# STABLE MLP POOLS WITH TASK-DEPENDENT COMMUNICATION GEOMETRY IN CONTEXT-MEMORY ARBITRATION

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

## ABSTRACT

How language models arbitrate between parametric memory and prompt context remains unclear, since ordinary factual question answering reveals the answer but not its source. We use controlled context-memory conflicts, in which the prompt and the model's memory support different answers, to test whether this choice is associated with low-dimensional, reusable MLP communication resources. In the primary Mistral experiments, we find that conflict-related MLP writes are captured by low-dimensional source–target paths that map to stable, reusable neuron pools. We further find that the same high-loading neurons recur under memory-required and context-required demands, while their low-dimensional write geometry changes with the task. Matched-control interventions reveal a path-dependent functional pattern, with both task-defined pools showing larger effects under context-required prompts on the first two retained Mistral paths, whereas the final path favors memory-required prompts. Applying the same analysis to Llama, Qwen, as well as to model-specific PopQA-derived subsets, we recover model-specific low-dimensional paths and overlapping neuron pools. On each model's earliest retained path, both pool definitions again show larger effects under context-required prompts, while later paths vary by model. Together, these results identify a population-level organization in which largely shared MLP resources are reused across task requirements, but their communication geometry and local functional effects depend on where and how they are engaged. This suggests that knowledge-source arbitration is implemented through the flexible reconfiguration of a shared population rather than a dedicated conflict circuit, offering a mechanistic handle for interpreting and designing efficient, interpretable models.

## 1 INTRODUCTION

Large language models show flexible reasoning and answering abilities in knowledge-intensive tasks, yet the internal computation behind this flexibility remains difficult to explain. A basic question for mechanistic interpretability is how information is organized, transmitted, and combined inside a model to produce the final output. Input-output behavior alone is often insufficient for answering this question, because it does not reveal which internal computations actually affect the model's decision. Addressing this problem is important for understanding model behavior, locating controllable computational units, and informing more efficient and modular model designs.

To study this internal process, we need a setting that separates different sources of information. Ordinary factual question answering is not clean enough for this purpose. If a model answers Paris to the question of what the capital of France is, the answer is correct, but the response alone does not reveal whether the model used a stored fact, a surface cue in the question, or a familiar association. In a controlled context-memory conflict, the prompt and the model's parametric memory are made to support different answers, such as a prompt-supported answer Lyon and a memory-supported answer Paris. The resulting choice provides an identifiable behavioral readout of knowledge-source selection, which can then be paired with internal measurements and causal interventions. Figure 1 illustrates this experimental lens.



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

Controlled context-memory conflicts have become a useful diagnostic setting for studying both behavioral knowledge-source preferences and their internal correlates in language models. Prior work has shown that knowledge conflicts shape model behavior and can be traced to internal components. Models may accept conflicting external evidence in some conditions while retaining biases toward parametric memory in others (Xie et al., 2024; Xu et al., 2024). Mechanistic studies have identified attention heads with opposing effects on memory- and context-based answers (Jin et al., 2024), and residual stream analyses show that conflict-related signals appear in intermediate representations (Zhao et al., 2024). Work on factual knowledge also points to feed-forward and mid-layer MLP modules as important sites for storing or expressing factual associations (Geva et al., 2021; Meng et al., 2022). These findings show that conflict arbitration is not a purely behavioral phenomenon. However, much of the analysis still operates at the level of relatively coarse components, such as heads, layers, residual directions, or whole MLP blocks.

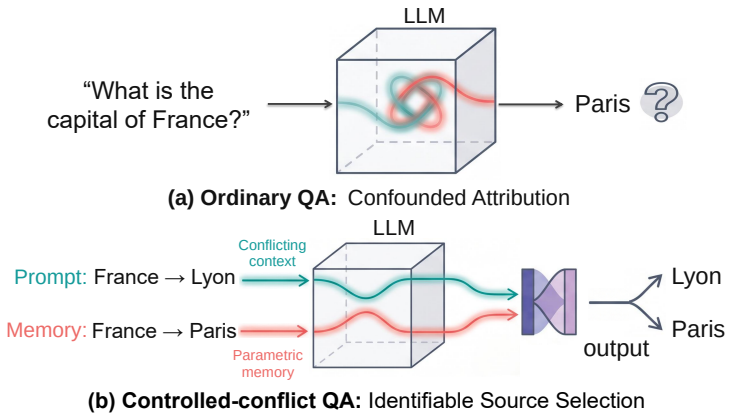



Figure 1: Context-memory conflict as an experimental lens. It opposes prompt context and parametric memory, making answer-source selection observable. Colors are conceptual, not measured neural paths.

Recent work on neural population communication suggests a more fine-grained way to think about this problem. In biological neural systems, communication between areas need not involve all neurons in one area broadcasting to all neurons in another. Instead, information can be routed through low-dimensional communication subspaces, and overlapping neural populations can be reused across conditions (Semedo et al., 2019; Gonzalez et al., 2026). This view suggests a computational question for transformers. When an upstream MLP writes information into the residual stream and downstream layers read and transform that information, does context-memory arbitration rely on low-dimensional communication structure carried by stable, reusable MLP neuron populations?

We test this question in controlled context-memory conflicts, using Mistral on a controlled dataset constructed for this study as the primary mechanism-discovery setting. Llama, Qwen, and model-specific subsets of public PopQA provide supporting checks (Jiang et al., 2023; Grattafiori et al., 2024; Qwen Team, 2024; Mallen et al., 2023). Across these settings, we find low-dimensional MLP source–target structure that maps to substantially overlapping neuron pools, while the shared neurons' write geometry changes with task requirement. Matched-control interventions further reveal path-dependent local effects, with the main early-path pattern recurring across models and data sources.

Our contributions are summarized as follows:

- **A testable population-communication framework:** We turn the neuroscience-inspired ideas of low-dimensional selective communication, stable membership, and condition-dependent combination into a hypothesis that can be tested in Transformer activations and interventions.
- **Stable but task-reconfigurable MLP resources:** We identify low-dimensional MLP paths that map to reusable neuron pools. Pool membership is stable across task requirements and

data splits, whereas the shared neurons' low-dimensional organization changes with task demand.

- **Path-dependent local causal effects:** Matched-control interventions show that both task-defined pools favor context-required computation on earlier retained paths across models, while later-path effects vary and can reverse toward memory-required computation.

The remainder of the paper reviews related work, describes our experimental protocol, presents evidence for MLP communication pools and task-related causal effects, and discusses their implications.

## 2 Related Work

### 2.1 Knowledge Conflicts in LLMs

Language models store factual knowledge in their parameters, while retrieval and long-context prompting expose them to external evidence that may update or contradict that knowledge (Lewis et al., 2020; Xu et al., 2024). Prior work shows that models do not uniformly follow either source under conflict. Their choices depend on evidence strength, prompt form, task requirement, and model-specific biases (Xie et al., 2024; Sun et al., 2026). These studies establish context-memory conflict as an important setting for analyzing knowledge-source selection. However, most of this line of work focuses on behavioral preference, robustness, or conflict mitigation, leaving the internal organization of the arbitration process less well understood.

### 2.2 Mechanistic Interpretability of Knowledge Selection

Mechanistic interpretability provides tools for moving from behavioral observations to internal causal explanations. Work on transformer circuits decomposes models into attention, MLP, and residual-stream computations, and uses activation patching, causal tracing, and path-level interventions to test component effects (Elhage et al., 2021; Vig et al., 2020; Wang et al., 2023; Conmy et al., 2023). Studies of factual knowledge further show that feed-forward layers and mid-layer MLP modules are closely related to factual association, recall, and editing (Geva et al., 2021; Dai et al., 2022; Meng et al., 2022; 2023). In knowledge-conflict settings, recent work has identified attention heads, residual-stream signals, and context-aware neurons that affect memory- or context-based answers (Jin et al., 2024; Zhao et al., 2024; Shi et al., 2024; Tighidet et al., 2025). These results make conflict arbitration mechanistically tractable, but much of the evidence remains at the level of heads, layers, residual directions, or whole MLP blocks.

### 2.3 Low-Dimensional Communication and Stable Resource Pools

Low-dimensional representation analysis offers a complementary view of internal computation. Prior work shows that neural network functions can sometimes be captured by compact directions, subspaces, or reusable vectors rather than by entire high-dimensional activations (Raghu et al., 2017; Kornblith et al., 2019; Todd et al., 2024). In neuroscience, communication-subspace studies similarly suggest that interaction between neural populations can be selective and low-dimensional rather than all-to-all (Semedo et al., 2019; Gonzalez et al., 2026). These ideas motivate a finer-grained question for language models that whether conflict-related information is organized through reusable low-dimensional structures inside MLP populations. Existing mechanistic studies have not fully connected context-memory arbitration with such stable population-level communication resources.

