CAUSAL AND INTERPRETABLE STRUCTURES IN LLM COMPOSITIONAL TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are able to solve tasks whose answers depend on not only individual input tokens, but also on relations among them. How is such relational information represented and processed across transformer layers? We study activations from ensembles of prompts that require inferring relationships between three tokens corresponding to a cyclic concept (months, hours, weekdays, and musical notes) to correctly predict the next token. Across model families (Llama, Qwen, Gemma, and Mistral) and cyclic concepts, we find a consistent layerwise progression in how the joint dependence among the tokens is geometrically organized and causally used: intermediate layers use a joint representation based on the inferred relationship between two tokens, while later layers use a joint representation associated with all three tokens to correctly complete the task. We also find other relationships between tokens that are geometrically structured but remain causally inert in the next-token prediction. Crucially, when taken together, these geometric and causal investigations reveal the representation-level mechanism that progressively organizes and composes the relational information to form the answer. More surprisingly, restricting the models to such causally relevant joint representations improves next-token prediction accuracy.

1 INTRODUCTION

Many tasks require a large language model (LLM) to use relationships between multiple input tokens. Consider the prompt, "*Hi John, the conference is scheduled between January 13 and March 13. This is the same time as between May 13 and*". The correct next-token prediction depends crucially on three tokens: *January*, *March*, and *May*. Let A, B, and C denote these three tokens, respectively, and define $\gamma = B - A \pmod{12}$. The correct answer is then $D = C + \gamma \pmod{12}$. The individual representations of A, B, and C lie on the low-dimensional manifold associated with calendar months (Engels et al., 2025). We ask whether computationally relevant quantities arising jointly from these tokens, such as γ or D, also have geometric structure, and whether that structure is used in the next-token prediction.

We vary A, B, C independently over calendar months while keeping everything else fixed. Across this ensemble of prompts, the last-token activation varies with A, B, and C individually and with the particular combinations in which they co-occur. These sources of variation are superposed in the full activation ensemble. We use the functional analysis of variance (ANOVA) decomposition to separate the resulting activation ensemble into terms associated with individual variables, pairwise interactions, and their three-way interaction (Hoeffding, 1948). We treat the interaction terms as geometric objects and study their organization and causal role across the layers of an LLM.

We study this question across several cyclic concepts: months, hours, weekdays, and musical notes. The cyclic structure gives us known relational quantities. In particular, $\gamma = B - A$ is a pairwise relation, while the answer $D = C + \gamma$ depends jointly on all three inputs. Importantly, γ is never supplied explicitly and must instead be inferred from A and B. This allows us to ask how the joint dependence associated with γ and D is represented and used across network depth.

Our contributions are: **(a) Geometric structure.** We find that the pairwise interaction between A and B is organized by γ in the middle layers. Other pairwise interactions are also organized by their corresponding pairwise differences. The three-way interaction between A, B, and C becomes organized by D at a later depth. **(b) Causality.** We show that the representation associated with

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.

