Knowledge Localization and Editability in Small Language Models: A Multi-Stage Experimental Study

Pranamya Nilesh Deshpande
Togo AI Labs
pranamyadeshpande14@gmail.com

Aiswarya Konavoor
Togo AI Labs
aiswarya@togolabs.ai

Sreedath Panat
Togo AI Labs
sreedath@togolabs.ai

Abstract

The internal mechanisms by which transformer-based language models encode and retrieve factual knowledge remain poorly understood, particularly for small language models (SLMs) operating in the 2–3 billion parameter range. This paper presents a systematic, multi-stage empirical investigation into knowledge localization, compression effects, and knowledge editability across four SLMs—Gemma-2B, Llama-3.2-3B-Instruct, Qwen-2.5-3B-Instruct, and Phi-2—with Meta-Llama-3-8B serving as a large-model baseline. Stage 1 employs causal tracing with activation patching on the CounterFact dataset (∼450–500 validated facts per model) to identify the layer or layers most causally responsible for factual recall. Stage 2 compares knowledge density, layer concentration, and redundancy between the 2–3B models and the 8B baseline to quantify the structural effects of model compression on knowledge storage. Stage 3 applies the Rank-One Model Editing (ROME) algorithm at the causally identified layers to assess whether localized knowledge can be reliably overwritten. Our results demonstrate that (i) factual knowledge in SLMs concentrates in upper-to-final transformer layers, with Llama-3B exhibiting extreme concentration in layer 28; (ii) compressed models store knowledge more densely per parameter but with substantially lower redundancy (Llama-3B: 0.047 vs. Llama-8B: 0.468); and (iii) editing success correlates strongly with architectural concentration rather than model size, with Llama-3B achieving 85.7% editing success versus 33% for Gemma-2B. These findings carry direct implications for interpretability, model editing, and the design of future small language model architectures.

1 Introduction

The past several years have witnessed the rapid proliferation of large language models (LLMs) capable of storing and retrieving vast quantities of world knowledge as implicit factual associations encoded in their parameters (Brown et al., 2020). As these models are deployed in increasingly consequential settings—question answering, technical reasoning, decision support—the ability to understand *where* and *how* factual knowledge is stored has become a central concern of the mechanistic interpretability research agenda (Olah et al., 2020).

However, the majority of mechanistic interpretability work has focused on large models containing tens of billions of parameters. The behavior of small language models (SLMs)—those in the 1–4B parameter range—has received comparatively less systematic attention, despite their growing practical importance. SLMs are frequently deployed at the edge, in resource-constrained environments, and as fine-tuning targets for domain-specific applications. Whether knowledge is stored in SLMs in qualitatively similar ways to large models, and whether it is equally amenable to surgical modification, are questions with both theoretical and practical significance.

Two central challenges motivate this research. First, **knowledge localization** in transformer models is non-trivial: factual associations are not stored in a single weight matrix but emerge from complex, high-dimensional interactions among attention heads, MLP sublayers, and the residual stream (Elhage et al., 2021a). Prior work has suggested that MLP blocks in the middle-to-late layers may function as "key-value stores" for factual associations (Meng et al., 2022), but the precise distribution of this storage—and how it scales with model size—has not been systematically studied across a diverse set of SLM architectures. Second, **knowledge editing**—the ability to modify individual factual associations without broadly disrupting model behavior—depends critically on knowledge being spatially concentrated (Meng et al., 2023a). If knowledge is distributed or redundant, local edits may fail to propagate effectively.

This paper addresses both challenges through a three-stage experimental framework: (1) **Stage 1 (Knowledge Localization):** For each of four SLMs, we apply causal tracing with activation patching to identify the transformer layers causally responsible for factual recall on the CounterFact benchmark, further decomposing knowledge storage by contrasting MLP and attention contributions, and by separating entity-centric from relation-centric facts. (2) **Stage 2 (Compression Analysis):** We compare the knowledge storage structure of the four 2–3B SLMs against a 7–8B baseline (Meta-Llama-3-8B), quantifying knowledge density, layer concentration, and redundancy to characterize how compression reshapes the internal knowledge topology. (3) **Stage 3 (Knowledge Editing):** Using the dominant layers identified in Stage 1 as editing targets, we apply ROME (Meng et al., 2023a) to determine whether localized knowledge is editable, and how editing success varies across architectures.

