Module Specialization in Transformer Factual Recall under Correlated Fact Distributions

Dong-Jae Lee **Junmo Kim**
KAIST, South Korea
{jhtwosun, junmo.kim}@kaist.ac.kr

Abstract

Recent studies have investigated how attention and MLP modules in Transformers store and recall individual facts. Real-world facts, however, exhibit shared attributes across subjects, overlap between relations, and uneven answer frequencies. To investigate how these structures affect factual recall, we conduct controlled experiments in two-layer Transformers, separately varying shared answer patterns across subjects, shared answers across relations, and concentration of answer frequencies on a small set. We measure how accurately linear classifiers predict answers from intermediate representations and how interventions on module outputs change the model's predictions. Under strong subject sharing, answers can be decoded almost as accurately before the final MLP as after it. In contrast, concentrating answer frequencies on a small set leaves a larger improvement in decoding accuracy after the final MLP. In separate intervention evaluations, altering the first attention module substantially changes answer predictions across independent, subject-sharing, and answer-concentration conditions. Additional interventions and a conditional cost analysis suggest a possible attention–MLP pathway for exploiting shared answer patterns across subjects. A comparison of three natural factual relations in pretrained language models also reveals relation-dependent differences in an intervention-based localization measure. These findings show that factual structure shapes patterns of answer decodability, while module interventions provide a complementary view of the computations underlying recall.

1 Introduction

Transformers can store factual associations in their parameters and retrieve them from subject–relation queries. Recent research has examined how attention and MLP modules support this ability. MLPs can act as key–value memories, while attention can extract relevant attributes and participate in storing factual associations [7, 8]. Associative-memory analyses further show that attention value matrices and MLPs can provide alternative ways to store facts in simplified settings [14]. These findings motivate a complementary question: *how does the structure of the facts being learned affect the roles these modules appear to play?*

Facts are rarely independent. Different subjects may share patterns of attributes, different relations may return the same answer for a given subject, and some answers may occur far more frequently than others. These structures can simplify factual recall in different ways. If several subjects share an entire answer pattern across relations, the model may be able to exploit that common pattern. By contrast, a small set of frequent answers reduces output diversity without necessarily creating shared patterns across subjects. Both structures introduce repetition, but they need not favor the same representations or computations. Understanding their effects therefore requires distinguishing what is shared, rather than treating all departures from independent facts as equivalent.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

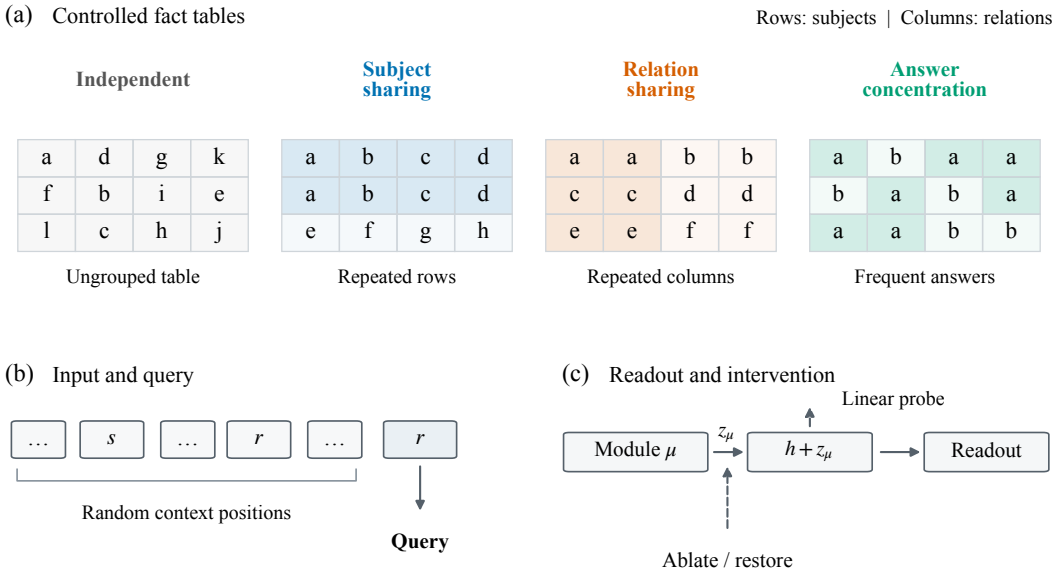

Figure 1: **Task structure and measurement sites.** Structured assignments create shared answer patterns across subjects, shared answers across relations, or frequent answer anchors; the correlation parameter controls the fraction assigned by each rule. The input repeats the contextual relation at the final query. Probes read post-module residuals; restoration replaces query-position module contributions before downstream computation. Fact tables are illustrative.

We investigate this question using a controlled synthetic factual-recall task adapted from **(author?)** [14]. We train two-layer Transformers on fixed subject–relation–answer tables and separately vary three forms of structure: shared answer patterns across subjects (**subject sharing**), shared answers across relations for each subject (**relation sharing**), and answer frequencies concentrated on a small set (**answer concentration**). This setting lets us manipulate the structure of the learned facts while retaining the same basic recall task (Figure 1). It isolates specific forms of statistical dependence rather than attempting to reproduce the full complexity of natural knowledge.

We analyze recall from two complementary perspectives. First, we train linear classifiers on intermediate representations to measure how accurately an answer can be decoded at each stage. Second, we intervene on module outputs and measure changes in the model's predictions. These measurements address different questions: a representation may make an answer easy to read without revealing which computation produced it, while a module intervention measures sensitivity to a particular change in the computation. This distinction is established in interpretability research, including direct comparisons of decodability and intervention effects [10, 11]. Here, we use it to examine how factual structure changes module-level measurement profiles.

Our clearest finding is a contrast between subject sharing and answer concentration. Under strong subject sharing, answers can be decoded almost as accurately before the final MLP as after it. Concentrating answer frequencies on a small set, however, leaves a larger improvement in decoding accuracy after that MLP. Thus, sharing answer patterns across subjects and repeatedly using the same answers produce different representational profiles. In separate intervention evaluations, the first attention module has the largest ablation and restoration effects across the independent, subject-sharing, and answer-concentration conditions tested. The small final-MLP decoding improvement under subject sharing therefore coexists with a substantial effect of intervening on early attention; it does not establish that the final MLP is dispensable.

To explore possible explanations, we examine how changes to a source module propagate through selected receiver inputs. These interventions, together with a conditional construction-cost analysis, motivate a candidate attention–MLP pathway that could exploit shared answer patterns across subjects. We also examine the scope of the observations through probe-fitting controls and width and depth comparisons. Finally, an evaluation of three natural factual relations in four frozen pretrained language models reveals relation-dependent differences in an intervention-based localization measure.

