclanthology.org/2023.findings-acl.286/.

Andrea E. Martin and Leonidas A. A. Doumas. Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1791):20190306, 2020. doi: 10.1098/rstb.2019.0306. Published online in 2019.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Akiyoshi Sannai, Makoto Kawano, and Wataru Kumagai. Invariant and equivariant reynolds networks. *Journal of Machine Learning Research*, 25(42):1–36, 2024. URL https://www.jmlr.org/papers/v25/22-0891.html.

# A Exchange spectrum and sharp constants

This appendix establishes the spectral facts used in Theorem A. The argument derives the exchange spectrum on one block, extends it to products of block slices, identifies the unit lower spectral gap and the upper endpoint $\kappa_\Pi$, and then shows that both constants and the forbidden merge factor are sharp.

## A.1 Up and down operators on one block

Let a block have size $b$, and let $V_k$ be the real function space on all $k$-subsets. We first use the unnormalized counting inner product; on one slice it differs from the uniform inner product only by a constant and therefore does not change eigenvalues. Define

$$\begin{aligned}(U_k f)(S) &= \sum_{\substack{T\subset S\\ |T|=k}} f(T), \quad |S| = k+1,\\ (D_k f)(T) &= \sum_{\substack{S\supset T\\ |S|=k}} f(S), \quad |T| = k-1.\end{aligned} \tag{44}$$

$U_{k-1}$ and $D_k$ are adjoint. Counting "add then delete" and "delete then add" paths gives

$$D_{k+1}U_k - U_{k-1}D_k = (b-2k)I. \tag{45}$$

The exchange-neighbor terms cancel; the paths returning to the original set occur $b-k$ and $k$ times, respectively.

Let $A_k$ be the Johnson adjacency operator. Every state has $k(b-k)$ neighbors. The operator $U_{k-1}D_k$ includes every exchange neighbor once and the original state $k$ times, so the unnormalized Laplacian is

$$\mathscr{L}_k = k(b-k)I - A_k = k(b-k+1)I - U_{k-1}D_k. \tag{46}$$

## A.2 Inductive spectrum

For $1 \le k \le b/2$, Equation equation 45 gives

$$D_k U_{k-1} = U_{k-2}D_{k-1} + (b-2k+2)I. \tag{47}$$

The first term on the right is positive semidefinite and the second is positive definite, so $U_{k-1}$ is injective. Hence

$$V_k = \operatorname{im} U_{k-1} \oplus \ker D_k. \tag{48}$$

Equations equation 46–equation 47 yield

$$\mathscr{L}_k U_{k-1} = U_{k-1}\mathscr{L}_{k-1}. \tag{49}$$

Thus old eigenvalues persist through the injective map. On $\ker D_k$, Equation equation 46 gives the new eigenvalue $k(b-k+1)$. Starting from the zero eigenvalue in $V_0$,

$$\mathrm{Spec}(\mathscr{L}_k) = \{d(b-d+1) : 0 \le d \le k\}, \qquad k \le b/2. \tag{50}$$

The new eigenspace has dimension $\binom{b}{k} - \binom{b}{k-1} > 0$, so each listed eigenvalue occurs. For $k > b/2$, the complement map identifies the graph with the slice of size $b-k$. Therefore the general range is $0 \le d \le \min(k, b-k)$.

### A.3 Multiple blocks and the full cube

Conditioned on all block counts, an orbit is a Cartesian product of slices. The block Laplacians act on different tensor factors and commute. Their joint eigenvalues are therefore

$$\lambda_{\boldsymbol{d}} = \sum_{B\in\Pi} \frac{d_B(|B| - d_B + 1)}{|B|}, \quad 0 \le d_B \le \min(k_B, |B| - k_B). \tag{51}$$

The all-zero index gives constants on the orbit. If any $d_B \ge 1$, its term is at least 1 because $d(b-d+1) \ge b$. Since the full cube contains every block-count pattern, the smallest positive eigenvalue is exactly 1 whenever a nontrivial block exists. Each term is maximized at $d_B = \lfloor |B|/2 \rfloor$, and suitable middle slices realize these maxima simultaneously, giving $\kappa_\Pi$.

The kernel consists exactly of functions constant on every orbit, with dimension $\prod_B(|B|+1)$. Splitting $v$ into the kernel and positive spectral subspace proves Theorem A without restricting the polynomial order, network architecture, or training process of $v$.

### A.4 Sharpness of the endpoints

The smallest-positive-eigenvalue eigenspace is nonempty. For $h \neq 0$ in this space, $P_\Pi h = 0$ and $J(h) = \|h\|_2^2$. The largest-eigenvalue eigenspace is also nonempty; for $h$ there, $J(h) = \kappa_\Pi \|h\|_2^2$. Because the space is finite, $\|h\|_\infty < \infty$; therefore $v = 1/2 + h/(2\|h\|_\infty)$ lies in $[0,1]$ and preserves the ratio, proving that neither endpoint constant can be uniformly improved.

For the merge lower bound, take $n = 2$ and $v = 1/2 + (x_1 - x_2)/2$. Then $q_{12} = 1$ and $E = 1/8$, attaining the factor in Equation equation 13.

## B Proof of the diffuse relational separation

This appendix proves Proposition 1, Theorem C, the aggregate count comparison in Equation equation 72, and the learned response stability bounds. The argument first computes the exact projection of the matched quadratic score through an arbitrary block quotient, then controls clipping and singleton substitutions.

