CONSTRUCTIVE STATE ABSTRACTION AND FINITE IDENTIFICATION IN AUTOREGRESSIVE TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding which internal information is computationally necessary in a Transformer is a basic question in mechanistic interpretability and state abstraction. We study the state information needed to preserve a fixed-weight, pre-normalized autoregressive Transformer's computation under a declared class of internal interventions. From the weights, we construct explicit updates for residual coordinates in the joint read–write subspace, the squared norm of its omitted complement when normalization requires it, and cache entries grouped by key with their multiplicities and observable value sums. This state reproduces every finite generation distribution without reconstructing the original residual vector. If the centered terminal readout is nonzero, finitely many single-forward-pass interventions identify the state from the probability ratio of one fixed token pair, even for a rank-one readout. Equality of the constructed states is therefore equivalent to equality of observations under all admissible controlled continuations. Controlled cache insertion separates the attention numerator and denominator. We show that residual resets are unnecessary and characterize the information retained when all executed queries lie in specified subspaces. These identification results yield minimum state dimensions for continuous and affine encoders on regular local domains, and a sharp first-order approximation bound under a specified observation metric. Checkpoint analyses examine ranks and sensitivity, while controlled FP64 experiments support fixed-token-pair reconstruction and its dependence on the available interventions, but reveal numerical instability in cache inversion. Quantized-checkpoint experiments on the Mistral-24B find equal candidate-scalar readings with unequal outputs and intervention effects, distinguishing causal influence from state sufficiency.

1 INTRODUCTION

Understanding a language model's computation requires distinguishing information decodable from its hidden states from information used by the computation. Probes can detect properties unnecessary for a task (Ravichander et al., 2021), and analyses in the neuron basis can miss distributed structure (Geiger et al., 2024). Causal abstraction instead requires correspondence under specified interventions (Geiger et al., 2026). We ask which state information preserves a fixed autoregressive Transformer's computation under such interventions, and when that information is necessary.

Causal abstraction formalizes correspondence under interventions, while distributed alignment search tests alignments with chosen high-level variables (Geiger et al., 2026; 2024). RASP describes Transformer algorithms, Tracr compiles programs into weights, and Transformer Programs trains a modified architecture that can be converted into discrete programs (Weiss et al., 2021; Lindner et al., 2023; Friedman et al., 2023). Model-extraction studies use output logits to recover information about hidden width and vocabulary projections (Finlayson et al., 2024). Orthogonal residual transformations support approximate width reduction (Ashkboos et al., 2024). Empirical KV-cache compression drops or merges entries while limiting accuracy loss (Zhang et al., 2023; Li et al., 2024), and Haris & Onak (2025) establish worst-case space lower bounds for attention-based generation. These approaches address different objects and access assumptions. Here the weights are known, the unknown object is an activation state, and identification uses internal controls while leaving the final readout unchanged.

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 seek an exactly updatable state and its minimum local dimension under those controls, without assuming a human-readable symbolic program.

General finite-experiment results already apply to analytic response maps, not only differential equations: Sontag (2003) proves that generic finite experiments preserve all distinguishability available within such a family. Unknown initial states can be included among the parameters. For language models, Golowich et al. (2026) characterize low extended-logit rank by time-varying input-switched affine networks and give learning guarantees with logit-query access. Their state is a linear recurrent realization of token-history behavior. Our object is instead an open domain of internal activations, with arbitrary continuous encoders, under specified interventions. The additional content is a concrete behavioral equivalence relation, fixed token-pair measurements, and an explicit inverse with full-rank differential; mere existence of a finite distinguishing family is not the novelty claim.

This question is closely related to bisimulation-based state abstraction (Givan et al., 2003; Li et al., 2006). Such work characterizes state equivalence through transition and observation behavior, whereas we construct an executable state directly from the weights of a fixed Transformer under specified internal interventions and show that it can be identified from finitely many final-probability observations. For any nonzero centered terminal readout, two states are behaviorally equivalent if and only if their constructed states agree. Our identification controls are also related to the internal readouts used by the logit lens and direct logit attribution (nostalgebraist, 2020; Elhage et al., 2021), but here they are used for finite state identification rather than only for interpreting intermediate representations.

More specifically, our method constructs an executable internal state directly from the model weights. For the residual state, we retain coordinates in the subspace E spanned jointly by all effective normalized reads and residual writes; when normalization also depends on the omitted orthogonal complement, we additionally retain its squared norm, which suffices to determine all normalization scales exactly. For each shared cache, we retain only the value components visible to at least one attention head and merge entries with identical keys into a key, its multiplicity, and the corresponding value sum. These variables admit closed updates, so the model can be executed without reconstructing the original residual vector or the full cache. To further show that this retained information is identifiable, we use insertions of known keys and values to separate the numerator and denominator coupled in an ordinary attention read, and recover the retained cache state from final probabilities using finitely many exponential-moment measurements.

Our central result is explicit finite identification from one fixed token-pair probability ratio and the resulting characterization of behavioral equivalence (Theorem 3). The construction also gives exact component updates and preserves finite generation distributions (Theorem 1). For an isolated-cache measurement family, we identify necessary and sufficient conditions, show that residual resets are unnecessary, and derive the state for restricted query spaces (Section 4.3). Minimum local state dimensions for continuous and affine encoders and a sharp first-order approximation bound follow from this identification result (Theorems 6–8). Experiments examine ranks, local sensitivity, numerical reconstruction, and the insufficiency of a causally influential scalar.

2 Model, state domain, and interventions

2.1 A fixed pre-normalized decoder

Fix a finite-depth causal Transformer (Vaswani et al., 2017) with residual width d, a vocabulary $\mathcal{V}$ of size M, and a generation horizon $H \geq 1$. We use exact real arithmetic with dropout disabled. All residual normalizations share the projection $P = \mathsf{I}_d$ or $P = \mathsf{I}_d - d^{-1}\mathbf{1}_d\mathbf{1}_d^{\mathsf{T}}$, for RMS normalization (Zhang & Sennrich, 2019) or layer normalization, respectively. At normalization site ι,

$$\mathsf{N}_\iota(x) = \frac{\Gamma_\iota P x}{\sqrt{\|Px\|_2^2/d + \epsilon_\iota}} + \beta_\iota, \tag{1}$$

with diagonal gain Γ_ι, offset β_ι, and stabilizer $\epsilon_\iota > 0$; allowing a zero stabilizer changes identifiability. The final site f produces logits $W_{\mathrm{U}}\mathsf{N}_f(x) + b_{\mathrm{U}}$, with $W_{\mathrm{U}} \in \mathbb{R}^{M\times d}$. Set $\Pi = \mathsf{I}_M - M^{-1}\mathbf{1}_M\mathbf{1}_M^{\mathsf{T}}$. Exact probabilities π and centered log probabilities $\mathcal{L} = \Pi \log \pi$ determine one another.

A cache group $c \in \mathcal{C}$ has key width d_c, raw value width w_c, and query heads $\mathcal{H}(c)$ sharing its keys and values. Head h writes through $O_h \in \mathbb{R}^{d \times w_c}$. Fixed score scaling and positional rotations are absorbed into the query q and stored key ϑ, giving score $q^{\mathsf{T}}\vartheta$. Attention reads all historical entries and the current appended entry; slot-dependent masks, KV cache eviction, and additional positional score terms are excluded.

Feed-forward branches use affine normalized reads, fixed pointwise nonlinearities, optional coordinatewise products, and affine residual writes. Sequential branches retain their original order; parallel attention and feed-forward branches read the same block input and sum their contributions. Attention-head, feed-forward, and constant-bias writes have independent gates. Native gates equal one; gating a residual contribution does not suppress its cache write. Appendix A specifies the component boundaries.

2.2 State domain and admissible controls

A token-entry state ξ contains the live residual and historical keys and values. Its discrete type p records current execution location, position, cache lengths, and remaining horizon. Residual and cache entries may vary arbitrarily at this token-entry boundary; prompt reachability is not assumed.

Controls allow native execution, forced continuations, independent residual gates, cache-write suppression, query replacement, replacement of the appended entry, whole-residual resets to declared constants, and translations by every vector in E, defined below. They neither replace individual historical slots nor inspect their order, and leave the terminal normalization and readout fixed. Policies depend only on emitted tokens and declared controls. Each pass has finitely many architecture-defined control sites; active heads must read nonempty caches.

Each identification intervention uses a separate copy of the token-entry state for one forward pass. It either bypasses residual branches with a declared translation, or resets the residual to a declared constant, isolates one head, and replaces its query and appended entry while disabling all other residual contributions and irrelevant cache writes. Compilation holds at the specified component boundaries, whereas identification and dimension bounds are stated for token-entry states.

3 Constructing an executable state from the weights

3.1 Observable values and the residual subspace

For each group define $S_c = \operatorname{span}\{\operatorname{row}(PO_h) : h \in \mathcal{H}(c)\}$. Let $B_c \in \mathbb{R}^{w_c \times s_c}$ have orthonormal columns spanning S_c, where $s_c = \dim S_c$, and remove groups with $s_c = 0$. Replace each raw value v by $\nu = B_c^{\mathsf{T}} v$. Since $PO_h = PO_h B_c B_c^{\mathsf{T}}$, discarded components do not affect any centered head write. Here observability refers to these maps; recovery from probabilities requires Section 4.

Let $\mathcal{R}$ contain $R\Gamma_\iota P$ for every normalized affine read $R\mathsf{N}_\iota(x) + b$, including queries, keys, projected value reads $B_c^{\mathsf{T}} W_{V,c}$, feed-forward inputs and gates, and the terminal read $\Pi W_{\mathrm{U}} \Gamma_f P$. Let $\mathcal{B}$ contain the centered head and feed-forward output matrices and centered write biases as columns. Define

$$E = \operatorname{span}\big(\{\operatorname{row}(R_0) : R_0 \in \mathcal{R}\} \cup \{\operatorname{col}(B_0) : B_0 \in \mathcal{B}\}\big), \tag{2}$$

and let $U \in \mathbb{R}^{d \times r}$ be an orthonormal basis, with $r = \dim E$. Thus $\operatorname{ran} P \cap E^{\perp}$ is neither read linearly nor written. For $r \geq 1$, retain

$$u = U^{\mathsf{T}} x,\ \rho = \left\|(P - UU^{\mathsf{T}})x\right\|_2^2,\ \eta = \mathbf{1}\{r < \operatorname{rank} P\}. \tag{3}$$

Omit ρ as a coordinate when $\eta = 0$, and set $\rho = 0$ in all update and readout formulas. When $r = 0$, vocabulary probabilities are constant and the empty state suffices.

Both reads and writes enter E because the state must remain sufficient after each update. A direction outside the current read space can still receive a residual write and affect later computation through normalization. By including every centered write, we keep the omitted vector fixed between resets. Its orientation is then irrelevant to all reads, while its norm contributes to every normalization denominator.

3.2 CACHE STATE AND DIRECT UPDATES

Let group c store n_c keys $\vartheta_{c,i}$ with raw values $v_{c,i}$. Merge equal keys into distinct keys $\kappa_{c,j}$, multiplicities $\mu_{c,j} \in \mathbb{N}$, and projected-value sums $\omega_{c,j} = \sum_{i:\vartheta_{c,i}=\kappa_{c,j}} B_c^{\mathsf{T}} v_{c,i}$. Retain the unordered collection $\mathcal{K}_c = \{(\kappa_{c,j}, \mu_{c,j}, \omega_{c,j}) : 1 \le j \le g_c\}$, $\sum_j \mu_{c,j} = n_c$. Multiplicities are discrete metadata. After appending the current key and value, head h contributes, with $G_h = U^{\mathsf{T}} P O_h B_c$,

$$u \longleftarrow u + G_h \frac{\sum_j \exp(q^{\mathsf{T}} \kappa_{c,j}) \omega_{c,j}}{\sum_j \mu_{c,j} \exp(q^{\mathsf{T}} \kappa_{c,j})}. \tag{4}$$

Insertion creates a new triple or increments the matching multiplicity and value sum. Token position remains in p, preserving later positional transformations.

Set $\sigma_\iota(u, \rho) = \sqrt{(\|u\|_2^2 + \rho)/d + \epsilon_\iota}$. Every retained normalized affine read becomes

$$R\mathsf{N}_\iota(x) + b = (R\Gamma_\iota U) \frac{u}{\sigma_\iota(u, \rho)} + R\beta_\iota + b. \tag{5}$$

Reduced matrices are precomputed. Queries and shared cache writes use the same incoming state; head contributions are summed before updating u. Parallel branches likewise read the same incoming (u, ρ) with their own normalization parameters. Feed-forward updates retain the original nonlinearities. Residual writes and translations $u \mapsto u + \delta$ leave ρ fixed; resets and token embeddings assign both statistics from the declared vector. Define

$$A = \Pi W_{\mathrm{U}} \Gamma_f U, \; b_f = \Pi(W_{\mathrm{U}} \beta_f + b_{\mathrm{U}}). \tag{6}$$

Then $\mathcal{L} = Au/\sigma_f(u, \rho) + b_f$ and $\pi = \mathrm{softmax}(\mathcal{L})$.

Theorem 1 (Exact constructive realization)**.** *For the architecture and controls of Section 2, $\Phi_p(\xi) = (u, \rho, \{\mathcal{K}_c\}_{c \in \mathcal{C}})$, with absent coordinates omitted, has explicit component updates F_e satisfying*

$$\Phi_{p'}(T_e \xi) = F_e(\Phi_p(\xi)), \; \pi(\xi) = \pi_{\mathrm{red}}(\Phi_p(\xi)). \tag{7}$$

Here T_e is any admissible component transition of type $p \to p'$, and the probability identity holds at terminal readout. Both decoders induce the same joint distribution on every finite generated sequence under the same admissible policy.

Proof sketch. Orthogonality gives $\|Px\|_2^2 = \|u\|_2^2 + \rho$. Effective reads annihilate the omitted vector, and centered writes leave it fixed. Merging preserves attention numerators and denominators, including insertion. Induction over components and token prefixes proves the result; Appendix B gives the full construction. □

Execution uses the reduced updates without reconstructing the original residual or calling the original Transformer. It need not reduce storage or computation when the retained dimensions are full. Changing orthonormal bases only rotates coordinates and conjugates updates.

4 FINITE IDENTIFICATION THROUGH FINAL PROBABILITIES

Assumption 2 (Nonzero terminal readout)**.** The matrix A in (6) is nonzero. The single-pass controls of Section 2 permit effective queries in an open ball about zero and all translations in E.

Since $\Pi W_{\mathrm{U}} \Gamma_f P = AU^{\mathsf{T}}$, the readout condition excludes only a constant terminal probability map. If $A = 0$, the empty behavioral state suffices under every declared continuation, regardless of r (Appendix B.1). If $A \neq 0$, even rank one suffices for identification below. When $E = \mathrm{ran}\, P$, $s_c = w_c$, and all historical keys are distinct, Theorem 6 precludes further exact continuous reduction after removing the residual mean, if present.

Checkpoint ranks. For GPT-2 small, Pythia-160M, and Llama-3.2-1B, the read–write span has numerical rank $\mathrm{rank}\, P$ at relative threshold 10^{-3}, and all 416 cache groups have full observable-value rank. GPT-2's terminal rank is 762 at this threshold and 767 at 10^{-5}, with smallest-to-largest singular-value ratio 6.14×10^{-5}. These measurements do not certify algebraic rank (Appendix I.1).

4.1 RECOVERING RESIDUAL STATISTICS

Choose tokens j_+, j_- such that $a_0 = A^\mathsf{T}(\mathbf{e}_{j_+} - \mathbf{e}_{j_-}) \neq 0$, and set $a = a_0 / \|a_0\|_2$. Such a pair exists because $\mathbf{1}_M^\mathsf{T} A = 0$. Bypass all residual branches and translate by $U\delta$. From this pair alone, with $c_f = d\epsilon_f$,

$$y_\delta = \frac{\log(\pi_{j_+}/\pi_{j_-}) - (b_{f,j_+} - b_{f,j_-})}{\sqrt{d}\ \|a_0\|_2} = \frac{a^\mathsf{T}(u+\delta)}{\sqrt{\|u+\delta\|_2^2 + \rho + c_f}}. \tag{8}$$

Extend a to an orthonormal basis $a, e_2, \ldots, e_r$ and write $u = ta + v$, $v \perp a$. Fix $s_{\mathrm{tr}}, h_{\mathrm{tr}} > 0$ and $\mathcal{B}_\perp = \{0, h_{\mathrm{tr}}e_2, \ldots, h_{\mathrm{tr}}e_r\}$. The $2r$ translations $\delta \in \{b, b + s_{\mathrm{tr}}a : b \in \mathcal{B}_\perp\}$ give, for $\mathcal{T}(z) = z/\sqrt{1-z^2}$,

$$R_b = \frac{s_{\mathrm{tr}}}{\mathcal{T}(y_{b+s_{\mathrm{tr}}a}) - \mathcal{T}(y_b)} = \sqrt{\|v+b\|_2^2 + \rho + c_f}. \tag{9}$$

Thus $t = R_0\mathcal{T}(y_0)$, $e_i^\mathsf{T} v = (R^2_{h_{\mathrm{tr}}e_i} - R_0^2 - h_{\mathrm{tr}}^2)/(2h_{\mathrm{tr}})$, and $\rho = R_0^2 - \|v\|_2^2 - c_f$. The denominators are strictly positive, including at $u = 0$ and $\rho = 0$. Appendix C proves smooth recovery and retains the shorter full-column-rank inverse used in the experiments. The token pair and translations are fixed before observing the state.

4.2 INSERTION AND FINITE CACHE RECONSTRUCTION

For one cache, write $Z(q) = \sum_j \mu_j e^{q^\mathsf{T}\kappa_j}$, $N(q) = \sum_j e^{q^\mathsf{T}\kappa_j}\omega_j$. Reading a nonempty historical cache with writes suppressed reveals only $N(q)/Z(q)$, which is invariant under a common key translation. Appending a zero key with known value b_{ins} instead gives $G_h \frac{N(q)+b_{\mathrm{ins}}}{Z(q)+1}$. Resetting to each $U\delta$ above and fixing the query and appended entry makes the isolated write independent of δ; its final residual coordinates are $w + \delta$. Equations (8)–(9), with $\rho = 0$, therefore recover w in $2r$ passes. Contrasting insertions with values zero and b_{ins}, chosen so $G_{h_0} b_{\mathrm{ins}} \neq 0$, identifies $(Z+1)^{-1}$. Zero-value insertions identify $G_h N$ for every head; the stacked G_h have column rank s_c, determining N. This uses at most $2r(|\mathcal{H}(c)| + 1)$ forward passes per query after zero-write heads are removed.