This natural-data comparison provides complementary evidence of variation across relations, while remaining distinct from the controlled synthetic manipulations.

Our contributions are a controlled comparison of three forms of factual structure, an analysis connecting their decoding profiles to module-intervention measurements, and an investigation of candidate pathways and their scope. Together, the results show that factual structure is an important axis for interpreting module specialization: what becomes readable at a given stage depends on how facts are shared, and understanding that pattern requires examining how the model's predictions respond to interventions.

2 Task and measurements

2.1 Controlled fact distributions

We represent facts as a table whose rows are subjects, columns are relations, and entries are answers. Sharing rows, sharing entries across columns, and concentrating the answer frequencies are distinct ways to introduce repetition. Let $\mathcal{S} = [n_s]$, $\mathcal{R} = [n_r]$, and $\mathcal{A} = [M]$ denote subjects, relations, and answers. A fixed fact map $f^* : \mathcal{S} \times \mathcal{R} \to \mathcal{A}$ contains $N = n_s n_r$ entries. We first sample this table, then repeatedly sample sequences querying its entries. Thus, the experiment tests recall of a learned table under new sequence realizations, rather than generalization to unseen facts.

For independent facts, each table entry is sampled uniformly from $\mathcal{A}$. The structured families use a class count or anchor-pool size K and a correlation probability c. At table construction, each entry follows the structured rule with probability c and otherwise receives an independent uniform answer. The default is $c = 0.9$.

- **Subject sharing:** subjects in the same class share a relation-indexed answer pattern. We partition subjects into K_s classes and draw one relation-indexed answer block $B_g \in \mathcal{A}^{n_r}$ per class, sampling its entries without replacement. The structured assignment is $f^*(s, r) = B_{\mathrm{cls}_{K_s}(s)}[r]$.
- **Relation sharing:** relations in the same class return the same answer for each subject. We partition relations into K_r classes and draw K_r distinct answers A_s without replacement for each subject. The structured assignment is $f^*(s, r) = A_s[\mathrm{cls}_{K_r}(r)]$.
- **Answer concentration:** structured entries draw from a common set of frequent answers, without requiring shared row or column patterns. We draw an anchor pool $\mathcal{P} \subseteq \mathcal{A}$ of size K_a and sample each structured entry independently from $\mathcal{P}$.

Smaller K_s or K_r means fewer input classes; smaller K_a means fewer frequent output anchors. These are different quantities. Answer concentration does not explicitly impose shared rows or columns, although repeated outputs can still make learning easier. Exceptions are sampled once into the fact table, not resampled as label noise during training. Full-class subject and relation references retain the within-row exclusion induced by sampling without replacement, so neither is identical to the independent-entry baseline.

Each length-16 input contains the subject and relation at distinct random non-final positions, with the relation repeated at the last position. The remaining positions contain noise tokens. Prediction is made at the last position. This implementation differs from the final EOS query in **(author?)** [14]; it also uses a common answer vocabulary rather than requiring disjoint answer sets across relations. In particular, subject and relation are not symmetric inputs in our architecture. Any explanation of a subject–relation difference must account for this query convention.

2.2 Models and fitted probes

The default is a two-layer, width-128 pre-LayerNorm transformer with unmasked bidirectional attention, four heads, fourfold-expansion ReLU MLPs, and attention temperature one. We use $n_s = n_r = 128$, $M = 256$, and 64 noise identities. Let $z_\mu(x)$ be module μ's contribution at the query, and $h_\mu(x)$ the residual immediately after that contribution is added. The four modules are $L_1^{\mathrm{attn}}, L_1^{\mathrm{mlp}}, L_2^{\mathrm{attn}}, L_2^{\mathrm{mlp}}$.

A fitted linear probe $(\widehat{W}_\mu, \widehat{b}_\mu)$ predicts an answer from h_μ. Its macro accuracy is

$$\widehat{P}_\mu = \frac{1}{|\mathcal{A}_{\text{eval}}|} \sum_{a \in \mathcal{A}_{\text{eval}}} \frac{1}{n_a} \sum_{i: y_i = a} \mathbf{1}\Big[\arg\max(\widehat{W}_\mu h_\mu(x_i) + \widehat{b}_\mu) = a\Big], \tag{1}$$

where $\mathcal{A}_{\text{eval}}$ contains the answer classes observed in evaluation. This is the accuracy of a trained readout, not an oracle maximization over all linear classifiers. During backbone training, we also fit probes to detached intermediate representations (joint-probe fitting), so the probe losses do not backpropagate into the backbone. A frozen-backbone control in Appendix C.1 checks sensitivity to this fitting procedure.

For two-layer models, we summarize the last MLP's probe increment by

$$\Delta_2 = 100\big(\widehat{P}_{L_2^{\text{mlp}}} - \widehat{P}_{L_2^{\text{attn}}}\big) \quad \text{percentage points (pp).} \tag{2}$$

A small Δ_2 means that the two fitted readouts have similar accuracy before and after the final MLP. It does not by itself imply that either accuracy is high, nor that the MLP is dispensable. Because these are consecutive residual states, the increment measures the change in fitted decodability across this module, not its causal contribution to the model's prediction. Residual updates can cancel or transform previously readable features. The architecture alone therefore guarantees neither increasing probe accuracy with depth nor a particular ordering of probe and intervention effects.

2.3 Ablation and restoration

Mean ablation estimates a query-position mean $\bar{z}_\mu$ and broadcasts that vector to replace the selected module's outputs at *all* token positions. Writing Z_μ for the full sequence of outputs and $\bar{Z}_\mu$ for this broadcast replacement, we report

$$D_\mu = 100\left(\text{Acc}(f) - \text{Acc}(f_{Z_\mu \leftarrow \bar{Z}_\mu})\right), \tag{3}$$

using example-weighted task accuracy. Positive values mean an accuracy decrease. Downstream activations are recomputed. This perturbation tests sensitivity to a particular replacement distribution and spatial scope; early ablations can affect many later computations. Baseline and ablation accuracy use separate sequence draws from the same fixed table.

Restoration uses a clean example x with answer y and a corrupt example x' sharing the relation and noise realization but differing in subject and answer. We run x', replace only $z_\mu(x')$ with $z_\mu(x)$, and recompute subsequent modules. The effect is

$$R_\mu = 100\mathbb{E}_{(x,y,x')}\left[p_f\big(y \mid x'; z_\mu \leftarrow z_\mu(x)\big) - p_f(y \mid x')\right]. \tag{4}$$

Patched top-1 accuracy to the clean answer is reported separately. Both quantities average over accepted pairs; they are not macro probe accuracy. Replacing a whole post-module residual would be a different intervention and is not the operation in Eq. (4).

