A Measurement Framework for Decomposing Factual Recall in Language Models: Storage, Access, and Competition

Anonymous authors
Paper under double-blind review

Abstract

Mechanistic analyses of factual recall often conflate whether a fact is encoded in model parameters (*storage*), whether a prompt activates an effective route to that fact (*access*), and whether alternative answers suppress the target during decoding (*competition*). We present a measurement framework that separates these components with prompt-family controls, matched competition controls, layerwise invariance measurements, and control-aware hidden-state interventions. We evaluate 6 locally hosted models from 3 families on a 5-relation benchmark of 90 facts rendered under 4 prompt families. The best prompt family for each model also maximizes matched-control selectivity, but the preferred family is model dependent rather than universal. In the descriptive full-benchmark analysis, same-fact prompts exhibit near-immediate subject-position invariance in 524 of 540 fact-model cases (97.0%, 95% CI 95.2–98.2%) with mean depth 0.19 layers, whereas pre-answer invariance is weaker and later, appearing in 454 of 540 cases (84.1%, 95% CI 80.7–86.9%) only after mean depth 18.72 layers. Control analyses show that this late effect survives the removal of chat-template prompts and weakens but remains deep under aggressive length matching. A null-controlled subspace test strengthens the causal picture: fact-component projection produces 0.0382 more alignment loss than relation-matched wrong-fact projection and 0.0491 more than random projection, while fact-component patching outperforms null patches but rarely restores strong retrieval on its own. These results support a two-stage view of factual recall and argue that explainability studies should measure storage-related support, access, and competition separately.

1 Introduction

Large language models often answer factual questions correctly while remaining mechanistically opaque. A correct output does not, by itself, reveal whether the fact was robustly represented, whether the prompt activated an effective retrieval route, or whether alternative answers simply failed to compete successfully. This ambiguity sits at the center of current work on parametric knowledge and explainability.

We argue that factual recall should be treated as a decomposition problem. At minimum, three components must be distinguished: *storage*, the extent to which a fact is encoded in model parameters; *access*, the prompt-conditioned ability to route computation toward that fact; and *competition*, the pressure exerted by distractor answers and relation-confusable alternatives during decoding. Existing analyses often entangle these components when they interpret salient neurons, influential layers, or successful generations as direct evidence of factual storage.

The present work does not claim to resolve that decomposition fully. Instead, it introduces a measurement framework intended to make the decomposition empirically tractable. We treat prompt families as controlled perturbations of access, matched controls as probes of competition, and prompt-invariant hidden-state structure as an initial proxy for storage-related support. Within that framing, the purpose of this paper is methodological: to establish a more reliable basis for studying factual recall before making stronger claims about where knowledge resides.

This framing yields two immediate benefits. First, it turns prompt variation from a nuisance factor into an experimental variable. Second, it clarifies why single-family analyses can be misleading: components highlighted under one surface form may reflect access-specific circuitry rather than fact-supporting structure that persists across paraphrastic variants. The experiments below show that this distinction is empirically meaningful. Prompt-family preference is model-dependent, matched-control selectivity tracks the same preference, and layerwise invariance reveals a consistent separation between early fact anchoring and later answer formation.

We make three contributions.

1. We introduce a storage-access-competition framing for factual recall and operationalize it with prompt families as controlled access perturbations, matched controls as competition probes, layerwise invariance gaps, and out-of-sample family selection.
2. We build and evaluate a reproducible local benchmark spanning 6 small language models, 5 relations, and 90 facts, showing that prompt-family preference is replicated but model-dependent and that prompt-invariant structure emerges in two distinct phases.
3. We add control-aware causal tests, grouped holdout ablations and null-controlled fact-subspace interventions, and release the full single-GPU pipeline, benchmark snapshots, analysis configs, and companion viewer as reproducibility artifacts.

The scope of the paper is intentionally narrower than storage localization. Our claim is not that factual storage has been identified, but that a serious empirical study of factual recall requires separate measurements of storage-related support, prompt-conditioned access, and answer-space competition.

2 Related Work

Transformers distribute computation across attention and feed-forward blocks, making simple local explanations difficult (Vaswani et al., 2017). Prior work has shown that feed-forward layers can behave like key-value memories (Geva et al., 2021), and factual associations can sometimes be localized, attributed, or edited with targeted interventions (Dai et al., 2021; Meng et al., 2022; Hernandez et al., 2023). At the same time, superposition and polysemanticity make single-neuron semantics unreliable in many cases (Elhage et al., 2022; Bricken et al., 2023). These tensions are especially relevant for factual recall, where a prompt-conditioned access route can easily be mistaken for the knowledge-bearing mechanism itself.

Another line of work studies relation structure and higher-level representational geometry rather than individual neurons. Relation decoding can behave linearly in transformer hidden states (Hernandez et al., 2024), and reusable function-like directions have been observed in latent space (Todd et al., 2024). These results motivate a subspace-oriented view of factual computation. Our work adopts that representational perspective, but shifts the experimental question: instead of asking only whether a relation or fact can be decoded from hidden states, we ask what structure remains stable when the same fact is expressed through different prompt families and whether that prompt-invariant structure survives matched competition controls and null-controlled interventions.