For a direction a_{key} separating the keys, $Z(ma_{\mathrm{key}}) = \sum_j \mu_j \lambda_j^m$, where $\lambda_j = e^{a_{\mathrm{key}}^\mathsf{T}\kappa_j}$. From samples with $0 \leq m < 2n_c$, the Hankel rank gives g_c, and a nonsingular Hankel system determines an annihilating polynomial. Its distinct positive roots give the nodes; a Vandermonde solve recovers their multiplicities (Vetterli et al., 2002). Coordinate-offset queries recover the keys (Cuyt & Lee, 2018), and $N(ma_{\mathrm{key}})$ recovers value sums. A fixed library of $J_c = 1 + (d_c - 1)\binom{n_c}{2}$ scaled moment-curve directions contains a separating direction for every cache of that length. All query grids fit inside the permitted ball; Appendix C supplies the non-adaptive construction, inversion formulas, and duplicate-key cases.

Theorem 3 (Finite identification and behavioral equivalence)**.** *Under Assumption 2, a fixed finite family identifies $\Phi_p(\xi)$ at every token-entry state of a fixed discrete type from the probability ratio of the same token pair in every pass, using at most*

$$2r\left[1 + \sum_{c:n_c>0} (|\mathcal{H}(c)| + 1) J_c\, n_c(d_c + 2)\right] \tag{10}$$

single-forward-pass interventions. Two such states have identical final-probability observations under every admissible continuation of length at most H, for any $H \geq 1$, if and only if their residual statistics and cache triples agree up to reordering.

Identification proves necessity; the closed updates prove sufficiency for future observations. The count establishes finiteness, not efficiency: the cache query grid alone can exceed 10^{12} queries for 10^3 entries of width 128.

Reconstruction selects a direction of maximum Hankel rank from the recorded outputs. A separating direction attains rank g_c, whereas a projection collision lowers it; no later intervention depends on earlier observations.

4.3 WHICH CONTROLS ARE NEEDED?

Residual resets can be omitted without increasing (10). First recover (u, ρ) by bypass measurements. Without resetting, a fixed isolated head write w followed by the same translations gives $(u+w+\delta, \rho)$. The scalar inverse recovers $u + w$, and subtraction gives w. Theorem 3 therefore remains valid for the reset-free control class (Appendix H).

For a precise necessity statement, fix a nonempty cache c with $d_c \geq 1$. Consider only isolated-head measurements with an accessible head set $\mathcal{H}_{\mathrm{obs}} \subseteq \mathcal{H}(c)$, queries in an open ball about zero in a linear subspace $\mathcal{Q}_c$, and one zero-key insertion whose projected value belongs to a nonempty set $\mathcal{B}_{\mathrm{ins}}$. Residual translations and bypass measurements remain available; there is no native suffix or additional insertion in a pass. Set $\mathsf{G}_{\mathrm{obs}} = (G_h)_{h \in \mathcal{H}_{\mathrm{obs}}}$.

Proposition 4 (Identification within an isolated-cache family)**.** *Under the nonzero-readout condition, a fixed finite family of these measurements identifies every cache of the prescribed length, up to its grouped triples, if and only if* $\mathcal{Q}_c = \mathbb{R}^{d_c},\ \operatorname{rank} \mathsf{G}_{\mathrm{obs}} = s_c,\ \exists h \in \mathcal{H}_{\mathrm{obs}},\ b_0, b_1 \in \mathcal{B}_{\mathrm{ins}} : G_h(b_1 - b_0) \neq 0.$

The value contrast recovers Z; the stacked writes then recover N. Conversely, restricted queries hide orthogonal key components, a deficient write stack hides value components, and without a value contrast a cache whose values all equal b_0 hides its keys. These necessities concern the specified measurement family, not every possible controlled computation.

A restricted query space also gives an exact smaller state when it constrains *every executed query*, including any native query. Let $V_{q,c}$ have orthonormal columns spanning $\mathcal{Q}_c$, with $t_c = \dim \mathcal{Q}_c$. Replace each key by $V_{q,c}^{\mathsf{T}}\vartheta$ and merge equal projected keys, retaining multiplicities and projected-value sums. Denote the resulting extractor by $\Phi_p^{\mathcal{Q}}$ and its group counts by $\bar{g}_c$.

Corollary 5 (State for restricted query spaces)**.** *Suppose* $A \neq 0$*. Keep* E*, its translations, all head gates, and zero/nonzero-value insertions as above. If query replacement covers a relative open ball in each* $\mathcal{Q}_c$ *and every executed query belongs to its declared subspace, equality of* $\Phi_p^{\mathcal{Q}}$ *is equivalent to behavioral equivalence under these controls. On regular projected-key charts the minimum continuous dimension is*

$$D_{\mathcal{Q}} = r + \eta + \sum_c \bar{g}_c(t_c + s_c). \tag{11}$$

For $t_c = 0$, a nonempty cache contributes only its value sum. Restricting query replacements alone is insufficient if native queries may leave $\mathcal{Q}_c$. Appendix H gives the update, reconstruction, and dimension proofs.

5 MINIMAL STATE AND OPTIMAL LOCAL APPROXIMATION

5.1 CONTINUOUS AND AFFINE STATE DIMENSIONS

On a regular chart, cache c has g_c distinct keys with fixed multiplicities and a fixed local ordering. Require $\rho > 0$ when $\eta = 1$. The continuous coordinates ζ have dimension

$$D = r + \eta + \sum_c g_c(d_c + s_c). \tag{12}$$

They vary on an open set because the token-entry state domain permits independent residual, key, and value variation. Let $\mathcal{E}(\zeta)$ concatenate the finite library's full centered log-probability observations; the identifying log ratios are fixed linear functions of this vector.

Theorem 6 (Local minimum continuous-state dimension)**.** *Under Assumption 2, any continuous encoder of the original activation state into* $\mathbb{R}^k$ *that reproduces every admissible intervention's final probabilities through arbitrary deterministic reconstruction maps satisfies* $k \geq D$ *on each regular chart. The constructed state attains this bound.*

Proof sketch. Choose a continuous section of the chart into original states, using a fixed unit direction for the omitted residual complement scaled by $\sqrt{\rho}$ when needed and linear sections for projected

values. Two chart points with the same code have identical reconstructed intervention observations, so finite identification makes the points equal. Composing the encoder with the section therefore gives a continuous injection from an open subset of $\mathbb{R}^D$ into $\mathbb{R}^k$. Invariance of domain requires $k \geq D$; the constructed coordinates attain equality. Appendix D gives the full argument. □

This local bound neither supplies global coordinates across key collisions nor constrains prompt-reachable states alone. Splitting collided keys can increase the required dimension.

For distinct historical keys, set $D_{\text{aff}} = \operatorname{rank} P + \sum_c n_c(d_c + s_c)$.

Theorem 7 (Minimum affine-state dimension)**.** *Under Assumption 2, any affine encoder of the full original token-entry state preserving every declared intervention output on a nonempty open set with distinct keys has rank at least D_{aff}. Retaining Px, every key, and every projected value attains the bound. On this stratum,*

$$D_{\text{aff}} - D = \operatorname{rank} P - r - \eta. \tag{13}$$

Proof sketch. A null direction of the affine encoder leaves its code constant along a short segment in the open domain. Identification forces every locally labeled key, projected value, and u to remain fixed. The omitted squared norm, when present, must also be constant; its quadratic coefficient is the squared length of the residual direction in that complement. This coefficient vanishes only when that direction is zero. The encoder can therefore discard no direction contributing to Px or the retained cache coordinates. Rank–nullity gives the bound (Appendix D.3). □

Thus one squared norm saves dimensions over every affine encoder exactly when $\operatorname{rank} P - r \geq 2$. A one-dimensional complement loses its sign but no local dimension. Appendix D.3 separates the savings from mean removal, normalization, value kernels, and duplicate keys.

5.2 A SHARP FIRST-ORDER APPROXIMATION BOUND

Choose a closed chart ball of radius $R_{\text{loc}} > 0$ about ζ_0. Define $e_k(R_{\text{loc}}) = \inf_{\psi,\mathcal{D}} \sup_{\|v\|_2 \leq R_{\text{loc}}} \|\mathcal{E}(\zeta_0 + v) - \mathcal{D}(\psi(v))\|_2$, where ψ is continuous into $\mathbb{R}^k$ and $\mathcal{D}$ is any deterministic decoder. Let $J_0 = D\mathcal{E}(\zeta_0)$, and assume $\mathcal{E}$ is C^2 near the ball with $\left\|D^2\mathcal{E}(\zeta)[v,w]\right\|_2 \leq \mathfrak{B}\,\|v\|_2\,\|w\|_2$ on the ball. Smoothness holds for these interventions (Appendix E).

Theorem 8 (Local approximation bound)**.** *Under Assumption 2 and these smoothness and Hessian conditions, for $0 \leq k < D$,*

$$\max\{0, R_{\text{loc}}\sigma_{k+1}(J_0) - \tfrac{1}{2}\mathfrak{B}R_{\text{loc}}^2\} \leq e_k(R_{\text{loc}}) \leq R_{\text{loc}}\sigma_{k+1}(J_0) + \tfrac{1}{2}\mathfrak{B}R_{\text{loc}}^2. \tag{14}$$

Moreover, $\operatorname{rank} J_0 = D$ *and* $\lim_{R_{\text{loc}}\downarrow 0} e_k(R_{\text{loc}})/R_{\text{loc}} = \sigma_{k+1}(J_0) > 0$.

Proof sketch. The smooth local inverse implies $\operatorname{rank} J_0 = D$ by differentiation. Taylor's theorem bounds the nonlinear remainder by $\frac{1}{2}\mathfrak{B}R_{\text{loc}}^2$. Retaining coordinates along the leading k right singular vectors and decoding affinely gives the upper bound. For any continuous encoder, Borsuk–Ulam supplies antipodal points with the same code on the sphere of radius R_{loc} in the span of the first $k+1$ right singular vectors. Their linear observations differ by at least $2R_{\text{loc}}\sigma_{k+1}(J_0)$. Their decoder outputs agree, so one error is at least half their observation distance. The Taylor bound then yields the lower bound (Appendix E). □

This continuous nonlinear-width result (DeVore et al., 1989) concerns the finite observation vector, not optimal compressed generation. Its metric uses orthonormal residual and value coordinates, stored key coordinates, the unscaled squared norm ρ, and centered log probabilities. Repeating every observation m_{rep} times multiplies the singular values and this Euclidean error by $\sqrt{m_{\text{rep}}}$; they are not observation-independent invariants. For an affine observation map, the error is exactly $R_{\text{loc}}\sigma_{k+1}(J_0)$.

Local sensitivity measurements. We use eight WikiText-103 validation states per checkpoint, two cached tokens, and all layers and groups. For each key width, a direction fixed independently of the states defines an empirical map $\mathcal{E}_{\text{exp}}$, distinct from the universal identification library. Its Jacobian J_{exp} concentrates 90% of its squared singular-value sum in 14.74–16.54% of the state dimensions (Table 1). At relative radius 10^{-4}, mean symmetric secant ratios along tested singular directions differ

from one by at most 0.040%. All sampled comparisons against raw cache values and checkpoint terminal modules pass; these are local sensitivity checks (Appendix I.2).

These spectra show that an exact dimension lower bound can coexist with substantial differences in sensitivity across state directions. Small singular values permit smaller first-order observation errors when the corresponding directions are omitted locally. The sum of all squared singular values is the squared Frobenius norm of this Jacobian; it does not give a fraction of preserved task accuracy or establish an executable compressed decoder.

Table 1: Full-state spectra with two-entry caches. k_p is the smallest number of leading singular values accounting for $p\%$ of their squared sum. Percentages are means and sample standard deviations over eight states.

Model	D	k_{90}/D	k_{95}/D	k_{99}/D
GPT-2 small	37,631	14.74 ± 0.40	18.28 ± 0.36	23.57 ± 0.31
Pythia-160M	37,631	15.79 ± 0.39	19.27 ± 0.35	25.33 ± 0.45
Llama-3.2-1B	34,816	16.54 ± 0.36	18.98 ± 0.37	24.00 ± 0.42

6 Stability and finite-horizon execution

Under Assumption 2, selecting a separating direction on a regular chart gives a smooth local inverse $\mathscr{I}$ with $\mathscr{I} \circ \mathcal{E} = \mathsf{I}$. Let $\chi = \sup \|D\mathscr{I}\|_2$ on a small convex observation neighborhood. Although finite away from singular configurations, χ can be large because small normalization stabilizers, weak token-pair contrasts, small translation scales, and nearly coincident exponential nodes can amplify reconstruction errors. Appendix F derives these factors and distinguishes moment-system conditioning from denominator growth under a fixed query ball.

For bounded controls whose continuations remain in a compact differentiability domain, let L_H bound derivatives of final centered-log-probability maps over continuations $\boldsymbol{e}$ of length at most H. When the relevant segments remain in these neighborhoods, $\sup_{\boldsymbol{e}:\,|\boldsymbol{e}|\le H} \|\mathcal{L}_{\boldsymbol{e}}(\zeta) - \mathcal{L}_{\boldsymbol{e}}(\zeta')\|_2 \le L_H \chi \|\mathcal{E}(\zeta) - \mathcal{E}(\zeta')\|_2$. Proposition 12 proves this local bound; it is not uniform over unbounded controls or key collisions.

Synthetic reconstruction uses the full-column-rank inverse with $d = 16$, $r = 8$, and caches of one, two, or four entries. In FP64 without added noise, all 150 runs complete and 147 meet the 10^{-3} recovery tolerance. The three inaccurate cases have four entries, $\epsilon_f = 10^{-5}$, and ill-conditioned moment systems. With observation-noise standard deviation 10^{-8}, no four-entry case meets tolerance (Appendix I.3). Exact identification therefore does not ensure stable numerical inversion.

7 Experiments and discussion

Full experimental protocols and results are provided in Appendices I and J.

Fixed-token-pair reconstruction. We test residual and cache reconstruction in FP64 with $d = 16$, $r = 8$, $\mathrm{rank}(A) = 1$, and ten seeds per condition. One state-independent cache direction is fixed in advance; the universal direction library is not evaluated. Across three stabilizers and four residual cases, all 600 noiseless runs complete, with accurate recoveries of $120/120$, $240/240$, and $216/240$ for cache lengths one, two, and four. At log-ratio noise standard deviation 10^{-8}, the corresponding counts fall to $120/120$, $102/240$, and $0/240$. The 24 noiseless failures involve four distinct keys, illustrating the distinction between exact identification and finite-precision recovery.

Intervention ablations. Over ten seeds per condition, reset-free reconstruction gives a mean write relative error of 1.18×10^{-13} without added noise and 1.02×10^{-6} at noise standard deviation 10^{-8}, compared with 1.16×10^{-8} with resets. Constructed cache pairs have mean observation gaps below 9×10^{-15} under a single insertion value, proper query subspaces, or a rank-three write stack; the corresponding positive controls give gaps of 0.588, 2.12–3.03, and 6.69. These results illustrate the

indistinguishability mechanisms in Proposition 4 and the numerical cost of removing resets; they do not establish end-to-end checkpoint identification.

Causal tests of candidate representations. Quantized-checkpoint experiments test whether a causally influential scalar determines a response; their complete protocol and results appear in Appendix J. The checkpoint is Mistral-Small-24B-Instruct-2501 with NF4 four-bit linear weights, double quantization, and FP16 compute on one RTX 4090. Attention is eager, batch size is one, and cache reuse is disabled. The response r_{num} is the original Yes-minus-No logit difference for tokens 13830 and 3501; no surrogate supplies the outputs. The task asks for membership in the least closure of two Horn rules,

$$\mathsf{P}(\mathsf{x}) \wedge \mathsf{R}(\mathsf{x},\mathsf{y}) \longrightarrow \mathsf{Q}(\mathsf{y}),\ \mathsf{Q}(\mathsf{y}) \wedge \mathsf{S}(\mathsf{y}) \longrightarrow \mathsf{Z}(\mathsf{y}). \tag{15}$$

The frozen candidate is a fitted direction at the layer-17 output, pooled over the relation's target-entity span. Its real-valued linear coordinate is c_{lin} ((59)); the implemented FP32 reading of FP16 activations is c_{num}. The fitting scale is 0.9344838042. Equal numerical readings below do not certify equal ideal-real projections.

The checkpoint answers 71.875% of 192 screening prompts correctly. Source interventions have positive average entity-match and downstream-premise interactions on 24 evaluation groups. Four selected layer-18 heads affect these contrasts on independent confirmation inputs, yet their ablation improves binary accuracy by 9.896 percentage points. These observations establish neither necessity for correct reasoning nor an identified intermediate logical fact (Appendices J.5–J.7).

Equal readings and unequal outputs. The baseline-coordinate preservation experiment holds the prompt and all other activation rows fixed while perturbing the candidate's orthogonal complement and correcting the actual FP16 rows to the baseline reading. Among 256 edits on 32 prompts, all reach tolerance 10^{-6}; 125 have exactly the baseline c_{num}. Of those, 58 change r_{num}, in 21 groups. All 32 no-op responses are unchanged. Thus a deterministic factorization of r_{num} through this numerical scalar fails on the tested states. The tolerance-filtered mean response range is 0.016602 with marginal 95% group-bootstrap interval $[0.013672, 0.019531]$; this separate statistic is not restricted to exact equality.

Equal targets and unequal effects. The shared-target intervention experiment uses 32 prompts, shared numerical targets at amplitudes $\gamma_{\mathrm{edit}} \in \{-0.5, 0.5\}$, and eight implementations per condition. At $\gamma_{\mathrm{edit}} = -0.5$ and 0.5, respectively, 132/256 and 72/256 implementations satisfy both exact target equality and the 10^{-4} perturbation-norm tolerance. At least two are accepted in 31 and 18 groups; 23 and 12 eligible groups have unequal effects. Mean effect ranges are 0.012601 with interval $[0.009577, 0.015625]$ and 0.012153 with interval $[0.007813, 0.016515]$. Intervals use 2,000 group-bootstrap resamples. These conditional finite-set results are not independent amplitude replications, global extrema, or counterexamples to preservation through the complete constructed state. Appendix J retains the numerical controls, coverage, and all stage-specific tables.