Ablation and restoration therefore differ in spatial scope as well as in replacement and outcome: the former changes every position, the latter only the query. Probing asks a representational question. We compare their profiles without treating any one of them as a complete measure of a module's role. Receiver-input interventions and their residual-route conventions are defined in Section 4 and Appendix A.

3 How factual sharing changes module measurements

3.1 Shared answer patterns differ from frequent answers

Does repeating entire answer patterns have the same effect as repeating individual answers? At fixed width $d = 128$ and $N = 16{,}384$, subject sharing produces the smallest final-MLP probe increments among the strong-sharing and concentrated-answer settings in Table 1. At $K_s = 2$, the increment is approximately 0.4 pp; the answer-pool setting $K_a = 4$ retains an increment of 17 pp. Thus, the final MLP changes fitted decodability much less when subjects share answer patterns than when answers are concentrated on a small pool. The absolute accuracies and intervention results below help interpret this contrast; Section 4 explores possible computations behind it.

Table 1: Layer-2 macro probe increments Δ_2 (pp), at $d = 128$, $N = 16{,}384$, after 8,000 steps in seed 0. Columns list the subject-class, relation-class, or answer-pool size. Reference settings with the full number of classes or answers differ in sampling constraints and final task accuracy. No seed uncertainty can be estimated from this sweep.

	Subject		Relation		Answer	
Setting	K_s	Δ_2	K_r	Δ_2	K_a	Δ_2
Full classes/pool	128	43.07	128	38.37	256	37.59
Moderate	16	0.98	16	11.98	32	22.28
Strong	4	0.40	4	1.49	8	21.65
Extreme	2	0.42	2	5.89	4	17.07

The relation increment is nonmonotone, rising from 1.5 pp at $K_r = 4$ to 6 pp at $K_r = 2$. The relation token's special query position and the independent exceptions are plausible contributors, but this comparison alone does not separate them. More generally, equal numerical values of K_s, K_r, and K_a do not match the effective difficulty or entropy of the three distributions.

3.2 Early attention remains influential across distributions

We next ask how the decoding profiles relate to the effects of changing module outputs. Figure 2 compares frozen-backbone probes, ablation, and restoration for independent facts, subject sharing ($K_s = 2$), and answer concentration ($K_a = 4$). The panels use separately trained cohorts and therefore compare protocol-level profiles, not identical checkpoints. First-attention probe accuracy averages 8.7%, 84.5%, and 47.9% for independent facts, subject sharing, and answer concentration, respectively; after the final MLP, the corresponding accuracies are 94.7%, 99.8%, and 98.6%.

Mean ablation produces the largest accuracy decrease at L_1^{attn} and the smallest at L_2^{attn} in all three evaluated settings. First-attention drops are approximately 97.5, 53.4, and 67.7 pp for independent facts, subject sharing, and answer concentration, respectively. Thus, the ranking is shared across these settings, but the magnitude varies.

The single-seed restoration evaluation shows the same ordering of the largest and smallest effects. Replacing L_1^{attn}'s query contribution increases clean-answer probability by 87.6, 97.8, and 82.6 pp for independent, subject, and answer data. At L_2^{attn}, the increases are 1.7, 2.0, and 5.5 pp. This restoration comparison uses one training seed, so it does not establish the seed stability of the ranking. Appendix B gives the complete numeric results and distinguishes the replication units.

These observations are consistent with early attention supplying information that becomes readable in later residual states, although the measurements do not uniquely identify this computation. In particular, the small final-MLP increment under subject sharing and the large first-attention restoration effect describe different aspects of recall: the former concerns readability across a late residual update, while the latter concerns sensitivity to an early module replacement. The independent condition also exhibits strong first-attention effects, so the general probe/intervention distinction should not be attributed exclusively to correlation. A relation-sharing intervention ranking, and a fully paired comparison of probes and interventions across the same checkpoints, remain outside the evidence reported here.

3.3 The contrast depends on model size and training conditions

We vary width at fixed N using the load proxy $\rho = N/d^2$. Table 2 shows that the subject increment remains small at $d = 64$ and 128, whereas the answer increment decreases substantially with width. Because depth, embeddings, and MLP dimensions also contribute to parameter count, ρ describes this width sweep rather than a universal capacity transition.

The depth sweep provides a related, but distinct, observation. At $d = 128$ and one layer, task accuracy is already 99.50% for independent facts, 99.55% for subject $K_s = 2$, and 95.40% for answer $K_a = 4$ in the seed-0 evaluation. These results do not support a claim that the answer or independent tasks require two layers. Differences at larger depths concern the tested optimization procedure and probe profiles, not an architectural necessity. The four-layer probe trajectories also differ by distribution;

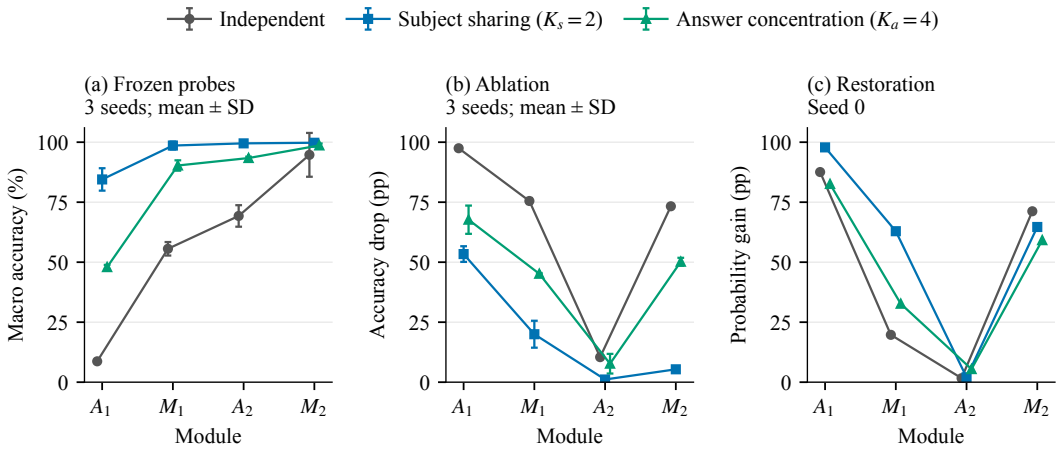

Figure 2: **Probe and intervention profiles in separate experimental cohorts.** Axis labels use $A_\ell = L_\ell^{\text{attn}}$ and $M_\ell = L_\ell^{\text{mlp}}$. (a) Frozen-backbone linear-probe macro accuracy. (b) Decrease in example-weighted accuracy after mean ablation replaces the selected module's outputs at all token positions with its query-position mean. (c) Query-only clean-answer restoration R_μ on 200 accepted pairs in seed 0. Panels (a,b) show means and sample standard deviations across three seeds; panel (c) has no seed uncertainty estimate. Checkpoints, spatial intervention scopes, and aggregation differ across panels. Relation sharing is not included.