Recent work also shows that semantically equivalent factual queries can yield language- and prompt-dependent internal effects (Chen et al., 2024). That observation is close to our starting point, but our aim is different. We do not treat prompt variability only as a robustness problem; we treat it as an experimental perturbation of *access*. This makes it possible to ask whether same-fact representations survive that perturbation and how prompt-conditioned failures interact with matched alternatives in answer space.

Our study also builds on recent small open models that are practical for repeated intervention on consumer hardware. We analyze Gemma 3 (Gemma Team, 2025), Qwen2.5 (Qwen Team, 2024), and SmolLM2 (Allal et al., 2025). Their scale is small enough for exhaustive prompt-family sweeps and repeated hidden-state intervention on a single RTX 4050 laptop GPU while still allowing cross-family and base-versus-instruction comparisons.

Relative to this literature, our contribution is not another neuron-localization or model-editing method. The distinctive elements here are: 1) treating prompt family as a controlled access perturbation rather than a nuisance variable, 2) pairing it with matched competition controls, 3) separating descriptive representational analyses from out-of-sample family selection and holdout causal tests, and 4) comparing fact-subspace interventions against relation-matched and random nulls. The framework is therefore aimed at measurement and decomposition, separating storage-related support, access, and competition, rather than direct editing or a claim that knowledge has already been localized.

3 Measurement Framework

Let a fact be represented as a subject-relation-object triple (s, r, o). Let $P = \{p_1, \ldots, p_k\}$ denote a set of prompt families that preserve the same fact while varying the surface form. In this study each fact is rendered as a raw question, a declarative completion, a Q/A template, and a single-turn chat instruction rendered with the tokenizer's native chat template when available. Appendix A gives the exact templates used for all 5 relations.

For each prompt instance we collect answer-level outputs together with hidden states at two positions: the last token of the subject span and the pre-answer position at the end of the prompt. Let $h_{p,l}^{(\mathrm{pos})}$ denote the hidden state for prompt p at layer l and position $\mathrm{pos} \in \{\mathrm{subject}, \mathrm{preanswer}\}$. For each fact and layer we compare two similarities:

$$\mathrm{WithinFact}_l^{(\mathrm{pos})} = \mathbb{E}_{p_i, p_j \in P_f} \left[\cos\left(h_{p_i,l}^{(\mathrm{pos})}, h_{p_j,l}^{(\mathrm{pos})} \right) \right], \tag{1}$$

$$\mathrm{SameRelation}_l^{(\mathrm{pos})} = \mathbb{E}_{p \in P_f, q \in P_r} \left[\cos\left(h_{p,l}^{(\mathrm{pos})}, h_{q,l}^{(\mathrm{pos})} \right) \right], \tag{2}$$

where P_f contains prompts for the same fact and P_r contains prompts for other facts with the same relation.

We define the *invariance gap*:

$$\mathrm{Gap}_l^{(\mathrm{pos})} = \mathrm{WithinFact}_l^{(\mathrm{pos})} - \mathrm{SameRelation}_l^{(\mathrm{pos})}. \tag{3}$$

A positive gap means that semantically equivalent prompts for the same fact are more similar to each other than to other prompts from the same relation class. We then define *invariance depth* as the first layer at which the gap becomes positive for two consecutive layers.

We treat this signal as *storage-related support* rather than as definitive storage localization. A positive and early subject-side gap suggests that prompt-specific surface form has been collapsed into a shared fact-conditioned state. A later pre-answer gap indicates that prompt-conditioned routing and answer-space competition remain active well after the queried entity has been internally stabilized. This interpretation is still operational rather than ontological: the framework measures when prompt-invariant structure appears and how it behaves under intervention, not where knowledge ultimately resides.

We also compute prompt-family selectivity, matched-control confusion, and grouped neuron ablations. These components are secondary in the present framing: they probe access and competition around the central representational question rather than serving as the entire explanation. Selection-based causal analyses use a deterministic analysis/holdout split, whereas representational invariance is reported descriptively on the full benchmark with additional holdout-only control variants.

4 Experimental Protocol

4.1 Models and Hardware

All experiments were run locally on a single NVIDIA GeForce RTX 4050 Laptop GPU. The system driver reported CUDA 12.6 with NVIDIA driver 560.35.03. Device memory was 6141 MiB as reported by `nvidia-smi`, while the local PyTorch 2.6.0+`cu124` runtime reported 5801 MiB addressable device memory and cuDNN 9.1. The evaluated checkpoints comprised six small models spanning three families: Gemma-3-270M and

Gemma-3-270M-IT, Qwen2.5-0.5B and Qwen2.5-0.5B-Instruct, and SmolLM2-360M and SmolLM2-360M-Instruct. This selection intentionally prioritizes repeated mechanistic experimentation over raw benchmark scale. The contribution of the paper is methodological: prompt-family control, not frontier-model accuracy.