### B.1 Exact projection of the matched quadratic score

Under the uniform coalition distribution, the variables $z_u$ are independent Rademachers. The degree two monomials $z_u z_v$ for unordered pairs $\{u, v\}$ are therefore orthonormal in $L^2(u)$. Write

$$Q_r = \frac{1}{\sqrt{r}} \sum_{e\in M_r} \chi_e, \qquad \chi_{\{u,v\}} = z_u z_v. \tag{52}$$

Consider a block $B_a$ of size $b_a$. If an edge $e = \{u, v\}$ is contained in $B_a$, averaging over permutations of $B_a$ gives

$$P_\Pi \chi_e = \frac{1}{\binom{b_a}{2}} \sum_{\{s,t\}\subset B_a} \chi_{\{s,t\}}, \qquad \|P_\Pi \chi_e\|_2^2 = \frac{1}{\binom{b_a}{2}}. \tag{53}$$

If $e$ joins distinct blocks $B_a$ and $B_b$, then

$$P_\Pi \chi_e = \frac{1}{b_a b_b} \sum_{s \in B_a, t \in B_b} \chi_{\{s,t\}}, \qquad \|P_\Pi \chi_e\|_2^2 = \frac{1}{b_a b_b}. \tag{54}$$

All matching edges with the same unordered pair of endpoint blocks have the same projection. Projections associated with different block pairs have disjoint support in the orthonormal degree two basis and are therefore orthogonal. If $m_{aa}$ and $m_{ab}$ denote the corresponding matching multiplicities, summing the squared coefficients gives Equation equation 22. Since $\|Q_r\|_2^2 = 1$ and $P_\Pi$ is an orthogonal projection,

$$\|Q_r - P_\Pi Q_r\|_2^2 = 1 - \|P_\Pi Q_r\|_2^2. \tag{55}$$

We next derive the simpler bound depending only on the number of blocks. For an internal matching count,

$$\frac{m_{aa}^2}{\binom{b_a}{2}} \le \frac{2m_{aa}}{b_a}, \tag{56}$$

because $m_{aa} \le b_a/2$. For $a < b$, $m_{ab} \le \min(b_a, b_b)$ gives

$$\frac{m_{ab}^2}{b_a b_b} \le \frac{1}{2}\left(\frac{m_{ab}}{b_a} + \frac{m_{ab}}{b_b}\right). \tag{57}$$

Every vertex belongs to exactly one matching edge, so

$$2m_{aa} + \sum_{b \ne a} m_{ab} = b_a. \tag{58}$$

Collecting the terms assigned to block $a$ in Equations equation 56–equation 57 gives at most

$$\frac{2m_{aa}}{b_a} + \frac{1}{2b_a}\sum_{b \ne a} m_{ab} = \frac{1}{2} + \frac{m_{aa}}{b_a} \le 1. \tag{59}$$

Summing over the $p$ blocks yields $\|P_\Pi Q_r\|_2^2 \le p/r$, which proves Equation equation 23.

### B.2 Clipping and fixed error lower bound

The summands $z_{f_i} z_{r_i}$ are independent Rademachers, so Hoeffding's inequality gives

$$\mathbb{P}(|Q_r| > t) \le 2e^{-t^2/2}. \tag{60}$$

Let $R_B = (|Q_r| - B)_+$. Then

$$\tau_B^2 = \mathbb{E}R_B^2 = \int_0^\infty 2s\mathbb{P}(|Q_r| > B + s)\, ds \tag{61}$$

$$\le 4e^{-B^2/2} \int_0^\infty s e^{-Bs}\, ds = \frac{4}{B^2} e^{-B^2/2}, \tag{62}$$

where $(B+s)^2 \ge B^2 + 2Bs$ was used in the inequality. Taking square roots gives

$$\tau_B \le \frac{2}{B} e^{-B^2/4}. \tag{63}$$

Let $\mathcal{H}_\Pi$ be the subspace of functions measurable with respect to the block quotient. Distance to a closed subspace is 1-Lipschitz, hence

$$\operatorname{dist}(\psi_B(Q_r), \mathcal{H}_\Pi) \ge \operatorname{dist}(Q_r, \mathcal{H}_\Pi) - \tau_B. \tag{64}$$

The constant $1/2$ belongs to $\mathcal{H}_\Pi$, so

$$\sqrt{E_{v_{r,B}}(\Pi)} \ge \frac{1}{2B}\left[\sqrt{\left(1 - \frac{p}{r}\right)_+} - \tau_B\right]_+. \tag{65}$$

For $B = 4$, $\tau_4 \le e^{-4}/2$. Squaring Equation equation 65 gives Equation equation 25. Suppose $E_{v_r}(\Pi) \le 1/256$. If $p \ge r$, then $p/r \ge 1 > c_0$ and the desired block-count bound is immediate. If $p < r$, then $\sqrt{E_{v_r}(\Pi)} \le 1/16$ and Equation equation 65 gives

$$\sqrt{1 - p/r} \le \frac{1}{2} + \frac{e^{-4}}{2}. \tag{66}$$

Rearranging yields Equation equation 26. Since each nonempty block $B$ contributes a factor $|B|+1 \ge 2$ to $K(\Pi)$, $K(\Pi) \ge 2^p \ge 2^{c_0 r}$.