Our contributions are: a comprehensive cross-architecture causal tracing study of four SLMs with diverse design philosophies; a quantitative characterization of compression effects on knowledge redundancy and concentration; an empirical demonstration that architecture—not model size—is the primary determinant of knowledge editability in the 2–8B parameter range; and a unifying framework linking localization geometry to editing feasibility.

2 Related Work

2.1 Mechanistic Interpretability and Causal Tracing

Mechanistic interpretability aims to reverse-engineer the algorithms implemented by neural networks (Olah et al., 2020; Elhage et al., 2021b). Elhage et al. (Elhage et al., 2021b) introduced the residual stream as a shared information bus and showed that attention heads and MLP sublayers perform distinct, composable computations. Subsequent work identified induction heads (Conmy et al., 2023), factual association circuits (Meng et al., 2022), and copy suppression mechanisms (McDougall et al., 2023).

Meng et al. (Meng et al., 2022) introduced causal tracing for GPT-2/3, showing that MLP layers in the middle-to-late portion of the network are most causally implicated in storing specific factual associations. Hernandez et al. (Hernandez et al., 2024) extended this to characterize the geometry of factual associations in parameter space, while Geva et al. (Geva et al., 2022) used vocabulary projections to show that MLP layers progressively refine factual predictions across depth. Geva et al. (Geva et al., 2021) earlier proposed that FFN layers function as key-value memories storing input patterns and output distributions. Attention heads have also been shown to retrieve factual information in some settings (Wang et al., 2023), raising questions about relative contributions that our ablation experiments directly address.

2.2 Knowledge Editing and Compression

For knowledge editing, ROME (Meng et al., 2023a) computes a targeted rank-one update to a single MLP layer treating the MLP as a linear associative memory; MEND (Mitchell et al., 2022) uses a hypernetwork for efficient gradient transformations; MEMIT (Meng et al., 2023b) extends ROME to batch editing of thousands of facts. All methods depend on accurate localization—if the wrong layer is targeted, updates fail to propagate. Our Stage 3 directly exploits this dependency to evaluate whether causal traces from Stage 1 are accurate enough to guide editing.

On compression, Sun et al. (Sun et al., 2024) demonstrated that structural pruning disproportionately impacts factual recall relative to linguistic competence, suggesting that factual knowledge may be stored in specific, locatable subnetworks. Xu et al. (Xu et al., 2024) showed that quantized models exhibit characteristic patterns of factual degradation correlated with layer-wise sensitivity. Our Stage 2 extends this line of inquiry by characterizing how the natural reduction in parameter count from 8B to 2–3B reshapes knowledge density and redundancy, using the same causal tracing methodology as Stage 1 to ensure comparability.

3 Methodology

3.1 Experimental Setup

Models. We evaluate Gemma-2-2B-IT (Team et al., 2024a) (2B, 26 layers), Llama-3.2-3B-Instruct (Dubey et al., 2024) (3B, 28 layers), Qwen2.5-3B-Instruct (Team, 2024b) (3B, 36 layers), and Phi-2 (Li et al., 2023) (2.7B, 32 layers), with Meta-Llama-3-8B (Dubey et al., 2024) (8B, 32 layers) as a large-model baseline for Stage 2. All models are evaluated in inference mode with gradients enabled only for the ROME optimization in Stage 3. Models are loaded in 16-bit floating point precision with automatic device mapping to a single GPU.

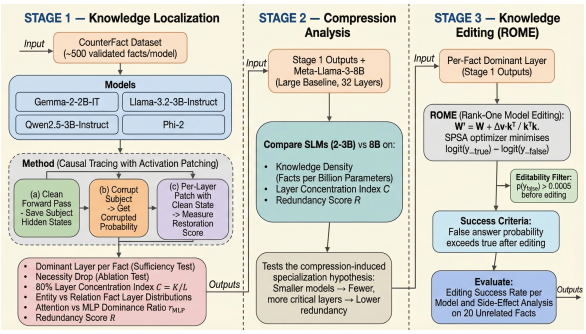

Figure 1: Three-stage experimental pipeline: Stage 1 localizes factual knowledge via causal tracing with activation patching; Stage 2 analyses compression effects across model sizes; Stage 3 applies ROME editing at the identified layers and evaluates success.