4.2 Benchmark and Matched Controls

We built a 5-relation factual benchmark with 90 facts total: capitals, official languages, currencies, company headquarters, and birth places. For capitals, official languages, and currencies, we filtered Rest Countries entries (REST Countries, 2026) to independent United Nations members with non-empty target fields, sorted the candidates deterministically by subject string, and retained the first 18 facts per relation. For headquarters and birth places, we used curated 18-entity seed sets and snapshot both the seed file and the filtered country payload in the release package. Each relation therefore contributes 18 facts. Each fact is rendered under 4 prompt families, giving 360 prompts per model and 2160 prompts across the 6-model comparison.

Each benchmark row also stores matched controls: same subject with a different relation, same relation with a different subject, and a lexical distractor answer. These controls allow us to distinguish target retrieval from three plausible alternative behaviors: drifting to another known fact about the same subject, substituting another entity with the same relation, or falling into a surface-level lexical trap. Missing controls are stored explicitly and contribute zero to the control maximum. We also store a deterministic 80/20 fact-level analysis/holdout split, created after subject sorting within each relation, which yields 70 analysis facts and 20 holdout facts overall (280 and 80 prompts per model, respectively). The analysis split is used only for selection steps such as choosing best/worst prompt families and ranking ablation candidates. Holdout prompts are reserved for out-of-sample prompt-family comparisons, grouped neuron ablations, and subspace interventions.

4.3 Scoring

The main results use answer-level scoring rather than only first-token rank. Each generation is scored with full exact match, answer containment, token-level answer F1 over valid aliases, and string similarity. All generations use greedy decoding (`do_sample=False`) with `max_new_tokens=10` and fixed seed 13. Text is lowercased, stripped of punctuation, and normalized by removing frequent answer wrappers such as declarative prefixes, `Q:/A:` markers, and chat-role markers before alias matching. Appendix A lists the exact templates and normalization rules. We also retain conventional next-token metrics such as target rank and target logit margin for comparison with prior mechanistic practice.

For matched controls we compute control-side answer F1 and define *target selectivity*:

$$\text{target_selectivity} = \text{answer_F1}_{\text{target}} - \max_{c} \text{answer_F1}_{c}, \tag{4}$$

where c ranges over the available matched controls. High selectivity indicates that the model favors the correct fact over the strongest control alternative.

4.4 Statistics and Causal Checks

For each relation and prompt-family pair we compute 400-sample bootstrap confidence intervals. Best and worst prompt families are chosen on the analysis facts for each model and then evaluated on the disjoint holdout facts. Paired prompt-family comparisons therefore align the same holdout facts across the analysis-selected best and worst families and report a two-sided sign-test p-value. We use these statistics as descriptive stability checks for a medium-scale local study rather than as formal population inference across model families.

To address the risk that late pre-answer effects are artifacts of chat wrappers or prompt length, we repeat the invariance analysis under two additional control variants: a *no-chat* condition that excludes chat-template prompts and a *length-matched* condition that only compares same-fact prompts whose rendered lengths differ by at most four tokens. These controls are reported on holdout facts.

Table 1: Prompt-family access summary for the revised 6-model study. Best and worst families are selected on the analysis split and evaluated on the holdout split. For each model, the best prompt family under answer F1 is also the best family under matched-control selectivity.

Model	Best family	Worst family	Best F1	ΔF1 [95% CI], p	ΔSelectivity [95% CI], p
Gemma-3-270M	Declarative	Q/A	0.1770	0.0183 [-0.0345, 0.0748], 1.0000	0.0402 [-0.0160, 0.1077], 0.5078
Gemma-3-270M-IT	Raw question	Declarative	0.2704	0.1074 [0.0289, 0.1925], 0.0129	0.0460 [-0.0639, 0.1400], 0.0923
Qwen2.5-0.5B	Q/A	Raw question	0.5033	0.3925 [0.1648, 0.5983], 0.0923	0.3454 [0.1652, 0.5510], 0.0386
Qwen2.5-0.5B-Instruct	Chat template	Declarative	0.5072	0.3517 [0.1804, 0.5222], 0.0129	0.2381 [0.0071, 0.4582], 0.0923
SmolLM2-360M	Q/A	Chat template	0.3235	-0.0078 [-0.1751, 0.1561], 0.4240	-0.0156 [-0.2027, 0.1667], 0.5811
SmolLM2-360M-Instruct	Chat template	Q/A	0.2908	0.2908 [0.1583, 0.4680], 4.88×10^{-4}	0.2390 [0.0815, 0.4401], 0.0117

We also perform two causal checks. First, we ablate grouped neuron sets on holdout prompts by zeroing top invariant and top prompt-sensitive units chosen from the analysis split, then re-measuring answer F1, target selectivity, and control confusion. Second, we estimate a fact-specific early component by subtracting the same-relation centroid from the same-fact subject representation at the earliest invariant layer. We then either project that component out of the model's best prompt family or patch it into the worst prompt family and measure both answer quality and late-layer pre-answer alignment. For the subspace study we add two nulls: a relation-matched wrong-fact component drawn from the same relation class and a norm-matched random direction, and we repeat the intervention at the adjacent early layer as a local robustness check.