### B.3 SINGLETON DISCREPANCIES AND COORDINATED RENAMING

For a true pair $f_i, r_i$, the comparison context contains neither slot. Adding either one makes the product $z_{f_i} z_{r_i} = -1$, so $q_{f_i,r_i} = 0$. For slots $u, v$ from different true pairs, let $u^*, v^*$ be their partners. Between $C \cup \{u\}$ and $C \cup \{v\}$,

$$\Delta Q_r = \frac{2}{\sqrt{r}}(z_{u^*} - z_{v^*}). \tag{67}$$

The two partner signs are independent Rademachers under the uniform context distribution, so $\mathbb{E}(\Delta Q_r)^2 = 8/r$. Since $t \mapsto 1/2 + \psi_B(t)/(2B)$ is $1/(2B)$-Lipschitz,

$$q_{uv} \le \frac{1}{4B^2}\frac{8}{r} = \frac{2}{B^2 r}. \tag{68}$$

At $B = 4$ this is $1/(8r)$, proving Equation equation 24. When $r \le 16$, $|Q_r| \le \sqrt{r} \le 4$, so clipping is inactive and the cross object value is exactly $1/(8r)$.

Every simultaneous permutation of the object index preserves the multiset of four local states $(x_{f_i}, x_{r_i}) \in \{0,1\}^2$ and therefore preserves $Q_r$. Two coalitions lie in the same orbit exactly when their four state counts $(n_{00}, n_{01}, n_{10}, n_{11})$ agree. The number of such histograms is $\binom{r+3}{3}$, proving Equation equation 29.

### B.4 AGGREGATE COUNT ERROR

For $r \le 16$, $v_r = 1/2 + Q_r/8$. Let $A = \sum_i z_{f_i} = 2a - r$ and $B' = \sum_i z_{r_i} = 2b - r$. Conditional on $a, b$, exchangeability of the two sign vectors gives

$$\mathbb{E}[Q_r \mid a, b] = \frac{AB'}{r\sqrt{r}}. \tag{69}$$

The two sign vectors are independent under the uniform distribution, so $\mathbb{E}A^2 = \mathbb{E}B'^2 = r$. Therefore

$$\mathbb{E}\big[\mathbb{E}(Q_r \mid a, b)^2\big] = \frac{\mathbb{E}A^2 \mathbb{E}B'^2}{r^3} = \frac{1}{r}. \tag{70}$$

Since $\mathbb{E}Q_r^2 = 1$, the conditional variance identity gives

$$\mathbb{E}\mathrm{Var}(Q_r \mid a, b) = 1 - \frac{1}{r}. \tag{71}$$

Since $v_r = 1/2 + Q_r/8$ in this range,

$$E_{(a,b)}(v_r) = \frac{1 - 1/r}{64}. \tag{72}$$

### B.5 STABILITY FOR LEARNED RESPONSES

For any closed subspace $\mathcal{H}$, the distance map $f \mapsto \mathrm{dist}(f, \mathcal{H})$ is 1-Lipschitz. Applying this with the block quotient subspace gives

$$|\sqrt{E_w(\Pi)} - \sqrt{E_{v_r}(\Pi)}| \le \eta. \tag{73}$$

Combining this with Equation equation 65 at $B = 4$ gives

$$\sqrt{E_w(\Pi)} \ge \left[\frac{\sqrt{(1 - p/r)_+} - e^{-4}/2}{8} - \eta\right]_+. \tag{74}$$

If $v_r$ is exactly invariant under $G_{\mathrm{bind}}$, then it belongs to the invariant subspace, and therefore

$$E_{G_{\mathrm{bind}}}(w) \le \eta^2. \tag{75}$$

Finally, suppose $E_w(\Pi) \le \varepsilon$. If $p \ge r$, Equation equation 32 is immediate whenever its right-hand side is at most one. If $p < r$, Equation equation 74 implies

$$\sqrt{1 - p/r} \le 8(\sqrt{\varepsilon} + \eta) + \frac{e^{-4}}{2}. \tag{76}$$

Squaring and rearranging gives Equation equation 32; if its right-hand side is nonpositive, the statement is vacuous.

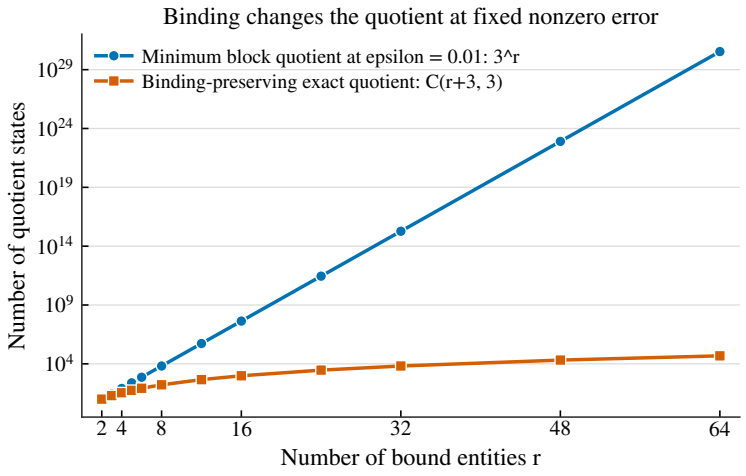



Figure 2: State budgets for the complementary exact-state construction at $\varepsilon = 0.01$. The minimum feasible block quotient has $3^r$ states, while coordinated binding preserving renaming has $\binom{r+3}{3}$ exact orbit states.

## C A COMPLEMENTARY EXACT-STATE BINDING SEPARATION

This appendix records a complementary construction with decisive local discrepancies and an exact minimum block state count. Let there be $r$ objects with fact slots $f_i$ and rule slots $r_i$, write $x_u = \mathbf{1}\{u \in S\}$, and define

$$a(S) = \sum_i x_{f_i}, \qquad b(S) = \sum_i x_{r_i}, \qquad m(S) = \sum_i x_{f_i} x_{r_i}. \tag{77}$$

Let $\Pi_{\text{obj}} = \{\{f_i, r_i\} : i \in [r]\}$. The construction below has minimum feasible block state count $3^r$ for every $1/(800r) \le \varepsilon < 49/1600$, while coordinated renaming is exact with $\binom{r+3}{3}$ orbit states. Unlike Theorem C, its cross-object discrepancies remain bounded away from zero.