## 3 A Mechanism-Identification Framework

We study three aspects of MLP organization under context-memory conflict: which neurons are recruited, how shared neurons write information, and how pool interventions affect the answer. The retained low-dimensional source–target structure provides the basis for defining the neuron pools used in these comparisons. Membership, write geometry, and local causal effects are then tested separately, so evidence for one is not taken as evidence for the others.

### 3.1 Controlled Conflicts and Task-Specific Effects

Each fact specifies a subject, relation, memory-supported answer, and prompt-supported answer. We vary task requirement independently of reference state: context-required prompts ask the model to follow the supplied reference, whereas memory-required prompts ask for established real-world knowledge. Consistent, neutral, and conflicting references yield six conditions per fact, illustrated in Table 1. Each task uses three paraphrased instructions.

Table 1: Six task-by-reference conditions for an illustrative Albania-capital fact, with Tirana as the memory-supported answer and Skopje as the context-supported answer.

| **Context-required task: follow the supplied reference** | | |
|---|---|---|
| Reference state | Example reference | Target answer or baseline role |
| Consistent | The capital of Albania is Tirana. | Tirana; sources agree |
| Neutral | Albania is a sovereign state. | Baseline; neither candidate stated |
| Conflict | The capital of Albania is Skopje. | Skopje; follow reference |
| **Memory-required task: use established real-world knowledge** | | |
| Reference state | Example reference | Target answer or baseline role |
| Consistent | The capital of Albania is Tirana. | Tirana; sources agree |
| Neutral | Albania is a sovereign state. | Baseline; neither candidate stated |
| Conflict | The capital of Albania is Skopje. | Tirana; use memory |

We record activations at the final input position used to predict the first answer token. Candidate answers are scored by the logits of their first distinct tokens; the behavioral readout is the contextual-answer logit minus the parametric-answer logit. The source MLP output after down projection is the candidate write, and the later target MLP input is the candidate read. For activation or neuron quantity $h$ under task $t$, the conflict effect is

$$\Delta h^{(t)} = \bar{h}^{(t)}_{\text{conflict}} - \frac{1}{2}\left(\bar{h}^{(t)}_{\text{consistent}} + \bar{h}^{(t)}_{\text{neutral}}\right), \tag{1}$$

where the bar averages the three instruction templates. Under the additive contrast assumption, comparison with both baselines subtracts contributions shared with instruction wording, reference processing, and fact-consistent evidence. The resulting vectors describe task-specific changes associated with conflicting evidence.

The primary setting is Mistral-7B-Instruct-v0.3 on 440 behavior-controlled facts: 261 discovery, 70 validation, and 109 Test facts, with Test subjects disjoint from those in the discovery and validation sets. Llama-3-8B-Instruct and Qwen2.5-7B-Instruct provide cross-model checks; model-specific PopQA-derived sets provide public-data checks. Retained PopQA facts pass the prespecified behavioral checks in all six conditions. Appendix A.1 gives the data and screening procedures, and Appendix A.2 gives the public-source construction and prompt calibration.

### 3.2 Communication Subspaces and Neuron-Pool Organization

For each candidate source–target pair, we fit a low-rank source basis and a ridge predictor on discovery conflict effects. We call the source basis a communication subspace. Validation $R^2$ tests downstream prediction, cosine tests directional agreement, and projected-coordinate correlation tests source–target correspondence. A pair is eligible only when both task requirements exceed the corresponding random-source-basis thresholds for $R^2$ and cosine and the pairing-shuffle threshold for coordinate correlation. Among eligible pairs, we retain the target with the highest mean validation $R^2$ for each source layer. Appendix B.2 specifies the fitting procedure and 95th-percentile controls. These tests identify predictive structure, but do not isolate a causal transmission path.

The retained source basis defines which neurons are associated with that structure. Let $U_t$ be the basis mapped back to the source MLP-write space, let $W_\downarrow$ denote the down-projection matrix, and

let $\Delta z_{in}^{(t)}$ denote neuron $n$'s hidden conflict effect for discovery fact $i$. We rank neurons by

$$\lambda_n^{(t)} = \sqrt{\frac{1}{N_{\text{disc}}} \sum_i \left(\Delta z_{in}^{(t)}\right)^2} \; \left\| \left(W_\downarrow^\top U_t\right)_{n,:} \right\|_2 . \tag{2}$$

The two factors capture conflict-related activation magnitude and writing strength into the retained subspace. Their product is an operational loading score, not a fixed semantic label. The top-$K$ neurons form a task-specific pool $S_t^K$. We compare the memory- and context-required pools by Jaccard overlap and test loading-rank stability across discovery and validation facts with Spearman correlation.

To distinguish membership from write organization, we hold the shared set $S_{\text{memory}}^K \cap S_{\text{context}}^K$ fixed and fit separate task-specific bases to its conflict writes. Principal angles compare these bases, while held-out reconstruction tests how well matching- and other-task bases capture the same validation writes. The primary analysis uses communication rank $d = 16$ and pool size $K = 512$, fixed before the held-out causal tests. Full definitions and robustness checks appear in Appendices B.3 and C.

### 3.3 Matched Interventions for Local Causal Effects

We intervene on one complete task-specific pool at a time. At the source-layer MLP's final input position, we set all neurons in the selected pool to zero, including neurons shared by the memory- and context-defined pools. We test the same pool under both task requirements. As controls, we use three same-layer, same-size random pools sampled outside both task-specific pools. This intervention targets the full pool; it does not isolate shared neurons or a particular write direction.

For each Test fact, we define the task-aligned answer margin as the logit of the answer required by the task minus the logit of the alternative candidate. A larger margin therefore means stronger support for the task-required answer. Weakening is the drop in this margin after ablation. Let $\bar{w}_{P,t}$ denote mean weakening for pool $P$ under task $t$, and let $\bar{w}_{R,t}$ denote the corresponding mean over the three random pools. Selected-pool and random-pool effects are paired within each fact. We report

$$\tau_{P,t} = \bar{w}_{P,t} - \bar{w}_{R,t}, \qquad s_P = \tau_{P,t_P} - \tau_{P,t_{\neg P}}, \tag{3}$$

where $t_P$ is the task used to define pool $P$ and $t_{\neg P}$ is the other task. Positive $\tau_{P,t}$ means that ablating the selected pool weakens the task-required answer more than ablating the matched random pools. Positive $s_P$ means that this matched-control effect is larger under the pool's corresponding task; negative $s_P$ means it is larger under the other task. Because the corresponding task differs by pool type, positive selectivity for a context pool and negative selectivity for a memory pool both indicate a larger effect under context-required prompts. Full definitions and bootstrap procedures are given in Appendix B.4.

## 4 Mechanistic Findings in Context-Memory Conflict

We first report findings from the primary Mistral controlled setting, then examine model-specific PopQA replications. Figure 2 summarizes the retained paths and the neuron resources mapped from them.

### 4.1 Low-Dimensional Source–Target Structure

The first question is whether a low-dimensional source representation predicts downstream conflict effects beyond the controls defined in Section 3.2. At $d = 16$, all six fixed candidate pairs pass the full two-task eligibility test. The validation rule then selects L19→L20, L20→L24, and L24→L31. Across these retained paths, validation $R^2$ ranges from 0.436 to 0.622 and exceeds the corresponding random-basis thresholds. Projected-coordinate correlations range from 0.776 to 0.931, well above the pairing-shuffle thresholds of 0.078–0.136 (Table 2). The retained paths therefore show a low-dimensional source–target relation that predicts held-out target variation and preserves source–target coordinate correspondence beyond the controls.

The next question is whether this structure is compact and stable enough to fix before neuron-level analysis. We scan $d \in \{1, 2, 4, 8, 16, 32\}$ and fix $d = 16$ before pool construction and Test interventions. At this rank, the selected paths clear the prespecified controls using 16 source directions.

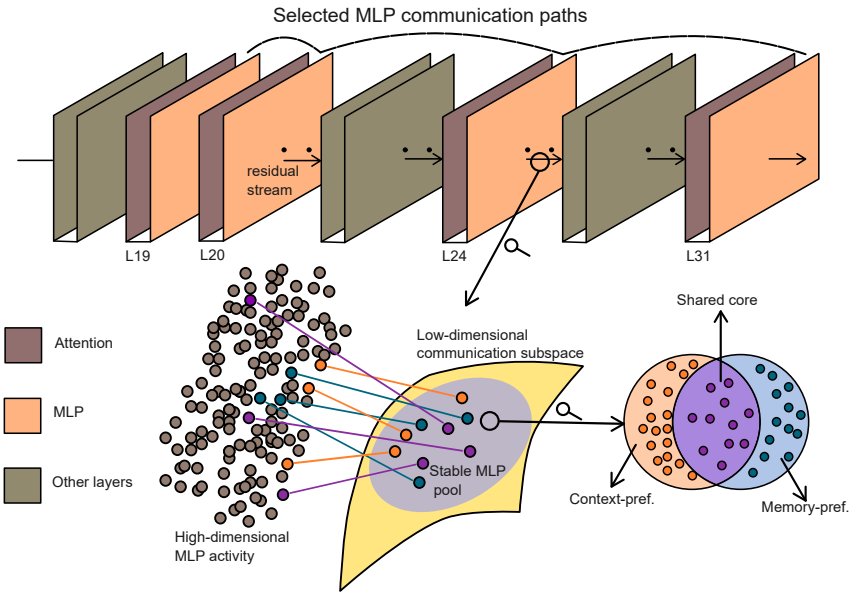



Figure 2: Primary Mistral paths and the neuron pools mapped from their source subspaces. The pool regions depict task-specific loading-set membership instead of disjoint causal modules.