4.5 Interactive System

Alongside the offline experiments, we implemented a FastAPI backend and React frontend that keep a selected checkpoint loaded on GPU and display prompt-conditioned layer and neuron activity in real time. The viewer is a companion artifact rather than a source of primary evidence, but it was useful for rapid iteration and will help external users reproduce the workflow. The full source code available in the github (*[viewer source anonymized for review]*).

5 Results

5.1 Prompt Families Behave as Model-Dependent Access Operators

We begin with the behavioral layer of the framework. Table 1 summarizes, for each checkpoint, the prompt family selected as best on the analysis facts and then evaluated against the analysis-selected worst family on the disjoint holdout facts. The table reports both answer quality and matched-control selectivity.

Three observations follow. First, prompt-family effects remain large out of sample. Qwen2.5-0.5B reaches 0.5033 answer F1 under Q/A prompts but only 0.1109 under raw questions, while SmolLM2-360M-Instruct reaches 0.2908 under chat templates but collapses to 0.0 under both raw-question and Q/A prompts. Second, the preferred access operator depends on both model family and tuning regime rather than following a universal template: Gemma-3-270M prefers declarative prompts, Gemma-3-270M-IT prefers raw questions, both Qwen variants prefer family-congruent instruction formats, SmolLM2-360M prefers Q/A, and SmolLM2-360M-Instruct prefers chat templates. Third, the winning prompt family under answer F1 is identical to the winning family under matched-control selectivity for all 6 models. The best prompt family therefore does not merely produce cleaner text; it also yields the clearest separation between the target fact and plausible alternatives.

The paired holdout statistics in Table 1 show that these access effects are not equally stable across models. Gemma-3-270M-IT, Qwen2.5-0.5B-Instruct, and SmolLM2-360M-Instruct show clear out-of-sample answer-

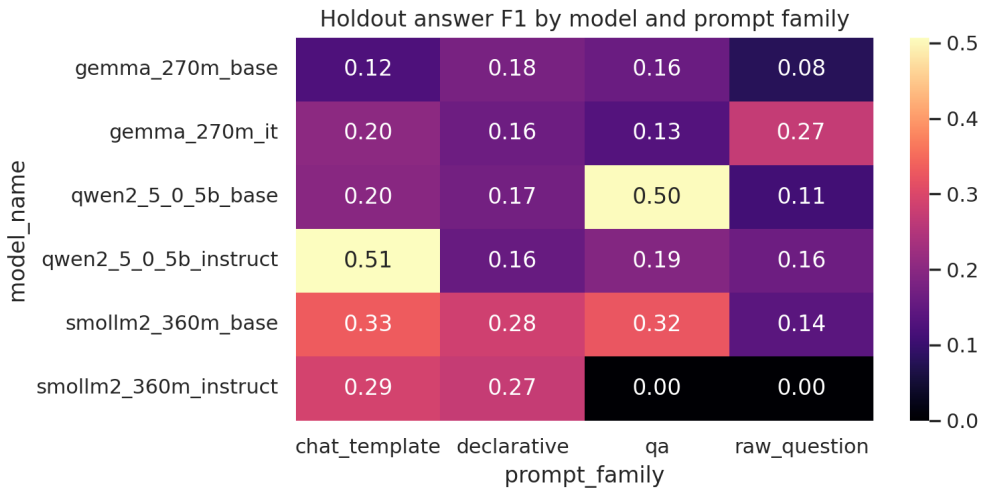

Figure 1: Model-by-prompt-family answer-F1 heatmap for the 90-fact benchmark. Warmer cells indicate stronger factual retrieval; the preferred access operator changes with both model family and instruction tuning.

F1 gains for the selected best family. Qwen2.5-0.5B shows the largest mean holdout F1 gap (0.3925), but its paired sign test is weaker because the advantage is concentrated unevenly across the 20 holdout facts. SmolLM2-360M is the clearest counterexample: the analysis split selects Q/A, but the holdout comparison between Q/A and chat-template prompts is effectively a tie.

5.2 Prompt-Invariant Structure Emerges in Two Phases

The principal representational finding is a two-stage pattern that persists on the expanded benchmark. Across all 540 fact-model cases in the descriptive full-benchmark analysis, subject-position invariance is detected in 524 cases (97.0%, 95% Wilson CI 95.2–98.2%) with a mean invariance depth of only 0.19 layers (95% bootstrap CI 0.15–0.23). At the *pre-answer* position, by contrast, invariance is weaker and much later: it appears in 454 of 540 cases (84.1%, 95% Wilson CI 80.7–86.9%) only after a mean depth of 18.72 layers (95% bootstrap CI 18.19–19.22).

This gap between early subject anchoring and later answer-stage convergence is consistent with the storage-access-competition framing. The subject-side signal suggests that prompt-specific surface form is collapsed into a shared fact-conditioned state very early in the network. The answer-supporting state, however, remains unresolved much deeper into the model, indicating that prompt-conditioned routing and answer-space competition persist well after the queried entity has been internally stabilized.

Table 2 shows that late convergence varies systematically across model families. Gemma reaches pre-answer invariance around layer 12 for both base and instruction-tuned variants. Qwen reaches it around layers 20–21. SmolLM2 reaches it only around layers 26–27, and SmolLM2-360M-Instruct does so for only 24.4% of fact-model cases. Early subject stabilization therefore appears nearly universal in this small-model regime, whereas answer-stage convergence is markedly deeper and less reliable.