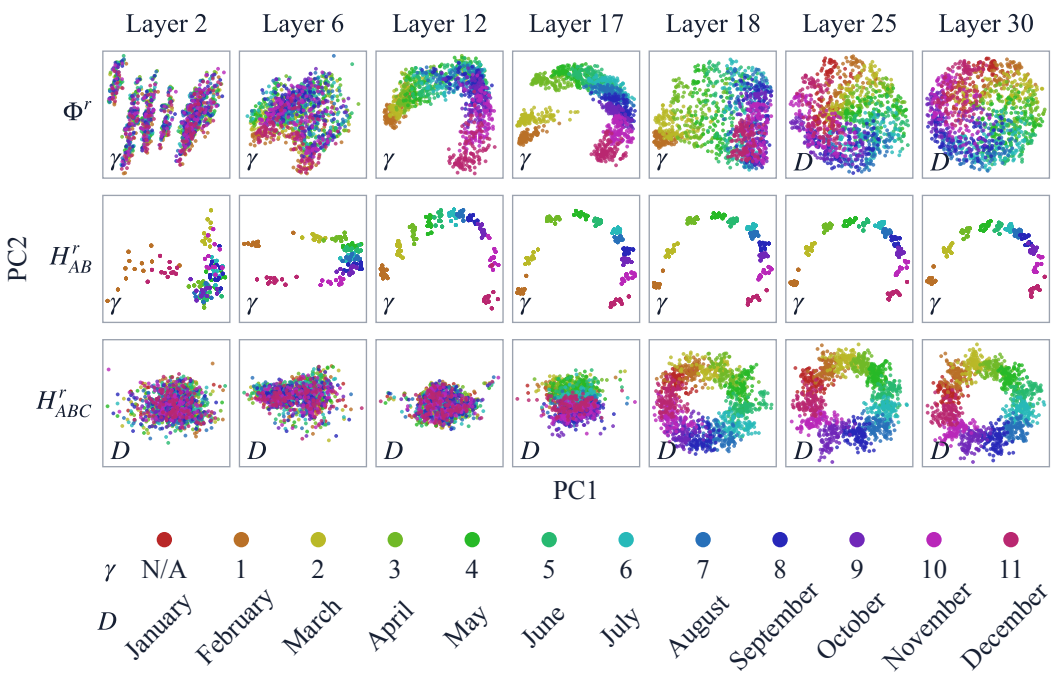

Figure 1: Two-dimensional visualization using the first two principal components for the activation ensemble Φ^r and its decomposed interaction ensembles H^r_{AB} and H^r_{ABC} for the prompt template in Section 2.1 using Llama-3.1-8B. Here, $\gamma = B - A \mod 12$ and the expected answer is $D = C + \gamma \mod 12$. Before fitting the principal components, we exclude prompts satisfying $A = B$, $B = C$, or $C = A$ (see Appendix B.3). The bottom left of each panel shows the coloring according to either γ or D. H^r_{AB} gets organized by γ in early layers. Until layer 17, H^r_{ABC} is unstructured but becomes organized by D at layer 18. See Section 3 for a quantitative analysis.

A, B is causally relevant in the middle layers, while the representation associated with A, B, C becomes causally relevant at later layers. **(c) Improved performance.** Retaining only the causally relevant three-way interaction improves the performance on the task. **(d) Relationship transfer.** We show that the relationship encoded by interaction terms transfers across cyclic domains, such as from hours to months.

1.1 Related work

Geometry of concepts. Prior work has identified low-dimensional geometric structure associated with individual concepts in latent activations (Engels et al., 2025; Modell et al., 2025; Park et al., 2025; Karkada et al., 2026; Hu et al., 2026; Prieto et al., 2026). Other work has studied how representational geometry evolves during computation and how interventions along that geometry affect model behavior (Gurnee et al., 2025; Wurgaft et al., 2026; Sarfati et al., 2026). We instead study the geometry and causality of joint dependence between multiple tokens.

Relational representations and arithmetic computation. Prior work has shown that latent activations contain structured representations of relations between inputs, including approximately linear transformations, as well as distributed representations that bind entities or variables to relational roles (Hernandez et al., 2024; Merullo et al., 2024; Feng & Steinhardt, 2024; Davies et al., 2023; Dai et al., 2026; Todd et al., 2026; Wang et al., 2024). Related work on arithmetic has identified structured numerical representations and computational mechanisms, including Fourier and periodic structure in transformers trained on modular arithmetic and in pretrained language models (Nanda et al., 2023; Furuta et al., 2024; Zhou et al., 2024; Kantamneni & Tegmark, 2025; Levy & Geva, 2025). A closely related work is that of Feucht et al. (2026), who study tasks involving an explicit offset γ and a concept C from a cyclic domain. They identify a universal base-10 addition mecha-

nism in Llama-3.1-8B by characterizing MLP neurons associated with the task's Fourier structure. In contrast, we work at the representation level. The relation γ in our setting is not supplied explicitly, but arises from the joint dependence of two independently varied inputs. We isolate the component of the residual-stream representation associated with this joint dependence and study its geometry and causal role across layers.