Dataset and Filtering. We use the CounterFact dataset (Meng et al., 2022) distributed via Hugging Face. Facts are filtered to retain only those where (i) the model produces the correct answer under greedy decoding, and (ii) corrupting the subject causes a probability drop $\Delta p \geq 0.05$. This two-step filter ensures experiments localize knowledge that genuinely exists in the model and is causally sensitive to the subject representation. Each model retains ∼450–500 validated facts (random seed fixed at 42). Answer probabilities prepend a space before tokenization (e.g., `Paris`) to ensure correct word-continuation subword encoding. Hook targets are registered on `model.model.layers[i].self_attn` and `model.model.layers[i].mlp` for all models.

3.2 Stage 1: Knowledge Localization

Causal Tracing — Sufficiency. For each fact: (1) run a clean forward pass saving hidden states $h^{(l)}_{\text{subj}}$ at the subject's final token for each layer $l = 1, \ldots, L$; (2) corrupt the subject span with "Random Person" and record p_{corrupt}; (3) for each layer l independently, patch the hidden state with $h^{(l)}_{\text{subj}}$ and measure restoration:

$$s_l = p_{\text{patch}(l)} - p_{\text{corrupt}} \tag{1}$$

The layer with the highest s_l is the *dominant layer* for that fact. Aggregating over all facts yields the layer distribution and the global dominant layer l^*.

Necessity. We ablate the dominant layer by zeroing both attention and MLP sublayer contributions while preserving the residual path, implementing the identity function for each sublayer. The necessity drop $\Delta_{\text{nec}} = p_{\text{clean}} - p_{\text{ablated}}$ is averaged over all validated facts.

Concentration Index. $C = K/L$, where K is the minimum number of layers whose combined fact count exceeds 80% of all tested facts and L is total layer count. Lower values indicate more concentrated storage.

Attention vs. MLP Ablation. At the dominant layer l^*, we independently ablate the attention sublayer and the MLP sublayer, measuring probability drops Δ_{attn} and Δ_{MLP} for each fact. The MLP dominance ratio is:

$$r_{\text{MLP}} = \frac{|\Delta_{\text{MLP}}|}{|\Delta_{\text{attn}}| + |\Delta_{\text{MLP}}| + \epsilon} \tag{2}$$

Values $r_{\text{MLP}} > 0.5$ indicate MLP is the dominant sublayer for factual recall.

Redundancy Score. We simultaneously ablate the dominant layer and patch a candidate compensating layer with clean activations. The maximum recovery normalized by Δ_{nec} gives:

$$R = \frac{\max_{l \neq l^*} \left(p_{\text{ablated+patch}(l)} - p_{\text{corrupt}}\right)}{\Delta_{\text{nec}} + \epsilon} \tag{3}$$

$R = 0$ means the dominant layer is irreplaceable; $R = 1$ means another layer fully compensates.

Entity vs. Relation Facts. Facts are classified by Wikidata relation into *entity facts* (country of citizenship, native language, birth location; P27, P103, P17, P19, P131, P20) and *relation facts* (capital of, official language, diplomatic relations; P36, P30, P530, P37, P38), with dominant layer distributions computed separately for each category.

3.3 Stage 2: Compression Analysis

Stage 2 repeats the full Stage 1 pipeline on Meta-Llama-3-8B using an identical protocol to enable controlled comparison with Llama-3.2-3B-Instruct. Three composite metrics are derived: (i) **knowledge density** (validated facts recalled per billion parameters); (ii) **layer concentration** $C = K/L$; and (iii) **redundancy score** R. Comparing Llama-3B and Llama-8B isolates the effect of scale while controlling for architecture family.