There is also a notable asymmetry across model families. Gemma exhibits the weakest mean subject-side invariance gaps (0.0608 and 0.0610 for its base and instruction-tuned variants) but the earliest pre-answer convergence. Qwen and SmolLM2 exhibit stronger subject-side gaps (0.1098–0.1322) yet substantially later answer-side convergence, and in the SmolLM2-360M-Instruct case that convergence often never materializes. Stronger early fact anchoring therefore does not automatically imply more efficient answer formation; some models stabilize the queried entity strongly while still requiring many additional layers to resolve access and competition.

Table 2: Mean invariance depth by model on the expanded 90-fact benchmark. Subject-side invariance is near-immediate across nearly all models, while pre-answer convergence is model-dependent and substantially later.

Model	Subject depth	Pre-answer depth	Pre-answer found
Gemma-3-270M	0.13	12.00	100.0%
Gemma-3-270M-IT	0.12	12.00	100.0%
Qwen2.5-0.5B	0.45	21.00	82.2%
Qwen2.5-0.5B-Instruct	0.47	20.73	100.0%
SmolLM2-360M	0.00	26.33	97.8%
SmolLM2-360M-Instruct	0.00	27.27	24.4%

Table 3: Control analyses for the invariance measurement, pooled across all 6 models. The no-chat condition addresses chat-wrapper effects, and the length-matched condition restricts within-fact comparisons to prompts whose lengths differ by at most four tokens.

Variant	Subject found	Subject depth	Pre-answer found	Pre-answer depth
All families	97.0%	0.19	84.1%	18.72
No chat	77.4%	0.06	80.6%	18.54
Length-matched	65.7%	0.74	41.7%	16.72

Table 3 addresses the main representational confounds. Excluding chat-template prompts leaves the late pre-answer effect largely intact: pre-answer invariance is still found in 80.6% of cases and still appears only after mean depth 18.54. The stricter length-matched condition lowers both subject-side and pre-answer detectability, which indicates that prompt-length variation does inflate the ease of measuring prompt invariance. However, when pre-answer invariance still appears under that control, it remains deep rather than shifting to the first few layers (mean depth 16.72). The two-stage pattern therefore weakens under aggressive length control but does not collapse into an early-layer artifact.

5.3 Relation Structure and Competition Remain Important

Relation type continues to modulate the invariance picture. Birth-place facts show the strongest subject-side invariance on average, with mean subject-gap 0.2024 and peak subject-gap 0.9152. Headquarters is the weakest relation, with mean subject-gap 0.0721 and subject-side invariance detected for 85.2% of fact-model cases. Some relations therefore provide cleaner prompt-invariant anchors than others even before the answer stage.

Competition becomes visible most clearly at failure time. Using a strict failure criterion on the holdout prompts, answer F1 equal to zero while at least one matched control receives nonzero answer F1, 26 of 480 prompt instances (5.4%) exhibit control attraction. Among those 26 cases, 18 match same-relation controls and 8 match lexical distractors. None of the holdout failures are dominated by same-subject controls.

This pattern suggests that outright retrieval failures are more often pulled toward another answer that fits the queried relation than toward a different fact about the same subject. Within the storage-access-competition framing, prompt families therefore appear to influence retrieval specificity and relation-level competition more strongly than subject-level semantic binding.

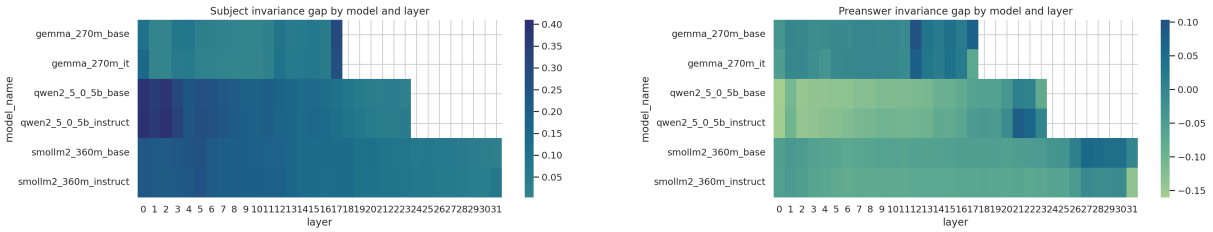

Figure 2: Layerwise invariance-gap heatmaps at the subject (left) and pre-answer (right) positions. Positive values indicate stronger same-fact than same-relation similarity.