8 Scope, Limitations, and Conclusion

The constructed state preserves computation and finite generation distributions through explicit updates derived from the weights. With any nonzero centered terminal readout, controlled cache insertion and residual translations identify its retained information from one fixed token-pair probability ratio, yielding local minimum-dimension results and a sharp first-order approximation bound. Resets are unnecessary, and restricting every executed query gives the projected-key state of Corollary 5. The necessary-and-sufficient criterion of Proposition 4 applies only to its isolated-cache measurement family, and weaker or stronger control classes may change the required state. The lower bounds apply to open activation domains and assign no semantic interpretation to the coordinates. A full read–write span, full observable-value ranks, and distinct keys can leave no exact continuous reduction; approximate reduction then depends on the observation metric and local spectrum. The experiments show concentrated local sensitivity, numerical conditioning failures, and a distinction between causal influence and state sufficiency.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this paper, to support the retrieval and discovery of related work, and to assist in checking and revising portions of the mathematical arguments and proofs. All AI-assisted writing was independently verified by the authors against the stated assumptions and derivations, and all references identified with AI assistance were checked against the original publications. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Technical definitions and complete proofs of the theoretical results are provided in Appendices A–H. Appendix I reports the checkpoint-rank, local-sensitivity, directional-validation, full-column-rank reconstruction, fixed-token-pair reconstruction, and intervention-ablation experiments, including their numerical settings and evaluation procedures. Appendix J provides the protocol, hardware and numerical implementation details, data partitions, random seeds, statistical procedures, and complete reported results for the quantized-checkpoint causal study. The supplementary material contains the source code used for the experiments and numerical evaluations reported in the paper.

REFERENCES

Saleh Ashkboos, Maximilian Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. Slicegpt: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, volume 2024, pp. 11682–11701, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Annie Cuyt and Wen-shin Lee. Multivariate exponential analysis from the minimal number of samples. *Advances in Computational Mathematics*, 44(4):987–1002, 2018. doi: 10.1007/s10444-017-9570-8. URL https://doi.org/10.1007/s10444-017-9570-8.

Ronald A. DeVore, Ralph Howard, and Charles A. Micchelli. Optimal nonlinear approximation. *manuscripta mathematica*, 63:469–478, 1989.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. Logits of API-protected LLMs leak proprietary information. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=oRcYFm8vyB.

Dan Friedman, Alexander Wettig, and Danqi Chen. Learning transformer programs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In Francesco Locatello and Vanessa Didelez (eds.), *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 01–03 Apr 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

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. *J. Mach. Learn. Res.*, 26(1), September 2026. ISSN 1532-4435.

Robert Givan, Thomas L. Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artif. Intell.*, 147:163–223, 2003.

Noah Golowich, Allen Liu, and Abhishek Shetty. Sequences of logits reveal the low rank structure of language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2af179e30c16f3a6d44311adbb5ec63f-Abstract-Conference.html.

Themistoklis Haris and Krzysztof Onak. Compression barriers in autoregressive transformers. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2757–2785. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/haris25a.html.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for mdps. In *AI&M*, 2006.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. Snapkv: Llm knows what you are looking for before generation. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

David Lindner, János Kramár, Sebastian Farquhar, Matthew Rahtz, Thomas McGrath, and Vladimir Mikulik. Tracr: compiled transformers as a laboratory for interpretability. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

nostalgebraist. interpreting GPT: the logit lens. August 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Eduardo D. Sontag. For differential equations with r parameters, $2r + 1$ experiments are enough for identification. *Journal of Nonlinear Science*, 12(6):553–583, 2003. doi: 10.1007/s00332-002-0506-0.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *Trans. Sig. Proc.*, 50(6):1417–1428, June 2002. ISSN 1053-587X. doi: 10.1109/TSP.2002.1003065. URL https://doi.org/10.1109/TSP.2002.1003065.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11080–11090. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/weiss21a.html.

Biao Zhang and Rico Sennrich. *Root mean square layer normalization*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2o: heavy-hitter oracle for efficient generative inference of large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

A NOTATION AND OPERATIONAL SPECIFICATION

The discrete type includes token position, remaining generation length, the current program location, and cache lengths. It is not a learned coordinate. All statements comparing states use the same discrete type. A cache group is a storage object; several query heads may read that one object. All matrix ranks in the exact results are algebraic ranks, not thresholded numerical ranks.

Symbol	Meaning
d, M, H	Residual width, vocabulary size, and finite generation horizon.
P, Π	Residual-centering projection and vocabulary-centering projection.
Γ_ι, β_ι, ϵ_ι	Gain, offset, and positive stabilizer at normalization site ι.
f, W_{U}, b_{U}	Final normalization site, vocabulary output projection, and vocabulary bias.
φ_ι	Fixed pointwise activation of the feed-forward sublayer at normalization site ι.
$c \in \mathcal{C}$	A shared key–value cache group.
$\mathcal{H}(c)$	Query heads reading group c, after zero centered-write heads are omitted.
d_c, w_c	Key width and raw value width for group c.
S_c	Joint row space of the centered head-output maps in cache group c.
O_h, B_c	Head output matrix and orthonormal basis of S_c.
s_c, n_c, g_c	Observable-value rank, number of stored slots, and number of distinct keys.
$\vartheta_{c,i}$	Key in original slot i of cache group c.
$\kappa_{c,j}$, $\mu_{c,j}$, $\omega_{c,j}$	Distinct key, its multiplicity, and its projected-value sum.
$\mathcal{R}$, $\mathcal{B}$	Finite families of effective residual reads and centered residual writes.
E, U, r	Constructed residual subspace, its orthonormal basis, and its dimension.
x, u, ρ	Original live residual, retained linear coordinates, and omitted squared norm.
$x_\perp$, $x_{\perp,0}$, $x_{\perp,*}$	Omitted residual component, fixed unit vector used in the chart section, and fixed nonzero vector used in the translation counterexample.
η	Indicator that ρ is a nonconstant coordinate.
G_h	Reduced residual-write matrix $U^{\mathsf{T}} P O_h B_c$.
A, b_f, $\mathcal{L}$	Effective centered terminal matrix, fixed terminal offset, and centered log probabilities.
a, a_{key}	Fixed terminal contrast direction and cache moment direction.
s_{tr}, h_{tr}	Positive translation lengths parallel and orthogonal to the terminal contrast.
$\mathcal{T}$, $\mathcal{T}_{\mathrm{vec}}$	Scalar inverse $z \mapsto z/\sqrt{1-z^2}$ and its vector radial extension, respectively.
ξ, ζ	Original activation state and local coordinates for its constructed state.
T_e, F_e, Φ_p	Original component transition, compiled transition, and state extraction.
Z, N	Cache exponential denominator and vector numerator before insertion.
$\mathcal{E}$	Concatenated centered log probabilities of the finite intervention family.
D, J_0, $\mathfrak{B}$	Local state dimension, intervention Jacobian, and Hessian bound.
$\mathcal{Q}_c$, $V_{q,c}$, t_c	Allowed query subspace, its orthonormal basis, and its dimension.
$\mathcal{H}_{\mathrm{obs}}$, $\mathsf{G}_{\mathrm{obs}}$	Accessible heads and their stacked reduced write maps in an isolated-cache family.
$\mathcal{B}_{\mathrm{ins}}$	Allowed projected values of the single zero-key insertion.
$\Phi_p^{\mathcal{Q}}$, $D_{\mathcal{Q}}$	Extractor and regular-chart dimension for restricted query spaces.
R_{loc}, $\mathcal{D}$	Local approximation radius and deterministic observation decoder.
$\mathscr{I}$, χ, L_H	Local reconstruction, its derivative bound, and a future-observation derivative bound.

A.1 ATOMIC COMPONENT BOUNDARIES

In a sequential block, the attention and feed-forward sublayers are separate atomic transitions. The attention transition reads its incoming normalized residual, computes all queries and shared key/value entries from that input, inserts each shared entry once unless insertion is disabled, and adds the gated head contributions and any separately gated attention output bias. The feed-forward transition evaluates its normalization at its incoming residual and adds its gated residual contributions.

A parallel block is one atomic transition. Its attention and feed-forward branches read the same block input through their respective normalization maps. Each shared cache entry is computed from

that input and inserted once unless insertion is disabled. The attention-head, feed-forward, and constant-bias contributions are combined using their independent gates before the residual is updated. Query and appended-entry replacements act as specified in Section 2.

Residual resets, translations, embedding initialization, and token emission are separate transitions. Resets and translations occur only at the declared state boundaries, not between branch evaluations of a parallel block. These are the component boundaries in Theorem 1. At finer implementation boundaries, any temporary queries, keys, activations, accumulators, or saved block inputs needed for subsequent computation must be included in the state. No state identity is asserted at such finer boundaries without these variables.

At each attention site, a head with zero gate may be skipped; an active head must read at least one entry. Native attention always inserts the current entry before reading. A pure cache insertion with declared key and projected value is also a well-defined state operation, with no residual change. No historical-slot replacement is included. Whole-cache initial states may nevertheless be arbitrary: the domain of a state abstraction and its allowed subsequent intervention operations are different specifications.

The results cover fixed positional linear transformations of queries and keys, provided the stored key is the one used in all later score computations. A causal prefix is then an unordered collection for attention purposes. The original position counter is retained even when keys coincide. A sliding window, a later position-dependent reweighting of old entries, or a slot-addressed intervention would require additional positional information and is outside this representation.

A.2 Native read and write matrices

For clarity, suppose an attention site has normalization index ι. Before positional transformation, its affine reads are

$$q_h^{\mathrm{pre}} = W_{Q,h}\mathsf{N}_\iota(x) + b_{Q,h}, \tag{16}$$

$$\vartheta_c^{\mathrm{pre}} = W_{K,c}\mathsf{N}_\iota(x) + b_{K,c}, \tag{17}$$

$$v_c^{\mathrm{new}} = W_{V,c}\mathsf{N}_\iota(x) + b_{V,c}. \tag{18}$$

The value read included in $\mathcal{R}$ is $B_c^{\mathsf{T}} W_{V,c}\Gamma_\iota P$, not the entire raw value read. The query and key reads are $W_{Q,h}\Gamma_\iota P$ and $W_{K,c}\Gamma_\iota P$. Omitted zero-write heads need not contribute query reads. The final matrix $\Pi W_{\mathrm{U}}\Gamma_f P$ is included directly.

Let φ_ι denote the fixed pointwise activation in the feed-forward sublayer at normalization site ι. For an ordinary feed-forward layer, include $W_1\Gamma_\iota P$ among the reads and PW_2 among the writes, where its residual contribution is $W_2\varphi_\iota(W_1\mathsf{N}_\iota(x) + b_1) + b_2$. A gated layer additionally includes the corresponding gate-read matrix. Centered output biases such as Pb_2 and an attention output bias are included among the write columns. The constant norm offsets enter the affine read constants and do not add residual directions by themselves.

The vocabulary readout is treated modulo a common logit shift. Thus a direction whose only effect is to add a scalar multiple of $\mathbf{1}_M$ does not enter the terminal effective read. This is necessary when the target observation is a probability vector rather than an absolute logit vector.

B Proof and implementation of the exact realization

Lemma 9 (Two elementary subspace identities)**.** *For every retained cache group c and every head $h \in \mathcal{H}(c)$, $PO_h = PO_h B_c B_c^{\mathsf{T}}$. Every effective read $R_0 \in \mathcal{R}$ satisfies $R_0(P - UU^{\mathsf{T}}) = 0$, and every projected write has image contained in E. The vertically stacked matrix $\mathsf{G}_c = (G_h)_{h\in\mathcal{H}(c)}$ has column rank s_c.*

Proof. The columns of B_c span the sum of the row spaces of the matrices PO_h. Orthogonal projection onto that subspace therefore leaves each of those matrices unchanged on the right. The identities for effective reads and projected writes follow from the definition of E and $E \subseteq \operatorname{ran} P$.

If $\mathsf{G}_c z = 0$, then $U^{\mathsf{T}} PO_h B_c z = 0$ for every head. Since $PO_h B_c z \in E$, it follows that $PO_h B_c z = 0$ for every head. Hence $B_c z$ belongs both to the common kernel of the stacked matrices PO_h and to

their joint row space. These two spaces are orthogonal, so $B_c z = 0$. Orthonormality gives $z = 0$. This proves full column rank of G_c. □

Proof of Theorem 1. Write $x = Uu + x_\perp + x_{\mathrm{mean}}$, where $x_\perp = (P - UU^\mathsf{T})x$ and $x_{\mathrm{mean}} = (\mathsf{I}_d - P)x$. These terms are mutually orthogonal. The last term vanishes for RMS normalization and is a scalar mean component for centered normalization. Therefore

$$Px = Uu + x_\perp,\ \|Px\|_2^2 = \|u\|_2^2 + \|x_\perp\|_2^2 . \tag{19}$$

For an affine normalized read, Lemma 9 gives $R\Gamma_\iota x_\perp = 0$, because the corresponding effective read is $R\Gamma_\iota P$. Substituting (19) in the normalizer proves (5). In particular, the compiled queries, keys, and projected values agree exactly with their original counterparts. Fixed positional transformations preserve this equality.

For a shared group, the native projected value is

$$\nu_c^{\mathrm{new}} = (B_c^\mathsf{T} W_{V,c} \Gamma_\iota U)\frac{u}{\sigma_\iota(u,\rho)} + B_c^\mathsf{T}(W_{V,c}\beta_\iota + b_{V,c}). \tag{20}$$

It is inserted only once, regardless of the number of query heads sharing that group. If the key is new, it contributes one exponential term to the denominator and its projected value to the corresponding numerator. If the key already occurs, increasing μ by one and ω by ν_c^{new} produces exactly the same two changes. The same reasoning applies to an externally replaced appended entry. Any projected value b_{ins} can be implemented by the raw value $B_c b_{\mathrm{ins}}$.

For any head, expansion of the grouped sums gives

$$G_h \frac{\sum_j e^{q^\mathsf{T}\kappa_{c,j}}\omega_{c,j}}{\sum_j \mu_{c,j} e^{q^\mathsf{T}\kappa_{c,j}}} = U^\mathsf{T} P O_h \frac{\sum_{i=1}^{n_c} e^{q^\mathsf{T}\vartheta_{c,i}} v_{c,i}}{\sum_{i=1}^{n_c} e^{q^\mathsf{T}\vartheta_{c,i}}}. \tag{21}$$

The sums on the right are over original slots and, at a native attention read, include the current inserted slot. Thus the compiled attention residual increment is the projection of the exact original increment, and every centered increment belongs to E. Consequently $x_\perp$ is unchanged by an attention residual write, even under independent head gates or constant-bias gates. The same is true of a feed-forward write, since its affine reads have already been shown equal and its centered output matrix and bias lie in E. Its fixed nonlinearities and any pointwise products therefore receive identical arguments.

For example, an ordinary feed-forward contribution compiles to

$$\begin{aligned} u' &= u + U^\mathsf{T} P W_2 \varphi_\iota\left((W_1\Gamma_\iota U)\frac{u}{\sigma_\iota(u,\rho)} + W_1\beta_\iota + b_1\right) + U^\mathsf{T} P b_2, \\ \rho' &= \rho. \end{aligned} \tag{22}$$

with each gate applied to its corresponding contribution. A gated feed-forward layer uses its two precontracted affine reads in the original coordinatewise product. These formulas never require Uu to be formed online.

For a parallel block, evaluate (5) separately at the attention and feed-forward normalization sites using the same incoming (u, ρ). Both branches therefore receive their original affine-read values, and the attention branch makes the same once-per-group cache update, subject to the declared controls. Let Δu_{attn} and Δu_{ff} denote the reduced residual increments of the two branches, including their separately gated output-bias terms. The whole-block update is

$$u' = u + \Delta u_{\mathrm{attn}} + \Delta u_{\mathrm{ff}},\ \rho' = \rho. \tag{23}$$

Each increment is the projection of the corresponding original branch contribution. Both centered original contributions lie in E, so their sum leaves $x_\perp$ unchanged. Thus (7) holds at the parallel-block boundaries without applying the feed-forward read to the attention-updated residual. The independent branch gates still permit the head-isolation interventions used in Section 4.

Residual translation by $U\delta$ preserves $x_\perp$ and changes u by δ. Because the admissible intervention class includes whole-residual resets, an intervention that resets the residual to a declared vector x_* assigns the reduced state $(U^\mathsf{T}x_*, \|(P - UU^\mathsf{T})x_*\|_2^2)$ and is therefore well defined in the reduced state. The same assignment applies to every new token embedding, including a fixed additive

positional embedding when present. Query replacements and insertion/write controls act on already retained variables or on declared constants. Cache multiplicities and the discrete position counter evolve according to the prescribed original operations.

Finally, centering the final logits eliminates their common scalar component and gives precisely $\mathcal{L} = Au/\sigma_f + b_f$. Because $\mathrm{softmax}(\Pi\ell) = \mathrm{softmax}(\ell)$ for every logit vector ℓ, the terminal probabilities agree. The component identities follow by induction over the operational specification in Appendix A.

For an arbitrary fixed token prefix, both systems have the same reduced state and next-token conditional probability. Multiplying these equal conditional probabilities proves equality for each finite token string. A policy depending only on previous tokens and declared controls chooses the same next control on equal prefixes, so the argument includes such policies. Randomized external policies are covered by conditioning on their shared randomness and then integrating. Early stopping can be padded with a deterministic terminal symbol. □

B.1 Constant terminal readout

The identity $\Pi W_{\mathrm{U}} \Gamma_f P = AU^{\mathsf{T}}$ follows because the row space of the left-hand side lies in E. Thus $A = 0$ if and only if the terminal probability map is the constant $\mathrm{softmax}(b_f)$. Every admissible execution ends at this same normalization and readout, so all states of a fixed discrete type have the same observations under every admissible continuation. The empty behavioral state is sufficient, even when internal reads or writes make $r > 0$. When $r = 0$, the readout is necessarily constant, which justifies the convention in Section 3. For $r > 0$ the constructed component state remains sufficient by Theorem 1, but need not be minimal when $A = 0$.

B.2 What the compilation computes

One may obtain B_c and U by rank-revealing linear algebra on finite matrices. With rational checkpoint entries, exact elimination is an alternative to numerical rank estimation. After choosing the bases, precompute every matrix in (5), (20), (22), and (6). This process does not require a prompt distribution, a guessed symbolic variable, a trained auxiliary network, or an oracle for the remaining decoder.

On a regular cache stratum the retained real storage is

$$r + \eta + \sum_c g_c(d_c + s_c), \tag{24}$$

plus discrete multiplicities, phase counters, and ordinary temporary registers. A normalized affine read with d_{read} output coordinates takes a matrix–vector product of shape $d_{\mathrm{read}} \times r$, rather than $d_{\mathrm{read}} \times d$. A query head over g_c grouped entries uses $O(g_c d_c + g_c s_c)$ scalar arithmetic for scores and weighted value sums, followed by its reduced output map. Feed-forward intermediate widths are not claimed to shrink. Producing all vocabulary probabilities still requires the vocabulary readout.

When $r = d$ and $s_c = w_c$ with distinct keys, the RMS-normalized realization retains the native continuous state dimension, so the construction yields no lossless dimensional reduction in this regime. The identification theorem may nevertheless rule out a smaller continuous state for the declared observations.