### C.1 WHY AN OR BASED WITNESS DOES NOT SEPARATE AT FIXED ERROR

A natural but insufficient construction is

$$v_r^{\text{OR}}(S) = \frac{a(S) + 2b(S)}{10r} + \frac{7}{10}\mathbf{1}\{m(S) > 0\}. \tag{78}$$

If only $(a, b)$ is retained, the linear term is exact. Let $H = \mathbf{1}\{m > 0\}$. Then

$$E_{\text{count}}(v_r^{\text{OR}}) = \frac{49}{100}\mathbb{E}\text{Var}(H \mid a, b). \tag{79}$$

If $p = \mathbb{P}(H = 1 \mid a, b)$, then $p(1-p) \le 1 - p$. Under uniform coalitions, each object forms a complete match with probability $1/4$ independently, so

$$\mathbb{E}[1 - p] = \mathbb{P}(m = 0) = (3/4)^r. \tag{80}$$

Therefore

$$E_{\text{count}}(v_r^{\text{OR}}) \le \frac{49}{100}(3/4)^r \to 0. \tag{81}$$

The OR witness still separates the exact quotient families, but it cannot support an asymptotic lower bound at a fixed positive error. A fixed-error separation therefore requires a relational witness whose contribution does not vanish with $r$; the parity term in Equation equation 82 provides that property.

### C.2 LOCAL DISCREPANCIES IN THE EXACT-STATE CONSTRUCTION

Consider

$$\widetilde{v}_r(S) = \frac{a(S) + 2b(S)}{10r} + \frac{7}{10}\mathbf{1}\{m(S) \text{ is odd}\}. \tag{82}$$

When comparing $f_i$ and $r_i$ from the same object, both are absent from the context. Adding either one alone cannot create a complete match, so the parity term is the same; the linear terms differ by $1/(10r)$. Hence

$$q_{f_i,r_i} = \frac{1}{100r^2}. \tag{83}$$

For same-type slots from distinct objects, say $f_i, f_j$, the context contains $r_i, r_j$ as independent Bernoulli$(1/2)$ variables. The parity of the added match differs exactly when these partner states differ, an event of probability $1/2$. The gate difference has magnitude $7/10$, giving

$$q_{f_i,f_j} = \frac{1}{2}\left(\frac{7}{10}\right)^2 = \frac{49}{200}. \tag{84}$$

The same holds for two rule slots. For $f_i, r_j$ with $i \neq j$, the gate difference is again $\pm 7/10$ with probability $1/2$, while the linear difference is always $\pm 1/(10r)$. Swapping the two partner states changes the sign of the gate difference and preserves the distribution, so the cross term has mean zero. Thus

$$q_{f_i,r_j} = \frac{49}{200} + \frac{1}{100r^2}. \tag{85}$$

### C.3 Error of the partition into object pairs

The partition $\Pi_{\mathrm{obj}}$ records only the participation count

$$c_i = x_{f_i} + x_{r_i} \in \{0, 1, 2\} \tag{86}$$

for each object. Given all $c_i$, the complete-match count is exactly $m = \#\{i : c_i = 2\}$, so the parity gate is preserved. The only remaining randomness is whether a singly occupied object contains its fact or its rule slot.

Let

$$L = \#\{i : c_i = 1\}. \tag{87}$$

Within the conditional orbit, each such object chooses fact or rule with probability $1/2$, and its linear contribution deviates from the orbit mean by $\pm 1/(20r)$. These deviations are conditionally independent, so

$$\mathrm{Var}(\widetilde{v}_r \mid t_{\Pi_{\mathrm{obj}}}) = \frac{L}{400r^2}. \tag{88}$$

Under a uniform coalition, each object has exactly one occupied slot with probability $1/2$, hence $\mathbb{E}L = r/2$. Therefore

$$E(\Pi_{\mathrm{obj}}) = \frac{1}{800r}. \tag{89}$$

### C.4 Lower bound for block quotients at fixed error

If a block contains two players from different objects,

$$q_{uv} \geq \frac{49}{200}. \tag{90}$$

The merge lower bound in Theorem A implies

$$E(\Pi) \geq \frac{q_{uv}}{8} \geq \frac{49}{1600}. \tag{91}$$

Thus, whenever $\varepsilon < 49/1600$, every feasible block partition must refine $\Pi_{\mathrm{obj}}$. If also $\varepsilon \geq 1/(800r)$, the partition into object pairs itself is feasible.

Refining one object pair changes its quotient-state factor from 3 to $2 \times 2 = 4$. Any strict refinement therefore strictly increases the total state count. Hence

$$\min_{\Pi : E(\Pi) \leq \varepsilon} K(\Pi) = 3^r \tag{92}$$

and the minimizing block partition is unique.

### C.5 Coordinated renaming

For every $\sigma \in \mathrm{Sym}(r)$, the simultaneous action

$$f_i \mapsto f_{\sigma(i)}, \qquad r_i \mapsto r_{\sigma(i)} \tag{93}$$

preserves $a, b, m$, hence preserves $\widetilde{v}_r$. Each object has one of four local states $(0,0), (1,0), (0,1), (1,1)$. Two coalitions are in the same coordinated-renaming orbit exactly when the four category counts agree. The orbit is therefore specified by

$$(n_{00}, n_{10}, n_{01}, n_{11}), \qquad \sum n_{uv} = r, \tag{94}$$

and the number of orbits is

$$\binom{r+3}{3}. \tag{95}$$

Behavior is constant on every such orbit, so the coordinated quotient has zero error.