Causal interventions. The presence of structured or decodable information in an activation does not by itself establish that the model uses that information to perform the task. Causal interventions and activation patching have been used in several works to test the functional role of such representations (Vig et al., 2020; Geiger et al., 2021; 2024; Zhang & Nanda, 2024; Arora et al., 2024). Our experiments similarly intervene directly on joint representations of multiple input tokens.

2 SETUP

In the main text, all results are reported using base Llama-3.1-8B (Llama Team, 2024a), which contains 32 layers indexed by $\ell \in \{0, \ldots, 31\}$. In the appendices, we reproduce all our results on base models: Llama-3.2-3B (Llama Team, 2024b), Qwen-2.5-7B (Qwen Team, 2024), Qwen-3-8B (Qwen Team, 2025), Gemma-2-9B (Gemma Team, 2024), Gemma-3-12B (Gemma Team, 2025), Mistral-small-24B-Base-2501 (Mistral AI Team, 2025). In each model, we analyze residual-stream activations after every transformer block.

2.1 CONSTRUCTION OF PROMPTS

Several other works have used controlled prompt variations to reveal manifolds associated with individual concepts (Engels et al., 2025; Kantamneni & Tegmark, 2025). We propose a similar experimental design, but we aim to elicit a relationship between concepts. We consider several cyclic concepts—months, weekdays, hours, musical notes—and let $\mathcal{X}$ be the associated vocabulary. In the main text, we discuss the months domain, $\mathcal{X} := \{\text{January}, \ldots, \text{December}\}$. Consider the variables $(A, B, C) \in \mathcal{X}^3$ along with the prompt

```
Hi {name}, the {noun} is scheduled between {A} {dates}
and {B} {dates}. This is the same time as between {C} {dates} and
```

Here, `name` is chosen from the most common names of the past century (Social Security Administration, n.d.), `noun` varies over synonyms of `conference`, while `dates` $\in \{1, \ldots, 28\}$ (see Appendix A.1 for a list). We call each combination of (`name`, `noun`, `dates`) a *replicate* and label it by r. These replicates define prompt variations that leave the underlying task unchanged. We further require that all replicate variables tokenize to the same length so that positional offsets do not introduce a confound. We randomly subsample 10 such replicates. Unless stated otherwise, we average over the replicates when we report any quantity.

We design prompts associated with each cyclic concept such that the answer to the task is the next token. We measure the performance by computing the top-1 and top-3 accuracy, where top-k accuracy is the fraction of prompts for which the correct answer features in the k largest logits. The top-1 accuracy of Llama-3.1-8B is $62.2\% \pm 5.1\%$, and the top-3 accuracy is $85.4\% \pm 3.3\%$, where the variation is reported as the standard deviation across replicates r[1]. Because top-3 accuracy varies less with larger offsets of $\gamma = B - A \gtrsim 6$ than top-1 accuracy (Appendix A), we use top-3 accuracy as the primary performance metric in the main text. The analogous results using top-1 accuracy are reported in the appendices.

2.2 ENSEMBLE DECOMPOSITION

Let $\boldsymbol{\phi}_\ell^r(A, B, C) \in \mathbb{R}^{d_{\text{model}}}$ denote the last-token activation at layer ℓ for replicate r. We define the corresponding activation ensemble as

$$\Phi_\ell^r := (\boldsymbol{\phi}_\ell^r(A, B, C))_{(A,B,C)\in\mathcal{X}^3} \,. \tag{1}$$

[1]The accuracies are computed over all prompts with distinct variables (A, B, C). See Appendix B.3.