Table 2: Macro probe increments Δ_2 (pp) in the single-seed width sweep, at fixed $K_s = 2$, $K_r = 4$, or $K_a = 4$, with 8,000 training steps. The load proxy keeps $N = 16{,}384$ fixed. No seed uncertainty is available for these probe comparisons; a separate intervention evaluation finds increased variability at larger widths.

d	N/d^2	Subject $K_s = 2$	Relation $K_r = 4$	Answer $K_a = 4$
64	4.00	0.76	14.77	37.68
128	1.00	0.42	1.49	17.07
256	0.25	2.40	2.52	15.96

under subject sharing, answers are substantially, but not perfectly, decodable after the first attention module (Appendix C.2).

A separate multi-seed evaluation of first-attention restoration finds substantial variability at $d \geq 256$, with standard deviations reaching approximately 41 pp in a $d = 512$ setting. This is variability in a patching effect, not in Δ_2. The strongest evidence for stable first-attention restoration effects under subject sharing therefore comes from the tested narrow models. The natural-language comparisons evaluate a different setting and do not establish stability across widths.

4 Candidate pathways and conditional explanations

4.1 A construction-cost hypothesis

How might shared answer patterns become readable before the final MLP? One candidate computation groups subjects before decoding a relation-conditioned answer. Another retains individual subject–relation distinctions until a later lookup. These are possible implementations of the structured table, not an exhaustive list of transformer algorithms.

In the deterministic limit $c = 1$, subject sharing leaves $K_s n_r$ table entries once the class map is known. A schematic representation cost can therefore be written as

$$C_{\text{group}} = \alpha K_s d + \beta K_s n_r, \qquad C_{\text{lookup}} = \gamma n_s n_r, \tag{5}$$

for positive, implementation-dependent constants. The first term represents a class-based interface of width d; the second counts class–relation associations. These expressions are a bookkeeping model.

They omit class assignment, token embeddings, routing, precision, and nonlinear decoding, and are not proven upper and lower bounds for the trained transformer.

Setting the constants equal suggests a comparison scale $K_s \sim n_s n_r/(d + n_r)$, which equals 64 for the default dimensions. We use this only as an order-of-magnitude hypothesis. It neither fixes a transition at 64 nor predicts the optimum of AdamW. For $c < 1$, exceptions may require retaining individual identities; their costs cannot simply be assumed to cancel between implementations. Relation grouping can also reduce table size, and our asymmetric query placement prevents a purely axis-independent capacity argument.

Optimization hypothesis (A1). When candidate implementations attain similar loss, training may favor one with a more compact effective representation. This is an empirical hypothesis, motivated broadly by implicit-bias results in simpler models [9, 12, 17]. Parameter count, rank, and weight norm are different quantities; a reduction in one does not prove a reduction in the others. Associative-memory constructions in **(author?)** [14] establish sufficiency under their stated assumptions, not the MLP lower bound or transformer optimization preference required to turn Eq. (5) into a theorem.

4.2 Receiver-input restoration

Module restoration changes a source module's output and allows the change to propagate through subsequent computations. To examine how a selected receiver uses that output, we restore the source contribution within the receiver's input and measure the resulting recovery of clean-answer predictions. We denote this effect by $R_{u \rightsquigarrow v}$ to distinguish it from a uniquely isolated graph edge. The implementation matters: later computations are recomputed, and for a final-MLP receiver the replaced source remains in the residual sent to the output head. The effect can therefore include a direct residual route as well as a change mediated by the receiver.

The subject setting gives a large $L_1^{\mathrm{attn}} \rightsquigarrow L_1^{\mathrm{mlp}}$ effect, approximately 63.2 pp against a 97.3 pp source-module baseline. Restoring L_1^{mlp} at the second MLP's input gives approximately 58.9 pp against a 63.2 pp module baseline. These results motivate the candidate sequence $L_1^{\mathrm{attn}} \to L_1^{\mathrm{mlp}} \to L_2^{\mathrm{mlp}}$. Because the restored contribution also follows the residual route, the measurements do not isolate the part of the latter effect mediated by L_2^{mlp}. The full receiver comparison appears in Appendix B.1.

For independent facts, the sum of the measured first-attention receiver effects falls approximately 62.8 pp below the module effect. The corresponding subject shortfall is approximately 17.1 pp. Such discrepancies show that the measurements are not an additive decomposition. They may reflect interactions and differences in intervention scope; they do not, without further ablations, identify LayerNorm or softmax as the cause of non-additivity.

We summarize two receiver-to-module comparisons with the descriptive index

$$I = \frac{R_{L_1^{\mathrm{attn}} \rightsquigarrow L_1^{\mathrm{mlp}}}}{R_{L_1^{\mathrm{attn}}}} \frac{R_{L_1^{\mathrm{mlp}} \rightsquigarrow L_2^{\mathrm{mlp}}}}{R_{L_1^{\mathrm{mlp}}}}. \tag{6}$$

The fixed-budget subject sweep shows a sharper increase than the answer sweep over the tested class and pool sizes (Figure 3c). Convergence also varies: one subject $K_s = 128$ run has only 71% clean accuracy. Thus, the entire sweep is not a comparison of equally successful solutions, and a sharp change cannot by itself validate A1 or the cost scale. Multiplication does not turn separately measured effects into a serial causal fraction: I need not be bounded by one in general. A direct joint intervention and a receiver-only implementation that removes unintended skip routes would be needed for a stronger path-specific interpretation.

4.3 Complementary checks and their scope

Hypothesis-consistent activation resampling provides a complementary test [3]. The reported subject result preserves $93.7 \pm 2.9\%$ accuracy under structured resampling, compared with $7.6 \pm 2.0\%$ under random resampling. This supports the utility of the chosen equivalence classes relative to that null. Since the final sampler conditions on the answer and all selected activations are replaced together, it does not establish that the proposed sequence is the unique algorithm. Appendix D states the protocol and its limitations.