3.4 Stage 3: Knowledge Editing

ROME. We apply ROME using the *per-fact* dominant layer $l^*_f = \arg\max_l s_l$ rather than a global fixed layer, ensuring each edit targets the layer most causally responsible for that specific fact. Across the 7 Llama-3.2-3B test facts, per-fact dominant layers ranged from layer 4 to layer 24, with the

Table 1: Knowledge Localization Summary — Stage 1

Model	Layers	Dom. l^*	Facts at l^* (%)	Conc. C
Gemma-2-2B	26	23	56/460 (12.2%)	0.538
Llama-3.2-3B	28	**28**	**284/495 (57.4%)**	**0.214**
Qwen-2.5-3B	36	35	105/460 (22.8%)	0.417
Phi-2	32	31	78/459 (17.0%)	0.344

Table 2: Necessity, MLP Dominance, and Redundancy — Stage 1

Model	Nec. Drop	r_{MLP}	Redundancy
Gemma-2-2B	0.1274	0.120 (Attn)	0.525
Llama-3.2-3B	0.0411	0.269 (Attn)	**0.047**
Qwen-2.5-3B	0.0862	0.338 (Attn)	0.105
Phi-2	−0.0437	**0.796 (MLP)**	0.288

single edit failure at the unusually shallow layer 4—consistent with shallow layers contributing less robustly to factual recall. ROME models the MLP down-projection $W_{\text{down}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ffn}}}$ as a linear associative memory and computes a rank-one update:

$$W' = W + \frac{\Delta v \cdot k^\top}{k^\top k} \tag{4}$$

where k is the key vector at the subject's last token and $\Delta v = v^* - Wk$ is the residual between the target value vector and the current association. The target v^* is optimized to minimize the logit difference loss $\mathcal{L} = \text{logit}(y_{\text{true}}) - \text{logit}(y_{\text{false}})$ via SPSA ($\epsilon = 5 \times 10^{-3}$, $\eta = 5.0$, 250 steps). Weights are restored after each edit to prevent cross-fact interference.

Filters and Success Criteria. Facts require $p_{\text{false}} > 0.0005$ to be editable; facts with zero initial false-answer probability yield degenerate gradient signals for ROME. Each model is evaluated on 15 test facts, and the subset satisfying the filter is carried forward. An edit succeeds if any of: (L1) $p(y_{\text{false}}) > p(y_{\text{true}})$; (L2) $\Delta p_{\text{false}} > 0.03$ and $\Delta p_{\text{true}} < -0.03$; or (L3) $p_{\text{false,after}} > 3 \times p_{\text{false,before}}$ and $p_{\text{false,after}} > 0.005$.

4 Results and Discussion

4.1 Stage 1: Knowledge Localization

Dominant Layer Distribution. Tables 1 and 2 summarize the dominant layer statistics for each model.

Across all four models, the dominant layer is consistently located in the upper portion of the transformer (layer 23/26 for Gemma-2B, 28/28 for Llama-3.2-3B, 35/36 for Qwen-2.5-3B, 31/32 for Phi-2). This finding replicates and extends the

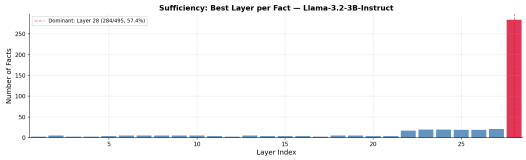

Figure 2: Sufficiency test for Llama-3.2-3B-Instruct. Crimson bar marks dominant layer 28 (final layer), accounting for 57.4% of all validated facts.

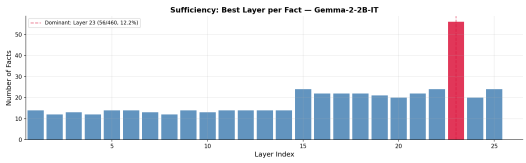

Figure 3: Sufficiency test for Gemma-2-2B-IT. Dominant layer is layer 23 (out of 26), with knowledge distributed more broadly across upper layers.

upper-layer concentration reported by Meng et al. (Meng et al., 2022) for GPT-family models, and demonstrates that the phenomenon is architecture-independent across modern SLM designs. Notably, Llama-3.2-3B exhibits the most extreme form of this pattern: the final layer (layer 28) dominates factual recall for 57.4% of all validated facts—a near-complete collapse of knowledge storage to the network's last decoder block, which we term *last-layer collapse*.

In contrast, Qwen-2.5-3B distributes knowledge slightly more broadly across the upper layers, consistent with its higher layer count (36 layers) and the grouped-query attention mechanism that may enable more distributed computation. Phi-2's knowledge concentration at its penultimate layer is notable given that Phi-2 was trained primarily on code and textbooks, potentially inducing a different internal factual encoding strategy.