### C.6 Exact error when only $(a, b)$ is retained

Let $F = \{i : f_i \in S\}$ and $R = \{i : r_i \in S\}$, and define

$$H = (-1)^{|F \cap R|}. \tag{96}$$

The odd-parity indicator is $(1 - H)/2$, while the linear term is exactly retained by $(a, b)$. Hence

$$E_{\mathrm{count}}(\widetilde{v}_r) = \frac{49}{400}\left(1 - \mathbb{E}[\mathbb{E}(H \mid a, b)^2]\right). \tag{97}$$

For fixed $b$, define

$$h_b(F) = \binom{r}{b}^{-1} \sum_{|R|=b} (-1)^{|F \cap R|}. \tag{98}$$

Expanding the square and averaging over uniform $F$, any $R \neq R'$ leaves a nonempty symmetric difference $R \triangle R'$ and hence a parity character with zero mean. Only $R = R'$ survives, so

$$\mathbb{E}_F h_b(F)^2 = \frac{1}{\binom{r}{b}}. \tag{99}$$

Since $b \sim \mathrm{Bin}(r, 1/2)$,

$$\mathbb{E}[\mathbb{E}(H \mid a, b)^2] = \sum_{b=0}^{r} \frac{\binom{r}{b}}{2^r} \frac{1}{\binom{r}{b}} = \frac{r+1}{2^r}. \tag{100}$$

Substituting into equation 97 gives

$$E_{\mathrm{count}}(\widetilde{v}_r) = \frac{49}{400}\left(1 - \frac{r+1}{2^r}\right). \tag{101}$$

## D Additional certification, stability, and distributional extensions

This appendix records extensions and technical consequences that support the main text: general transformation groups, cross-instance diagnostics, finite-query qualifications, independent validation, perturbation stability, nonuniform query distributions, stochastic response noise, and finite-state learning details.

### D.1 General transformation groups

For a finite group $G$ with inverse closed generating set $\mathcal{A}$ and symmetric positive weights, define $L_G = \sum_{h \in \mathcal{A}} a_h (I - U_h)$ and $J_G = \langle v, L_G v \rangle$. If $\lambda_G$ and $\Lambda_G$ are the smallest and largest eigenvalues on the noninvariant subspace, then

$$J_G/\Lambda_G \leq E_G \leq J_G/\lambda_G. \tag{102}$$

Thus the same local-to-global principle applies once a transformation group and a controlled generating set are declared. This statement certifies a supplied family; it does not discover an arbitrary relational group.

### D.2 Cross-instance decomposition and four-query diagnostics

For Theorem D, write

$$Y - F(T) = [Y - F_A(T)] + [F_A(T) - F_*(T)] + [F_*(T) - F(T)]. \tag{103}$$

The first term has zero conditional mean given $(A, T)$, while the second has zero conditional mean given $T$. The three terms are therefore pairwise orthogonal in $L^2$, proving Equation equation 41.

For the diagnostic in Equation equation 42, condition on $T = t$. Independence within one instance gives

$$\mathbb{E}[(Y_1 - Y_2)^2 \mid A, t] = 2\mathrm{Var}(Y \mid A, t), \tag{104}$$

so averaging the two within-instance terms yields $E_{\mathrm{id}}$. For the cross-instance product,

$$\mathbb{E}[(Y_1 - Z_1)(Y_2 - Z_2) \mid A, A', t] = [F_A(t) - F_{A'}(t)]^2. \tag{105}$$

Averaging independent $A, A' \sim \mathcal{L}(A \mid T = t)$ and multiplying by $1/2$ gives the conditional variance of $F_A(t)$, hence $E_{\mathrm{temp}}$. Shared support is essential: a state observed in only one instance supplies no testable cross-instance comparison.

### D.3 Thresholded pairwise rates and fixed-sample qualification

For $p_{ij}(\tau) = \mathbb{P}_C(|D(i, j \mid C)| > \tau)$ and $v \in [0, 1]$,

$$\tau^2 p_{ij}(\tau) \le q_{ij} \le \tau^2 + (1 - \tau^2) p_{ij}(\tau). \tag{106}$$

The finite-query certificate in Theorem B assumes a fixed sample size $M$. Adaptive stopping requires a concentration guarantee that remains valid over time rather than reinterpreting the fixed-sample interval after observing the data.

### D.4 Independent orbit validation

The local certificate $J$ can be conservative. For a finite candidate family $\mathcal{H}$, independent validation data can directly estimate each orbit error. For a candidate group $G$, draw $S \sim u$ and $g \sim \mathrm{Unif}(G)$ and set

$$Z_G = \frac{1}{2}[v(S) - v(gS)]^2 \in [0, 1/2]. \tag{107}$$

Its expectation is $E_G$. With $M$ independent samples per candidate, define

$$a_M = \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{8M}}. \tag{108}$$

Hoeffding plus a union bound gives, with probability at least $1-\delta$, simultaneous bounds $|\widehat{E}_G - E_G| \le a_M$ for all candidates. One may therefore generate a finite candidate set adaptively on discovery data and, on independent validation data, choose the smallest-budget candidate satisfying $\widehat{E}_G + a_M \le \varepsilon$.

### D.5 Role separation and perturbation stability

Suppose a target partition $\Pi_*$ satisfies $q_{ij} \le a$ within blocks and $q_{ij} \ge b$ across blocks, with $b > a$. If the simultaneous estimation error obeys $h < (b - a)/2$, then the threshold graph $\widehat{q}_{ij} \le (a + b)/2$ exactly recovers the blocks of $\Pi_*$. Without a separation margin, connected components of a threshold graph cannot be interpreted unconditionally as role classes.