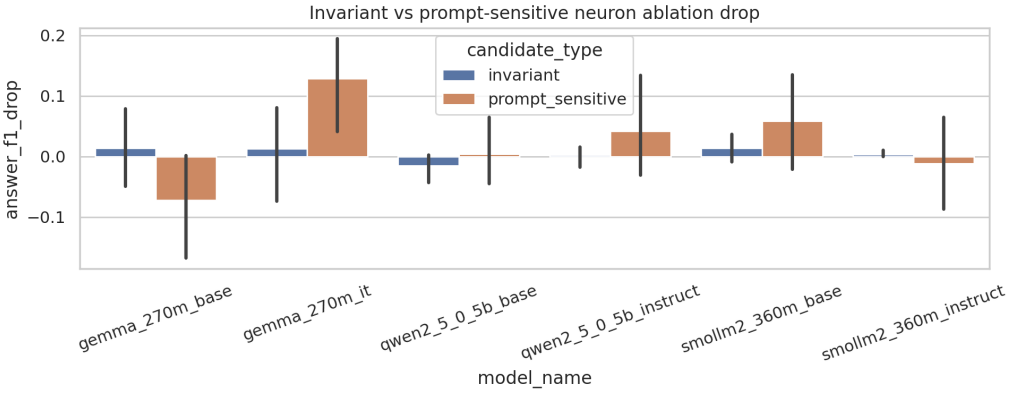

Figure 3: Holdout answer-F1 drop after zeroing invariant and prompt-sensitive neuron groups selected on the analysis split. The contrast is directionally informative but remains too coarse for a strong localization claim.

5.4 Coarse Causal Checks Remain Preliminary

The grouped-ablation results are informative but not decisive. Candidate neurons are ranked on the analysis split and zeroed only on the holdout prompts. Under that protocol, mean answer-F1 drop is higher for prompt-sensitive neuron groups than for invariant groups overall (0.0253 versus 0.0052), and mean target-selectivity drop is also larger (0.0214 versus 0.0113). Several cases are directionally consistent with the access interpretation: for Gemma-3-270M-IT under chat-template prompts, prompt-sensitive ablation causes a 0.2033 drop in answer F1 and a 0.1617 drop in selectivity, and for SmolLM2-360M under chat-template prompts, prompt-sensitive ablation causes a 0.1494 answer-F1 drop and a 0.1127 selectivity drop.

However, the pattern is not universal. Invariant groups sometimes induce larger selectivity degradation, and some ablations mildly improve performance or reduce distractor competition. The causal conclusion from these grouped ablations is therefore intentionally modest: they support the access interpretation as a sanity check, but they are not sufficiently fine-grained to identify a storage mechanism.

5.5 Subspace Interventions Suggest Necessity More Than Sufficiency

The strongest causal evidence comes from the fact-subspace interventions. For each model, we take the analysis-selected best and worst prompt families from the access study, estimate an early fact component at layer 0 from the subject representation, and use the model-specific pre-answer convergence layer as the late measurement point. We then apply two complementary interventions: projecting the early component out of the strongest access route and patching that same component into the weakest route. Each fact-component intervention is compared against two nulls: a relation-matched wrong-fact component and a norm-matched

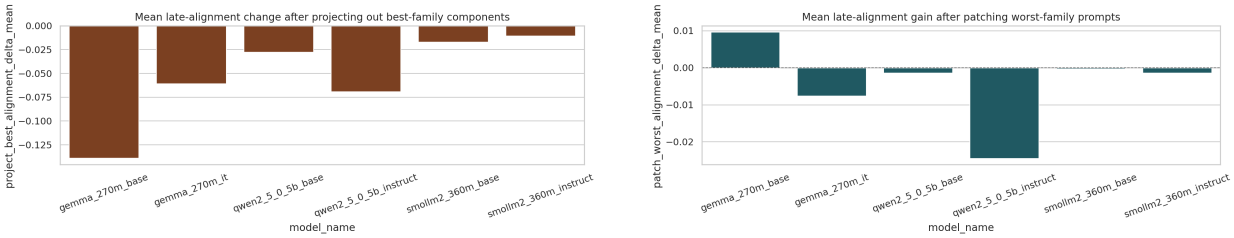

Figure 4: Per-model late-alignment change for fact-component projection (left) and patching (right). Relation-matched wrong-fact and random-null comparisons are reported in the text.

random direction. As a local robustness check, we repeat the same procedure at the adjacent early layer when available.

The resulting asymmetry is consistent across the study and becomes clearer under the null comparisons. Under the primary estimator, projecting out the fact component degrades late pre-answer alignment in all 6 models, with mean alignment delta -0.0542, whereas relation-matched wrong-fact projection yields only -0.0108 and random projection only -0.0004. Across all fact-model instances, fact projection causes 0.0382 additional alignment loss over matched-control projection (95% CI 0.0285–0.0475, sign-test $p = 1.20 \times 10^{-9}$) and 0.0491 additional loss over random projection (95% CI 0.0391–0.0591, $p = 7.82 \times 10^{-13}$). The same pattern appears in answer quality when evaluated relative to the nulls: fact projection causes 0.0364 additional F1 loss relative to matched-control projection (95% CI 0.0107–0.0658, $p = 0.0038$) and 0.0614 additional F1 loss relative to random projection (95% CI 0.0283–0.1017, $p = 6.42 \times 10^{-6}$).