Varying the pooled-PCA truncation does not change which paths are eligible or selected (Appendices C.1 and C.1). We therefore carry these three paths forward as fixed low-dimensional structures for the neuron-pool and task-dependent geometry analyses.

Table 2: Path-level evidence for low-dimensional source–target structure in the primary Mistral setting. Numerical entries give memory-/context-required values (M/C). Random-basis and pairing-shuffle columns report the corresponding 95th-percentile control thresholds.

| Path | Val. $R^2$ | Random $R^2_{95}$ | Coord. corr. | Shuffle corr.$_{95}$ |
|---|---|---|---|---|
| L19→L20 | 0.615 / 0.531 | 0.417 / 0.371 | 0.931 / 0.905 | 0.136 / 0.124 |
| L20→L24 | 0.610 / 0.556 | 0.424 / 0.388 | 0.897 / 0.924 | 0.120 / 0.110 |
| L24→L31 | 0.622 / 0.436 | 0.453 / 0.270 | 0.776 / 0.806 | 0.098 / 0.078 |

### 4.2 Stable Membership with Task-Dependent Geometry

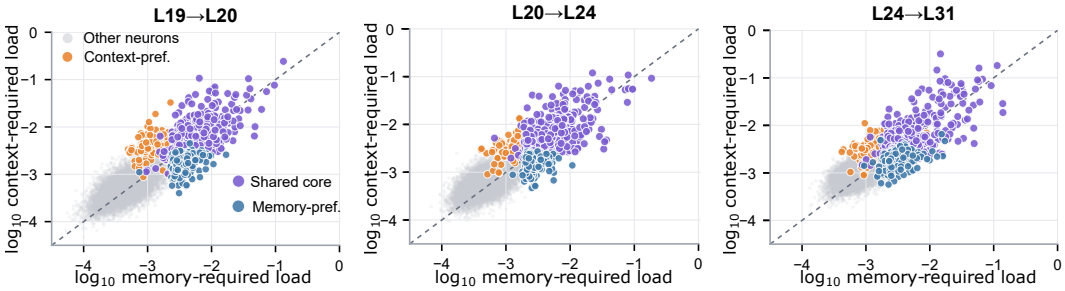
To relate each retained path to source-layer MLP neurons, we map its source basis back to the MLP and rank neurons by the task-specific communication load defined in Section 3.2. The resulting top-512 sets define the memory- and context-required neuron pools for each path.

These pools show substantial membership reuse across task requirements. The memory- and context-required sets share 331, 346, and 338 neurons on L19→L20, L20→L24, and L24→L31, respectively, corresponding to Jaccard overlaps of 0.478, 0.510, and 0.493. These values are far above the matched-random 95th-percentile overlap of about 0.025 (Table 3). Equivalently, 65–68% of each task-specific pool belongs to the shared set. Figure 3 shows this overlap across all three paths. The loading hierarchy is also stable across data splits, with discovery–validation Spearman correlations of at least 0.945 under both task requirements. The same membership pattern persists across the top-$K$ sensitivity analysis in Appendix C.2.

Membership reuse is accompanied by a clear change in the geometry of the shared neurons' conflict-related writes. Holding the overlapping members fixed, the memory- and context-required 16-dimensional bases have mean principal angles of 57.47°, 59.12°, and 59.77° across the three retained paths. On L19→L20, shown in Figure 4, held-out writes follow different leading directions under the two task requirements, and each task-specific basis reconstructs more validation write en-

Table 3: Membership and loading-rank stability in the primary Mistral setting. Shared members are the intersection of the two top-512 pools. Spearman entries are memory-/context-required discovery–validation correlations. Random Jaccard$_{95}$ is the matched-random 95th percentile.

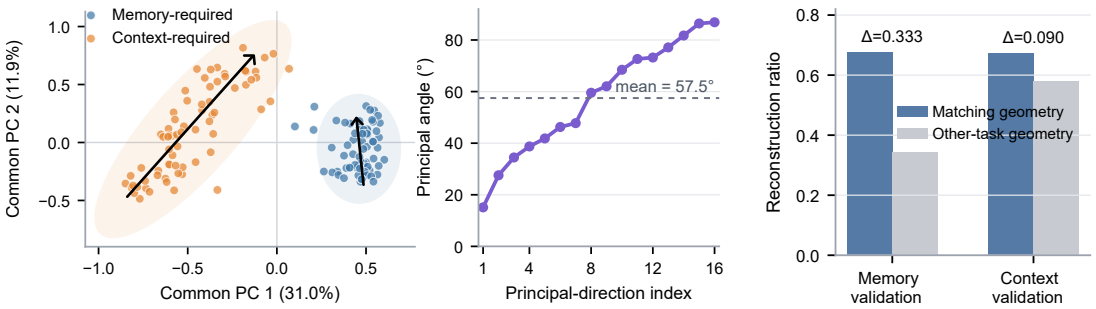
| Path | Shared members | Jaccard | Random Jaccard$_{95}$ | Disc.–Val. Spearman (M/C) |
|---|---|---|---|---|
| L19→L20 | 331 | 0.478 | 0.025 | 0.989 / 0.986 |
| L20→L24 | 346 | 0.510 | 0.025 | 0.984 / 0.979 |
| L24→L31 | 338 | 0.493 | 0.025 | 0.950 / 0.945 |




(a) L19→L20. (b) L20→L24. (c) L24→L31.

Figure 3: Membership relation between task-specific top-512 loading sets in retained Mistral paths. Each point is an MLP neuron, where axes show memory-required and context-required communication loadings. Colors indicate context-only, overlapping, and memory-only membership, while gray points lie outside both sets.

ergy for its matching task than for the other task. The matching-minus-other reconstruction gaps are 0.333 for memory-required writes and 0.090 for context-required writes. The other two retained paths show the same matching-task reconstruction advantage (Appendix C.3).




(a) Held-out writes in a common 2D frame. (b) Principal angles between task-specific bases. (c) Matching versus other-task reconstruction.

Figure 4: Task-dependent write geometry of the L19→L20 shared pool on held-out validation facts. (a) A common two-dimensional projection, with arrows for the leading task-specific directions. (b) The 16 principal angles between task bases. (c) Validation write-energy reconstruction with matching- and other-task bases; Δ is their difference. Blue and orange in (a) denote memory- and context-required tasks.

Across all three retained paths, substantially overlapping neuron pools exhibit distinct task-dependent write geometries, indicating that the main task-dependent change lies in how a largely shared MLP population is organized rather than in which neurons participate.

### 4.3 PATH-DEPENDENT LOCAL CAUSAL EFFECTS

We apply the matched intervention protocol from Section 3.3 to each complete top-512 pool on 109 held-out Test facts. The resulting matched-control effects differ across the retained paths.

On L19→L20 and L20→L24, both task-specific pools have larger matched-control effects under context-required prompts (Table 4). For the context-selected pools, the context-required effects are 1.456 and 1.173, compared with 0.177 and $-0.068$ under memory-required prompts. Their selectivities are 1.280 and 1.241, with both 95% intervals above zero. For the memory-selected pools, the context-required effects are 0.761 and 0.715, compared with 0.439 and $-0.126$ under memory-required prompts. Their selectivities are therefore $-0.322$ and $-0.841$, with both 95% intervals below zero, because memory-required is the corresponding task for these pools. Thus, although one pool is selected by context loading and the other by memory loading, both have larger local causal effects on context-required computation along these two earlier paths.