For the basis-change statement in the main text, substituting UQ_E and $B_c Q_c$ gives $G'_h = Q_E^{\mathsf{T}} G_h Q_c$, $A' = AQ_E$, and the stated coordinate transformations. The projector UU^{T} is unchanged. Substitution in the compiled read, attention, and feed-forward formulas proves the conjugacy of updates. This is ordinary change-of-basis invariance, not an additional identifiability claim for semantic coordinates.

C Finite reconstruction and proof of behavioral equivalence

C.1 Recovering the residual statistics

If $A \neq 0$ and every pair of its rows were equal, vocabulary centering would force every row to be zero. Hence the pair j_+, j_- in Section 4 exists and depends only on the weights; $\mathbf{e}_j$ denotes the jth standard basis vector of $\mathbb{R}^M$. Subtracting its fixed bias contrast from the observed log probability ratio gives (8). No inverse of A is used.

Lemma 10 (Recovery from a fixed token pair)**.** *Under Assumption 2, the $2r$ translations specified in Section 4 identify (u, ρ) on $\rho \geq 0$, with ρ omitted when $\eta = 0$. At every exact observation their reconstruction extends to a smooth function on an open observation neighborhood.*

Proof. Write $u = ta + v$ with $a^\top v = 0$, and put $c_f = d\epsilon_f > 0$. For every $b \in \mathcal{B}_\perp$, orthogonality gives

$$y_b = \frac{t}{\sqrt{t^2 + \|v + b\|_2^2 + \rho + c_f}}, \; y_{b+s_{\mathrm{tr}}a} = \frac{t + s_{\mathrm{tr}}}{\sqrt{(t + s_{\mathrm{tr}})^2 + \|v + b\|_2^2 + \rho + c_f}}. \tag{25}$$

Both scalars lie strictly between -1 and 1. Applying $\mathcal{T}$ yields t/R_b and $(t + s_{\mathrm{tr}})/R_b$, respectively. Their difference is $s_{\mathrm{tr}}/R_b > 0$, proving (9). Taking squared differences then gives the explicit inverse

$$\begin{aligned} t &= R_0 \mathcal{T}(y_0), \quad v = \sum_{i=2}^{r} \frac{R_{h_{\mathrm{tr}} e_i}^2 - R_0^2 - h_{\mathrm{tr}}^2}{2h_{\mathrm{tr}}} \, e_i, \\ u &= ta + v, \qquad \rho = R_0^2 - \|v\|_2^2 - c_f. \end{aligned} \tag{26}$$

For $r = 1$ the sum is empty. These identities hold also when $t = 0$, $t = -s_{\mathrm{tr}}$, or $\rho = 0$; no division by an observed numerator is required. All square-root arguments and differences used in the inverse are strictly positive at an exact observation. They remain so on a sufficiently small open neighborhood, on which the displayed reconstruction is smooth. An extension off the exact image need not return a nonnegative ρ; this does not affect its left-inverse identity on actual states. When $\eta = 0$, simply omit the reconstructed last coordinate. The translations are fixed in advance and have norm at most $\sqrt{s_{\mathrm{tr}}^2 + h_{\mathrm{tr}}^2}$, so they also fit any prescribed positive translation bound after choosing $s_{\mathrm{tr}}, h_{\mathrm{tr}}$ sufficiently small. □

Unlike the generic finite-experiment existence result of Sontag (2003), Lemma 10 gives a fixed token-pair measurement family together with an explicit smooth inverse for the retained residual statistics. We make no optimality claim for the number of passes.

C.2 The full-column-rank special case

When A has column rank r, the full output vector permits a shorter inverse. In a residual-bypass intervention, set

$$\bar{y} = \frac{A^\dagger(\mathcal{L} - b_f)}{\sqrt{d}} = \frac{u}{\sqrt{\|u\|_2^2 + \rho + d\epsilon_f}}, \; \mathcal{T}_{\mathrm{vec}}(\bar{y}) = \frac{\bar{y}}{\sqrt{1 - \|\bar{y}\|_2^2}}. \tag{27}$$

Here $\mathcal{T}_{\mathrm{vec}}$ is the vector radial inverse, distinct from the scalar function $\mathcal{T}$. If $\eta = 0$, one pass recovers $u = \sqrt{d\epsilon_f}\, \mathcal{T}_{\mathrm{vec}}(\bar{y})$. Otherwise, put $\tau = \rho + d\epsilon_f$ and use one additional fixed translation $\delta \neq 0$. Then $d_y = \mathcal{T}_{\mathrm{vec}}(\bar{y}_\delta) - \mathcal{T}_{\mathrm{vec}}(\bar{y}) = \delta/\sqrt{\tau}$, giving

$$\sqrt{\tau} = \frac{\|\delta\|_2^2}{\delta^\top d_y}, \; u = \sqrt{\tau}\, \mathcal{T}_{\mathrm{vec}}(\bar{y}), \; \rho = \tau - d\epsilon_f. \tag{28}$$

The denominator is strictly positive. This special case is used in Appendix I.3; the fixed-token-pair reconstruction in Appendix I.4 instead uses Lemma 10.

C.3 Recovering an isolated head write from final probabilities

Fix the target head, effective query, appended key, and appended projected value. For each of the $2r$ translations δ, start from a fresh copy of the same token-entry state and reset the residual to $U\delta$. Bypass every residual contribution except the selected head, including constant output biases, and suppress irrelevant cache writes. The target query and appended entry are replaced by their fixed values, so its reduced write w depends only on the original historical cache and these constants, not on δ. No earlier branch changes the residual or that historical cache. After the head, bypass all remaining contributions. The final residual therefore has reduced coordinates $w + \delta$ and omitted squared norm zero.

Lemma 10 recovers w from these $2r$ observations of the same token pair. This also covers $Aw = 0$: the translated denominators still reveal the components perpendicular to the chosen contrast. A centered-normalization head may write a residual mean, but P removes it from every later normalization. The argument applies to parallel blocks because the target query and appended entry are fixed and all other branch writes are gated off; no intervention between parallel branch evaluations is needed.

When A has full column rank, a single zero-reset pass instead gives the shorter formula

$$w = \sqrt{d\epsilon_f}\,\mathcal{T}_{\mathrm{vec}}\left(\frac{A^{\dagger}(\mathcal{L} - b_f)}{\sqrt{d}}\right). \tag{29}$$

In both constructions the final normalization and vocabulary readout remain unchanged. The intervened suffix is bypassed by the declared gates; no claim of injectivity is made for an arbitrary native suffix.

C.4 Recovering the numerator and denominator for a shared cache

For one group, choose h_0 and b_{ins} such that $g_{\mathrm{ins}} = G_{h_0} b_{\mathrm{ins}} \neq 0$. Such a choice exists because the group is retained. Denote by $w_h^{(0)}(q)$ the write recovered from a zero-key, zero-value insertion in head h, and by $w_{h_0}^{(\mathrm{ins})}(q)$ the write recovered when the appended projected value is b_{ins}. Then

$$\alpha(q) = \frac{g_{\mathrm{ins}}^{\mathsf{T}}\big(w_{h_0}^{(\mathrm{ins})}(q) - w_{h_0}^{(0)}(q)\big)}{\|g_{\mathrm{ins}}\|_2^2} = \frac{1}{Z(q)+1}. \tag{30}$$

With $w^{(0)}(q)$ the vertical stack of $w_h^{(0)}(q)$, full column rank of G_c yields

$$Z(q) = \alpha(q)^{-1} - 1,\ N(q) = \alpha(q)^{-1}\mathsf{G}_c^{\dagger} w^{(0)}(q). \tag{31}$$

The same stored cache is used in every intervention; each intervention starts from a fresh copy and inserts just its declared new entry. Each query requires one zero-value head-write reconstruction per active head and one additional nonzero-value reconstruction for the reference head. Each reconstruction uses $2r$ passes in the general case, or one pass in the full-column-rank special case.

C.5 A finite separating direction library

For $n = n_c > 0$ and key width d_c, let $J = 1 + (d_c - 1)\binom{n}{2}$. Choose J distinct real numbers $t_1, \ldots, t_J$ in $[0, 1]$ and directions

$$a_{\mathrm{key},\ell} = \gamma_0(1, t_\ell, \ldots, t_\ell^{d_c-1}),\ 1 \le \ell \le J, \tag{32}$$

where $\gamma_0 > 0$ is fixed. A nonzero difference of two keys gives a nonzero polynomial in t of degree at most $d_c - 1$. It can vanish at no more than $d_c - 1$ of the chosen points. There are at most $\binom{n}{2}$ pairs of distinct keys. A union bound on these finite sets proves that one $a_{\mathrm{key},\ell}$ separates every pair. For $d_c = 1$, one nonzero direction suffices.

Let $e_{\mathrm{key},i}$ be the ith standard basis vector of $\mathbb{R}^{d_c}$. For each direction include all queries

$$\begin{aligned} &\{m a_{\mathrm{key},\ell} : m \in \{0, \ldots, 2n-1\}\} \\ &\cup \{m a_{\mathrm{key},\ell} + \gamma e_{\mathrm{key},i} : m \in \{0, \ldots, n-1\},\ i \in \{1, \ldots, d_c\}\}. \end{aligned} \tag{33}$$

There are at most $n(d_c + 2)$ queries per direction. If the permitted query ball has radius $R_q > 0$, it is enough, for example, to choose $2n\gamma_0\sqrt{d_c} + \gamma < R_q$. The choices are independent of the actual cache entries. The library is therefore fixed for the discrete type and does not require searching over prompts.

C.6 Moment inversion, including duplicate keys

Suppress the group index and write a_{key} for one library direction. If some distinct keys have equal projection onto this direction, group together keys with the same projection. Let g_{proj} be the number of distinct projected nodes, write them as $\lambda_1, \ldots, \lambda_{g_{\mathrm{proj}}} > 0$, and let their positive multiplicities

be $\widetilde{\mu}_j$. The moments are $s_m = Z(ma_{\text{key}}) = \sum_j \widetilde{\mu}_j \lambda_j^m$, with $m \in \{0, \dots, 2n-1\}$. For each $m_{\text{H}} \in \{1, \dots, n\}$, form

$$\mathsf{H}_{m_{\text{H}}} = (s_{i+j})_{0 \le i,j < m_{\text{H}}}. \tag{34}$$

When $m_{\text{H}} \ge g_{\text{proj}}$, its factorization as a Vandermonde matrix times $\operatorname{diag}(\widetilde{\mu}_1, \dots, \widetilde{\mu}_{g_{\text{proj}}})$ times the transpose shows that its rank is g_{proj}. Positivity of the multiplicities is crucial for this rank assertion. Thus $\operatorname{rank} \mathsf{H}_n = g_{\text{proj}}$ exactly. Select a direction maximizing this rank. The separating-direction argument proves that the selected rank is g_c and that its nodes correspond to genuinely distinct keys, not to projection collisions.

Put $g = g_c$ for that direction. The leading matrix H_g is positive definite. Solve

$$\mathsf{H}_g \begin{bmatrix} c_0 & \cdots & c_{g-1} \end{bmatrix}^{\mathsf{T}} = - \begin{bmatrix} s_g & \cdots & s_{2g-1} \end{bmatrix}^{\mathsf{T}}. \tag{35}$$

For $p_*(X) = X^g + \sum_{i=0}^{g-1} c_i X^i$, this equation is equivalent to $\sum_j \mu_j \lambda_j^m p_*(\lambda_j) = 0$ for $m \in \{0, \dots, g-1\}$. The Vandermonde matrix and the positive diagonal multiplicity matrix are invertible, so $p_*(\lambda_j) = 0$ for every j. The degree and the number of distinct nodes are both g, hence these are exactly its roots and all are simple.

Choose the increasing order of the positive roots and let $\mathsf{V} = (\lambda_j^m)_{0 \le m < g,\, 1 \le j \le g}$. Recover multiplicities and key coordinates from

$$(\mu_j)_{j=1}^{g} = \mathsf{V}^{-1}(s_m)_{m=0}^{g-1}, \tag{36}$$

$$z^{(i)} = \mathsf{V}^{-1}\big(Z(ma_{\text{key}} + \gamma e_{\text{key},i})\big)_{m=0}^{g-1}, \tag{37}$$

$$\kappa_{j,i} = \gamma^{-1} \log\big(z_j^{(i)} / \mu_j\big). \tag{38}$$

Every logarithm has a positive argument on exact observations. Inverting V separately in each coordinate of $\big(N(ma_{\text{key}})\big)_{m=0}^{g-1}$ recovers the vectors ω_j. These are the sums over identical original keys, not the individual raw values. The multiplicities recovered from an exact original cache are positive integers and sum to n.

This procedure uses moment values up to $2n-1$. The polynomial degree is the number of distinct keys rather than the number of slots, so duplicate keys do not require treating repeated roots as simple ones. When $n = 0$, the cache is empty and no reconstruction step is necessary. Polynomial root operations are an exact algebraic specification in the ideal observation model; numerical recovery to finite precision also depends on the conditioning discussed below. No finite-bit exact-recovery claim is made.

Proof of Theorem 3. Lemma 10 uses $2r$ passes to recover the live residual statistics. For each nonempty group, the fixed query library has at most $J_c n_c (d_c + 2)$ queries. At most $2r(|\mathcal{H}(c)| + 1)$ head-isolation passes per query recover its Z and N, and the finite reconstruction above identifies its cache triples. The same token pair is used throughout. This gives (10). When A has full column rank and full probability vectors are recorded, Appendix C.2 reduces the count to $1 + \eta + \sum_{c: n_c > 0} (|\mathcal{H}(c)| + 1) J_c n_c (d_c + 2)$.

If two extracted states agree, Theorem 1 gives identical outputs under every admissible controlled continuation, including all finite generation distributions. Conversely, suppose two original states have equal observations for every admissible continuation of length at most $H \ge 1$. They therefore have equal observations for the finite single-pass intervention library. Its deterministic reconstruction recovers the same residual statistics and cache triples from both. Hence their extracted states agree, up to the irrelevant ordering of distinct-key triples. This proves both implications. □

D Dimension minimality and its domain

D.1 A regular state chart

Fix positive multiplicities summing to the prescribed cache length in each group. For each group choose a separating direction and restrict to an open set where the order of the projected distinct keys is fixed. The coordinates consist of $u \in \mathbb{R}^r$, the scalar $\rho > 0$ when present, and one key and

one value-sum vector for each distinct key. They form an open set in $\mathbb{R}^D$, with D as in (12). A key collision lies outside this chart.

There is a smooth section from this chart to original activation states. If $\eta = 1$, fix a unit vector $x_{\perp,0} \in \operatorname{ran} P \cap E^{\perp}$ and set

$$x = Uu + \sqrt{\rho}\, x_{\perp,0}. \tag{39}$$

If $\eta = 0$, set $x = Uu$. Fix a partition of the original cache slots into groups of sizes $\mu_{c,j}$. Fill every slot in part j with key $\kappa_{c,j}$ and raw value $B_c\omega_{c,j}/\mu_{c,j}$. These values sum to the required projected value. Unobservable value components and discarded groups can be fixed to zero. Thus every chart point is represented by an admissible original activation state, and the section is continuous.

Proof of Theorem 6. Let S denote this section and ψ the proposed continuous original-state encoder into $\mathbb{R}^k$. If $\psi(S(\zeta)) = \psi(S(\zeta'))$, every deterministic reconstruction map applied to that code produces the same observation. Exact faithfulness for the finite intervention library therefore gives $\mathcal{E}(\zeta) = \mathcal{E}(\zeta')$. Theorem 3 and the fixed key order imply $\zeta = \zeta'$. Hence $\psi \circ S$ is a continuous injection from an open subset of $\mathbb{R}^D$ into $\mathbb{R}^k$.

For completeness, the required standard consequence of invariance of domain is the following. If $k < D$ and such an injection existed, appending $D - k$ zero coordinates would give a continuous injective map into $\mathbb{R}^D$. Invariance of domain would make its image open in $\mathbb{R}^D$. The image is contained in a proper coordinate subspace, which has empty interior, a contradiction. Thus $k \geq D$.

The constructed cache triples and residual statistics themselves supply D coordinates on the fixed chart. Their transitions are already specified, with a change of type and chart when a new key or a collision changes the stratum. This establishes the matching local upper bound without claiming a single global Euclidean coordinate system for all multisets. □

D.2 What the lower bound does and does not exclude

The lower bound applies to continuous encoders on an open regular chart and permits arbitrary deterministic decoders. It does not apply to discontinuous encodings of arbitrary real precision or to finite sets of prompt states, which need not contain an open D-dimensional subset. Additional finite discrete tags do not evade the bound locally, since any continuous map into a discrete space is locally constant. At exact key collisions, only grouped value sums are required, so the corresponding stratum may have smaller dimension than nearby strata in which the keys separate.

D.3 Proof of the affine-state bound

Proof of Theorem 7. Vectorize the original live residual and all stored key and raw-value slots, counting a shared group once. Write an affine encoder as $L\xi + b_0$, and take an arbitrary direction v in $\ker L$. Choose a point ξ_* in the nonempty open domain. For sufficiently small real t, the segment $\xi_* + tv$ remains in that domain and all keys remain distinct. The encoder is constant along this segment, so exact reproduction of the finite library and Theorem 3 force the extracted state to be constant as well.

Distinct keys can be labeled consistently on this small segment. Their equality as unordered collections implies that each labeled key is fixed: a continuous permutation of a finite set of separated keys cannot change locally. Thus every key component of v is zero. Every projected value component is also zero, since each key has multiplicity one. If v_x is the live-residual component of v, residual-coordinate constancy gives $U^{\mathsf{T}} v_x = 0$. If $\eta = 1$, the polynomial

$$t \longmapsto \left\| (P - UU^{\mathsf{T}})(x_* + tv_x) \right\|_2^2 \tag{40}$$

is constant on an interval. Its quadratic coefficient is $\left\|(P - UU^{\mathsf{T}})v_x\right\|_2^2$, so that vector must be zero. When $\eta = 0$ it is already zero identically. Combining these statements gives $Pv_x = 0$.