Patching tells a different story. In absolute terms, patching the fact component into the weakest access route produces little aggregate recovery: mean alignment change is -0.0043, mean answer-F1 change is -0.0221, only 1 of 6 models shows positive alignment gain, and only 1 of 6 shows positive mean answer-F1 gain. Gemma-3-270M is the clearest positive case, with a $+0.0096$ alignment gain and a $+0.0488$ answer-F1 gain under patching. Even so, fact patching is still more helpful than the nulls: it yields 0.0178 more alignment gain than matched-control patching (95% CI 0.0140–0.0221, $p = 1.42 \times 10^{-11}$) and 0.0329 more than random patching (95% CI 0.0271–0.0396, $p = 1.04 \times 10^{-18}$). For answer F1, the advantage over random patching remains robust at 0.0395 (95% CI 0.0179–0.0609, $p = 1.25 \times 10^{-6}$), whereas the advantage over matched-control patching is smaller at 0.0210 and its bootstrap interval slightly overlaps zero.

The adjacent-layer robustness check supports the same interpretation. The projection-induced alignment change has the same sign under the primary and plus-one early-layer estimators in all 6 models, and the mean difference between the two strategies is only -0.0011 (95% bootstrap CI -0.0108 to 0.0106). This does not eliminate estimator sensitivity, but it reduces the risk that the causal effect is an artifact of a single hand-picked early layer.

Taken together, these results provide the strongest causal evidence in the current study. The early prompt-invariant component behaves more like a *necessary* ingredient of a strong retrieval path than a *sufficient* ingredient of factual recovery. Put differently, the strongest access route appears to rely on a fact-specific early state, and that state is intervention-specific above matched wrong-fact and random nulls. However, weak access routes usually fail for additional reasons that remain in place even after that state is inserted. Those reasons are likely to include prompt-conditioned routing choices and downstream competition in answer space. Because the subspace estimate remains coarse, the result should not be interpreted as proof of storage localization. It is better understood as evidence that early prompt-invariant structure participates causally in successful retrieval while constituting only one component of the overall mechanism.

6 Discussion

The results support four conclusions. First, prompt family is more productively interpreted as an access operator than as superficial prompt engineering. Changing prompt family alters both answer quality and

matched-control selectivity, often by large margins, which indicates that different surface forms route retrieval through materially different internal computations.

Second, single-family analyses can over-attribute access circuitry to storage. If Q/A is the preferred route for one model while chat templates are preferred for another, then neurons or layers highlighted under a single prompt family cannot automatically be interpreted as the locus of factual storage.

Third, factual computation appears to unfold in two stages. Subject-position invariance emerges early and robustly across models, whereas pre-answer invariance is later, weaker, and more model-dependent. The contrast between Gemma and the Qwen/SmolLM2 family further indicates that stronger early anchoring does not guarantee earlier answer convergence. Early fact stabilization and successful answer formation are related, but they are not identical processes.

Fourth, the subspace interventions reveal a consistent asymmetry between necessity and sufficiency. Removing the early prompt-invariant fact component harms strong retrieval routes more than removing relation-matched wrong-fact or random null components, but inserting it into weak routes rarely repairs them fully even though it outperforms those null patches. This result moves beyond the behavioral statement that prompts matter. It suggests that access failure is not merely the absence of an early fact trace; downstream routing and answer-space competition must also be modeled explicitly.

Taken together, these observations motivate a more ambitious research program. A natural next step is to identify prompt-invariant internal objects more directly, measure how they emerge across layers, and test whether they transfer across prompt families, relations, and model families. The current intervention asymmetry suggests a particularly concrete agenda: characterize which parts of the early fact state transfer under patching, which parts fail to transfer, and how those failures interact with answer-space competition. In that sense, the present study is best viewed as a measurement paper: it isolates an early prompt-robust stage, a later answer-convergence stage, and an initial causal asymmetry between necessity and sufficiency strongly enough to make deeper storage questions experimentally tractable.

7 Limitations

This study has important limitations.

All checkpoints are small. That is essential for single-GPU reproducibility and repeated intervention, but it limits direct generalization to much larger frontier models.

The benchmark remains modest by standard evaluation scales, even after expansion. It covers a limited set of relation types and a limited number of facts per relation, so relation-conditioned estimates should still be interpreted as medium-scale evidence rather than benchmark-saturating measurement.

The headquarters and birth-place relations are curated rather than harvested from a large external factual benchmark. They are included because they let us test non-country relations locally, but they make the benchmark less standardized than a purely public source.

The matched controls are useful but not exhaustive. Some relations offer richer same-subject structure than others, and the lexical-distractor heuristic is intentionally simple.

The current causal interventions remain approximate. Grouped neuron ablations operate on coarse sets of units, and the subspace interventions use a centroid-difference estimate of the early fact component rather than a fully learned or causal-tracing-derived feature basis. The resulting necessity-versus-sufficiency asymmetry is informative, but it should still be treated as evidence for a measurable fact-support signal rather than as definitive circuit localization.