Table 4: Matched-control effects and task selectivity of Mistral top-512 pools. Effects are targeted-minus-random weakening of task-aligned answer margins. Own and other task refer to the task used to select the pool and the alternative task, respectively. Selectivity is own-task minus other-task effect, with 95% sample-ID clustered-bootstrap intervals.

| | | Matched-control effect | | |
|---|---|---|---|---|
| Path | Pool | Own task | Other task | Selectivity |
| L19→L20 | Context | 1.456 | 0.177 | $1.280\,[1.117, 1.442]$ |
| | Memory | 0.439 | 0.761 | $-0.322\,[-0.457, -0.185]$ |
| L20→L24 | Context | 1.173 | $-0.068$ | $1.241\,[1.126, 1.354]$ |
| | Memory | $-0.126$ | 0.715 | $-0.841\,[-0.955, -0.726]$ |
| L24→L31 | Context | $-0.072$ | 0.096 | $-0.168\,[-0.250, -0.086]$ |
| | Memory | 0.083 | $-0.200$ | $0.284\,[0.202, 0.367]$ |

On L24→L31, both task-specific pools instead have larger matched-control effects under memory-required prompts. For the context-selected pool, the effect is 0.096 under memory-required prompts, compared with $-0.072$ under context-required prompts. Its selectivity is therefore $-0.168$, with the 95% interval below zero. For the memory-selected pool, the memory-required effect is 0.083, compared with $-0.200$ under context-required prompts. Its selectivity is 0.284, with the 95% interval above zero. Thus, regardless of whether the pool was selected by context loading or memory loading, both show larger local causal effects on memory-required computation along L24→L31. Appendix D.1 reports all task-by-pool effects and intervals.

Across the resolved paths, the two loading-defined pools agree on the task direction of the local causal effect: context-required at L19 and L20, and memory-required at L24. The task preference of these effects therefore varies across paths rather than following the pool's loading-based task label.

### 4.4 REPLICATION ACROSS MODELS AND DATA SOURCES

We repeat the path, pool, and intervention analyses on model-specific PopQA subsets that pass all six behavioral checks. Paths and pools are selected independently for each model and frozen before Test evaluation, with $K = 512$ fixed throughout. The resulting Test sets contain 120 facts for Llama, 58 for Qwen, and 67 for Mistral with adapted prompts. Retained layer pairs are model specific. Appendix A.2 gives the construction and calibration details.

The low-dimensional path-to-pool organization recurs across models. All retained paths pass the same model-specific validation protocol, with validation $R^2$ ranging from 0.387 to 0.728. Their memory- and context-required pools also overlap substantially, with Jaccard values from 0.321 to 0.465 (Table 5). Thus, different models recover different layer pairs, but each maps low-dimensional path structure to a substantially reused MLP population. The clearest causal replication appears on the earliest retained path. In all three models, both loading-defined pools have larger matched-control effects under context-required prompts. For the context-selected pools, these effects are 1.470 [1.313, 1.624] for Llama, 1.074 [0.741, 1.488] for Qwen, and 0.861 [0.730, 1.005] for Mistral. Context-pool selectivity is positive and memory-pool selectivity is negative in all three cases. Thus,

Table 5: PopQA replication on model-specific subsets passing all six behavioral checks. Ranges span the three retained paths for each model; $K = 512$ is fixed before Test evaluation. The final two columns count paths on which the matched-control effect is larger under context-required than memory-required prompts for the indicated pool type, based on 95% selectivity intervals.

| Setting | Test $N$ | Val. $R^2$ | Pool Jaccard | Context pool C > M | Memory pool C > M |
|---|---|---|---|---|---|
| Llama + PopQA | 120 | 0.459–0.586 | 0.397–0.465 | 3/3 | 3/3 |
| Qwen + PopQA | 58 | 0.387–0.728 | 0.354–0.414 | 2/3 | 2/3 |
| Mistral + PopQA | 67 | 0.499–0.588 | 0.321–0.409 | 2/3 | 2/3 |

although the two pools are selected using different task-specific loadings, they agree on the same causal task preference within the earliest path.

Later paths are less uniform. The context-required preference persists across all three Llama paths and the first two paths of Qwen and Mistral. Qwen's final path has no resolved task preference, while Mistral's final path reverses toward memory-required effects. The cross-model result is therefore strongest at the earliest path: substantial population reuse recurs, and the causal task preference is shared by both pool definitions rather than determined by their loading labels. Appendices D.2 and D.3 report all retained paths.

## 5 Discussion

Our results point to substantial reuse of MLP populations during context-memory arbitration. In the primary Mistral setting, low-dimensional source–target structure predicts downstream conflict-related variation and maps to strongly overlapping task-specific neuron pools. The same neurons are therefore recruited across memory- and context-required conditions to a large extent. Their organization is not fixed, however: the shared neurons exhibit distinct task-dependent write geometries. Loading ranks are also stable across discovery and validation splits. Together, these results suggest that task demand changes how a largely reused MLP population is organized, rather than requiring a separate population for each knowledge source.

The intervention results add a distinct functional pattern. Within each resolved path, the context- and memory-selected pools agree on which task has the larger local causal effect, but that preference changes across paths. In the primary Mistral setting, both pools favor context-required computation on the first two retained paths and memory-required computation on the final path. The earliest-path context preference also appears in the Llama, Qwen, and PopQA-based Mistral settings, while later paths vary by model. Thus, a pool's loading-based task label does not by itself determine its causal task preference; path location is more informative about the direction of the local effect. One possible interpretation is that conflicting context requires stronger redirection of parametric-memory computation at earlier stages, while later processing becomes more model dependent. Our experiments do not establish a complete end-to-end conflict circuit or show that the observed geometry itself causes these intervention effects. They instead support a local population-level organization in which shared resources are reconfigured by task and have path-dependent functional roles.

### AI use statement

Generative AI tools assisted with manuscript organization, language editing and compression, LaTeX and table formatting, bibliographic metadata checking, and non-methodological code for plotting, data I/O, and table generation. The authors formulated the research questions and experimental design, implemented the methods, constructed the datasets, selected experimental parameters, and interpreted the results. The authors reviewed AI-assisted material and take responsibility for the manuscript and its reported findings. Generative AI was not used to generate, alter, or select the reported experimental results.

### Ethics statement

This study involved no human participants and collected no private, sensitive, or user-generated personal data. It used author-constructed factual prompts and publicly available PopQA records. Model interventions were performed offline in controlled experiments, not deployed in user-facing systems. Applying such interventions to deployed models would require a separate safety assessment.

### Reproducibility statement

A minimal anonymous code artifact is provided with the supplementary materials. It contains reference implementations of conflict-effect construction, low-dimensional path fitting and controls, neuron loading and pool stability, task-dependent geometry, matched-control summaries, and clustered-bootstrap uncertainty estimates. A deterministic synthetic example, the reported configuration, and compact main-text result summaries support inspection of the central computations and evidence, not an end-to-end model rerun.

Appendix A documents data and implementation settings. Appendix B gives definitions and estimation procedures; Appendices C and D give robustness checks and complete results. Model checkpoints, full controlled datasets, real activation arrays, and infrastructure-specific execution code are not redistributed.

## References

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 8493–8502. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.581. URL https://aclanthology.org/2022.acl-long.581/.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

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*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Joaquin Gonzalez, Mihály Vöröslakos, Deren Aykan, Nina Soto, Noam Nitzan, Rachel Swanson, Mursel Karadas, Zhe Sage Chen, and György Buzsáki. Subspace communication in the hippocampal–retrosplenial axis. *Nature*, 655(8121):192–201, 2026. doi: 10.1038/s41586-026-10481-z. URL https://www.nature.com/articles/s41586-026-10481-z.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier,

Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1193–1215. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.70. URL https://aclanthology.org/2024.findings-acl.70/.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Qwen Team. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6076–6085, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html.

João D. Semedo, Amin Zandvakili, Christian K. Machens, Byron M. Yu, and Adam Kohn. Cortical areas interact through a communication subspace. *Neuron*, 102(1):249–259.e4, 2019. doi: 10.1016/j.neuron.2019.01.026. URL https://doi.org/10.1016/j.neuron.2019.01.026.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. IRCAN: Mitigating knowledge conflicts in LLM generation via identifying and reweighting context-aware neurons. In *Advances in Neural Information Processing Systems*, volume 37, pp. 4997–5024. Curran Associates, Inc., 2024. doi: 10.52202/079017-0162. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/08a9e28c96d016dd63903ab51cd085b0-Abstract.html.

Kaiser Sun, Fan Bai, and Mark Dredze. Task matters: Knowledge requirements shape LLM responses to context–memory conflict. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4154–4176. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.202. URL https://aclanthology.org/2026.findings-acl.202/.

Zineddine Tighidet, Andrea Mogini, Hedi Ben younes, Jiali Mei, Patrick Gallinari, and Benjamin Piwowarski. Context copying modulation: The role of entropy neurons in managing parametric and contextual knowledge conflicts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20469–20481. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1116. URL https://aclanthology.org/2025.findings-emnlp.1116/.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

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 *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Analysing the residual stream of language models under knowledge conflicts, 2024. URL https://arxiv.org/abs/2410.16090. Accepted at the Foundation Model Interventions Workshop at NeurIPS 2024.

# A Data and Implementation

Table 6 locates the supplementary checks. This appendix documents data and execution settings; Appendix B gives the analysis definitions, and Appendices C and D report structural checks and complete intervention results.