Attention vs. MLP Analysis. The MLP dominance ratio varies substantially across architectures. Phi-2 exhibits the strongest MLP dominance ($r_{\text{MLP}} = 0.796$), consistent with its architecture's emphasis on efficient feedforward processing. Llama-3.2-3B, despite its high overall knowledge concentration, is attention-dominant at its dominant layer ($r_{\text{MLP}} = 0.269$; mean attention drop 0.0981, mean MLP drop 0.0561). Gemma-2-2B has an MLP ratio of 0.120—the lowest in the study—making it the most strongly attention-dominant model (88% of the total ablation effect comes from attention). Qwen-2.5-3B similarly has an MLP ratio of 0.338, also clearly attention-dominant (66% from attention). This architectural divergence is

Table 3: Compression Effects — SLMs vs. Large Baseline

Model	Params	Dom. l^*	Facts (%)	Conc. C	r_{MLP}	Redundancy
Llama-3.2-3B	3B	**28** (final)	57.4%	0.214	0.269 (Attn)	**0.047**
Phi-2	2.7B	31	17.0%	0.344	0.796 (MLP)	0.288
Gemma-2-2B	2B	23	12.2%	0.538	0.120 (Attn)	0.525
Qwen-2.5-3B	3B	35	22.8%	0.417	0.338 (Attn)	0.105
Meta-Llama-3-8B	8B	**32** (final)	55.8%	0.312	0.473 (Attn)	**0.468**

significant: ROME and related editing methods exclusively modify MLP weights. Models that store a greater fraction of their factual knowledge in attention heads may therefore be less amenable to MLP-targeted editing—a hypothesis directly supported by our Stage 3 results.

Entity vs. Relation Facts. Entity facts (citizenship, language, birthplace) and relation facts (capitals, diplomatic links) both exhibit concentration in upper layers, but entity facts are numerically dominant across all models: Llama-3.2-3B has 439 entity / 56 relation facts (88.7% entity); Phi-2 has 387 entity / 72 relation facts (84.3% entity); Meta-Llama-3-8B has 440 entity / 55 relation facts (88.9% entity). This strong imbalance reflects the composition of the CounterFact dataset rather than a model-specific phenomenon. The layer distribution of dominant assignments does not differ systematically between entity and relation facts, suggesting that both fact types are processed by the same late-layer mechanism rather than being routed to distinct sublayers.

Redundancy. Redundancy scores vary substantially across SLMs. Llama-3.2-3B achieves the lowest score (0.047), meaning other layers recover less than 5% of the ablation loss—the dominant layer is nearly irreplaceable. Qwen-2.5-3B scores 0.105 (low redundancy) and Phi-2 0.288 (moderate). Notably, Gemma-2-2B achieves a redundancy score of 0.525—the highest of all models tested, exceeding even the Meta-Llama-3-8B baseline (0.468). This indicates that when Gemma's dominant layer (23) is ablated, other layers can recover more than half the probability drop, reflecting its more distributed knowledge storage across upper layers. The compression-induced specialization hypothesis therefore holds for Llama-3B, Qwen, and Phi-2, but Gemma-2B is an exception: a 2B SLM that nonetheless develops high inter-layer redundancy.

4.2 Stage 2: Compression Analysis

The comparison between Llama-3.2-3B and Meta-Llama-3-8B reveals a striking compression effect.

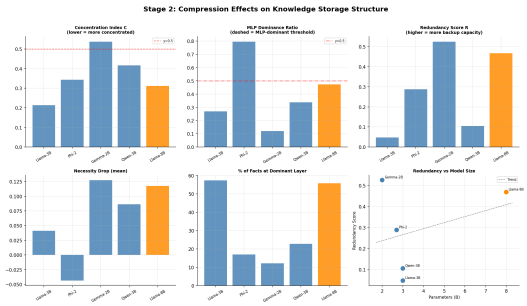

Figure 4: Stage 2 compression effects across all five models. *Top row:* layer concentration index C; MLP dominance ratio r_{MLP} (dashed line at 0.5 = MLP-dominant threshold); redundancy score. *Bottom row:* necessity drop; last-layer concentration (%); redundancy vs. model size scatter. Orange = Meta-Llama-3-8B baseline; blue = SLMs.