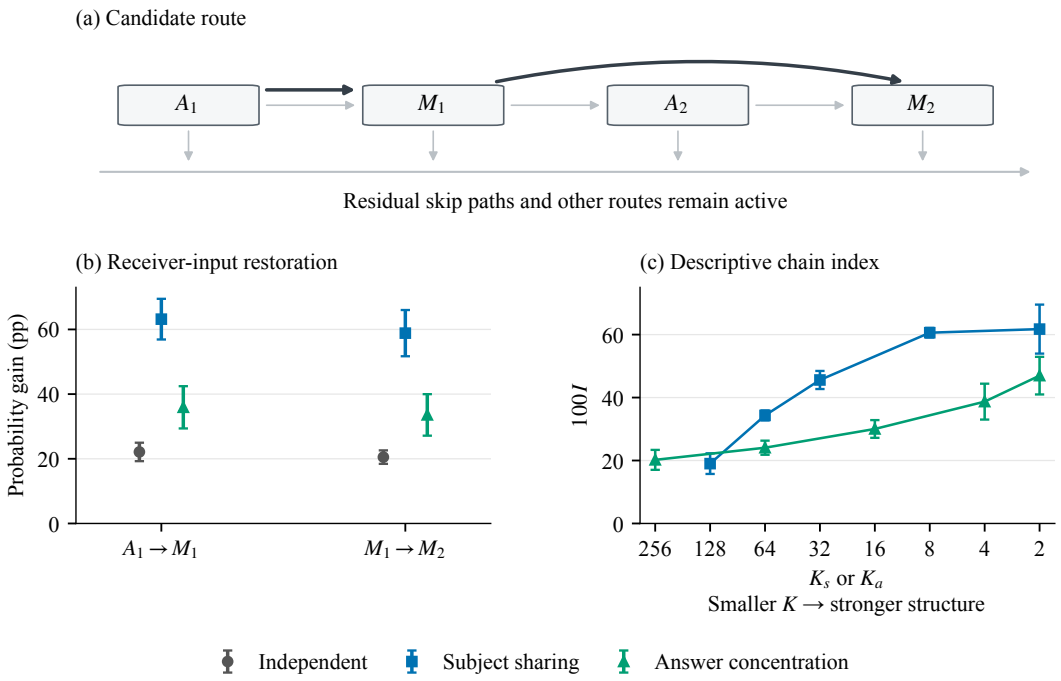

Figure 3: **Testing a candidate pathway.** (a) The attention–MLP–MLP sequence is a hypothesis about information flow. (b) Receiver-input restoration effects for independent, subject $K_s = 2$, and answer $K_a = 4$ data: mean and sample standard deviation across three seeds, 200 pairs per seed. Labels A_ℓ, M_ℓ follow Figure 2; arrows denote receiver-input restoration, including downstream and residual routes. (c) The descriptive index I in Eq. (6) across subject-class and answer-anchor sweeps, computed within each seed before aggregation. It is neither a joint-patch effect nor an identified fraction of causal influence. Smaller class counts strengthen sharing; smaller pools concentrate answers; the two axes represent different manipulations.

A combined subject/answer sweep also suggests that the two manipulations do not simply add. With a sufficiently broad answer pool, stronger subject sharing reduces the final probe increment; at a very small answer pool, a nonzero increment persists. This descriptive interaction does not establish which compression mechanism dominates. Details and the limits of the available summaries are in Appendix C.3.

5 Factual structure in pretrained language models

The controlled experiments establish a contrast between shared subject patterns and concentrated answers in small models. We extend the analysis to pretrained models through two complementary evaluations. Fine-tuning Pythia-160M on synthetic identities retains control over the fact table but changes the backbone and tokenizer (Appendix E.1). Evaluating natural facts in frozen models removes that training intervention but also removes control over the data-generating distribution. Neither evaluation directly replicates the chain-index analysis in the controlled models.

Natural relations and selection. We compare three CounterFact relations [13]: mother tongue (P103), continent (P30), and place of birth (P19). These were used as subject-sharing, concentrated-answer, and comparatively diverse-answer proxies. We retain their relation names in the results because a relation-level sample is not an intervention on a generative structure. In particular, low empirical $H(A \mid S)$ is uninformative about subject sharing when most sampled subjects appear only once.

For each frozen model, the evaluation is restricted to correctly answered facts and includes a search over candidate source–receiver layer pairs. The reported peak edge/module ratio is the maximum ratio of the edge effect to its corresponding source-module effect across these pairs, excluding

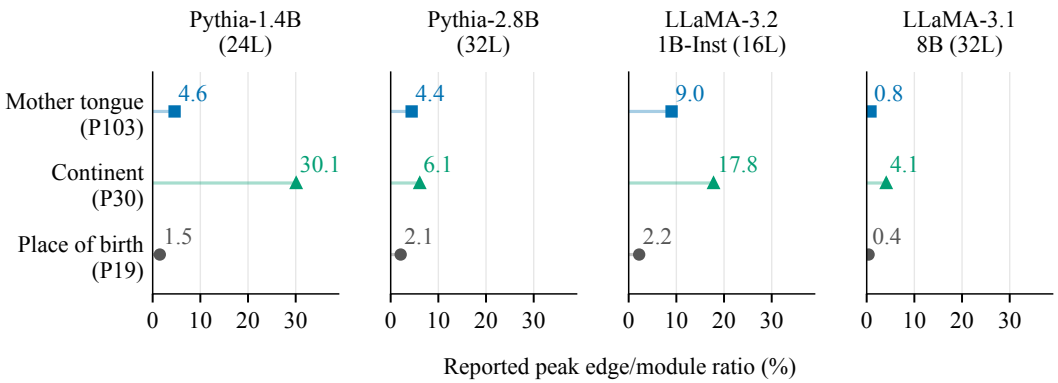

Figure 4: **Relation-level localization in frozen pretrained models.** Reported peak edge/module ratios for mother tongue (P103), continent (P30), and place of birth (P19), with selected source layers listed in Table 7. Each point is a reported aggregate from a layer search with module effects below 1 pp excluded. The available summary does not support seed-level error bars or relation-level confidence intervals. Relations are empirical proxies, not controlled realizations of the synthetic distributions.

source-module effects below 1 pp. This is a selected localization statistic on model-known facts. The selection and denominator cutoff can change the value, so it is not an unbiased estimate of how much of all factual recall resides in one edge.

Results and external scope. Continent has the largest reported peak ratio in each of the four models: 30.1% in Pythia-1.4B, 6.1% in Pythia-2.8B, 17.8% in LLaMA-3.2-1B-Instruct, and 4.1% in LLaMA-3.1-8B (Figure 4). Mother tongue and place of birth have smaller ratios in these comparisons. This is consistent with concentrated answers yielding a more localized signal under the evaluated protocol, but three relations do not isolate answer concentration from relation semantics, prompt structure, or selection effects.