Therefore $\ker L$ is contained in the kernel of the linear map that retains Px, every key, and every $B_c^{\mathsf{T}} v_{c,i}$. The latter map has rank D_{aff}, since its blocks act on separate raw state components and B_c has rank s_c. Rank–nullity implies $\operatorname{rank} L \geq D_{\mathrm{aff}}$. Conversely, retaining exactly these linear coordinates permits computation of u and ρ and of every cache triple, hence is sufficient by Theorem 1. This proves attainability and (13). □

The reduction can be attributed to architectural sources. Let $D_{\mathrm{raw}} = d + \sum_c n_c(d_c + w_c)$ be the original continuous storage after output-null groups are removed, counting each shared cache once. On a fixed-multiplicity chart,

$$D_{\mathrm{raw}} - D = (d - \operatorname{rank} P) + (\operatorname{rank} P - r - \eta) + \sum_c \big[n_c(w_c - s_c) + (n_c - g_c)(d_c + s_c)\big]. \tag{41}$$

The terms are the removed residual mean, the reduction of an unread residual complement to its norm, the discarded value kernels, and the merged equal-key entries; all are nonnegative. The identity shows when no exact reduction occurs, namely when every term vanishes, and it prevents shared keys and values from being counted once per query head. When $r = \operatorname{rank} P$, $s_c = w_c$, and $g_c = n_c$ for every retained group, all terms except the residual-mean term vanish.

In (41), D_{raw} is understood after entirely output-null cache groups are removed; if such groups are retained in a raw implementation, their full storage is an additional removable term. The budget follows by expanding its right-hand side: $n_c(w_c - s_c) + (n_c - g_c)(d_c + s_c) = n_c(d_c + w_c) - g_c(d_c + s_c)$. Since η equals one exactly when $\operatorname{rank} P - r > 0$, every residual term is nonnegative. The formula is a dimension identity on the specified stratum, not a bit-storage or parameter-count theorem.

E Proof of the dimension–error relation

The finite intervention maps used here bypass the feed-forward sublayers. Their remaining operations are affine maps, exponentials, positive denominators, final normalization with positive stabilizer, and softmax. Centered log probabilities eliminate the log-softmax normalizer. Thus $\mathcal{E}$ is real analytic on each regular chart. In particular it is C^2 on a neighborhood of every sufficiently small closed chart ball. The C^2 hypothesis in the main text is automatic for these interventions; it is stated explicitly to define the approximation argument.

Under Assumption 2, work on a neighborhood of a fixed regular state and retain a particular separating direction and the fixed numbers of distinct keys. The inverse in Appendix C uses strictly positive paired translation contrasts, nonsingular Hankel matrices, and polynomials with simple positive roots. These conditions persist near the exact observations. Together with Lemma 10 and (26) this gives a smooth local reconstruction $\mathscr{I}$ satisfying

$$\mathscr{I}(\mathcal{E}(\zeta)) = \zeta. \tag{42}$$

Differentiation at ζ_0 gives $D\mathscr{I}(\mathcal{E}(\zeta_0))J_0 = \mathsf{I}_D$, so J_0 has column rank D. The extra interventions in the full library do not harm this conclusion; the reconstruction may simply ignore them.

Proof of Theorem 8. The Hessian bound and Taylor's theorem give

$$\mathcal{E}(\zeta_0 + v) = \mathcal{E}(\zeta_0) + J_0 v + R(v),\ \|R(v)\|_2 \le \tfrac{1}{2}\mathfrak{B}\,\|v\|_2^2\,. \tag{43}$$

Choose a singular-value decomposition of J_0, with right singular vectors $v_1, \dots, v_D$ ordered by decreasing singular values. For the upper bound, let V_k contain the first k vectors and use

$$\psi(v) = V_k^{\mathsf{T}} v,\ \mathcal{D}(t) = \mathcal{E}(\zeta_0) + J_0 V_k t. \tag{44}$$

When $k = 0$, this denotes the constant decoder. The residual linear term has operator norm $\sigma_{k+1}(J_0)$, so (43) gives the upper bound in (14).

For any continuous competing encoder ψ, restrict it to the radius-R_{loc} sphere in $\operatorname{span}(v_1, \dots, v_{k+1})$. Borsuk–Ulam supplies antipodal points v_* and $-v_*$ with equal codes. The case $k = 0$ is immediate, since $\mathbb{R}^0$ consists of a single point. Any deterministic decoder therefore returns the same vector b_* on these two inputs. By the triangle inequality,

$$\begin{aligned}\max\{\|\mathcal{E}(\zeta_0 + v_*) - b_*\|_2\,, \|\mathcal{E}(\zeta_0 - v_*) - b_*\|_2\} &\ge \tfrac{1}{2}\,\|\mathcal{E}(\zeta_0 + v_*) - \mathcal{E}(\zeta_0 - v_*)\|_2 \\ &\ge \|J_0 v_*\|_2 - \tfrac{1}{2}\mathfrak{B} R_{\mathrm{loc}}^2 \\ &\ge R_{\mathrm{loc}}\sigma_{k+1}(J_0) - \tfrac{1}{2}\mathfrak{B} R_{\mathrm{loc}}^2. \end{aligned} \tag{45}$$

Taking the infimum over encoders and decoders, and using nonnegativity of error, proves the lower bound. Dividing by R_{loc} and taking $R_{\mathrm{loc}} \downarrow 0$ gives the limit. Positivity follows from the full column rank of J_0 established above under Assumption 2. □

Corollary 11 (Exact affine observation width)**.** *For a linear observation map $K : \mathbb{R}^D \to \mathbb{R}^{d_{\mathrm{obs}}}$, continuous encoding into $\mathbb{R}^k$, arbitrary deterministic decoding, and a closed Euclidean ball of radius R_{loc}, the optimal worst-case Euclidean error equals $R_{\mathrm{loc}}\sigma_{k+1}(K)$, with singular values beyond the rank taken as zero.*

Proof. The preceding proof applies with zero Taylor remainder. Truncated singular-value encoding attains this bound; if k is at least the rank, the observation itself factors exactly through that many coordinates. □

The ball assumption is essential. On a finite dataset, or a nonconvex set of smaller intrinsic dimension, the antipodal argument may not apply. The continuity condition concerns the encoder, while the decoder in the lower bound need not be continuous. This is a worst-case reconstruction statement for the finite intervention vector, not a distribution-dependent accuracy guarantee for natural language.

F Local stability of the finite reconstruction

Under Assumption 2, we make explicit why a local inverse exists and what may make its norm large. Fix a regular chart and a separating direction for each nonempty group at a reference state. The observation neighborhood can be chosen small enough that all of the following hold: the scalar arguments in $\mathcal{T}$ remain strictly between -1 and 1; every paired translation contrast in (9) remains positive; the appended-value contrast remains positive; the selected Hankel matrices remain invertible; their annihilating polynomials have distinct positive real roots in the same order; and the recovered multiplicities and offset coefficients remain positive. These are strict conditions at the reference observation and therefore persist on a neighborhood.

The scalar reconstruction is a composition of explicit differentiable operations. With $\Delta_b = \mathcal{T}(y_{b+s_{\mathrm{tr}}a}) - \mathcal{T}(y_b)$, its first derivatives satisfy

$$\mathcal{T}'(z) = (1-z^2)^{-3/2},\ dR_b = -\frac{s_{\mathrm{tr}}}{\Delta_b^2}\left[\mathcal{T}'(y_{b+s_{\mathrm{tr}}a})\,dy_{b+s_{\mathrm{tr}}a} - \mathcal{T}'(y_b)\,dy_b\right]. \tag{46}$$

The map from a centered observation to y_δ has norm $\sqrt{2}/(\sqrt{d}\,\|a_0\|_2)$. Equation (26) further contributes factors h_{tr}^{-1} and the recovered radii and residual coordinates. These expressions explain why a nonzero readout and positive translation scales ensure local smoothness without providing uniform conditioning.

For the full-column-rank special case of Appendix C.2, the vector radial inverse has derivative

$$D\mathcal{T}_{\mathrm{vec}}(y) = \frac{\mathsf{I}_r}{\sqrt{1-\|y\|_2^2}} + \frac{yy^{\mathsf{T}}}{(1-\|y\|_2^2)^{3/2}},\ \|D\mathcal{T}_{\mathrm{vec}}(y)\|_2 \le (1-\|y\|_2^2)^{-3/2}. \tag{47}$$

When A has full column rank and $\eta = 0$, the complete residual inverse is $u = \sqrt{d\epsilon_f}\,\mathcal{T}_{\mathrm{vec}}(y)$ with $y = A^\dagger(\mathcal{L} - b_f)/\sqrt{d}$. Hence

$$\begin{aligned} D_{\mathcal{L}}u &= \sqrt{\epsilon_f}\,D\mathcal{T}_{\mathrm{vec}}(y)A^\dagger, \\ \|D_{\mathcal{L}}u\|_2 &\le \sqrt{\epsilon_f}\,(1-\|y\|_2^2)^{-3/2}\left\|A^\dagger\right\|_2. \end{aligned} \tag{48}$$

The derivative of $\mathcal{T}_{\mathrm{vec}}$ alone should therefore not be interpreted as the sensitivity of the complete residual reconstruction. In this special case the terminal inverse contributes $\left\|A^\dagger\right\|_2 = \sigma_r(A)^{-1}$. For $\epsilon_f = 10^{-5}$ and $\|u\|_2^2/d$ of order one, the norm bound in (48) contains a factor of order 10^5 multiplying $\left\|A^\dagger\right\|_2$. In either reconstruction, the appended-value contrast in (30) has a fixed coefficient norm proportional to $\|G_{h_0}b_{\mathrm{ins}}\|_2^{-1}$, and the derivative of α^{-1} has magnitude α^{-2}. The recovery of the shared numerator also contributes $\left\|\mathsf{G}_c^\dagger\right\|_2$.

For a linear system $Lz = b$, differentiation gives

$$dz = L^{-1}(db - (dL)z). \tag{49}$$

This applies to the Hankel system and the Vandermonde systems. If $p_*(X) = X^g + \sum_{i=0}^{g-1} c_i X^i$ has a simple root λ_j, its coefficient derivative is

$$d\lambda_j = -\frac{\sum_{i=0}^{g-1} \lambda_j^i \, dc_i}{p'_*(\lambda_j)}, \ |p'_*(\lambda_j)| = \prod_{l \neq j} |\lambda_j - \lambda_l|. \tag{50}$$

Key recovery in (38) adds the positive reciprocals of the offset coefficients and multiplicities, and the factor γ^{-1}. These formulas also apply on a fixed duplicate-key stratum, since its annihilating polynomial has one simple root per distinct key. A chart with fixed multiplicities need not remain identifiable with the same dimension if arbitrary perturbations are allowed to split those keys.

The fixed query ball also changes how moment magnitudes should be assessed. For a nonempty group, suppose $\|\kappa_j\|_2 \leq K_{\max}$ on the state set under consideration and $\|q\|_2 \leq R_q$. Since the multiplicities sum to n_c,

$$n_c e^{-R_q K_{\max}} \leq Z(q) \leq n_c e^{R_q K_{\max}}. \tag{51}$$

The ratio of the largest to the smallest denominator on this ball is therefore at most $e^{2R_q K_{\max}}$, independently of n_c when R_q and $K_{\max}$ are fixed. This bound does not control the condition numbers of the moment systems: reducing the query scale can bring the projected exponential nodes closer together and make those systems ill-conditioned.

Taking operator-norm bounds through these finite compositions gives a finite bound χ for $D\mathscr{I}$ on any compact observation neighborhood satisfying the stated regularity conditions. On such a compact set, χ can be bounded from the explicit denominators and matrix inverses appearing in the reconstruction. We do not report a numerical value for any checkpoint.

Proposition 12 (Local control by finite interventions)**.** *Let an observation neighborhood be convex and let $\|D\mathscr{I}\|_2 \leq \chi$ there. Suppose a family of future centered-log-probability maps $\mathcal{L}_{\boldsymbol{e}}$, with $|\boldsymbol{e}| \leq H$, has derivative norm at most L_H on the segment between two reconstructed states. Then those two states satisfy equation*

$$\sup_{\boldsymbol{e}:\, |\boldsymbol{e}| \leq H} \|\mathcal{L}_{\boldsymbol{e}}(\zeta) - \mathcal{L}_{\boldsymbol{e}}(\zeta')\|_2 \leq L_H \chi \, \|\mathcal{E}(\zeta) - \mathcal{E}(\zeta')\|_2 . \tag{52}$$

Proof. Integrating $D\mathscr{I}$ along the segment joining the two intervention vectors yields $\|\zeta - \zeta'\|_2 \leq \chi \, \|\mathcal{E}(\zeta) - \mathcal{E}(\zeta')\|_2$. For each future intervention, integration along the reconstructed-state segment gives $\|\mathcal{L}_{\boldsymbol{e}}(\zeta) - \mathcal{L}_{\boldsymbol{e}}(\zeta')\|_2 \leq L_H \, \|\zeta - \zeta'\|_2$. Taking the supremum over $\boldsymbol{e}$ proves the result. □

For bounded controls, finite horizon, and a compact collection of states whose controlled continuations remain in a compact differentiability domain, such an L_H exists. This does not establish an efficient algorithm for a useful small bound. Unbounded replacements can destroy any proposed uniform constant, and near-coincident keys can make the finite inverse arbitrarily ill-conditioned. Exact identification alone is consequently not a numerical robustness guarantee.

G Boundary cases and consequences for interpretability

G.1 Mean shifts and positive normalization stabilizers

For centered pre-normalization, adding a multiple of $\mathbf{1}_d$ to the live residual does not affect any normalized read. Residual updates can carry a changed mean forward, but the next normalization again removes it, and the final normalization removes it before vocabulary readout. This verifies the removal of the mean component in the exact state. It does not justify discarding arbitrary other coordinates.

For RMS normalization with positive stabilizer, scaling is not an exact symmetry. Indeed, for $x \neq 0$ and a positive scalar c,

$$\frac{cx}{\sqrt{c^2 \|x\|_2^2 / d + \epsilon}} = \frac{x}{\sqrt{\|x\|_2^2 / d + \epsilon}} \tag{53}$$

holds only when $c = 1$. The same radial distinction applies on the centered subspace for layer normalization. This is why the inverse of the final radial map is well defined when its omitted squared norm is known. At $\epsilon = 0$, radial scale would be lost and the identification proof would have to change.

G.2 Why translations outside the constructed subspace are excluded

Suppose $\operatorname{ran} P \cap E^{\perp}$ contains a nonzero vector $x_{\perp,*}$ and choose u with $Au \neq 0$, which exists under Assumption 2. The original states $Uu + x_{\perp,*}$ and $Uu - x_{\perp,*}$ have the same extracted state. Translating both by $x_{\perp,*}$ gives omitted squared norms $4\,\|x_{\perp,*}\|_2^2$ and zero. Under Assumption 2, their bypass observations differ because $Au \neq 0$ and their normalization scales are different. Hence a state storing only the omitted norm is not closed under arbitrary translations in the omitted complement. The declared intervention restriction is necessary for this abstraction, not a technicality that can be removed without changing the result.

G.3 Reading a cache is not the same as extending it

For a vector a_*, replacing every historical key by $\kappa_j + a_*$ multiplies both $Z(q)$ and $N(q)$ by $e^{q^{\top} a_*}$. A single unmodified attention read is unchanged. Insertion of a fixed zero key instead yields denominators $Z(q)+1$ and $e^{q^{\top} a_*} Z(q)+1$. The difference between zero and nonzero appended values identifies these denominators through (30). Thus the same key translation is generally distinguishable under insertion. When keys are variable, a dimension argument that treats this translation as an enduring symmetry of the whole autoregressive state would be incorrect for our controls.

If queries are restricted to a proper linear subspace, key components perpendicular to that subspace may remain unobservable even after fixed-key insertion. Our intervention library requires effective query replacement in a full-dimensional ball. The resulting lower bound must not be transferred to a restricted native query family without an additional argument establishing comparable distinguishability.

G.4 Readout degeneracy and interpretation

A rank-deficient but nonzero A is sufficient for the identification result because residual translations expose directions in its kernel through the normalization denominator. The case $A = 0$ is treated in Appendix B.1. The identified coordinates are computational state variables rather than semantic variables; no correspondence with human-interpretable concepts is implied.

H Identification under weaker control families

The results here retain the original weight-derived subspace E. They do not recompute a smaller read–write span after restricting the controls. Every comparison uses the same discrete type and the nonzero terminal readout of Assumption 2.

H.1 Removing residual resets

Let the reset-free class be the class of Section 2 with whole-residual resets removed. Embedding initialization after emission is still the native token transition, not an extra available reset.

Proposition 13 (Reset-free finite identification)**.** *The finite identification count in* (10)*, the behavioral equivalence statement in Theorem 3, and the local dimension bounds remain valid for the reset-free class. The corresponding finite observation map has a smooth local left inverse on each regular chart.*

Proof. Use the original $2r$ bypass translations to recover the incoming (u, ρ). For a target head, fix its effective query and appended key and value, suppress earlier writes to the relevant historical cache, gate off every other residual contribution, and suppress irrelevant cache writes. Translate the incoming residual by $U\delta$ at the token-entry boundary. These translations are permitted without resets. The target head write w is independent of δ and of the live residual because both its query and its new cache entry are replaced by declared constants. Every other residual contribution, including output biases, is zero. The final residual statistics are therefore $(u + w + \delta, \rho)$, also for parallel blocks.

Apply Lemma 10 to the same $2r$ token-pair measurements. It recovers $(u + w, \rho)$; subtracting the bypass estimate of u gives w. This recovers exactly the isolated writes needed in (30)–(31), with the same number of passes. The query library and the moment reconstruction are unchanged. All

experimental choices are fixed in advance; only the subsequent reconstruction uses the recorded baseline.

Thus equality of observations in the reset-free class forces equality of Φ_p. Conversely, equality of Φ_p preserves every reset-free continuation because these controls form a subset of the original class and Theorem 1 applies. The reconstructed differences, residual inverse, and fixed separating-direction inverse are smooth on a regular observation neighborhood. The dimension proofs use only this finite identification and the original continuous section, so they continue to apply. The approximation proof applies to the reset-free observation map as well, but its Jacobian and singular values are those of that new map, not necessarily the original reset-based map. □

H.2 A NECESSARY-AND-SUFFICIENT ISOLATED-CACHE CRITERION

We specify the measurement family of Proposition 4 to separate a uniform identification statement from identification at particular caches. The unknowns are the keys and projected values in one retained cache with fixed length $n_c \geq 1$ and key width $d_c \geq 1$, on the full domain allowed in Section 2. The incoming residual and all other caches are fixed across comparisons; the incoming residual may be unknown but is recovered by bypass measurements. Each non-bypass pass reads only one head $h \in \mathcal{H}_{\mathrm{obs}}$, at a replaced query in a fixed relative open ball about zero in $\mathcal{Q}_c$, after one zero-key insertion with declared projected value $b \in \mathcal{B}_{\mathrm{ins}}$. No other residual branch, cache insertion, native suffix, or second head read is performed. Translations use the scalar reconstruction library. Equivalently, the cache-dependent vectors available through this family are

$$w_h^b(q) = G_h \frac{N(q) + b}{Z(q) + 1}. \tag{54}$$

The preceding reset-free construction shows that describing these vectors does not assume direct access to internal writes.