Since we report all quantities as a function of layer, we drop the label ℓ for brevity. We decompose each activation as

$$\phi^r(A,B,C) = \underbrace{\boldsymbol{\mu}^r}_{\text{mean}} + \underbrace{\boldsymbol{h}^r_A + \boldsymbol{h}^r_B + \boldsymbol{h}^r_C}_{\text{additive terms}} + \underbrace{\boldsymbol{h}^r_{AB} + \boldsymbol{h}^r_{BC} + \boldsymbol{h}^r_{CA} + \boldsymbol{h}^r_{ABC}}_{\text{interaction terms}}, \tag{2}$$

where every term is defined as an average with respect to the ensemble Φ^r:

$$\begin{aligned} \boldsymbol{\mu}^r &:= \mathbb{E}_{ABC}(\boldsymbol{\phi}^r), \\ \boldsymbol{h}^r_A &:= \mathbb{E}_{BC}(\boldsymbol{\phi}^r) - \boldsymbol{\mu}^r, \\ \boldsymbol{h}^r_{AB} &:= \mathbb{E}_{C}(\boldsymbol{\phi}^r) - \boldsymbol{\mu}^r - \boldsymbol{h}^r_A - \boldsymbol{h}^r_B, \\ \boldsymbol{h}^r_{ABC} &:= \boldsymbol{\phi}^r - \boldsymbol{\mu}^r - \boldsymbol{h}^r_A - \boldsymbol{h}^r_B - \boldsymbol{h}^r_C - \boldsymbol{h}^r_{AB} - \boldsymbol{h}^r_{BC} - \boldsymbol{h}^r_{CA}. \end{aligned} \tag{3}$$

Analogously to Equation 1, we define the ensemble associated with decomposed terms as

$$H^r_A = (\boldsymbol{h}^r_A)_{A\in\mathcal{X}}\,, \qquad H^r_{AB} = (\boldsymbol{h}^r_{AB})_{(A,B)\in\mathcal{X}^2}\,, \quad \text{and} \quad H^r_{ABC} = (\boldsymbol{h}^r_{ABC})_{(A,B,C)\in\mathcal{X}^3}\,.$$

The remaining terms are defined by cyclic permutation of the expressions above. Intuitively, $\boldsymbol{\mu}^r$ is the center of the activation ensemble, H^r_A captures the dependence on A alone, H^r_{AB} captures the joint dependence on A and B that remains after averaging over C and removing their individual effects, and H^r_{ABC} contains the residual three-way dependence after accounting for all lower-order terms. We refer to H^r_{AB}, H^r_{BC}, H^r_{CA}, and H^r_{ABC} as the interaction ensembles.

We emphasize that Equation 2 is an exact identity rather than an approximation. It is known as the functional ANOVA decomposition (Hoeffding, 1948; Owen, 2013). The decomposed components are pairwise orthogonal under the ensemble-averaged inner product (Appendix B). We do not apply the inferential machinery of ANOVA, since the associated questions of statistical significance are orthogonal to the goal of this paper. Figure 1 shows the first two principal components of Φ^r, H^r_{AB}, and H^r_{ABC}. The remainder of the paper analyzes the interaction ensembles as geometric objects to study their structure and causal role across network depth.

3 Geometric structure of interaction terms

The ANOVA decomposition described in Equation 2 breaks the ensemble into its constituent parts. Figure 1 visually suggests that H^r_{AB} can be parametrized by $\gamma = B - A$ and similarly H^r_{ABC} can be parametrized by $D = C + \gamma$. In this section, we first rigorously quantify the dependence of interaction terms on γ or D by exploiting the cyclic nature of A, B, C using the discrete Fourier transform in Section 3.1. Second, we analyze the norm of vectors in each interaction ensemble in Section 3.2 to check for ensembles that are highly organized but have negligible norm overall.