The two larger models have smaller continent ratios than the smaller model in the same family. This is a descriptive model comparison rather than a causal scaling law; the LLaMA comparison also changes model version and instruction-tuning status. A fixed source-layer choice can change the apparent ordering, further motivating full layer profiles and held-out layer selection. Appendix E.2 lists the values and methodological limits. The synthetic fine-tuning result supports a similar relation between answer concentration and this particular localization statistic, whereas the chain index in the controlled models is higher under subject sharing. Because these statistics measure different aspects of restoration, the findings do not establish a common ordering of factual structures across settings.

6 Discussion and limitations

The central contrast is between sharing answer patterns and concentrating answer frequencies. Under strong subject sharing, the representation before the final MLP already supports nearly the same decoding accuracy as the representation after it; answer concentration leaves a larger improvement across that module. Repetition in a fact table therefore cannot be characterized solely by the number of frequent answers. Separate intervention evaluations reveal strong first-attention effects in both settings and in independent data, providing a complementary view of the computations supporting recall. A probe profile alone does not locate a computation, and a restoration profile alone does not identify its algorithm.

The first set of limitations concerns what the tasks and measurements represent. The asymmetric relation query and duplicated contextual relation constrain subject–relation comparisons. Fitted probes depend on optimization and macro averaging, while interventions use different example or pair weighting. Receiver interventions include residual and downstream routes, making strong path-specific claims premature.

The second concerns replication and coverage. Some aggregate sweeps lack recoverable per-cell uncertainty, the principal restoration comparison has one seed, and the width evaluation reveals substantial variability. The natural-data comparison uses selected known facts, three proxy relations, peak layer ratios, and a denominator cutoff. Broader relation coverage, matched checkpoint evaluations, and explicitly isolated joint interventions would strengthen the conclusions. An exploratory two-hop result is included in Appendix E.3; it does not distinguish sequential retrieval from a learned flat lookup.

7 Related work

Factual recall and associative memory. Prior work links MLP representations to factual associations and studies the stages of recall [7, 8, 13]. Associative-memory analyses provide constructions and capacity results under explicit assumptions [1, 2, 14]. Our question concerns how the structure of the fact table changes module-level observations: shared patterns across subjects, shared answers across relations, and frequent answers need not have the same effect. We investigate these distinctions experimentally, while separating table compression from a proof about a trained transformer's parameter requirements.

Representations, interventions, and circuits. Probe methodology distinguishes readable information from the capacities and biases of the readout [10, 16, 18]. Work directly comparing decodability and causal effects reinforces this distinction [11]. Activation patching, circuit discovery, and causal scrubbing study intervention-dependent explanations [3, 5, 20]. Recent studies distinguish structural from functional circuit similarity and examine causal communication and circuit completeness [4, 6, 15]. We combine probes and interventions to examine how controlled changes in factual structure affect decodability and module influence. Analyses of two-hop retrieval likewise motivate testing an explicit computational hypothesis before attributing a failure or recovery to a particular stage [19].

8 Conclusion

Different forms of factual repetition produce different decoding profiles in two-layer Transformers. When subjects strongly share answer patterns, answers are nearly as decodable before the final MLP as after it; concentrating answer frequencies leaves a larger improvement across that module. Separate interventions show that early attention remains influential in the evaluated independent, subject-sharing, and answer-concentration conditions. Receiver-input tests and a conditional cost model suggest a possible computation that exploits shared subject patterns, while leaving its uniqueness and optimization origin unresolved. The three-relation comparison in frozen language models also reveals variation in selected localization statistics, without isolating its cause. Together, these results motivate interpreting module specialization in relation to both the structure of learned facts and the distinct questions answered by probes and interventions.

Acknowledgements

This work was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea) & Gwangju Metropolitan City.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *International Conference on Learning Representations (ICLR)*, 2025.

[2] Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, 2023.

[3] Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A

method for rigorously testing interpretability hypotheses. *Alignment Forum / Redwood Research Tech Report*, 2022.

[4] Hang Chen, Jiaying Zhu, Xinyu Yang, and Wenya Wang. Rethinking circuit completeness in language models: AND, OR, and ADDER gates. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[5] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *NeurIPS*, 2023.

[6] Gabriel Franco and Mark Crovella. Pinpointing attention-causal communication in language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[7] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *EMNLP*, 2023.

[8] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *EMNLP*, 2021.

[9] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *ICML*, 2018.

[10] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP*, 2019.

[11] Lianghuan Huang and Yingshan Chang. Causality $\neq$ decodability, and vice versa: Lessons from interpreting counting ViTs. In *NeurIPS 2025 Workshop on Interpreting Cognition in Deep Learning Models (CogInterp)*, 2025. Workshop paper; not a NeurIPS main-conference paper.

[12] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *ICLR*, 2020.

[13] Kevin Meng, David Bau, Alex J. Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

[14] Eshaan Nichani, Jason D. Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. In *International Conference on Learning Representations (ICLR)*, 2025.

[15] Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in VLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[16] Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *ACL*, 2020.

[17] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19:1–57, 2018.

[18] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *ACL*, 2019.

[19] Constantin Venhoff, Ashkan Khakzar, Sonia Joseph, Philip H. S. Torr, and Neel Nanda. Too late to recall: Explaining the two-hop problem in multimodal knowledge retrieval. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[20] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *ICLR*, 2023.

A Implementation and evaluation conventions

A.1 Data, optimization, and evaluation units

The synthetic table is fixed within a run. Training and evaluation sample its entries with freshly generated noise tokens and random non-final subject/relation positions. The relation also appears at the final query. The small-model implementation uses unrestricted self-attention over the sequence and predicts only at this query; the task contains no answer token in the input. Learned token and positional embeddings feed pre-LayerNorm attention and ReLU MLP blocks, followed by a linear output head.

The central module-intervention runs use $8{,}000$ training steps, batch size 128, and backbone learning rate 10^{-3} with AdamW. Their configurations specify a probe learning rate of 5×10^{-3} where joint probes are fitted. Probe inputs are detached. A probe optimizer cannot change the backbone through this detached path; differences between joint and frozen protocols can nevertheless arise from readout optimization, checkpoints, or evaluation sampling. We do not attribute those differences to a demonstrated change in backbone geometry.