Proof of Proposition 4. For sufficiency, choose h_* and b_0, b_1 with $g_* = G_{h_*}(b_1 - b_0) \neq 0$. Recover $w_h^{b_0}(q)$ for every accessible head and $w_{h_*}^{b_1}(q)$ by fixed token-pair measurements. Then

$$\begin{aligned} \alpha(q) &= \frac{g_*^{\mathsf{T}}(w_{h_*}^{b_1}(q) - w_{h_*}^{b_0}(q))}{\|g_*\|_2^2} = \frac{1}{Z(q)+1}, \\ G_h N(q) &= \alpha(q)^{-1} w_h^{b_0}(q) - G_h b_0. \end{aligned} \tag{55}$$

Full column rank of $\mathsf{G}_{\mathrm{obs}}$ recovers $N(q)$ and $Z(q) = \alpha(q)^{-1} - 1$. A full-dimensional query ball contains the finite library of Appendix C, which reconstructs every grouped triple. At most $2r(|\mathcal{H}_{\mathrm{obs}}| + 1)$ passes per query and the shared $2r$ bypass passes suffice.

For necessity of the query condition, suppose $\mathcal{Q}_c$ is proper. Choose a nonzero $z_q \in \mathcal{Q}_c^{\perp}$ and translate every historical key by z_q. All dot products with allowed queries, hence all vectors in (54), remain identical. The two finite key multisets differ: a nonempty finite set cannot equal its translate by a nonzero vector. The grouped triples are therefore different and cannot be identified.

For necessity of the value-rank condition, choose $0 \neq z_v \in \ker \mathsf{G}_{\mathrm{obs}}$. Keep all keys fixed and change the projected value in one slot by z_v. This changes its grouped value sum, but changes $G_h N(q)$ by zero for every accessible head and every query. No allowed measurement distinguishes the two caches.

Finally, suppose no insertion contrast exists. Fix any $b_0 \in \mathcal{B}_{\mathrm{ins}}$. The failure of the contrast condition means $G_h b = G_h b_0$ for every allowed b and every accessible h. Assign projected value b_0 to each historical slot, so $N(q) = Z(q) b_0$. Equation (54) then gives $w_h^b(q) = G_h b_0$, independently of all historical keys. Two different key multisets of the same length give identical observations. This disproves uniform identification, even though some particular caches could remain identifiable. The three arguments establish necessity on the stated unrestricted cache domain. □

Without insertion at all, the ordinary writes $G_h N(q)/Z(q)$ are invariant under a common key translation for every query. Thus absolute historical keys cannot be uniformly recovered by those reads. In contrast, Proposition 4 concerns exactly one inserted entry with a controlled value set. It does

not assert that its value-contrast condition is necessary when multiple insertions, variable insertion counts, other residual branches, or multi-step native continuations are added to the measurement family.

H.3 Executable state for restricted query spaces

For each cache group fix a linear space $\mathcal{Q}_c \subseteq \mathbb{R}^{d_c}$ and an orthonormal basis matrix $V_{q,c} \in \mathbb{R}^{d_c \times t_c}$. The restricted execution class contains only operations for which every active effective query belongs to $\mathcal{Q}_c$. It includes query replacements in a relative open ball about zero, all original residual gates and translations in E, and the zero-key zero/nonzero-value insertions used in the main theorem. Resets may be retained or omitted. A native query is allowed only when it belongs to its declared space throughout the state domain under consideration. Imposing a projection on a native query that does not satisfy this condition changes that computation; no equivalence to such an unrestricted native decoder is asserted.

Project each stored key to $\bar{\vartheta} = V_{q,c}^{\mathsf{T}} \vartheta \in \mathbb{R}^{t_c}$. Merge slots with the same projected key, not merely the same original key. Write their triples as $(\bar{\kappa}_{c,j}, \bar{\mu}_{c,j}, \bar{\omega}_{c,j})$, with $1 \le j \le \bar{g}_c$. For a query $q \in \mathcal{Q}_c$, let $\bar{q} = V_{q,c}^{\mathsf{T}} q$. Then

$$u \longleftarrow u + G_h \frac{\sum_j e^{\bar{q}^{\mathsf{T}} \bar{\kappa}_{c,j}} \bar{\omega}_{c,j}}{\sum_j \bar{\mu}_{c,j} e^{\bar{q}^{\mathsf{T}} \bar{\kappa}_{c,j}}}. \tag{56}$$

The formula uses the same reduced residual state as (4).

Proof of Corollary 5. For every admissible query, $q^{\mathsf{T}} \vartheta = \bar{q}^{\mathsf{T}} \bar{\vartheta}$. Therefore grouping by projected keys preserves each attention numerator and denominator. A new key is computed from the existing retained residual statistics or supplied by a declared replacement, then projected by $V_{q,c}^{\mathsf{T}}$ before insertion. Projected coincidences update the multiplicity and value sum. All other reads, residual updates, translations, resets when permitted, positions, and emission probabilities compile exactly as before. This proves a closed realization $\Phi_p^{\mathcal{Q}}$ and sufficiency for all finite restricted continuations.

For necessity, residual bypass measurements still identify (u, ρ) because E and its translations have not changed. Head isolation and a visible inserted-value contrast recover the restricted functions $Z(V_{q,c}\bar{q})$ and $N(V_{q,c}\bar{q})$, using all the original heads in that group. For $t_c \ge 1$, apply the finite moment construction in dimension t_c. It reconstructs every projected-key triple, including original keys whose projections coincide. One may use

$$\bar{J}_c = 1 + (t_c - 1)\binom{n_c}{2} \tag{57}$$

directions and at most $\bar{J}_c n_c (t_c + 2)$ queries for that group. For $t_c = 0$, the only query is zero, $Z(0) = n_c$ is known, and $N(0)$ is the total projected-value sum. The stacked zero-value-insertion writes equal $\mathsf{G}_c N(0)/(n_c + 1)$ and recover that sum by full column rank. An empty cache requires no observations. Thus equality of all restricted observations is equivalent to equality of $\Phi_p^{\mathcal{Q}}$.

On a regular chart, keep projected multiplicities and a local order of the distinct projected keys fixed, and take $\rho > 0$ when present. The coordinates vary on an open subset of dimension $D_{\mathcal{Q}}$ in (11). A continuous section into original states is obtained by lifting each projected key as $V_{q,c}\bar{\kappa}_{c,j}$, distributing its value sum equally among its slots, and using the residual section in (39). Components orthogonal to $\mathcal{Q}_c$ can be fixed. Finite identification forces any sufficient continuous encoder restricted to this section to be injective. Invariance of domain gives the lower bound $D_{\mathcal{Q}}$; the displayed coordinates and their compiled updates attain it locally. This does not assert a fixed-dimensional global chart across projected-key collisions. □

H.4 Internal-intervention minimality versus native behavior

Even a full read–write span need not imply that native behavior requires all residual coordinates. Consider an RMS-normalized decoder with no retained caches, zero residual contributions, and a centered terminal readout depending only on the first residual coordinate. Include a feed-forward branch with $W_1 = \mathsf{I}_d$, $W_2 = 0$, zero output bias, and identity normalization gain. Its read makes

Table 2: Checkpoint rank measurements. The two numerical-rank entries in each column use relative thresholds 10^{-2} and 10^{-3}, respectively. The terminal ratio is $\sigma_{\operatorname{rank} P}/\sigma_1$ for $\Pi W_{\mathrm{U}} \Gamma_f P$. The final column counts cache groups with full observable-value rank at threshold 10^{-3}.

Model	$\operatorname{rank} P$	Read–write ranks	Terminal ranks	Terminal ratio	Full-rank groups
GPT-2 small	767	767 / 767	758 / 762	6.14×10^{-5}	144 / 144
Pythia-160M	767	767 / 767	767 / 767	4.54×10^{-2}	144 / 144
Llama-3.2-1B	2,048	2,048 / 2,048	2,046 / 2,048	2.48×10^{-3}	128 / 128

$E = \mathbb{R}^d$ by (2), although its native write is zero. The current native distribution depends only on

$$s_{\mathrm{native}}(x) = \frac{x_1}{\sqrt{\|x\|_2^2/d + \epsilon_f}}. \tag{58}$$

With no cache, subsequent emitted-token embeddings reset the native residual to a known vector, so this scalar also suffices for the native continuation law from the current token entry. Yet the declared translations in E make every residual coordinate identifiable through Lemma 10. This example does not contradict the dimension theorem: the observation classes differ. Full checkpoint matrix ranks alone cannot transfer the internal-intervention lower bound to native generation or prompt-reachable states.

I Numerical experiments

The experiments examine checkpoint ranks, local sensitivity of the intervention observations, and finite-precision reconstruction. They do not evaluate a compressed language model or the local future-observation bound of Proposition 12.

I.1 Checkpoint ranks

The LayerNorm checkpoints are `gpt2` and `EleutherAI/pythia-160m`; the RMSNorm checkpoint is `meta-llama/Llama-3.2-1B`. Pythia uses parallel attention and feed-forward branches (Biderman et al., 2023); the theoretical specification treats each such block as one atomic transition (Appendix A). For each model, we construct every effective read and residual write in (2), including normalization gains, projected value reads, residual-write biases, and the centered terminal readout. Shared key–value groups are counted once. Weight-derived matrix operations, Gram accumulation, and spectral decompositions are performed in FP64. Conversion to FP64 preserves the stored checkpoint values; it does not recover precision absent from the checkpoint.

The Gram matrix of the read–write span is accumulated in blocks, without materializing the full stacked matrix. Its eigenvalues give the squared singular values. We also construct the terminal Gram matrix separately. Observable-value bases are obtained by an SVD of the stacked matrices PO_h within each cache group, using relative threshold 10^{-8}. A reported numerical rank counts singular values strictly larger than the specified relative threshold times the largest singular value of the matrix being analyzed.

The read–write span fills the normalized residual space at both reporting thresholds. Every cache group retains all 64 value coordinates, including at the construction threshold. GPT-2 has five terminal singular directions below the 10^{-3} threshold but above 10^{-5}. Thus a thresholded rank deficit should not be interpreted as an exact kernel. Equality of thresholded ranks does not certify full column rank, which is needed only for the shorter inverse used below. Assumption 2 requires merely a nonzero readout, not equality of these ranks. The measurements are consistent with the full-dimensional regime described in Section 4, while revealing substantial differences in terminal conditioning.

I.2 Full-state spectra and directional checks

Reference states. We shuffle the WikiText-103 validation split with seed 20260911 and select eight eligible passages, excluding texts shorter than 80 characters. Each checkpoint uses its own

Table 3: Directional checks. Entries are percentage deviations of mean ratios from one, maximized over the tested selectors. One-sided refers to the larger distance to the reference; symmetric refers to the secant between the two perturbed states.

	Relative radius 10^{-4}		Relative radius 3×10^{-3}	
Model	One-sided	Symmetric	One-sided	Symmetric
GPT-2 small	0.0425	0.00134	0.756	1.145
Pythia-160M	1.0839	0.03917	9.206	17.526
Llama-3.2-1B	0.2501	0.00119	6.376	1.049

tokenizer without added special tokens. The first two tokens form the cache, and the third defines the token-entry residual, including the position embedding for GPT-2. We analyze the full residual state after removal of the LayerNorm mean direction, together with every key and observable value in every layer and cache group. The resulting dimensions are 37,631 for GPT-2 and Pythia and 34,816 for Llama. Extracted states and the subsequent analytic linear-algebra calculations use FP64.

For each key width, we generate a Gaussian direction a_{key} using seed $314159 + d_c$ and normalize it to length 0.20. This direction is reused across all reference states, layers, and cache groups of that width. We use the line and coordinate-offset grids of (33) with this single direction, $n_c = 2$, and $\gamma = 0.05$, giving 132 queries per group. The direction is not selected using the measured keys. Its minimum projected separation is checked against 10^{-8} at each evaluated state; the implementation does not replace a direction that fails this check. Each query uses a zero-key insertion with zero value for each head and an additional nonzero-value insertion for the reference head. The reference head and value are selected from the fixed output matrices, not from the activations.

The empirical observation map $\mathcal{E}_{\mathrm{exp}}$ concatenates these head-isolation observations and one residual-bypass observation. It is fixed across states, but it is not the universal direction library of Theorem 3. If A has full column rank and this direction separates the keys, the inverse in Appendix C.2 gives a smooth local reconstruction for this smaller family; the residual has no ρ coordinate in these checkpoint configurations. These experiments do not use the $2r$ translated passes of the general construction. We compute $J_{\mathrm{exp}} = D\mathcal{E}_{\mathrm{exp}}(\zeta_0)$ analytically. Since each observation depends on either the residual block or one cache group, $J_{\mathrm{exp}}^{\mathsf{T}} J_{\mathrm{exp}}$ is block diagonal. We diagonalize every block and merge the spectra, retaining all layers and cache groups. Squared singular-value sums describe the Frobenius energy of this particular Jacobian; they are not fractions of preserved task accuracy.

Directional validation. We test $k = 0$, 64, 128, $\lfloor 0.05D \rfloor$, $\lfloor 0.10D \rfloor$, k_{90}, k_{95}, and k_{99}. Let v_{k+1} be the corresponding unit right singular vector and set $\zeta_\pm = \zeta_0 \pm R_{\mathrm{loc}} v_{k+1}$. The relative radii are 10^{-4}, 3×10^{-4}, 10^{-3}, and 3×10^{-3}. Each is multiplied by the larger of one and the norm of the state block containing the selected singular vector. The perturbation is then embedded in the full state with all other blocks fixed. Consequently, a relative radius is not the same as the absolute radius R_{loc}, and different selected blocks can have different absolute radii.

We divide the larger of the two output distances from $\zeta_\pm$ to ζ_0 by $R_{\mathrm{loc}} \sigma_{k+1}(J_{\mathrm{exp}})$. We also report the symmetric secant ratio $\|\mathcal{E}_{\mathrm{exp}}(\zeta_+) - \mathcal{E}_{\mathrm{exp}}(\zeta_-)\|_2 / (2R_{\mathrm{loc}} \sigma_{k+1}(J_{\mathrm{exp}}))$. Both approach one in the first-order limit. Table 3 reports the largest absolute deviation from one among the eight selector-specific mean ratios; each mean is taken over eight reference states. It is not a maximum over individual states. These checks evaluate the local linear prediction underlying Theorem 8; they do not numerically solve the minimax problem in $e_k(R_{\mathrm{loc}}) = \inf_{\psi,\mathcal{D}} \sup_{\|v\|_2 \le R_{\mathrm{loc}}} \|\mathcal{E}(\zeta_0 + v) - \mathcal{D}(\psi(v))\|_2$.

Checkpoint-based forward checks. On the first reference state, we check cache group 0 in layers 0, 6, and 11 for GPT-2 and Pythia, and layers 0, 8, and 15 for Llama. The sampled queries are 0, a_{key}, $3a_{\mathrm{key}}$, and $\gamma e_{\mathrm{key},1}$. A fixed-seed random cache perturbation has relative radius 10^{-3} under the same block-norm convention. We form the isolated head write directly from raw cached values and the checkpoint output matrix, then evaluate the checkpoint's final-normalization and vocabulary-readout modules. The Euclidean distance between the resulting centered logit vectors is compared with the analytic distance computed in projected value coordinates. Centered logits equal centered log probabilities, so the observations match the metric used in the analysis.

All sampled comparisons satisfy the implemented tolerance, an absolute error of 10^{-6} plus 5×10^{-3} times the larger distance. The maximum observed relative discrepancies are much smaller: 1.82×10^{-13} for GPT-2, 7.63×10^{-11} for Pythia, and 6.33×10^{-7} for Llama. These checks compare output distances obtained through raw and projected coordinates and directly exercise the terminal modules. They are not an independent end-to-end implementation of all intervention gates, and they do not validate multi-step compiled generation.

I.3 Finite-precision reconstruction

Synthetic systems and paired conditions. We use the full-column-rank reconstruction of Appendix C.2 and its algebraic observation model, with $d = 16$, $r = 8$, vocabulary size 64, key and observable-value widths 4, and two heads sharing one cache. The centered terminal matrix A and the stacked head-write matrix have orthonormal columns. The residual satisfies $\|u\|_2^2 = d$, the omitted norm is fixed at $\rho = 0.016$, and the translation has norm 0.50. Keys and raw values are sampled from zero-mean Gaussian distributions with standard deviation 0.50; distinct keys are required to have Euclidean separation at least 0.20. This is a synthetic test of the inverse formulas, not a trained decoder experiment.

We use cache lengths 1, 2, and 4. For lengths 2 and 4, we test both distinct keys and one duplicated key; the duplicated slots contribute their summed values. There are ten random instances for each length and multiplicity pattern. Within an instance, the system and state are held fixed while ϵ_f varies over 10^{-1}, 10^{-3}, and 10^{-5}. Observation noise is sampled separately from the instance, centered across vocabulary coordinates, and added to the centered logits before softmax. Its configured standard deviations are 0, 10^{-12}, 10^{-10}, and 10^{-8}. The noise generator is reset to the same instance-specific seed in each condition. The design contains 50 distinct system–state instances and 600 condition-specific runs, all computed in FP64.

Fixed direction and reconstruction. A Gaussian direction generated with seed $314159 + d_c$ is normalized to length 0.50 and reused in every condition. Coordinate offsets have magnitude $\gamma = 0.40$, and the query-ball radius is 10. The smallest projected separation among distinct keys is 2.48×10^{-3}, above the 10^{-8} check threshold. No direction is selected or replaced using a tested state. The implementation first evaluates the $2n_c$ line queries and estimates the Hankel rank using relative threshold 10^{-10}. It then requests the offset moments required by that estimated rank. Thus the direction is fixed, while the evaluated subset of offset queries can depend on the rank estimate. We do not claim to execute the full non-adaptive library of Theorem 3.

Accuracy criteria and results. Numerical completion and accurate recovery are recorded separately. Accurate recovery requires relative errors at most 10^{-3} for u, ρ, the keys, the value sums, and the concatenated continuous state. It also requires a maximum absolute multiplicity error at most 10^{-3} and exact integer multiplicities after rounding. All runs, including exceptions, enter the success-rate denominator. Residual errors remain recorded when subsequent cache recovery fails; cache-error summaries are conditional on completed reconstruction.

Without added noise, 150/150 runs complete and 147/150 meet the accuracy criteria. The three inaccurate cases all have $n_c = 4$ and $\epsilon_f = 10^{-5}$; their Hankel condition numbers range from 7.84×10^8 to 3.48×10^9. Across all no-added-noise runs, the relative residual errors remain below 1.3×10^{-12} for u and 5.4×10^{-12} for ρ, so these three failures are in cache recovery rather than residual reconstruction. At noise level 10^{-8}, the accurate-recovery counts pooled over stabilizers and multiplicity patterns are 30/30, 20/60, and 0/60 for lengths 1, 2, and 4, respectively. In duplicate-key cases, noise can also alter the estimated Hankel rank; the reported failures therefore include rank-estimation errors as well as errors in subsequent inversion. The measured RMS error ratios describe the full-column-rank reconstruction, not the operator-norm bound χ for the translated scalar-probe inverse in Section 6.