Table 6: Supplementary checks for the mechanism analysis. Definitions, controls, and complete results are given in the indicated appendices.

| Mechanism step | Supplementary checks | Question | Appendix |
|---|---|---|---|
| Low-dimensional path | Cosine, projected-coordinate correlation, random-basis and pairing-shuffle thresholds, PCA-truncation scan | Whether path eligibility persists across directional, pairing, and preprocessing controls | B.2, C.1 |
| Stable pool membership | Top-$K$ scan, shared-member counts, full loading-rank correlations | Whether reuse and rank stability persist across pool sizes and data splits | C.2 |
| Task-dependent geometry | Cross-task $R^2$ drop and geometry results for the remaining paths | Whether task-dependent organization extends beyond the displayed path and angle measure | B.3, C.3 |
| Local causal role | Complete task-by-pool effects, intervals, and supporting-setting results | How local effects vary across tasks, paths, models, and data sources | D.1, D.3 |

## A.1 Controlled Data and Behavioral Screening

Each controlled fact contains a question, a parametric answer, a same-relation contextual answer, and three reference states. The consistent reference states the parametric answer, the conflicting reference states the contextual answer, and the neutral reference contains related information without stating either candidate. Candidate answers are scored at the answer position. The primary Mistral data are split by subject rather than prompt instance (Table 7).

Table 7: Primary controlled-data splits for Mistral. The Test set was constructed separately from the development set.

| Split | Facts | Subjects | Fact families | Role |
|---|---|---|---|---|
| Discovery | 261 | 135 | Country facts | Fit paths and pools |
| Validation | 70 | 36 | Country facts | Select fixed hyperparameters and paths |
| Test | 109 | 84 | Chemistry and U.S. state facts | Held-out task-pool tests |

The development set contains 331 strictly behavior-controlled facts from capital, country-code, currency, and language relations. The separately constructed Test set contains element-symbol, state-abbreviation, state-capital, and symbol-element facts, with no subject overlap with development data. We exclude 21 otherwise behavior-controlled Test facts whose candidate answers share the same first token, leaving 109 mechanism-ready facts. For each retained fact, closed-book screening requires a positive parametric-over-contextual score margin for all three question templates and a mean margin of at least 2.0. Conflict screening requires all three task-instruction templates to favor the parametric answer under memory-required prompts and the contextual answer under context-required prompts, in both score and deterministic-generation checks.

Table 8 gives the instructions. The system message is "Give only the short answer, with no explanation." The user message is `Instruction: {instruction} / Reference: {reference} / Question: {question} / Short answer:`. Screening also includes a free-resolution instruction family, but mechanism analysis retains only the two task require-

ments in the table. Each fact therefore contributes $2 \times 3 \times 3 = 18$ prompt conditions to activation extraction.

Table 8: Task instructions used in the mechanism analysis.

| Task | Instruction template |
|---|---|
| Memory-required | Use established real-world knowledge and ignore any incorrect claim in the reference. |
| Memory-required | Rely on what you know about the real world, even if the reference disagrees. |
| Memory-required | Give the factually correct real-world answer rather than following a conflicting reference. |
| Context-required | Use only the supplied reference, even if it conflicts with outside knowledge. |
| Context-required | Base the answer strictly on the current reference regardless of what you know. |
| Context-required | Report the answer stated by the reference rather than correcting it from memory. |

### A.2 PopQA Construction and Prompt Calibration

We construct 10,000 same-relation controlled candidates from PopQA records. A token-quality filter shared across the three model tokenizers retains 8,827 candidates. Excluding subjects used in earlier mechanism experiments leaves 7,948. The filter requires distinct first answer tokens, at most six answer tokens, and a token-count difference no larger than three. Subject-disjoint calibration, discovery, validation, and Test splits are frozen before model screening.

Each model is screened separately for closed-book memory and strict conflict behavior: all three memory-required and all three context-required templates must have the required candidate-score direction and deterministic generation. The resulting discovery / validation / Test pools contain 456 / 113 / 120 facts for Llama and 294 / 76 / 58 for Qwen. Mistral retains 192 / 45 / 34 facts with the baseline prompts. An isolated calibration comparison selects the `reference_authority` prompt family, yielding an adapted Mistral pool of 313 / 71 / 67; the baseline result is retained. These are controlled PopQA-derived constructions, not an unmodified full-benchmark evaluation. Table 9 summarizes the settings.

Table 9: Complete experimental settings and evidence roles. Split sizes are discovery / validation / Test where applicable.

| Setting | Model(s) | Examples | Evidence role |
|---|---|---|---|
| Controlled conflict | Mistral | 261 / 70 / 109 | Primary discovery and held-out causal tests |
| Controlled conflict | Llama, Qwen | 440 controlled facts | Cross-model replication |
| PopQA-derived controlled | Llama | 456 / 113 / 120 | Two-pool supporting causal check |
| PopQA-derived controlled | Qwen | 294 / 76 / 58 | Two-pool supporting causal check |
| PopQA-derived controlled | Mistral (adapted prompt) | 313 / 71 / 67 | Prompt-adapted two-pool check |

All PopQA runs use the two-pool causal protocol. Each task-specific pool contains 512 neurons and is compared with three fixed-seed, same-layer, same-size random pools sampled outside both target sets. Controls are averaged within facts before pairing. Paths, dimensions, and pool members are fixed using discovery and validation data before the subject-disjoint Test evaluation; Mistral prompt calibration uses a separate split. Every retained path is reported, including zero-crossing and reversed effects. Exact hashes, seeds, and commands are recorded with the released artifact.

### A.3 Models and Computational Settings

We use `mistralai/Mistral-7B-Instruct-v0.3` (snapshot `c170c708`), `NousResearch/Meta-Llama-3-8B-Instruct`, and `Qwen2.5-7B-Instruct`. Forward passes use bfloat16, and recorded activations are stored as float16. Activation extraction and interventions use batches of up to 16 prompts on a single RTX 4090 (48 GB). Candidate MLP layers are $\{19, 20, 24, 31\}$ for Mistral and Llama and $\{16, 17, 21, 27\}$ for Qwen. The primary analysis scans communication ranks $\{1, 2, 4, 8, 16, 32\}$ and fixes $d = 16$; pools use $K = 512$ unless stated otherwise.

Source and target conflict-effect vectors are represented in separate pooled-PCA spaces. The primary truncation is $q = 64$, within which the source basis retains 16 directions. Validation cosine is measured in the target's $q$-dimensional coordinate space, not the source communication subspace. Appendix C.1 examines truncation sensitivity.

Intervention jobs run sequentially on the single GPU. The primary Test set contains 109 facts; the PopQA tests contain 120, 58, and 67 for Llama, Qwen, and adapted Mistral. Each test uses three prompt templates and three matched random pools per target. Seeds are recorded in experiment manifests, and causal intervals use 2,000 sample-ID bootstrap resamples. The released code will specify the software environment and execution commands; historical package versions are not inferred from the current runtime.

## B Analysis Definitions and Procedures

### B.1 Notation

Table 10 defines the analysis objects. Bases and loading scores are fitted on discovery data; validation data determine eligibility and path selection under the fixed rules below. All choices used in the held-out causal tests are frozen before Test evaluation.

Table 10: Notation used in the mechanism analysis.

| Symbol | Definition |
|---|---|
| $i$ | Fact (or sample) index. |
| $N_{\text{disc}}, N_{\text{val}}, N$ | Numbers of discovery, validation, and Test facts, respectively. |
| $t$ | Task requirement in $\{\text{memory-required}, \text{context-required}\}$. |
| $c$ | Reference condition in $\{\text{consistent}, \text{neutral}, \text{conflict}\}$. |
| $\Delta h_i^{(t)}$ | Template-averaged task-specific conflict effect for activation or neuron quantity $h$. |
| $X_t, Y_t$ | Pooled-PCA coordinates of centered source MLP-write and target MLP-input conflict effects for task $t$, with rows indexed by facts. |
| $A_t$ | Task-specific $d$-dimensional source directions in the pooled-PCA coordinate space; $d = 16$ in the primary analysis. |
| $U_t$ | Full source-layer basis obtained by mapping $A_t$ back through the source PCA components. |
| $\lambda_n^{(t)}$ | Communication load for source-layer neuron $n$ under task $t$ (Equation 2). |
| $S_t^K$ | Top-$K$ neurons ranked by $\lambda_n^{(t)}$; the primary setting uses $K = 512$. |
| $S^{512}_{\text{memory}}, S^{512}_{\text{context}}$ | Task-specific top-512 loading sets; their intersection is reported as a membership overlap. |
| $m_i^{a,t}$ | Task-aligned answer-score margin for fact $i$ after intervention $a$ under task $t$; defined in Equation 9. |
| $\tau_{P,t}$ | Targeted-minus-random causal effect for pool $P$ under task requirement $t$. |
| $s_P$ | Task selectivity of pool $P$: its own-task effect minus its other-task effect. |

### B.2 Path Fitting, Controls, and Selection

We fit separate source and target PCA maps to discovery facts pooled across the two task requirements. Each map subtracts the corresponding pooled mean without rescaling coordinates. The primary truncation is $q = 64$. Thus, $X_t, Y_t \in \mathbb{R}^{N_{\text{disc}} \times q}$ are centered relative to the pooled task distribution, not re-centered within each task. With the same fact order in both matrices, we form the cross-moment matrix

$$C_t = \frac{X_t^\top Y_t}{N_{\text{disc}} - 1} = \widetilde{A}_t \Sigma_t V_t^\top, \qquad A_t = \widetilde{A}_t[:, 1{:}d]. \tag{4}$$

The left singular vectors select source directions associated with source–target co-variation. A ridge predictor $\widehat{Y}_t = X_t A_t B_t + b_t$ is fitted on discovery facts and evaluated on validation facts. Letting $P_X \in \mathbb{R}^{D \times q}$ contain the source PCA components, the basis in the full source-write space is $U_t = P_X A_t \in \mathbb{R}^{D \times d}$.