3.1 Fourier Transform

We perform a discrete Fourier transform (DFT) of every interaction term for every replicate r. Consider H^r_{AB} as an example. Its DFT is given by

$$\widehat{\boldsymbol{h}}^r_{k_A k_B} = \sum_{A,B} \boldsymbol{h}^r_{AB} e^{-2\pi i(k_A A + k_B B)/12}, \tag{4}$$

where we identify the months with $A, B \in \{0, \ldots, 11\}$ and k_A, k_B are the associated Fourier modes. Let $\gamma = B - A$. If $\boldsymbol{h}^r_{AB}$ depends only on γ, that is, $\boldsymbol{h}^r_{AB} \equiv \boldsymbol{f}(\gamma)$, then its DFT

$$\widehat{\boldsymbol{h}}^r_{k_A k_B} = \sum_{\gamma} \boldsymbol{f}(\gamma) e^{-2\pi i k_B \gamma/12} \sum_{A} e^{-2\pi i(k_A + k_B)A/12}$$

vanishes unless $k_A + k_B = 0 \mod 12$. Thus concentration on modes satisfying $k_A = -k_B$[2] quantifies the extent to which H^r_{AB} is organized according to γ. Therefore we define *mode fraction* as

$$m(H^r_{AB}) := \frac{\sum_{k_A=-k_B} \left\| \widehat{\boldsymbol{h}}^r_{k_A k_B} \right\|_2^2}{\sum_{k_A,k_B} \left\| \widehat{\boldsymbol{h}}^r_{k_A k_B} \right\|_2^2}. \tag{5}$$

[2]Strictly, the condition is $k_A + k_B = 0 \mod 12$; all frequency equalities are understood modulo 12.

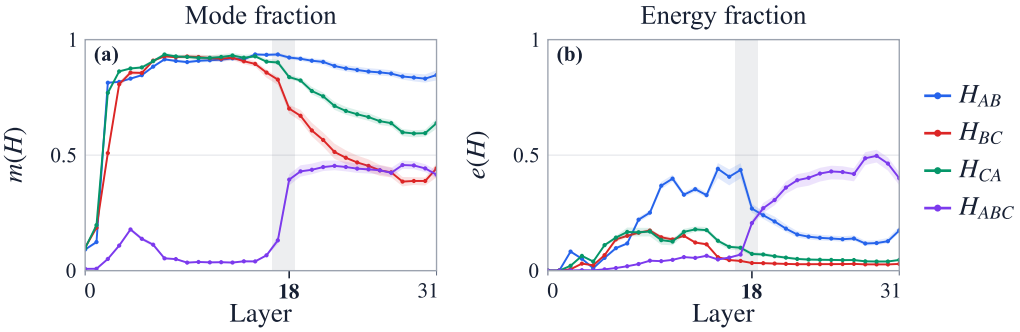

Figure 2: Mode fraction and energy fraction of the interaction ensembles across layers of Llama-3.1-8B. We zero out vectors for which A, B, and C are not pairwise distinct before computing both fractions (Appendix B.3). **(a)** Mode fraction measures the organization of H^r_{AB}, H^r_{BC}, and H^r_{CA} by their corresponding pairwise differences, and of H^r_{ABC} by $D = C + B - A$. Under a random null model, the expected mode fraction for second-order interactions is 0.09 and that for third-order interaction is 0.008. **(b)** Energy fraction measures the fraction of mean-subtracted activation energy attributable to each interaction ensemble. Throughout this work, colored bands show the central 80% interval across 10 replicates, and vertical gray bands highlight the layers discussed in the text.

The mode fraction $m(H^r_{AB}) \in [0, 1]$ measures the degree of organization of H^r_{AB} by γ, with $m = 1$ when the ensemble depends on γ alone. We define an analogous mode fraction for H^r_{BC}, H^r_{CA}, and H^r_{ABC} to probe for their dependence on $\gamma' = C - B$, $\gamma'' = A - C$, and $D = C + B - A$ respectively. For the latter, the modes under investigation are $(k_A, k_B, k_C) = (-k, k, k)$.