I.4 Fixed-token-pair reconstruction

Systems and sampling. We evaluate the scalar inverse of Lemma 10 using NumPy FP64 calculations with $d = 16$, $r = 8$, vocabulary size 64, key and value widths four, and two heads sharing one cache. A centered Gaussian vocabulary vector and a normalized Gaussian residual direction form

Table 4: Finite-precision recovery. Each cell gives accurate recoveries / completed runs, out of ten trials. Noise columns give the configured standard deviation before vocabulary centering. "Duplicate" means that one key appears twice. A cell $0/0$ means that none of the ten runs completed, not that the condition was omitted.

ϵ_f	n_c	Keys	0	10^{-12}	10^{-10}	10^{-8}
10^{-1}	1	Distinct	10/10	10/10	10/10	10/10
	2	Distinct	10/10	10/10	9/10	8/10
	2	Duplicate	10/10	10/10	4/4	0/0
	4	Distinct	10/10	3/10	1/9	0/3
	4	Duplicate	10/10	9/9	3/4	0/0
10^{-3}	1	Distinct	10/10	10/10	10/10	10/10
	2	Distinct	10/10	10/10	9/10	8/10
	2	Duplicate	10/10	10/10	5/5	1/1
	4	Distinct	10/10	3/10	0/9	0/3
	4	Duplicate	10/10	9/10	3/5	0/0
10^{-5}	1	Distinct	10/10	10/10	10/10	10/10
	2	Distinct	10/10	9/10	8/10	3/9
	2	Duplicate	10/10	4/4	0/0	0/0
	4	Distinct	8/10	1/10	0/4	0/1
	4	Duplicate	9/10	4/5	0/0	0/0

a rank-one terminal matrix A. The two tokens are selected by the largest and smallest vocabulary coefficients and remain fixed throughout each system's measurements. The terminal bias is sampled from a Gaussian with standard deviation 0.10 before centering. Head-write matrices have Gaussian entries with standard deviation 0.25 and a stacked column rank of four. The insertion value is a Gaussian direction scaled to norm 0.50 and checked for a nonzero first-head write.

We use master seed 20260917 and trial seeds 1000–1009. Cache lengths are one, two, and four; lengths two and four include both distinct keys and one duplicated key. Keys and individual values have Gaussian standard deviation 0.50, with minimum Euclidean separation 0.20 between distinct keys. Values belonging to a duplicated key are summed. Four residual cases are evaluated: a Gaussian direction rescaled to norm $\sqrt{8}$ with $\rho = 0.016$; $u = 0$ with $\rho = 0.016$; a residual of norm $\sqrt{8}$ with $\rho = 0$; and a residual whose component orthogonal to a has norm two and whose coefficient along a is 0.05, with $\rho = 0.016$. Each case is evaluated at $\epsilon_f \in \{10^{-1}, 10^{-3}, 10^{-5}\}$ and four observation-noise levels. The 50 system–cache instances therefore produce 2,400 condition-specific runs. Random weights, caches, and residual coordinates are held fixed across stabilizer and noise settings; observation noise uses separate condition-specific seeds.

Measurements and inversion. Each scalar log-probability ratio is evaluated as the corresponding logit difference. Independent Gaussian noise with standard deviation 0, 10^{-12}, 10^{-10}, or 10^{-8} is added to this scalar before inversion. This differs from the vocabulary-centered noise in Appendix I.3. Residual recovery uses the $2r = 16$ translations with $s_{\mathrm{tr}} = h_{\mathrm{tr}} = 0.30$. For cache recovery, the same inverse recovers each isolated write after a residual reset. Zero-value insertions are measured for both heads, and a nonzero-value insertion is additionally measured for the first head. Equations (30)–(31) then recover $Z(q)$ and $N(q)$ using 48 scalar observations per query.

A single Gaussian cache direction, generated with seed 314163 and normalized to length 0.50, is reused in every condition. Coordinate offsets have magnitude $\gamma = 0.40$. All $2n_c$ line queries and $4n_c$ coordinate-offset queries are fixed before observing the state; their number does not depend on the estimated Hankel rank. The universal separating-direction library is not evaluated. Reconstruction uses the Hankel and Vandermonde systems from Appendix C, with relative Hankel-rank threshold 10^{-10} and write-stack pseudoinverse cutoff 10^{-14}. Polynomial roots must have positive real parts and imaginary parts no larger than 10^{-6} times the larger of one and the absolute real part. Nonpositive offset ratios or invalid scalar inverses are recorded as numerical failures.

Evaluation and results. Accurate recovery requires the correct number of distinct keys, exact multiplicities after rounding, and maximum absolute multiplicity error at most 10^{-3}. Relative errors for nonzero residuals, keys, value sums, and the concatenated diagnostic vector must also be at most 10^{-3}. Residual error is absolute when $u = 0$, and error in ρ is always absolute. The concatenated vector includes u, ρ, keys, multiplicities, and value sums; it is not the continuous chart vector used in the dimension theorem. Keys and associated quantities are matched by their ordering along the fixed cache direction.

All attempted runs enter the accuracy denominator. Numerical completion is recorded separately and can include an incorrect estimated group count; key, value, and concatenated errors are recorded only when that count is correct. The CSV summaries report means, sample standard deviations, medians, and maxima over available values within each condition.

Table 5: Fixed-token-pair recovery at the two endpoint noise levels. Entries are accurate recoveries over all attempted runs, pooled over stabilizers, residual cases, and applicable multiplicity patterns. These are repeated conditions on shared system–cache instances, not independent model draws.

Log-ratio noise s.d.	$n_c = 1$	$n_c = 2$	$n_c = 4$
0	120/120	240/240	216/240
10^{-8}	120/120	102/240	0/240

All 600 runs without added noise complete, but 24 fail the accuracy criteria; these cases have four distinct keys. The deterioration under noise is consistent with the conditioning of moment-based cache inversion. The experiment evaluates the scalar reconstruction formulas in a controlled algebraic system, rather than end-to-end identification of a pretrained decoder.

I.5 Intervention ablations

Common settings. We use the same residual, vocabulary, key, and value dimensions as above, with $\epsilon_f = 10^{-3}$, $s_{\mathrm{tr}} = h_{\mathrm{tr}} = 0.30$, and rank-one terminal readouts. Cache tests use three entries. Master seed 20260917 and trial seeds 2000–2009 define ten runs per setting, giving 100 runs across four reset-noise settings, one insertion test, four query-space dimensions, and one write-rank test. All controlled calculations use NumPy FP64.

For a fixed write, the observation vector contains the 16 translated token-pair log ratios. A reported observation gap is the Euclidean distance between two such vectors. In query-space tests, we maximize this distance over the tested queries; in write-rank tests, we maximize over queries and both heads. Table 6 reports means and sample standard deviations across ten seeds after these within-instance maxima.

Residual resets. We sample residual and write vectors with $\|u\|_2 = 2$, $\|w\|_2 = 0.40$, and $\rho = 0.05$. The reset-based inverse reconstructs w at omitted norm zero. The reset-free inverse separately reconstructs u and $u + w$ at the original ρ, then subtracts the two estimates. Both methods use the same vectors within a seed–noise condition, with separate observation-noise draws; the vectors are resampled across noise levels. This isolates the residual-reconstruction step rather than running a complete reset-free cache reconstruction.

Without added noise, mean write relative errors are 2.02×10^{-15} with resets and 1.18×10^{-13} without resets. At noise standard deviation 10^{-8}, they become 1.16×10^{-8} and 1.02×10^{-6}. Reset-free recovery uses 32 measurements including the baseline, or 16 additional measurements once u is known. Its larger errors reflect both the conditioning of reconstruction at the nonzero residual and subtraction of two estimates.

Insertion values. All historical values are set to the same reference value b_{ins}, of norm 0.50. For a query of norm 0.50, the second cache is obtained by shifting its three keys along the query direction by 0.25, 0.50, and 0.75. Inserting b_{ins} makes the write independent of these keys because $N(q) = Z(q)b_{\mathrm{ins}}$. A zero-value insertion supplies the second measurement condition and distinguishes the constructed pair.

The mean denominator difference is 0.885, whereas the mean token-pair observation gap with only the reference insertion is 4.33×10^{-15}. The zero-value insertion gives a mean gap of 0.588. An additional algebraic check applies the insertion-contrast formula directly to the forward-computed write vectors, giving mean denominator relative errors 2.50×10^{-16} and 4.60×10^{-16} for the two caches. These last errors do not include terminal-output inversion; that complete reconstruction is evaluated in Appendix I.4.

Query spaces. Random orthonormal bases define query subspaces of dimensions zero, one, three, and four. For each proper subspace, the paired caches have identical values, while their keys differ by 0.30, 0.60, and 0.90 times a unit vector in its orthogonal complement. We test 12 queries of norm 0.50 in each nonzero subspace and only the zero query in the zero-dimensional case. A query of norm 0.50 along the hidden direction supplies an unrestricted witness. The full-dimensional positive control instead shifts one key by 0.60 along an accessible direction; its mean observation gap is 1.95. The test compares projected coordinates and output differences, rather than reconstructing or updating projected-key triples.

Write rank. A rank-three stacked write map is constructed as the product of Gaussian matrices of sizes 16×3 and 3×4, then normalized to operator norm one. One historical value is changed by 0.50 times its numerically computed null direction. The paired caches are evaluated at 12 queries of norm 0.50 using both heads. An independently sampled full-column-rank stack provides the positive control. Relative singular-value thresholds 10^{-8}, 10^{-10}, and 10^{-12} return ranks three and four, respectively, in every trial. These are ranks of the write maps, not of the observation Jacobian.

Table 6: Observation gaps in constructed cache pairs, reported as mean and sample standard deviation over ten seeds. Positive controls use the zero-value insertion, an unrestricted witness query, or the independently sampled full-rank write stack, respectively.

Restricted measurement	Observation gap	Positive-control gap
Reference insertion only	$(4.33 \pm 2.09) \times 10^{-15}$	0.588 ± 0.239
Query dimension zero	0	2.25 ± 0.97
Query dimension one	$(5.72 \pm 2.64) \times 10^{-15}$	2.12 ± 1.18
Query dimension three	$(7.49 \pm 2.24) \times 10^{-15}$	3.03 ± 1.91
Rank-three write stack	$(8.74 \pm 1.67) \times 10^{-15}$	6.69 ± 2.11

All 100 runs complete. The paired constructions illustrate the indistinguishability mechanisms of Proposition 4; their finite observation gaps do not establish uniform identification or a minimum state dimension. Trial-level results and condition-level statistics are saved separately by `exp1_fixed_pair_reconstruction.py` and `exp2_intervention_ablation.py`.

J Experimental protocol and complete results

This appendix gives the protocol and reported results for the quantized-checkpoint causal experiments. Direction interventions and component ablations provide candidate and component evidence, while the baseline-coordinate preservation and shared-target intervention experiments examine the candidate-state and partial-intervention questions of Section 7. The tables specify the estimators, sampling units, and confidence intervals for each stage. Unless stated otherwise, confidence intervals use 2,000 percentile-bootstrap resamples at the problem-group level, with all within-group conditions resampled together. No model was trained for these experiments.

J.1 Relation to the theoretical results

The causal study evaluates the sufficiency of a fitted scalar candidate rather than the complete constructed state Φ_p. In particular, equal numerical readings of c_{num} with unequal responses can reject factorization through that numerical scalar on the tested states, but they do not imply equality of the ideal linear coordinate c_{lin} or of the complete constructed state.

J.2 Model, hardware, and numerical implementation

The Mistral experiments use one NVIDIA RTX 4090 with 24 GB GPU memory, an Intel Xeon Gold 6136 CPU, and 512 GB system RAM. The checkpoint identifier is `mistralai/Mistral-Small-24B-Instruct-2501`. Loading uses bitsandbytes NF4 four-bit linear weights with double quantization and an FP16 compute dtype. The model remains in evaluation mode. Eager attention is used, the batch size is one, and forward calls set `use_cache=False`. The code disables TF32 for matrix multiplication and cuDNN and fixes the analysis random seeds. These settings do not assert that every primitive accumulates in FP16.

The task instruction is inserted into the user message, followed by the facts, rules, and question. The checkpoint's chat template is applied with a generation prompt, and the assistant prefix is `Answer:`. The answer strings `␣Yes` and `␣No` each resolve to one token, with IDs 13830 and 3501. Their logits are read from the original final vocabulary output; no new classifier replaces it. The pairwise decision is Yes exactly when their difference is positive. Full-vocabulary top-token decisions are reported only for stages where those outputs were saved.

For an input prompt x_{prompt}, let $\mathcal{I}_{\mathrm{src}}(x_{\mathrm{prompt}})$ be the selected token positions, $n_{\mathrm{span}}(x_{\mathrm{prompt}})$ their count, and H_{act} the activation matrix at this boundary. The corresponding real-valued coordinate is

$$c_{\mathrm{lin}}(H_{\mathrm{act}}) = v_{\mathrm{cand}}^{\mathsf{T}} \frac{1}{n_{\mathrm{span}}(x_{\mathrm{prompt}})} \sum_{i_{\mathrm{tok}} \in \mathcal{I}_{\mathrm{src}}(x_{\mathrm{prompt}})} (H_{\mathrm{act}})_{i_{\mathrm{tok}},:}^{\mathsf{T}}. \tag{59}$$

The candidate is loaded from the saved direction CSV, cast to FP16, normalized using an FP32 norm, and cast back to FP16. For candidate measurements, the selected activation rows are converted to FP32, averaged in FP32, and dotted with the FP32 conversion of the stored direction. Python floating-point export preserves the returned reading; it does not make the inner product an exact real-arithmetic operation. The scalar in the exact-reading subsets is therefore c_{num}, not c_{lin}. This distinction is also why the smooth rank and dimension results are not inferred from these subsets.

The fixed scale in the baseline-coordinate preservation and shared-target intervention experiments is 0.9344838042, while the source-selection record gives 0.9344838042364072. The experiments verify agreement within the prescribed scale tolerance and do not refit the direction. Statistics are computed from exported scores using FP64 arrays; model inference itself is not performed in FP64.

J.3 Logical inputs, source location, and data partitions

The rules in (15) use one substitution for all occurrences of a variable within each application. Negative labels mean nonmembership in the finite least deductive closure, not an explicitly negated proposition. With distinct $\mathtt{a}, \mathtt{b}, \mathtt{c}$, the facts $\mathsf{P}(\mathtt{a}), \mathsf{R}(\mathtt{a}, \mathtt{b}), \mathsf{S}(\mathtt{b})$ support the query. Replacing $\mathsf{P}(\mathtt{a})$ by $\mathsf{P}(\mathtt{c})$ breaks the first-rule binding. Removing the first or downstream premise prevents the two-rule derivation. Four distractor facts use predicates absent from the rule bodies. Generated fact strings are checked by a separate deterministic closure computation.

The predicates $\mathsf{P}, \mathsf{R}, \mathsf{Q}, \mathsf{S}, \mathsf{Z}$ are rendered as *glowing*, *visits*, *calm*, *steady*, and *ready*, respectively. Entity aliases are drawn from a fixed list of 32 names; distractors alternate between unary adjectives and binary relations outside the rule bodies. Aliases can recur across partitions. Evaluation concerns new generated combinations, not an unseen-name vocabulary.

Direction fitting uses 40 pairs. In the first member the upstream binding is valid, and in the second it is broken; neither member contains a steady fact, so both final labels are No. Equal task labels do not force equal model predictions. The direction is the normalized mean difference of pooled relation-target activations. This fitting procedure does not by itself identify a logical state variable on the intervention domain. Its scale is the square root of the average of the two unbiased within-condition projection variances. Zero or nonfinite directions are rejected. The fitted scale is nondegenerate; the code's small-variance fallback is not used for the selected direction.

A source test group has six contexts, shown in Table 7. Each has a broken first-rule binding and logical label No. The two target roles exchange the relevant entity names. Entity aliases and distractor facts are shared within a group, but fact order is shuffled independently across conditions. Deleting a premise also changes sequence length. Thus the difference between matched and mismatched

intervention effects is not a fully order-controlled intervention on entity identity alone. Each individual source effect still compares edited and unedited execution of the same prompt.

Table 7: The six source-intervention contexts in a group. Every row has a broken first-rule binding and logical answer No. Entity names and distractors are shared; fact order is independently shuffled across rows.

Target	Upstream facts	Downstream fact	Context
b	$\mathsf{P}(\mathtt{c}), \mathsf{R}(\mathtt{a}, \mathtt{b})$	$\mathsf{S}(\mathtt{b})$	Matched
b	$\mathsf{P}(\mathtt{c}), \mathsf{R}(\mathtt{a}, \mathtt{b})$	$\mathsf{S}(\mathtt{c})$	Mismatched
b	$\mathsf{P}(\mathtt{c}), \mathsf{R}(\mathtt{a}, \mathtt{b})$	None	Absent
c	$\mathsf{P}(\mathtt{b}), \mathsf{R}(\mathtt{a}, \mathtt{c})$	$\mathsf{S}(\mathtt{c})$	Matched
c	$\mathsf{P}(\mathtt{b}), \mathsf{R}(\mathtt{a}, \mathtt{c})$	$\mathsf{S}(\mathtt{b})$	Mismatched
c	$\mathsf{P}(\mathtt{b}), \mathsf{R}(\mathtt{a}, \mathtt{c})$	None	Absent

The natural-input evaluation instead supplies $\mathsf{P}(\mathtt{a})$, so the first rule is enabled. Two matching contexts have label Yes and the other four have label No. No source edit is applied in that stage. The baseline-coordinate preservation and shared-target intervention experiments each use 32 separately generated prompts with $\mathsf{P}(\mathtt{c}), \mathsf{R}(\mathtt{a}, \mathtt{b}), \mathsf{S}(\mathtt{b})$ and distractors. These prompts all have label No. Within a prompt, every alternative state or edit keeps the exact token sequence and downstream context fixed. These experiments do not use cross-prompt fact-order changes to establish output differences.

The source span is located through character offsets in the relation fact "a visits b", not through the query or final-answer position. In each experiment, 28 prompts have a one-token source span and four have a two-token span. The edits and intended projection shifts use the recorded span count, as in the general multi-token formula. The span occurs in the fact list before the supplied rules and query, and it may occur before the relevant glowing premise. Causal attention cannot read those later tokens at this position. Predictability from the prefix or pretrained weights is different from access to their current occurrences; the source reading is not thereby an already derived $\mathsf{Q}(\mathtt{b})$.