Probe macro accuracy averages over answer classes present in the evaluation sample. Module-ablation tables use overall, example-weighted accuracy, even though the evaluation code also records macro accuracy. Counterfactual restoration averages clean-answer probabilities over accepted pairs; its top-1 endpoint is also pair-weighted. Frequent anchors can therefore contribute differently to these metrics. Seed standard deviations, when available, describe variability across training runs; they are not confidence intervals over facts.

A.2 Exact intervention locations

For module restoration, the clean run supplies only the query-position output of the selected module. All other positions retain their corrupt activations at the intervention site, and subsequent modules run normally. A clean/corrupt pair shares relation, noise tokens, and subject position; pairs with identical subjects or answers are rejected. The reported main module-restoration evaluation uses 200 accepted pairs for each distribution in one seed. The mean-ablation and receiver-input comparisons have three seeds.

For $L_1^{\mathrm{attn}} \rightsquigarrow L_1^{\mathrm{mlp}}$, the source's clean contribution is inserted into the first MLP's query input; the updated MLP output is then added to the residual containing the corrupt attention contribution, and downstream modules are recomputed. For an attention receiver, the query input to that receiver is changed and its output feeds the later computation. For $u \rightsquigarrow L_2^{\mathrm{mlp}}$, the implementation instead retains the modified receiver input in the final residual before adding the new MLP output. Consequently, the clean source also reaches the output head directly. This last operation is a receiver-input restoration with a skip route, not a receiver-mediated edge effect alone.

The "direct" control changes only the source contribution in the final residual, without recomputing either second-layer module. Module baselines restore the source once and let every later computation respond. These conventions explain why the sum of receiver-input effects need not equal the module effect, even before considering nonlinear interactions. A strict path-specific mediation analysis would require additional intervention definitions and controls.

B Complete module and receiver results

Table 3: Mean-ablation accuracy decreases (pp). Entries are rounded means across three seeds, with columns identifying the ablated module; positive values indicate an accuracy decrease. Figure 2b shows variability across seeds.

Distribution	L_1^{attn}	L_1^{mlp}	L_2^{attn}	L_2^{mlp}
Independent	97.5	75.5	10.4	73.3
Subject $K_s = 2$	53.4	20.0	1.1	5.4
Answer $K_a = 4$	67.7	45.2	7.7	50.2

Table 4: Single-seed module restoration, 200 accepted pairs per distribution. The two blocks separate clean-answer probability gain (pp) from patched top-1 clean-answer accuracy (%). These seed-0 measurements do not provide a between-seed uncertainty estimate.

Distribution	L_1^{attn}	L_1^{mlp}	L_2^{attn}	L_2^{mlp}
Probability gain (pp)				
Independent	87.6	19.7	1.7	71.2
Subject $K_s = 2$	97.8	62.9	2.0	64.7
Answer $K_a = 4$	82.6	32.7	5.5	59.2
Patched accuracy (%)				
Independent	91.0	22.5	2.0	75.0
Subject $K_s = 2$	98.5	65.0	3.0	69.5
Answer $K_a = 4$	85.5	34.0	6.0	62.5

B.1 Receiver-input comparison

Table 5 gives the source and receiver breakdown. The module baselines in this experiment need not equal those in Table 4, which uses a different set of runs. The gap column reports the sum of receiver and direct effects minus the module baseline. Because these interventions differ in scope, this statistic does not represent a residual from an additive decomposition.

Table 5: Receiver-input restoration effects (pp), three-seed means. "Direct" changes only the final residual; "Module" restores the source for downstream computation. "Gap" is the sum of the displayed receiver and direct effects minus the module effect. The MLP2 receiver includes the skip route described in Appendix A. Computing the gap from the rounded entries may give a slightly different value.

Source / distribution	$\rightsquigarrow L_1^{\text{mlp}}$	$\rightsquigarrow L_2^{\text{mlp}}$	$\rightsquigarrow L_2^{\text{attn}}$	Direct	Module	Gap
L_1^{attn} source						
Independent	22.1	6.2	0.1	0.1	91.3	−62.8
Subject $K_s = 2$	63.2	14.2	1.3	1.5	97.3	−17.1
Answer $K_a = 4$	35.9	23.0	1.6	2.1	86.0	−23.3
L_1^{mlp} source						
Independent	—	20.5	0.1	1.0	22.1	−0.5
Subject $K_s = 2$	—	58.9	0.1	14.3	63.2	10.1
Answer $K_a = 4$	—	33.6	0.5	9.1	35.9	7.2

The descriptive index I is calculated separately per seed before averaging in Figure 3c. A ratio of aggregate means would be a different estimator. The denominators must be nonzero, and ratios become unstable when module effects are small. No causal composition identity is assumed. A comparison of a joint chain intervention with a sum of individually restored edges must use the same edge set, source distribution, and denominator; otherwise the difference cannot be assigned to non-additivity of that chain.

C Probe controls, depth, and distribution sweeps

C.1 Frozen-backbone probe control

A two-layer backbone is trained for 8,000 steps using task cross-entropy, then frozen while fresh probes are trained for 4,000 steps, with three seeds per distribution. The final-MLP increments average 0.251 pp for subject $K_s = 2$, 5.249 pp for answer $K_a = 4$, and 25.464 pp for independent data. The single-seed joint-probe width sweep at $d = 128$ gives 0.42 pp for subject $K_s = 2$, 17.07 pp for answer $K_a = 4$, and a 43.07 pp full-class subject reference. The protocols use separate backbones and replication units, so their numerical differences are not paired estimates of a training-method effect. Thus, the subject-sharing versus answer-concentration contrast persists under a second probe-

fitting procedure, although its magnitude changes substantially. This comparison does not establish that every probe reached its optimum or that intervention rankings are stable.

C.2 Depth and width

The depth study varies $L \in \{1, 2, 3, 4\}$ at $d = 128$, attaching a probe after each attention and MLP residual update. In the four-layer profile, subject $K_s = 2$ is already substantially decodable after the first attention module (approximately 82% in the evaluated profile), while answer and independent trajectories rise later. The independent probe endpoint is approximately 70%. These profiles should not be conflated with the nearly converged two-layer intervention runs; comparisons must account for final task accuracy and training budget.

The width evaluation measures first-attention restoration at widths 128, 256, and 512 with three seeds. At 128, reported standard deviations are at most approximately 7 pp. At larger widths, independent and subject settings become more variable, reaching approximately 41 pp in this evaluation; the answer-pool setting at 512 remains substantially more stable (approximately 2 pp). A separate single-seed width-768 probe comparison does not establish a scaling trend. The probe and intervention statistics concern different observables and are kept separate.