If $\|v - v_0\|_2 \le \eta$, unitarity of the swap operator gives

$$|\sqrt{q_{ij}(v)} - \sqrt{q_{ij}(v_0)}| \le 2\sqrt{2}\,\eta, \tag{109}$$

and

$$|\sqrt{E_v(\Pi)} - \sqrt{E_{v_0}(\Pi)}| \le \eta. \tag{110}$$

Thus exact role structure is stable under small $L^2$ perturbations when the relevant margins are large enough.

### D.6 NONUNIFORM QUERY DISTRIBUTIONS

If every player is retained independently with a common probability $p \in (0,1)$ and the contexts are sampled with the same $p$, every block-count orbit remains internally uniform. The corresponding certificate becomes

$$J_p(\Pi) = \sum_{B \in \Pi} \frac{p(1-p)}{|B|} \sum_{i<j \in B} q_{ij,p}, \tag{111}$$

with the same spectral sandwich constants.

For a general empirical distribution $\nu$ that is not group invariant, the best quotient approximation is a weighted conditional expectation rather than a simple group average. If $d\nu/du \le R$, then for a fixed approximation $\widehat{v}$,

$$\mathbb{E}_\nu(v-\widehat{v})^2 \le R\,\mathbb{E}_u(v-\widehat{v})^2. \tag{112}$$

Without coverage or density-ratio control, small error under uniform masking does not imply small error on a natural-text distribution.

### D.7 GENERATION NOISE

The finite sample bounds in the main text assume that $v(S)$ is evaluated deterministically. If generation is itself stochastic, or if $v(S)$ is estimated from finitely many completions, naively squaring noisy differences adds an additional variance contribution and may bias $q_{ij}$ upward. An empirical study of a language model must distinguish uncertainty from coalition sampling and uncertainty from response estimation, using independent repetitions for debiasing or explicitly incorporating the second layer of uncertainty into the confidence budget.

### D.8 FINITE-STATE LEARNING DETAILS AND THRESHOLDED BEHAVIOR

For the lookup regressor described in the main text, if $N_t$ is the number of training samples observed in abstract state $t$ with probability $p_t$, the proof uses

$$\mathbb{E}\frac{1}{N_t+1} = \frac{1-(1-p_t)^{m+1}}{(m+1)p_t}. \tag{113}$$

After summing over the $K$ states, this identity gives the estimation term $K/[2(m+1)]$ used in the main text.

A standard margin argument also translates mean squared prediction error into agreement with a thresholded teacher. If a teacher thresholds $Y$ at $b$ and the student uses $\widehat{Y}$, then for every $\gamma > 0$,

$$\mathbb{P}(\mathbf{1}\{Y \ge b\} \ne \mathbf{1}\{\widehat{Y} \ge b\}) \le \mathbb{P}(|Y-b| \le \gamma) + \frac{\mathbb{E}(Y-\widehat{Y})^2}{\gamma^2}. \tag{114}$$

This concerns preservation of teacher behavior away from the decision margin, not correctness with respect to an external ground truth.

## E COUNTEREXAMPLES AND IDENTIFIABILITY BOUNDARIES

This appendix also records two interpretation identities used to separate quotient structure from thresholded sensitivity and interaction order.

### E.1 THRESHOLDED DISAGREEMENT AND INTERACTION ORDER

For $\delta_\Pi(\tau) = \mathbb{P}(|v(S)-v(T)| > \tau)$ with $S, T$ drawn independently from one quotient orbit and $v \in [0,1]$,

$$\frac{\tau^2}{2}\delta_\Pi(\tau) \le E(\Pi) \le \frac{\tau^2 + (1-\tau^2)\delta_\Pi(\tau)}{2}, \qquad E(\Pi) = \int_0^1 \tau\,\delta_\Pi(\tau)\,d\tau. \tag{115}$$

Thus a thresholded asymmetry rate is an observable of quotient error, not a reasoning complexity measure. Likewise, if $I_v$ denotes the Harsanyi/Mobius coefficients, then for every permutation group $G$,

$$v(gS) = v(S)\ \forall g, S \quad \Longleftrightarrow \quad I_v(gT) = I_v(T)\ \forall g, T. \tag{116}$$

For a singleton substitution, $D(i, j \mid C) = \sum_{U \subseteq C}[I_v(U \cup \{i\}) - I_v(U \cup \{j\})]$, so the local query aggregates differences across all interaction orders. Exact quotient symmetry therefore does not imply low interaction order.

Each counterexample below rules out an inference that the behavioral quotient framework does not by itself license. Together they separate symmetry from interaction order, behavioral equivalence from semantic equivalence, exact distinguishability from random-query efficiency, approximate optimality from uniqueness, per-instance compression from cross-instance reuse, and independent from coordinated substitutions.

### E.2 Full symmetry can coexist with a highest-order interaction

The function

$$v(S) = \mathbf{1}\{S = N\} \tag{117}$$

is fully symmetric in all players, but its only nonzero nonempty Möbius joint term is the top-order set $N$. Full symmetry therefore does not imply the absence of high-order conjunctions or gates.

### E.3 Complete asymmetry can be purely additive

For

$$v(S) = \sum_{i \in S} \frac{2^{i-1}}{2^n - 1}, \tag{118}$$

different subsets have different values. For every positive threshold below $1/(2^n - 1)$, distinct equal-cardinality coalitions are distinguishable, yet all second- and higher-order Möbius interactions vanish. Identity asymmetry is therefore not a reasoning-complexity measure.