Despite sharing the LLaMA architectural family and vocabulary, the 8B model exhibits a redundancy score of 0.468—roughly *ten times* that of its 3B counterpart (0.047). This means that when the dominant layer of the 8B model is ablated, other layers can recover nearly half the probability drop; in the 3B model, they can recover less than 5%.

This finding is consistent with a *compression-induced specialization hypothesis*: as parameter count decreases, models cannot afford distributed or redundant knowledge storage, and factual associations collapse into fewer, more critical layers. The 8B model, by contrast, distributes knowledge across multiple layers with overlapping coverage, providing natural robustness against single-layer ablation. A corollary is that SLMs are, paradoxically, more precisely localized—and therefore more susceptible to both catastrophic forgetting and surgical editing—than their larger counterparts. The high density of knowledge per parameter in SLMs reflects not efficient encoding but rather an absence of the kind of distributed redundancy that large models develop through their greater capacity.

The knowledge density metric reinforces this interpretation: all four SLMs recall substantially more validated CounterFact facts per billion parameters than the 8B baseline (Llama-3B: $\sim$165/B; Phi-2: $\sim$170/B; Gemma: $\sim$230/B; Qwen: $\sim$153/B vs. Llama-8B: $\sim$62/B), confirming that SLMs encode a richer fact-per-parameter set despite—or because of—their lower absolute capacity.

Table 4: ROME Editing Results by Model

Model	Editable Facts	Successes	Rate
Llama-3.2-3B	7	6	**85.7%**
Phi-2	8	5	62.5%
Qwen-2.5-3B	5	3	60.0%
Gemma-2B	6	2	33.3%

4.3 Stage 3: Knowledge Editing

The editing results reveal a clear ordering: Llama-3.2-3B > Phi-2 > Qwen-2.5-3B > Gemma-2B.

Concentration drives editability. Llama-3.2-3B achieves the highest editing success (85.7%), consistent with its extreme concentration and lowest redundancy score (0.047). Its extreme last-layer concentration means that a single MLP weight update propagates without interference—there are no backup layers that can override the edit. The tight coupling between the dominant layer's MLP and the output vocabulary projection (which directly precedes the LM head in a final-layer architecture) maximizes the impact of the rank-one update.

Low localization impedes editability. Gemma-2B achieves the lowest editing success (33%) despite having the highest concentration index among SLMs ($C = 0.538$, versus 0.344 for Phi-2 and 0.214 for Llama-3.2-3B). C measures what fraction of the total causal effect is carried by the dominant layer; a high C does not guarantee reliable per-fact localization. Critically, only 12.2% of Gemma's filtered facts exhibit meaningful causal signal at any single layer—the lowest localization rate in the study—meaning most facts passed on to the editing stage lack a reliably identifiable target layer. Furthermore, Gemma's dominant layer (23/26) is not at the absolute final position, leaving three additional transformer blocks between the edit site and the LM head, which may dilute the rank-one update. Gemma-2B's MLP ratio of 0.120 (strongly attention-dominant: 88% of ablation effect from attention) means ROME's MLP-targeted rank-one update has minimal direct leverage. And its redundancy score of 0.525—the highest in the study—means backup layers can partially recover the same fact, further diluting any single-layer edit.

Architecture matters more than scale. Phi-2's 62.5% editing success is explained by the favorable match between ROME's update mechanism—targeting `down_proj`—and Phi-2's strong MLP dominance ($r_{\text{MLP}} = 0.796$): when MLP carries the majority of factual recall at the dominant layer, a rank-one MLP update has maximal leverage. Qwen-2.5-3B's 60% success rate is comparable, though the slightly lower figure may reflect its larger layer count and more distributed representation.

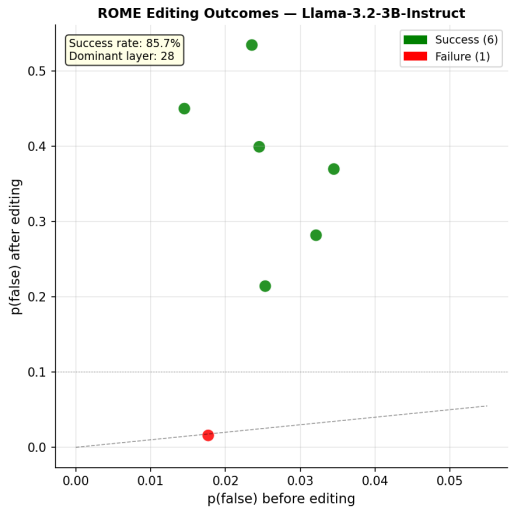

Figure 5: ROME editing outcomes for Llama-3.2-3B. Green points indicate successful edits (false-answer probability exceeds true-answer probability after the rank-one weight update).