C.3 Combined subject and answer structure

A reported 7×8 subject-class/answer-pool grid, with three seeds per cell, explores the interaction between input sharing and output concentration. For answer pools of at least 32, the probe increment tends toward zero as the subject class count decreases. At answer-pool size 2, a residual increment near 10 pp persists across much of the subject sweep. These observations motivate a competition hypothesis. The interaction remains descriptive because no fitted model comparison or approximation-error analysis is available. Interpreting the full grid requires accounting for task accuracy and the combined fact-table generation procedure.

C.4 Correlation strength and optimization robustness

The central reported experiments use $c = 0.9$. At $c < 1$, exceptions prevent an exact class-level factorization of the full fact map. As c changes, both the structure and the learning difficulty change, so a fixed training budget need not give comparable convergence. We do not infer a c-independent crossover from the cost model. Likewise, stability across weight-decay or initialization choices, if established, would support robustness but would not prove A1 or distinguish low norm from low rank. A full-rank initialization does not constrain the rank of a trained map, and an orthogonal initialization has condition number one when square.

D Hypothesis-consistent resampling

For the subject-sharing hypothesis, let $g = \mathrm{cls}_{K_s}(s)$ for an example (s, r, a^*). The reported sampler draws the first-attention activation from a fresh example in class g with a different relation, the first-MLP activation from the same subject-class/relation block (g, r), and the second-MLP activation from an example with answer a^*. The null draws each activation without these restrictions. All three sampled activations are applied simultaneously, and preservation is top-1 accuracy on a^*. The reported evaluation uses 200 pairs and three seeds.

For subject $K_s = 2$, clean accuracy is 99.8%, hypothesis-consistent preservation is $93.7 \pm 2.9\%$, and null preservation is $7.6 \pm 2.0\%$ (mean and standard deviation across seeds). The large difference shows that these restricted donor distributions preserve more predictive behavior than the unrestricted null. However, answer-conditioned resampling of the final activation supplies target-aligned information, and simultaneous replacements do not determine which serial edges are necessary. A more specific algorithmic claim would require component-wise and alternative-hypothesis controls.

The reported summaries also list preservation values of $73.7 \pm 4.8\%$ versus $5.9 \pm 2.0\%$ for independent data and $69.5 \pm 3.0\%$ versus $25.9 \pm 0.6\%$ for answer concentration. Their equivalence-class conventions are not specified as fully as the subject sampler. We therefore do not use them to establish one common chain mechanism across all distributions. In particular, the answer-pool null has a much higher baseline, which changes how preservation gaps should be interpreted.

E Pretrained-model analyses

E.1 Synthetic fine-tuning

Pythia-160M is fine-tuned for $12{,}000$ steps with separate body and output-head learning rates of 5×10^{-5} and 5×10^{-3}. Subject and relation identities map to fixed disjoint subsets of the pretrained tokenizer's vocabulary. This experiment tests synthetic recall in a pretrained backbone after adaptation, rather than natural factual knowledge in a frozen model. Table 6 gives the reported intervention summaries for three seeds and 200 counterfactual pairs per distribution.

Table 6: Reported synthetic fine-tuning results for Pythia-160M: mean±standard deviation across three seeds. Accuracy is in percent and restoration effects in pp. The source/receiver convention is specific to this pretrained-model experiment and is not assumed to match the receiver-input intervention in the two-layer models.

Distribution	Task acc.	L_3 module	$L_3 \to L_{11}$	L_5 module	$L_5 \to L_{11}$
Independent	86.9 ± 0.5	87.7 ± 0.5	0.9 ± 0.9	87.6 ± 0.4	0.7 ± 0.2
Subject $K_s = 2$	93.8 ± 2.0	93.2 ± 1.6	4.9 ± 1.2	93.4 ± 1.6	1.3 ± 1.1
Answer $K_a = 4$	78.4 ± 2.6	82.1 ± 3.0	4.4 ± 0.4	82.3 ± 2.9	10.5 ± 1.8

The answer setting has a larger $L_5 \to L_{11}$ effect relative to its source module than the other settings in this comparison. Task accuracy is also lower, so localization and task competence should not be conflated. Reported probe summaries show early subject decodability, with approximately 38% at the embedding, 78% by L_3, and an 88% peak. The embedding result can reflect the relation-conditioned answer marginal because the relation is already available at the query. It is not evidence that an embedding has retrieved a particular subject's answer. A terminal-layer probe decline remains uninterpreted because its tensor and normalization conventions are unresolved.

E.2 Natural-fact peak ratios

Table 7: Reported natural-fact peak edge/module ratios (%) and selected source layers. Values summarize a search over candidate layer pairs after excluding module effects below 1 pp. They describe model-known facts from three CounterFact relations; they do not measure a population fraction of all stored facts.

Model	P103: mother tongue	P30: continent	P19: birthplace
Pythia-1.4B	4.6 (L_{19})	30.1 (L_{14})	1.5 (L_{19})
Pythia-2.8B	4.4 (L_{19})	6.1 (L_{19})	2.1 (L_{19})
LLaMA-3.2-1B-Inst	9.0 (L_9)	17.8 (L_9)	2.2 (L_{12})
LLaMA-3.1-8B	0.8 (L_{19})	4.1 (L_{25})	0.4 (L_{19})

Peak selection can overstate localization relative to a prespecified layer, particularly when the denominator is small. The aggregate table does not provide the accepted fact counts per model/relation, all receiver layers, uncertainty for the selected maximum, or sensitivity to the cutoff. These missing details limit statistical and mechanistic interpretation. Assessing the stability of the peak ratios would additionally require answer-frequency distributions and known-fact counts for each retained relation, complete layer profiles, and held-out or prespecified layer selection. Comparing entropy proxies also requires repeated observations per subject; an empirical zero conditional entropy from unique subjects is not evidence for the synthetic subject-sharing distribution.

E.3 Exploratory two-hop recall

A separate Pythia-1.4B experiment fine-tunes a 24-layer model on two-hop synthetic queries and evaluates candidate layer pairs with three seeds. The reported peak receiver/module ratio is 29.0% at $L_9 \to L_{10}$, with another ratio of 19.4% at $L_{13} \to L_{14}$. These localized sensitivities motivate more targeted composition tests. They do not assign either effect to a specific hop, and they do not rule out memorization of the composed input–output map. Even a decodable intermediate entity would need hop-specific counterfactual tests to distinguish its causal use from correlated representation.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claim is accurately reflect it.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper dicuss the limitations.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provide the full set.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification:Paper shows full details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provide sufficient details.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provide sufficient details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper provide sufficient details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provide sufficient details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: yes.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: No societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not available.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM conducts and verify experiments, however core idea is based on the human research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.