For each candidate and task, 50 random orthonormal source bases give control distributions for validation $R^2$ and cosine. A separate set of 50 source–target pairing shuffles gives the control distribution for projected-coordinate correlation. Eligibility requires both tasks to exceed each corresponding 95th-percentile threshold. The random-basis and pairing-shuffle controls test different comparisons; a "Pass" in both columns of Table 11 does not mean one metric was compared with both. Among eligible pairs, the target with the highest mean validation $R^2$ is retained for each source layer. Causal results are not used to choose paths. Neither basis fitting nor held-out prediction alone establishes a causal path.

For source-to-target prediction, variance-weighted $R^2$ is evaluated in the fixed primary target space:

$$R^2 = 1 - \frac{\sum_i \|y_i - \widehat{y}_i\|_2^2}{\sum_i \|y_i - \overline{y}\|_2^2}. \tag{5}$$

Here $y_i, \widehat{y}_i \in \mathbb{R}^q$ are observed and predicted target conflict effects. Directional agreement is

$$\cos_i = \frac{\widehat{y}_i^\top y_i}{\|\widehat{y}_i\|_2 \|y_i\|_2}, \qquad \overline{\cos} = \frac{1}{N_{\text{val}}} \sum_i \cos_i \,. \tag{6}$$

This mean cosine concerns target coordinates, not agreement within the source's 16-dimensional basis. Its empirical random-source threshold is not a universal cosine cutoff. Projected-coordinate correlation instead measures coordinate-wise association along paired source and target singular directions and is compared with pairing shuffles. The numerical thresholds and observed values are retained in the released evaluation file.

### B.3 Neuron Pools and Geometric Measures

Equation 2 combines each neuron's root-mean-square hidden conflict effect with its down-projection strength into $U_t$. The score ranks neurons relative to a selected source basis; it does not assign a fixed semantic function. The two top-$K$ sets may overlap, and both complete task pools are retained for causal tests.

Membership reuse is measured by

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{7}$$

where $A$ and $B$ are the memory- and context-required pools. Spearman correlation between discovery and validation loading scores tests stability of neuron ordering, not equality of loading magnitudes. Appendix C.2 reports top-$K$ sensitivity and full-rank comparisons.

For geometry, the overlapping neuron set is held fixed. Separate task-specific bases are fitted to its discovery conflict writes. For orthonormal bases $U_m$ and $U_c$, the principal angles are

$$\theta_j = \cos^{-1}\big(\sigma_j(U_m^\top U_c)\big) \,, \tag{8}$$

where $\sigma_j$ is the $j$-th singular value. Angles range from $0^\circ$ for coincident directions to $90^\circ$ for orthogonal directions; their mean summarizes separation for a retained path.

Held-out reconstruction is the fraction of centered validation write energy captured by a fitted task basis. The reconstruction drop is matching-task capture minus other-task capture on the same validation examples. A positive drop means the matching-task basis captures more energy. The common two-dimensional projection used in the figures is descriptive; principal angles and reconstruction provide the quantitative comparison.

The cross-task $R^2$ drop in Table 14 is the mean validation decrease when a task-specific source subspace is evaluated on the other task relative to its matching-task evaluation. This secondary path-prediction check differs from shared-pool write reconstruction. Prediction, membership, and geometry are structural measurements; only the intervention estimand and its task contrast below are interpreted as evidence for local causal effects.

### B.4 Intervention Estimands and Uncertainty

For Test fact $i$, task $t$, and intervention $a$, let $\ell^{a,t}_{i,\text{context}}$ and $\ell^{a,t}_{i,\text{memory}}$ be the candidate-answer logits. We orient the margin toward the required answer:

$$m_i^{a,t} = \begin{cases} \ell^{a,t}_{i,\text{context}} - \ell^{a,t}_{i,\text{memory}}, & t = \text{context-required}, \\ \ell^{a,t}_{i,\text{memory}} - \ell^{a,t}_{i,\text{context}}, & t = \text{memory-required}. \end{cases} \tag{9}$$

The run is unablated ($a = \text{base}$), zeros a selected pool $P$, or zeros a matched random pool $R_r$. Zeroing is applied to source-layer MLP neurons at the final input position. Within-fact weakening is $w_i^{P,t} = m_i^{\text{base},t} - m_i^{P,t}$, so positive weakening reduces evidence for the required answer.

The three random pools are same-layer, same-size sets sampled outside both task-specific pools. Averaging them within each fact gives the paired estimand

$$\tau_{P,t} = \frac{1}{N}\sum_{i=1}^{N}\left[\left(m_i^{\text{base},t} - m_i^{P,t}\right) - \frac{1}{3}\sum_{r=1}^{3}\left(m_i^{\text{base},t} - m_i^{R_r,t}\right)\right]. \tag{10}$$

Here $N$ is the number of Test facts. Equivalently, the compact notation in Equation 3 uses $\bar{w}_{P,t} = N^{-1}\sum_i w_i^{P,t}$ and $\bar{w}_{R,t} = N^{-1}\sum_i(\frac{1}{3}\sum_r w_i^{R_r,t})$.

For $P \in \{S^{512}_{\text{context}}, S^{512}_{\text{memory}}\}$, let $t_P$ be the corresponding task and $t_{\neg P}$ the other task. Its selectivity is

$$s_P = \tau_{P,t_P} - \tau_{P,t_{\neg P}}. \tag{11}$$

Positive selectivity means a larger targeted-minus-random effect under the corresponding task. Negative selectivity means a larger effect under the other task, not necessarily a negative own-task effect.

We obtain 95% intervals from 2,000 bootstrap resamples of Test-fact IDs. The three random draws are averaged within each fact before pairing, and resampling retains the paired target and random effects. Selectivity intervals also preserve pairing between the two task requirements. The pool intervention includes shared and nonshared members; it does not isolate the shared-pool geometry or a source–target edge.

## C Structural Results and Robustness

### C.1 Candidate Paths and Dimension Sensitivity

Table 11 reports all six candidate pairs, their two-task validation measurements, and the selection outcome. All six satisfy the eligibility criteria; the mean-$R^2$ rule in Appendix B.2 retains one target per source layer. The "Retained" and "Causal test" columns therefore record a further selection among eligible paths.

**PCA-truncation sensitivity.** The primary pooled-PCA truncation is $q = 64$. We repeat path evaluation at $q \in \{32, 64, 128, 256\}$, holding communication rank $d = 16$, the discovery–validation split, and random-basis and pairing-shuffle controls fixed. Across truncations, all evaluated path–task checks exceed their corresponding controls, and the selected paths remain L19→L20, L20→L24, and L24→L31 (Table 12). Absolute $R^2$ and cosine values change with the target space and should not be treated as a common performance scale. This scan tests eligibility and selection stability rather than selecting $q$ for the largest prediction score.

Table 11: All candidate source–target pairs in the primary Mistral setting. Eligibility requires the fixed two-task controls; retention additionally selects the highest mean validation $R^2$ target for each source layer.

| Path | Context $R^2$ | Memory $R^2$ | Mean cosine | Random basis | Pair shuffle | Retained | Causal test |
|---|---|---|---|---|---|---|---|
| L19→L20 | 0.531 | 0.615 | 0.847 | Pass | Pass | Yes | Yes |
| L19→L24 | 0.535 | 0.599 | 0.842 | Pass | Pass | No | No |
| L19→L31 | 0.364 | 0.566 | 0.768 | Pass | Pass | No | No |
| L20→L24 | 0.556 | 0.610 | 0.849 | Pass | Pass | Yes | Yes |
| L20→L31 | 0.379 | 0.581 | 0.776 | Pass | Pass | No | No |
| L24→L31 | 0.436 | 0.622 | 0.784 | Pass | Pass | Yes | Yes |

Table 12: Sensitivity of path identification to the pooled-PCA truncation level $q$, with communication rank fixed at $d = 16$. Mean values aggregate the six evaluated candidate paths.

| $q$ | Mean val. $R^2$ | Mean val. cosine | Eligible paths | Selected paths |
|---|---|---|---|---|
| 32 | 0.610 | 0.851 | 6/6 | L19→L20; L20→L24; L24→L31 |
| 64 | 0.533 | 0.811 | 6/6 | L19→L20; L20→L24; L24→L31 |
| 128 | 0.469 | 0.775 | 6/6 | L19→L20; L20→L24; L24→L31 |
| 256 | 0.418 | 0.744 | 6/6 | L19→L20; L20→L24; L24→L31 |

**Communication-rank selection.** Communication rank is varied separately over $\{1, 2, 4, 8, 16, 32\}$. Figure 5 reports the scan. Validation $R^2$ improves from 1 to 16 dimensions, and the retained paths exceed the random 95th-percentile baseline at 16. Rank 32 gives further gains on some paths with smaller returns and a weaker compactness constraint. The primary choice $d = 16$ retains a compact source representation that separates the selected paths from the controls. It is fixed before the pool and held-out causal analyses.