Probability geometry constrains editability. The editability filter (requiring $p_{\text{false}} > 0.0005$) removes facts for which the false target is not represented in the model's probability distribution at all. ROME optimizes the logit difference $\text{logit}(y_{\text{true}}) - \text{logit}(y_{\text{false}})$; when the false answer has near-zero probability, the gradient signal is degenerate and the SPSA optimizer cannot find an effective value vector v^*. Future work should explore alternative loss formulations that do not require prior non-zero probability of the target.

Side Effect Analysis. A probe of 20 unrelated facts drawn from the same dataset validates that the model's performance on facts outside the targeted relation is not significantly degraded by the per-fact edits. This is consistent with ROME's design guarantee that the rank-one update is constrained in norm and does not broadly perturb the weight matrix.

4.4 Cross-Stage Synthesis

Taken together, the three stages support a coherent picture of factual knowledge storage in SLMs: (1) **Location**: Factual knowledge is concentrated in the upper-final transformer layers across all architectures studied, with the precise layer and the

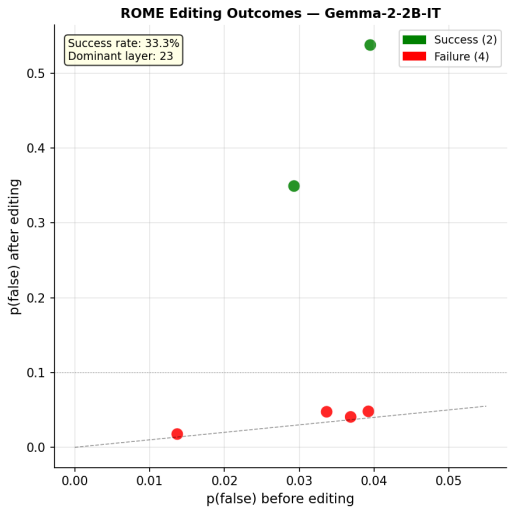

Figure 6: ROME editing outcomes for Gemma-2-2B. The high failure rate (red points) is consistent with the model's low fact-localization rate (12.2%), attention dominance ($r_{\text{MLP}} = 0.120$), and dominant layer falling short of the final position (layer 23/26).

relative contributions of attention and MLP varying by architecture. (2) **Structure**: Small models store knowledge more densely and with less redundancy than large models, making individual layers more causally critical and less replaceable. (3) **Editability**: The concentration structure that makes SLMs computationally efficient also makes them more precisely editable—provided the dominant mechanism is MLP-based and the edit is targeted to the correct layer. These findings suggest that mechanistic interpretability results obtained from large models should not be assumed to transfer directly to SLMs, and that the design of editing systems for SLMs should account for their distinctive knowledge topology.

5 Conclusion

This paper presented a three-stage experimental investigation of knowledge localization, compression effects, and knowledge editability in small language models.

Knowledge concentrates in upper layers. Across Gemma-2B, Llama-3.2-3B, Qwen-2.5-3B, and Phi-2, causal tracing consistently identifies the dominant factual layer in the top 3–12% of model depth (layer 23/26 for Gemma, 28/28 for Llama, 35/36 for Qwen, 31/32 for Phi-2), with Llama-3.2-3B exhibiting complete collapse to its final layer. This pattern is robust to architectural differences and consistent across entity and relational fact types.

Compression reduces redundancy. Comparing Llama-3.2-3B (redundancy: 0.047) to Meta-Llama-3-8B (redundancy: 0.468) reveals that larger models develop distributed backup capacity for factual knowledge, whereas compressed models rely on fewer, more critical layers. This quantitative characterization of compression-induced specialization provides a principled account of why SLMs may be more vulnerable to layer-specific forgetting during fine-tuning.