8 Conclusion

We presented a measurement framework for decomposing factual recall into storage, access, and competition using prompt families as access perturbations, matched controls as competition probes, layerwise invariance gaps as a representational signal, and fact-subspace interventions as an initial causal test. Across 6 small

language models and 90 facts, prompt-family preference is structured, replicated, and model-dependent, and the best prompt family for a given model is also the one that most cleanly separates the target fact from matched alternatives. More importantly, same-fact prompts exhibit a two-stage internal pattern: near-immediate subject-position invariance in 524 of 540 fact-model cases, followed by much later and less reliable pre-answer convergence in 454 of 540 cases. The control-aware subspace interventions add an important asymmetry to this picture: removing an early prompt-invariant fact component harms strong retrieval routes more than removing matched wrong-fact or random null components, whereas inserting it into weak routes rarely repairs them fully even though it outperforms those null patches.

These results do not yet support a claim of storage localization. Instead, it shows why that goal requires separate measurement of storage-related support, prompt-conditioned access, and answer-space competition. It also identifies a more concrete target for future work: the prompt-invariant internal state that survives early access variation and eventually supports answer convergence. Advancing beyond the present results will require moving from neuron rankings to fact subspaces, relation-conditioned operators, cross-family transfer tests, and scaling-style regularities in where convergent states emerge. That is the broader empirical program that this framework is intended to support.

References

Loubna Ben Allal et al. Smollm2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. doi: 10.48550/arXiv.2502.02737. URL https://arxiv.org/abs/2502.02737.

Trenton Bricken et al. Towards monosemanticity: Decomposing language models with dictionary learning. Anthropic, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. One mind, many tongues: A deep dive into language-agnostic knowledge neurons in large language models. *arXiv preprint arXiv:2411.17401*, 2024. doi: 10.48550/arXiv.2411.17401. URL https://arxiv.org/abs/2411.17401.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021. doi: 10.48550/arXiv.2104.08696. URL https://arxiv.org/abs/2104.08696.

Nelson Elhage et al. Toy models of superposition. Transformer Circuits Thread, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. doi: 10.48550/arXiv.2503.19786. URL https://arxiv.org/abs/2503.19786.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Evan Hernandez, Belinda Li, Jacob Andreas, Andrew McCallum, Antonio Torralba, and David Bau. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.751/.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*, 2024. doi: 10.48550/arXiv.2308.09124. URL https://arxiv.org/abs/2308.09124.

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

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Table 4: Exact prompt templates used in the 5-relation benchmark. The Q/A family is formed by wrapping the raw-question template as `Q: <question>\nA:`.

Relation	Raw question	Declarative completion	Chat-template user message
Capital	`What is the capital of {subject}?`	`The capital of {subject} is`	`What is the capital of {subject}? Answer with only the capital city.`
Official language	`What is an official language of {subject}?`	`An official language of {subject} is`	`What is an official language of {subject}? Answer with only one language.`
Currency	`What is the currency of {subject}?`	`The currency of {subject} is`	`What is the currency of {subject}? Answer with only the currency name.`
Headquarters	`Where is the headquarters of {subject}?`	`The headquarters of {subject} is in`	`Where is the headquarters of {subject}? Answer with only the city.`
Birth place	`Where was {subject} born?`	`{subject} was born in`	`Where was {subject} born? Answer with only the place name.`

REST Countries. REST Countries API v3.1. Online, 2026. URL https://restcountries.com/.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*, 2024. doi: 10.48550/arXiv.2310.15213. URL https://arxiv.org/abs/2310.15213.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

A Exact Prompt Families and Answer Normalization

Table 4 lists the exact per-relation prompt templates used in the benchmark. The Q/A family applies a universal wrapper to the raw-question template: `Q: <question>\nA:`. For chat-template prompts, the user message is first rendered from the relation-specific chat instruction and then passed through the tokenizer's native `apply_chat_template` method when available; otherwise the fallback wrapper is `User: <instruction>\nAssistant:`.

Before answer scoring, model output is normalized by lowercasing, removing punctuation and chat special tokens, and collapsing repeated whitespace. We then strip common answer wrappers and repeated prompt stems, including the relation-specific declarative prefix, the `Q:/A:` prefix, standalone prefixes such as `A`, `Answer`, and `Assistant`, and the fallback chat wrapper. The normalized answer is matched against the alias set stored with each fact.

B Release Artifacts and Reproducibility

The code, experiment configurations, and paper-relevant result summaries are publicly available under the MIT license at *[code repository anonymized for review]*. Model weights are not redistributed; they are obtained from their original Hugging Face sources through an included download script. The large raw activation artifacts (approximately 2.5 GB) are archived as a companion Hugging Face dataset at *[dataset anonymized for review]*.

The release package contains: 1) the filtered `Rest Countries` snapshot used for the country-based relations (capitals, official languages, and currencies); 2) the curated headquarters and birth-place seed file; 3) the rendered benchmark JSONL files for the prompt-family, storage-access, and subspace studies; 4) the exact YAML configuration files used for the revision experiments; 5) the per-model and aggregate multi-model summary tables (CSV), figures, and Markdown reports; and 6) the scripts that rebuild the aligned paper statistics.

The main pipeline runs in three stages, prompt-family analysis, storage-access analysis, and subspace intervention analysis, after which the paper statistics are regenerated from the result artifacts. All experiments

use the fixed seed (13) and greedy decoding settings reported above, and the six models are loaded sequentially so that the full comparison fits on a single NVIDIA GeForce RTX 4050 Laptop GPU 6 GB VRAM GPU.