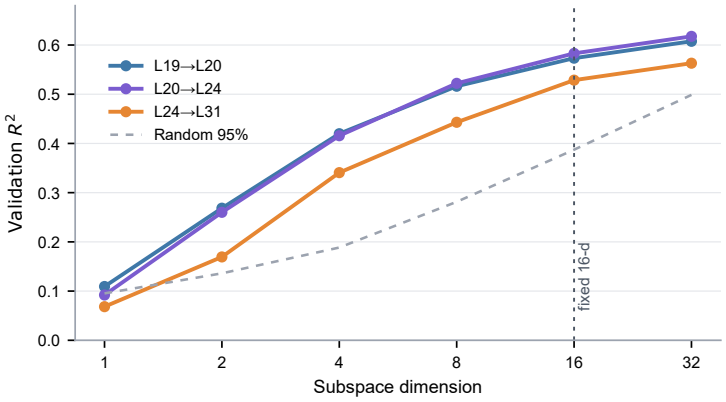



Figure 5: Validation $R^2$ across subspace dimensions for retained Mistral paths. The dashed line is the random 95% baseline; the vertical line marks the main 16-dimensional setting.

### C.2 Pool Membership and Rank Stability

The primary pool size is 512 neurons per task, following a scan over $K \in \{64, 128, 256, 512, 1024\}$. Across top-64 to top-1024, observed memory–context overlap on all three paths remains above the matched-random 95th-percentile bounds (Figure 6). The membership result is therefore not confined to the primary threshold.

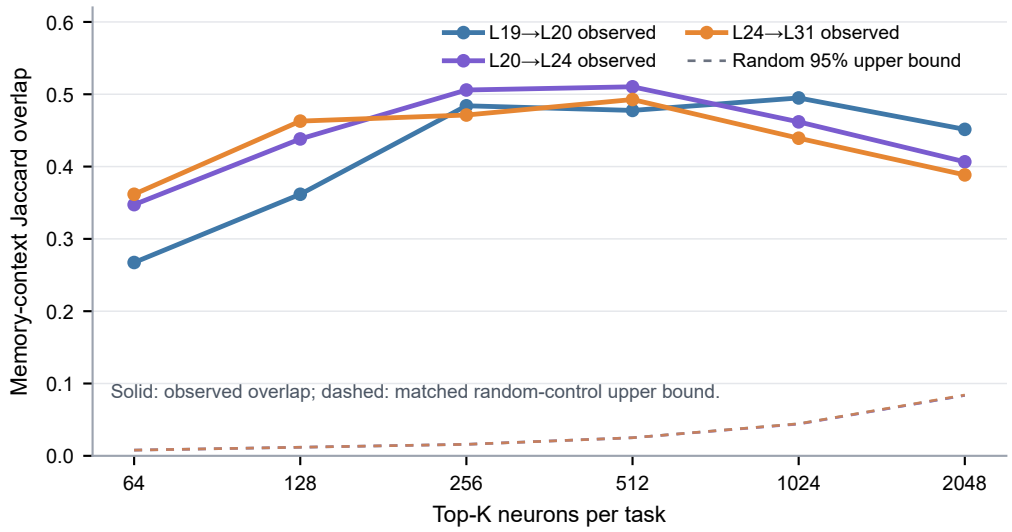



Figure 6: Top-$K$ membership overlap for retained Mistral paths. Solid lines show observed Jaccard overlap; dashed lines show matched random-control 95% upper bounds.

Figure 7 compares discovery and validation percentile ranks across all source-layer neurons. Near-diagonal density and correlations of 0.945–0.989 show that stability extends beyond selected members to the full loading order. Table 13 gives shared-pool size and the lower split correlation across tasks for each path.

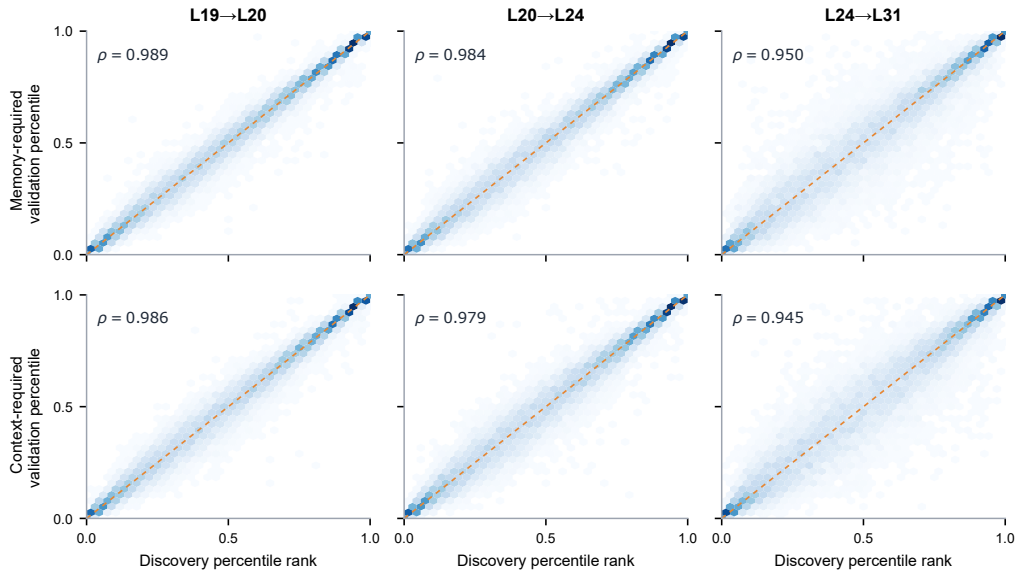



Figure 7: Discovery–validation loading-rank stability. Hexagons aggregate MLP neurons; the dashed diagonal denotes identical percentile ranks.

### C.3 Geometry on Additional Paths

Figures 8 and 9 extend the main-text L19→L20 comparison to L20→L24 and L24→L31. Both have large mean principal angles and lower other-task reconstruction. As in Figure 4, the common-

Table 13: Detailed membership stability in the primary Mistral setting. Shared size is the intersection of the two task-specific top-512 sets; Disc.–Val. Spearman is the lower loading-rank correlation across task requirements.

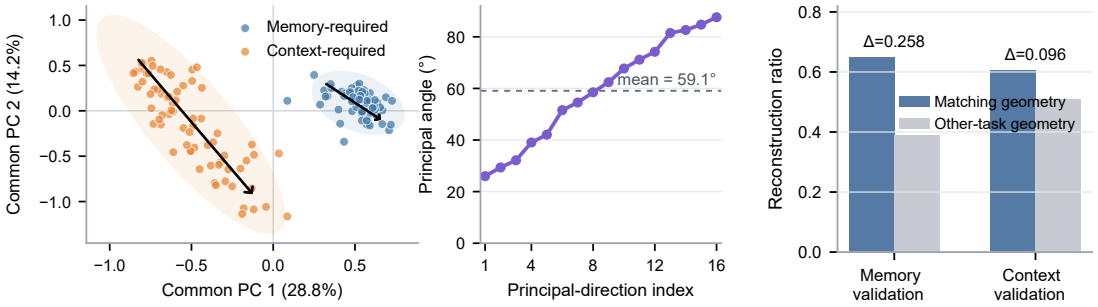
| Path | Shared size | Disc.–Val. Spearman |
|---|---|---|
| L19→L20 | 331 | $\geq 0.986$ |
| L20→L24 | 346 | $\geq 0.979$ |
| L24→L31 | 338 | $\geq 0.945$ |

coordinate projection is descriptive; the angle spectrum and held-out reconstruction quantify geometry on the shared neuron set.

Table 14 reports reconstruction drops alongside the secondary cross-task $R^2$ check. The latter evaluates task-specific source subspaces on the other task and complements the shared-pool geometry measurements.