### E.4 A scalar task can hide semantic roles

Suppose a task reports only whether a fact and a rule are jointly present, with $v(\{f, r\}) = 1$ and all other inputs mapped to zero. Then $f$ and $r$ are behaviorally interchangeable even if an external semantic interpretation assigns them different meanings.

**Proposition 2** (Observational non-identifiability)**.** *If two semantic worlds return identical outputs for every allowed query, then any algorithm that accesses only those queries has the same output distribution in both worlds. Under equal priors, it cannot distinguish the worlds with probability greater than* $1/2$*.*

Couple an adaptive query algorithm in both worlds with the same random seed. The first answer agrees by assumption; induction makes every later query, answer, and final output identical. Recovering finer semantic roles therefore requires a richer task or output family, not a new statistic applied to the same responses.

### E.5 Exact distinguishability need not be easy to observe randomly

Fix a pair $i, j$ and let

$$v(S) = \mathbf{1}\{S = \{i\}\}. \tag{119}$$

The only distinguishing context is the empty context, so

$$q_{ij} = 2^{-(n-2)}. \tag{120}$$

With $M$ uniform contexts, the probability of missing that witness completely is

$$(1 - 2^{-(n-2)})^M. \tag{121}$$

Thus the existence of a distinguishing context does not imply efficient recovery under random querying; recovery from finitely many random queries needs probability mass or $L^2$ separation. This is a statement about the declared random-query protocol, not an identical lower bound for every adaptive value-query algorithm.

### E.6 Approximate minimum block quotients need not be unique

Let $n = 3$ and

$$v(S) = \tfrac{1}{2} + \tfrac{1}{2}(x_3 - x_1). \tag{122}$$

At $\varepsilon = 1/16$, both $\{\{1,2\},\{3\}\}$ and $\{\{1\},\{2,3\}\}$ are feasible and have the same minimum state budget, while the fully merged partition is infeasible. Without a separation condition such as Equation equation 18, one should report the optimal set or Pareto frontier rather than assert a canonical approximate role partition.

### E.7 Per-instance compressibility need not imply a shared template

Take two equally likely instances with

$$v_0(S) = 0, \qquad v_1(S) = 1. \tag{123}$$

Each instance is fully symmetric and has $E_{\mathrm{id}} = 0$. A shared template that depends only on an aligned state $T$ must predict $1/2$, giving $E_{\mathrm{temp}} = 1/4$. Cross-instance consistency is therefore a substantive requirement rather than a decorative assumption.

### E.8 Independent and coordinated substitutions are different questions

Theorem C and Appendix C show two complementary regimes. In the diffuse regime, no individual pair is decisive at the target tolerance even though independent block compression remains globally expensive. In the exact-state construction, cross object pairs are individually decisive. Both admit much smaller coordinated quotients because the admissible transformation preserves the binding correspondence. Any interpretation of a quotient must therefore specify whether it uses block exchange, coordinated renaming, or another structure preserving transformation.

## F Controlled Transformer experiment and numerical audits

This appendix records the full training protocol, seed level results, exhaustive quotient analysis, order control, and numerical theorem audits. All reported learned quotient quantities are computed after model selection on fixed trained models.

### F.1 Architecture, optimization, and split

Each of the $2r = 10$ clauses is represented as the sum of a field type embedding, an object identity embedding, a binary value embedding, and a learned positional embedding. A pooled token is prepended. The encoder has two Transformer layers, hidden width 48, four attention heads, feed forward width 96, and no dropout. The classifier predicts the six mismatch counts $0, \dots, 5$. Training uses AdamW with learning rate $3 \times 10^{-3}$, weight decay $10^{-4}$, and batch size 256. Clause order is resampled during training.

For each random seed, the 1024 complete states are randomly split into 716 training, 153 validation, and 155 test states. Epoch selection uses validation loss only. The untouched test split is evaluated once after model selection. The scalar response in Equation equation 35 is then evaluated on all 1024 states for post-training quotient analysis.

### F.2 Seed level results

| Seed | Test acc. | $\|v_\theta - v_r\|_2$ | $K_{\min}$ at $1/256$ |
|---|---|---|---|
| 0 | 1.000 | $3.75 \times 10^{-4}$ | 243 |
| 1 | 1.000 | $2.63 \times 10^{-4}$ | 243 |
| 2 | 1.000 | $2.22 \times 10^{-4}$ | 243 |
| 3 | 1.000 | $3.26 \times 10^{-4}$ | 243 |
| 4 | 1.000 | $4.86 \times 10^{-4}$ | 243 |

Across seeds, mean response RMSE is $(3.34 \pm 1.03) \times 10^{-4}$. The mean object-pair quotient error is $(1.65 \pm 1.68) \times 10^{-8}$, the mean coordinated binding quotient error is $(3.13 \pm 3.24) \times 10^{-8}$, the

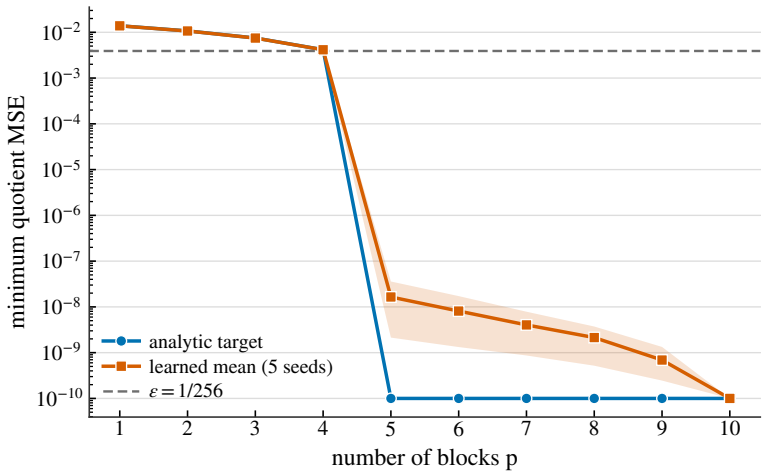



Figure 3: Exact block quotient frontier at $r = 5$. The analytic target and five learned responses remain above $\varepsilon_0 = 1/256$ with four blocks and become nearly exact with five, although every pairwise forbidden merge test is silent.