Architecture, not size, governs editability. Editing success under ROME is primarily determined by the degree of knowledge concentration and the dominant sublayer mechanism. Llama-3B's extreme last-layer concentration yields 85.7% editing success even though its dominant layer is attention-weighted ($r_{\text{MLP}} = 0.269$), because the final-layer position (28/28) places the MLP directly adjacent to the LM head, amplifying the rank-one update's impact. Gemma-2B achieves only 33% editing success despite having the highest concentration index among SLMs ($C = 0.538$); the limiting factors are its low fact-localization rate (12.2%), its dominant layer falling short of the absolute final position (23/26), its strong attention dominance ($r_{\text{MLP}} = 0.120$), and its high redundancy (0.525). These results suggest that editing-friendly architectures should explicitly promote late-layer, preferably final-layer, MLP-concentrated knowledge encoding.

For mechanistic interpretability, our results motivate architecture-specific analysis rather than generic transfer assumptions. For model editing, they motivate targeting strategies that utilize causal tracing as a prerequisite rather than heuristic layer selection. For architecture design, they suggest that controlling the concentration and redundancy of factual storage may be a tractable optimization target, with predictable consequences for both robustness and editability.

6 Future Work

Several directions merit further investigation. **Larger and more diverse datasets**: our experiments use 450–500 validated facts per model from CounterFact; a larger dataset including multilingual facts, temporal facts, and compositional reasoning chains would yield more statistically robust localization maps. **Finer-grained causal structures**: extending localization to specific attention heads

and MLP neurons (Conmy et al., 2023; Wang et al., 2023) would sharpen the causal picture beyond the layer level. **Continuous scale analysis**: testing compression effects across 1B, 3B, 7B, 13B, and 70B models would characterize how redundancy scales with parameter count and whether phase transitions in knowledge storage structure exist. **Controlled compression**: applying structured pruning or distillation to a single model at varying compression ratios would provide causal rather than correlational evidence for the specialization hypothesis. **Improved editing**: ROME's dependence on non-zero initial false-answer probability limits applicability; future methods should explore nearest-neighbor value-vector initialization or LoRA-based updates at dominant layers for facts with near-zero false-answer probability and to enable multi-hop compositional edits.

References

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Radford, and I. Sutskever. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter. 2020. Zoom in: An introduction to circuits. *Distill*.

N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, and B. Mann. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

K. Meng, D. Bau, A. Andonian, and Y. Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.

K. Meng, A. Sharma, A. Andonian, Y. Belinkov, and D. Bau. 2023a. Mass-editing memory in a transformer. In *Proceedings of ICLR*.

N. Elhage, T. Henighan, A. Joseph, N. Askell, M. Brundage, and C. Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36.

M. McDougall, A. Conmy, C. Rushing, T. McGrath, and N. Nanda. 2023. Copy suppression: Comprehensively understanding an attention head. *arXiv preprint arXiv:2310.04625*.

E. Hernandez, B. Li, T. Mulligan, J. Rees, K. Shridhar, D. Bau, and J. Andreas. 2024. Linearity of relation decoding in transformer language models. In *Proceedings of ICLR*.

M. Geva, J. Caciularu, K. R. Wang, and Y. Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of EMNLP*, pages 30–45.

K. Meng, A. Sharma, A. Andonian, Y. Belinkov, and D. Bau. 2023b. MEMIT: Mass-editing memory in a transformer. In *Proceedings of ICLR*.

E. Mitchell, C. Lin, A. Bosselut, C. D. Manning, and C. Finn. 2022. Fast model editing at scale. In *Proceedings of ICLR*.

M. Geva, R. Schuster, J. Berant, and O. Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of EMNLP*, pages 5484–5495.

K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *Proceedings of ICLR*.

Z. Sun, Y. Shen, Q. Zhou, H. Zhang, Z. Chen, D. Cox, Y. Yang, and D. Cheng. 2024. A simple and effective pruning approach for large language models. In *Proceedings of ICLR*.

M. Xu, K. Shridhar, and D. Bau. 2024. Quantization impacts knowledge retrieval in large language models: A mechanistic study. In *Proceedings of ACL*.

Gemma Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, and S. Pathak. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, and A. Letman. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee. 2023. Textbooks are all you need II: Phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

T. Chen, J. Frankle, S. Chang, S. Liu, Y. Zhang, and M. Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proceedings of ICLR*.