Table 14: Additional task-dependent measures for retained Mistral paths. Cross-task $\Delta R^2$ is the validation decrease when a task-specific source subspace is evaluated on the other task. Reconstruction drops are matching-task minus other-task validation write-energy capture.

| Path | Cross-task $\Delta R^2$ | Memory recon. drop | Context recon. drop |
|---|---|---|---|
| L19→L20 | 0.160 | 0.333 | 0.090 |
| L20→L24 | 0.221 | 0.258 | 0.095 |
| L24→L31 | 0.282 | 0.331 | 0.198 |



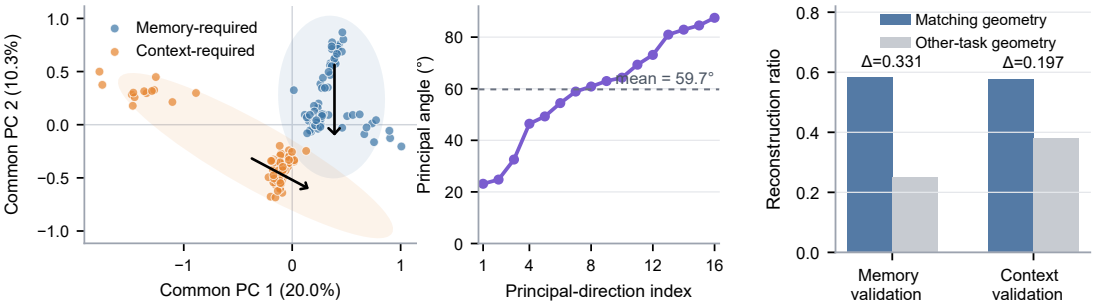

(a) Common-coordinate validation projection. (b) Principal-angle spectrum. (c) Held-out reconstruction.

Figure 8: Task-dependent geometry of the L20→L24 overlapping pool. Panels follow Figure 4.

# D Local Interventions and Cross-Model Results

## D.1 Complete Primary-Setting Effects

Table 15 gives every held-out Mistral task-by-pool effect. Positive values mean greater weakening of the task-aligned margin than under the matched random control. On L19→L20 and L20→L24, both the context- and memory-selected pools have larger matched-control effects under context-required prompts. On L24→L31, both pools instead have larger effects under memory-required prompts. Thus, the two pool definitions agree on the task-effect direction within each retained path, while that direction changes at the final path. The complete effects under both tasks are reported regardless of sign or magnitude.




(a) Common-coordinate validation projection. (b) Principal-angle spectrum. (c) Held-out reconstruction.

Figure 9: Task-dependent geometry of the L24→L31 overlapping pool. Panels follow Figure 4.

Table 15: Complete held-out Mistral task-pool effects on 109 Test facts. Each cell gives targeted-minus-random aligned-margin weakening above its 95% clustered-bootstrap interval. Context and memory columns denote the corresponding complete top-512 pools.

| | Context-required | | Memory-required | |
|---|---|---|---|---|
| Path | Context pool | Memory pool | Context pool | Memory pool |
| L19→L20 | 1.4561.320,1.597 | 0.7610.664,0.868 | 0.1770.072,0.276 | 0.4390.341,0.544 |
| L20→L24 | 1.1731.078,1.266 | 0.7150.624,0.804 | -0.068-0.126,-0.010 | -0.126-0.199,-0.053 |
| L24→L31 | -0.072-0.125,-0.018 | -0.200-0.252,-0.149 | 0.0960.039,0.153 | 0.0830.016,0.147 |

### D.2 Cross-Model Structural Results

The supporting settings differ in paths, prompt calibration, and Test size, so effect estimates are not pooled. Table 16 gives structural measurements for every retained PopQA-derived path. Each passed the random-source and pairing-shuffle criteria during model-specific validation. Discovery–validation rank correlations are reported as the lower of the two task values.

Table 16: Structural measures for all retained paths on PopQA-derived data. $\rho$ is the lower discovery–validation loading-rank Spearman across the two task requirements.

| Setting | Path | Val. $R^2$ | Val. cosine | Pool Jaccard ($K = 512$) | Disc.–Val. $\rho$ |
|---|---|---|---|---|---|
| Llama | L19→L20 | 0.586 | 0.866 | 0.403 | 0.953 |
| Llama | L20→L24 | 0.496 | 0.810 | 0.397 | 0.936 |
| Llama | L24→L31 | 0.459 | 0.783 | 0.465 | 0.853 |
| Qwen | L16→L17 | 0.728 | 0.901 | 0.414 | 0.988 |
| Qwen | L17→L21 | 0.613 | 0.868 | 0.354 | 0.986 |
| Qwen | L21→L27 | 0.387 | 0.694 | 0.382 | 0.951 |
| Mistral-adapted | L19→L20 | 0.588 | 0.852 | 0.409 | 0.969 |
| Mistral-adapted | L20→L24 | 0.581 | 0.837 | 0.347 | 0.953 |
| Mistral-adapted | L24→L31 | 0.499 | 0.775 | 0.321 | 0.877 |

### D.3 Cross-Model Effects and Task Selectivity

Tables 17 and 18 report all four task-by-pool effects and paired selectivities, respectively. No retained path was removed because of its causal result. On all three Llama paths and the first two Qwen and adapted Mistral paths, both pool definitions indicate larger matched-control effects under context-required prompts: context-pool selectivity is positive and memory-pool selectivity is negative. On the final Qwen path, neither pool has a resolved task preference, while the final adapted

Mistral path reverses toward larger memory-required effects for both pools. Thus, the two pool definitions agree on the task-effect direction within each resolved path, while the direction of later-path effects varies across models.

Table 17: Complete PopQA task-pool effects. Each cell gives targeted-minus-random aligned-margin weakening above its 95% sample-ID clustered-bootstrap interval. C and M denote context and memory pools. Settings are grouped without pooling their estimates.

| | Context-required | | Memory-required | |
|---|---|---|---|---|
| Path | C pool | M pool | C pool | M pool |
| **Llama** | | | | |
| L19→L20 | 1.4701.313,1.624 | 1.3091.159,1.457 | 0.3410.240,0.438 | 0.4850.357,0.608 |
| L20→L24 | 0.3580.256,0.461 | 0.4490.328,0.570 | 0.1770.081,0.276 | 0.2170.065,0.367 |
| L24→L31 | 0.4600.322,0.592 | 0.4270.274,0.556 | 0.1690.025,0.303 | 0.1800.042,0.322 |
| **Qwen** | | | | |
| L16→L17 | 1.0740.741,1.488 | 0.8290.578,1.102 | -0.227-0.341,-0.113 | -0.173-0.295,-0.044 |
| L17→L21 | 0.209-0.036,0.493 | 0.7550.514,1.047 | -0.191-0.358,-0.017 | -0.521-0.730,-0.319 |
| L21→L27 | -0.037-0.282,0.174 | 0.4260.177,0.643 | 0.069-0.238,0.351 | 0.285-0.027,0.621 |
| **Mistral-adapted** | | | | |
| L19→L20 | 0.8610.730,1.005 | 0.3820.277,0.499 | -0.084-0.203,0.039 | -0.225-0.365,-0.077 |
| L20→L24 | 0.6740.543,0.802 | 0.7210.592,0.861 | -0.030-0.157,0.101 | -0.110-0.235,0.019 |
| L24→L31 | -0.249-0.392,-0.102 | -0.205-0.334,-0.060 | 0.4470.274,0.612 | 0.4640.301,0.625 |

Table 18: Task selectivity on PopQA-derived data. Entries are own-task minus other-task effects with 95% sample-ID clustered-bootstrap intervals. Because the own task differs by pool type, positive context-pool selectivity and negative memory-pool selectivity both indicate a larger matched-control effect under context-required prompts.

| Path | Context pool | Memory pool |
|---|---|---|
| **Llama** | | |
| L19→L20 | $1.129\,[0.952, 1.299]$ | $-0.824\,[-1.016, -0.640]$ |
| L20→L24 | $0.181\,[0.036, 0.317]$ | $-0.232\,[-0.441, -0.034]$ |
| L24→L31 | $0.291\,[0.089, 0.482]$ | $-0.247\,[-0.440, -0.034]$ |
| **Qwen** | | |
| L16→L17 | $1.301\,[0.977, 1.666]$ | $-1.002\,[-1.258, -0.742]$ |
| L17→L21 | $0.400\,[0.115, 0.683]$ | $-1.276\,[-1.584, -0.987]$ |
| L21→L27 | $-0.106\,[-0.450, 0.227]$ | $-0.141\,[-0.478, 0.200]$ |
| **Mistral-adapted** | | |
| L19→L20 | $0.945\,[0.767, 1.136]$ | $-0.607\,[-0.807, -0.399]$ |
| L20→L24 | $0.703\,[0.559, 0.852]$ | $-0.831\,[-1.016, -0.662]$ |
| L24→L31 | $-0.695\,[-0.882, -0.518]$ | $0.669\,[0.510, 0.831]$ |