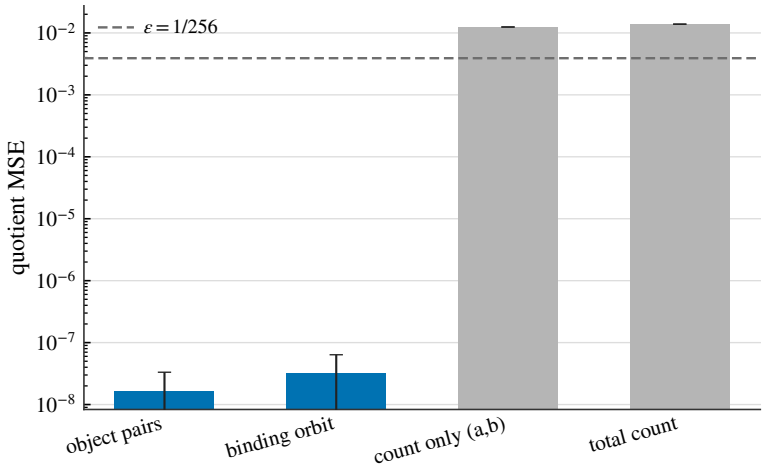



Figure 4: Quotient errors for the learned Transformer responses at $r = 5$. Object-pair block counts and the smaller binding preserving orbit representation have near zero error, while aggregate count representations discard substantial relational information. Error bars summarize five random seeds.

mean $(a, b)$ quotient error is $0.0124641 \pm 8.59 \times 10^{-6}$, and the mean total-count quotient error is $0.0138490 \pm 9.52 \times 10^{-6}$.

### F.3 Exact quotient enumeration and pairing control

For ten labeled slots there are 115,975 set partitions. We enumerate every partition and compute its exact conditional mean projection under the uniform distribution on all 1024 states. At $\varepsilon = 1/256$ and at $\varepsilon = 10^{-3}$, the minimum feasible state count is 243 for the analytic target and for every learned response. At $\varepsilon = 10^{-3}$, all 945 perfect pairings of the ten slots are also enumerated. The true object pairing is the unique feasible pairing for the target and for every trained model.

### F.4 Order control

The architecture includes learned positional embeddings, so permutation invariance is not built in. For the final seed 0 model, all 1024 states are reevaluated under 32 independently sampled clause

orders. Classification accuracy is 100% for every order. The mean per-state standard deviation of the scalar response is $6.64 \times 10^{-5}$, the maximum is $1.55 \times 10^{-3}$, and the mean order induced response variance is $1.45 \times 10^{-8}$.

### F.5 THEORY AND NUMERICAL AUDITS

The exact projection formula in Proposition 1 is independently checked by exhaustive partition enumeration at $r = 2, 3, 4$, covering 15, 203, and 4,140 partitions. The maximum absolute discrepancy between the direct quotient calculation and Equation equation 22 is $2.22 \times 10^{-16}$, and no violation of the $p/r$ bound is observed above numerical tolerance.

We also retain the earlier audits of the local to global certificate and cross-instance diagnostics. These include exhaustive checks of all 65,536 Boolean functions in the small Boolean setting, exchange spectrum checks for $n \leq 6$, 400 repetitions of the finite query selection procedure, an independent exact rational audit of the complementary binding construction, and 200,000 repetitions of the four query diagnostic. These numerical checks do not replace the analytical proofs.

## G PROOF AND NUMERICAL VERIFICATION STATUS

This appendix maps the main mathematical and empirical statements to their proof or numerical verification status. The controlled Transformer settings, data splits, quotient analysis, order control, and numerical checks are specified in Appendix F.

### G.1 PROOF AND CLAIM STATUS

The following table records which statements are established mathematically and which remain empirical questions.

| Item | Status |
|---|---|
| Exact contextual substitutability and block quotient | Complete proof |
| Best projection and within-orbit pairwise identity | Complete proof |
| Sharp local-to-global spectral bounds and forbidden merge | Complete proof + low-dimensional spectral audit |
| Finite-query selection-after-certification | Complete proof |
| Conditional unique minimum block quotient | Complete proof; explicitly sufficient |
| Exact arbitrary-partition relational projection | Complete proof + exhaustive low-dimensional audit |
| Diffuse relational separation | Complete proof + exhaustive low-dimensional audit |
| Learned-response stability bound | Complete proof |
| Complementary exact-state binding separation | Complete proof + exact rational audit |
| Controlled Transformer relational quotient | Five trained models; exhaustive $r = 5$ quotient analysis |
| Identity/template/fit decomposition | Complete proof |
| Four-query unbiased diagnostic | Complete proof |
| Finite-state learning bound | Complete proof; not an LLM optimization theorem |
| Low-residual relational quotients in pretrained LLMs on natural text | Not yet established |
| Priority or "first" claims | Require a systematic priority audit; not asserted |