Table 8: Data partitions in the Mistral analysis. Counts are problem groups, not repeated model fits. The two shared-target amplitudes use the same 32 prompts.

Stage	Groups	Seed	Purpose
Behavioral screen	48	1103	Four conditions per group
Source fitting	40	3101	Two upstream donor conditions
Source validation	12	3102	Layer selection at amplitude one
Source dose test	24	3103	Five-point responses
Whole-block screen	24	4101	Later attention and feed-forward blocks
Single-head screen	24	4102	Heads in eligible attention layers
Joint-head discovery	24	5201	Restricted ranked-prefix search
Joint-head confirmation	32	5202	Selected set and fixed random controls
Natural-input evaluation	32	5203	No source edit
Baseline-coordinate preservation	32	6101	Baseline-coordinate preservation
Shared-target intervention	32	6102	Shared nonzero coordinate targets

Master seed 1729 controls the analysis routines, and the eight joint-head random sets use seed 5204. The candidate and component selections are frozen before each subsequent evaluation stage. The two experiments use fixed numerical protocols and distinct evaluation samples from the preceding diagnostic analyses.

J.4 Behavioral screening results

The Mistral screen contains 192 prompts. Every full-vocabulary top token is one of the two declared answer tokens, so pairwise and full-vocabulary accuracies agree in this screen. Overall accuracy is 71.875%, below the 75% attained by always answering No under the one-positive/three-negative composition. Only eight of the 48 groups have all four answers correct. Nevertheless, the complete-chain score exceeds every disrupted-condition score within each group. Mean paired gaps are 1.80501

Table 9: Mistral behavioral screening: 48 groups, four conditions per group. Accuracy is the original full-vocabulary top-token accuracy. The score is the un-oriented Yes-minus-No logit difference.

Condition	Correct / 48	Accuracy (%)	Mean score
Complete chain	48	100.00	2.14974
Binding conflict	14	29.17	0.34473
Missing first premise	28	58.33	-0.20443
Missing downstream premise	48	100.00	-1.39648

for binding conflict, 2.35417 for the missing first premise, and 3.54622 for the missing downstream premise. These score differences motivated the continuous-intervention analysis; they do not establish exact rule execution.

The relation-target study selected layer 17, with fitted scale 0.93448 and within-fitting-sample standardized separation 1.56485. The projection means in the two fitting conditions were -0.0269029438 and -1.4892345667. This separation is computed on the same data used to fit the direction and is not a held-out decoding score. Layer selection uses the smallest of the mean matched effect, the two role-specific entity differences, and the matched–absent difference at amplitude one, minus the absolute mean random-direction effect; no significance threshold is imposed.

J.5 Source responses and finite-dose interactions

At amplitude γ_{edit}, a source intervention adds $\gamma_{\text{edit}}\sigma_{\text{edit}}v_{\text{cand}}/\sqrt{n_{\text{span}}(x_{\text{prompt}})}$ to every selected activation row.

For group i, target role `b` or `c`, and a matching, mismatching, or absent-premise context, subtract the unedited score from the edited score. The five-point ordinary least-squares slope includes an intercept and uses the symmetric amplitude grid $\mathcal{D}_{\text{dose}} = \{-1, -1/2, 0, 1/2, 1\}$. For a fixed prompt, writing its effect as $e(\gamma_{\text{edit}})$, the slope is

$$\beta_{\text{dose}} = \frac{\sum_{\gamma_{\text{edit}} \in \mathcal{D}_{\text{dose}}} \gamma_{\text{edit}} e(\gamma_{\text{edit}})}{\sum_{\gamma_{\text{edit}} \in \mathcal{D}_{\text{dose}}} \gamma_{\text{edit}}^2}. \tag{60}$$

This summarizes finite edits and is not a measured derivative. Average the two target-role slopes within each context. Their matching minus mismatching difference is the entity-match interaction; matching minus absent is the premise interaction. The random-direction control is a normalized Gaussian direction projected off the candidate, using the same nominal norm and positions. It is fixed within a group across amplitudes and target roles. It does not exhaust all answer-associated directions.

Table 10: Source-direction effects on 24 evaluation groups. Values are logit units per unit amplitude. Intervals are marginal 95% group-bootstrap intervals. Role-specific rows are retained because the aggregate does not establish each role separately.

Endpoint	Mean	95% interval
Matched-context slope	0.00742	$[0.00391,\ 0.01113]$
Mismatched-context slope	0.00358	$[0.00182,\ 0.00547]$
Absent-premise slope	0.00013	$[-0.00150,\ 0.00169]$
Random-direction matched slope	0.00098	$[-0.00065,\ 0.00247]$
Target-`b` entity difference	0.00378	$[-0.00117,\ 0.00872]$
Target-`c` entity difference	0.00391	$[0.00052,\ 0.00768]$
Aggregate entity interaction	0.00384	$[0.00046,\ 0.00723]$
Downstream-premise interaction	0.00729	$[0.00325,\ 0.01133]$
Matched candidate minus random	0.00645	$[0.00345,\ 0.00944]$

The matching effect is larger on average than the mismatching and absent-premise effects. Mismatching effects are themselves positive, so the result is relative selectivity rather than exclusive response to a matching entity. The target-`b` interval crosses zero. A random-direction interval containing zero is

not evidence of equivalence to zero, and the average interaction does not establish both target-specific interactions separately.

Table 11: Source dose means and marginal 95% intervals on 24 groups. Zero effects are defined from the baseline; they are not independent no-op measurements.

Amplitude	Endpoint	Mean	95% interval
-1	Matched effect	-0.00781	$[-0.01172, -0.00358]$
-1	Entity interaction	-0.00618	$[-0.01172, -0.00098]$
-1	Premise interaction	-0.00553	$[-0.01074, -0.00033]$
-0.5	Matched effect	-0.00553	$[-0.00911, -0.00195]$
-0.5	Entity interaction	-0.00651	$[-0.01139, -0.00098]$
-0.5	Premise interaction	-0.00326	$[-0.00781, 0.00098]$
0	Matched effect	0.00000	$[0.00000, 0.00000]$
0	Entity interaction	0.00000	$[0.00000, 0.00000]$
0	Premise interaction	0.00000	$[0.00000, 0.00000]$
0.5	Matched effect	0.00098	$[-0.00195, 0.00391]$
0.5	Entity interaction	-0.00358	$[-0.00716, 0.00001]$
0.5	Premise interaction	0.00326	$[-0.00065, 0.00716]$
1	Matched effect	0.00749	$[0.00293, 0.01205]$
1	Entity interaction	0.00195	$[-0.00456, 0.00716]$
1	Premise interaction	0.00944	$[0.00326, 0.01530]$

The mean matching response increases across the displayed amplitudes, but the entity interaction does not. Its amplitude-one interval crosses zero. The positive fitted slope therefore cannot be described as a positive interaction at every positive amplitude or as an exact monotone logical gate. More generally, different slopes exclude a common additive amplitude response plus context offsets, while output nonlinearities can also produce context-dependent slopes without a Boolean conjunction. These finite-dose responses therefore provide evidence for context-dependent causal effects but do not establish the state-sufficiency property tested subsequently.

J.6 Downstream screening and four-head confirmation

The whole-block screen evaluates 44 later attention and feed-forward blocks at zero-based indices 18–39. Only attention blocks 18 and 19 pass the rule requiring positive mean reductions and positive marginal bootstrap lower endpoints for the matched response and both interactions. Their 64 individual heads are then screened; none passes the same criterion. This does not establish redundancy or synergy. Weak distributed effects, nonlinearity, and limited power remain alternatives.

Table 12: Eligible whole-attention blocks in the discovery screen (24 groups). These observations determined the subsequent search; they are not an independent confirmation result. No feed-forward block or individual head passed the same three-endpoint rule.

Block	Endpoint reduction	Mean	95% interval
Attention 18	Matched response	0.00798	$[0.00586, 0.01025]$
	Entity interaction	0.00749	$[0.00326, 0.01188]$
	Premise interaction	0.00846	$[0.00586, 0.01156]$
Attention 19	Matched response	0.00488	$[0.00244, 0.00716]$
	Entity interaction	0.00407	$[0.00081, 0.00749]$
	Premise interaction	0.00602	$[0.00277, 0.00944]$

The joint-head study tests ranked prefixes of sizes $2, 4, 8, 12, 16, 24$ within the eligible layers and equal-sized prefixes across both layers. Heads are ranked by the minimum of the three reduction means, using their sum as a secondary key. Whole-layer sets are benchmarks but are excluded from selection. Among passing tested sets, the procedure chooses the smallest, then the largest minimum mean. The selected set is the four layer-18 heads $\{5, 22, 29, 14\}$. Minimality is only within the tested candidate list, not among arbitrary subsets.

A head is intercepted after attention-weighted value aggregation and before the output projection; its slice is set to zero at every token position. This is not edge-specific path patching or restoration to a clean activation. Query, key, value, mask, and other projection operations remain in the computation. Unlike the source study, component tests use the two-point response

$$\widetilde{\beta}_{\mathrm{dose}}(\mathcal{H}_{\mathrm{abl}}) = \frac{r_{\mathrm{num}}(1, \mathcal{H}_{\mathrm{abl}}) - r_{\mathrm{num}}(-1, \mathcal{H}_{\mathrm{abl}})}{2}, \tag{61}$$

where the first argument is source amplitude and $\mathcal{H}_{\mathrm{abl}}$ is the set of ablated heads. The same role averages and context differences define the two interactions. A reported reduction is the signed unablated statistic minus the ablated statistic. It can cross zero, reverse sign, or exceed the unablated effect; it is not necessarily an absolute-magnitude decrease or a fraction of causal computation.

Table 13: Four-head joint ablation on 32 confirmation groups. Random means average eight fixed layer- and size-matched sets. The recorded confirmation statistics contain marginal intervals, not the paired selected-minus-random interval; the last column is descriptive.

Endpoint reduction	Selected mean	95% interval	Random mean
Matched response	0.00549	[0.00293, 0.00830]	0.00165
Entity interaction	0.00586	[0.00293, 0.00879]	0.00188
Premise interaction	0.00513	[0.00208, 0.00818]	0.00229

All three selected-set confirmation intervals exclude zero, and the selected means exceed those of the eight matched random sets. The random sets match layer and head count and are fixed across groups. However, the available summary does not provide a paired selected-minus-random interval or the corresponding unablated two-point interaction magnitudes. We therefore report the random-set comparison descriptively and do not estimate a mediated proportion.

J.7 Natural-input score contrasts and accuracy

On 32 natural-input groups, the first rule is satisfied and no source shift is used. Matching-minus-mismatching and matching-minus-absent differences are computed from unedited answer scores rather than intervention slopes. Selected and random-set effects are paired inside each problem group. Random controls are averaged within the group before resampling; their eight runs are not eight independent model replications.

Table 14: Natural-input ablation on 32 groups. Each value is a mean with a marginal 95% interval from the paired role-contrast export. Accuracy is pairwise Yes/No accuracy; a negative drop means improvement after ablation.

Endpoint	Selected signed reduction	Selected minus random
Entity gap	$0.52295\ [0.44702, 0.60059]$	$0.36514\ [0.29167, 0.43272]$
Premise gap	$1.08350\ [1.01976, 1.14551]$	$0.86154\ [0.80334, 0.91788]$
Accuracy drop (pp)	$-9.896\ [-13.021, -6.771]$	$-10.026\ [-13.021, -6.901]$

The mean selected-minus-random reductions are 0.36514 for the entity gap and 0.86154 for the premise gap, with the saved paired intervals shown in the table. Thus the heads affect natural task contrasts, not only the response to artificial edits. At the same time, the selected ablation improves pairwise accuracy by 9.896 percentage points. These two facts are compatible: score separation and correctness at a zero decision threshold are different quantities. The result does not identify the cause of the improvement and does not support describing the heads as necessary for correct task performance. The natural-ablation script did not save full-vocabulary outputs, so the pairwise improvement is not reported as greedy full-vocabulary accuracy.

J.8 Baseline-coordinate preservation: construction, controls, and exact-reading witnesses

Each of the 32 prompts supplies one source state and eight nominally coordinate-preserving edits: four random directions in the candidate orthogonal complement and both signs. Before rounding,

each selected row receives the same orthogonal vector scaled by $\sigma_{\text{edit}}/\sqrt{n_{\text{span}}(x_{\text{prompt}})}$. The total Frobenius perturbation norm is therefore nominally σ_{edit}. The script also executes a no-op and two candidate-aligned controls per group. These control rows are not included in the eight-edit response ranges.

Projection correction is performed on the actual FP16 edited rows. Up to 12 global iterations test correction multipliers $0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 3, 4$. If needed, up to 64 additional steps change one coordinate to an adjacent FP16 value in a direction that improves the projection residual. The target is the original numerical reading and the stopping tolerance is 10^{-6}. The correction depends on the activations and the candidate reading, not on an output-effect optimization. No gradients of the downstream score are used to select the directions or clamp corrections.

Table 15: Baseline-coordinate-preservation diagnostics. Intervals are the saved marginal 95% group-bootstrap intervals. Range statistics use all eight tolerance-passing edits per group; exact-reading counts are a separate analysis of the raw rows.

Quantity	Result
Prompts / orthogonal edits	32 / 256
Edits with candidate error at most 10^{-6}	256/256
Edits with exactly the baseline numerical reading	125/256
Equal-reading edits with a response change	58/125
Groups containing an equal-reading edit	31/32
Groups containing an equal-reading response difference	21/32
No-op response changes	0/32
Mean eight-edit response range	$0.016602\ [0.013672, 0.019531]$
Mean maximum absolute response change	$0.015625\ [0.013184, 0.018066]$
Mean absolute response change	$0.006714\ [0.005188, 0.008301]$
Mean group-maximum candidate error	2.7101×10^{-7}
Largest individual candidate error	9.5367×10^{-7}
Mean final total perturbation norm	0.934489545

The mean group-maximum projection drift decreases from 1.0735×10^{-4} before correction to 2.7101×10^{-7} after correction, with mean correction norm 0.000341988. Final perturbation norms range from 0.934395015 to 0.934598565. Twenty-eight groups have a nonzero response range among the eight tolerance-passing edits; these ranges are not restricted to edits with exactly equal numerical readings.

For the exact-reading claim, the raw records are filtered independently by equality of the exported pre-edit and post-edit readings. There are 125 such edits, and 58 change the answer score, in 21 distinct groups. For example, the exact equality concerns the same FP32 measurement procedure on two FP16 activation states; a nonzero ideal-real dot-product difference could still have rounded to the same reading. Consequently the conclusion is failure of output sufficiency for c_{num} on the evaluated states, not a certified fiber counterexample for c_{lin}.

J.9 Shared-target interventions: norm control and coverage

For each source prompt and $\gamma_{\text{edit}} \in \{-0.5, 0.5\}$, a corrected aligned reference edit first targets an intended candidate shift of $\gamma_{\text{edit}}\sigma_{\text{edit}}/\sqrt{n_{\text{span}}(x_{\text{prompt}})}$. Its attained numerical reading becomes the common target for all subsequent implementations in that prompt–amplitude condition. This target is fixed without selecting it for a downstream output effect. The reference has a nominal perturbation norm of $|\gamma_{\text{edit}}|\sigma_{\text{edit}}$ and is not included in the equal-norm implementation ranges.

For a nominal unit orthogonal direction u_{edit} and sign $\varsigma \in \{-1, 1\}$, each selected row initially receives

$$\frac{\sigma_{\text{edit}}}{\sqrt{n_{\text{span}}(x_{\text{prompt}})}} \left(\gamma_{\text{edit}} v_{\text{cand}} + \varsigma \sqrt{1 - \gamma_{\text{edit}}^2}\, u_{\text{edit}} \right). \tag{62}$$

Four directions and both signs produce eight implementations. The equation gives equal target-coordinate shifts and equal total norms in ideal arithmetic under unit orthogonality. The code checks the actual FP16 outcomes instead of assuming these ideal identities survive rounding. The same

projection-correction procedure is applied to reach the common numerical target. Norm control is checked after correction; it does not assert exact equality of total norms.

All 512 implementations reach the shared target within 10^{-6}, and every implementation meets the 10^{-4} norm tolerance. Exact target equality holds for 132 of 256 implementations at $\gamma_{\text{edit}} = -0.5$ and 72 of 256 at $\gamma_{\text{edit}} = 0.5$. The strict subset combines exact equality of the implemented reading with the norm tolerance. It has at least two implementations in 31 and 18 groups, respectively. Groups lacking two exact matches are excluded from exact-subset range statistics, not assigned range zero. All-group tolerance results remain available below to show this coverage difference.

Table 16: Shared-target intervention results. All-group ranges use the eight tolerance-passing implementations; strict ranges use exact numerical readings and post-correction norm control. Confidence intervals are the corresponding saved marginal intervals.

Quantity	$\gamma_{\text{edit}} = -0.5$	$\gamma_{\text{edit}} = 0.5$
Implementations within projection tolerance	256/256	256/256
Exact numerical target and norm-controlled	132/256	72/256
Groups with at least two strict implementations	31/32	18/32
Eligible groups with different effects	23/31	12/18
Mean all-group effect range	0.018066	0.016113
95% interval	$[0.015625, 0.020996]$	$[0.011719, 0.020020]$
Mean strict effect range	0.012601	0.012153
95% interval	$[0.009577, 0.015625]$	$[0.007813, 0.016515]$
Mean strict half-range	0.006300	0.006076
95% interval	$[0.004788, 0.007813]$	$[0.003906, 0.008247]$
Largest candidate-target error, all runs	8.9407×10^{-7}	9.9838×10^{-7}
Largest norm error, all runs	8.9206×10^{-5}	8.8253×10^{-5}
Mean final total perturbation norm	0.934487253	0.934485672

The strict negative-amplitude mean range uses 31 groups; the positive-amplitude mean uses 18. They are not estimates over an identical eligible subset, and their difference is not tested as an amplitude comparison. Likewise, the 49 eligible prompt–amplitude conditions are not 49 independent prompts. Confidence intervals resample eligible problem groups for each amplitude separately. Eligibility is determined by numerical target and norm conditions, not by requiring a nonzero effect, but the statistics remain conditional on numerical success.

The equal-reading endpoints leave the activation orthogonal complement unspecified, and matching perturbation norms do not fix the remaining activation state. The result therefore demonstrates non-uniqueness under the tested partial specification. It neither contradicts preservation through the complete constructed state nor excludes a richer retained state that determines the intervention effect. Each endpoint is constructed from its corresponding source activation and is therefore certified only for that source state.