Dependence on γ cannot occur in terms lower-order than H^r_{AB}, and dependence on D cannot occur in terms lower-order than H^r_{ABC}. Thus, if the residual stream contains structure organized by γ or D, that structure must reside in H^r_{AB} or H^r_{ABC}, respectively. Consistently, we find large mode fractions for both interaction ensembles but at different depths. Importantly, we also find that all second-order interaction terms are strongly organized by their corresponding pairwise differences, and that H^r_{ABC} undergoes a sharp increase in organization by D around layers 17–18 (Figure 2(a)). At the same depth, the mode fractions of H^r_{BC} and H^r_{CA} begin to decrease, while $m(H^r_{AB})$ remains large throughout the depth of the transformer. Results across models and domains appear in Appendix C.

Control experiments detailed in Appendix E suggest that the pairwise-difference organization of the second-order interaction ensembles can arise from the co-occurrence of the corresponding tokens. This is in agreement with Karkada et al. (2026). In contrast, H^r_{ABC} is not organized by D when the task structure is removed from the prompt, even when all three tokens A, B, C are present.

3.2 Energy Fraction

An ensemble can have a high mode fraction even if all of its vectors are close to zero. Directly comparing the norms of interaction terms across layers is misleading because $\boldsymbol{\phi}^r$ grows substantially with depth, with most of this growth coming from the mean $\boldsymbol{\mu}^r$. Therefore we need a measure of relative magnitude for each ensemble across layers. For $S \in \{AB, BC, CA, ABC\}$, we define energy fraction $e(H^r_S)$ as[3]

$$e(H^r_S) = \frac{\mathbb{E}_S\left(\|\boldsymbol{h}^r_S\|_2^2 \right)}{\mathbb{E}_{ABC}\left(\|\boldsymbol{\phi}^r(A, B, C) - \boldsymbol{\mu}^r\|_2^2 \right)} \tag{6}$$

Figure 2(b) shows the energy fraction of each interaction ensemble. The energy fraction of H^r_{AB} peaks at layer 15 and starts decreasing around layers 17–18. Around the same depth, the energy fraction of H^r_{ABC} increases dramatically. This coincides with the sharp increase in organization of H^r_{ABC} by D, as found in the previous Section 3.1. Together, these observations hint at a transition from a pairwise representation of the input relation $B - A$ to a three-way representation aligned with the answer $C + B - A$, which we test causally in the following section.

[3]See Appendix D for a relation between ANOVA decomposition and our definition of energy fraction.

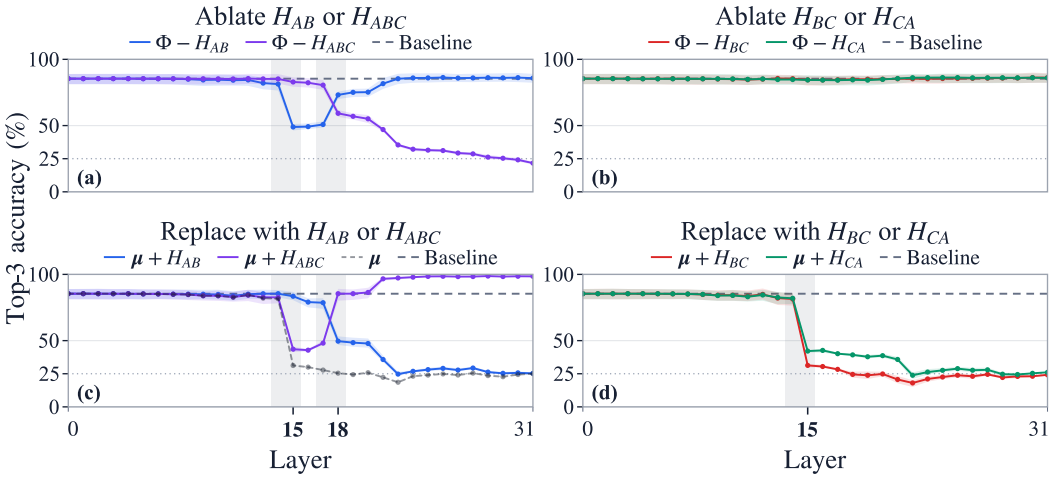

Figure 3: Top row shows the result of ensemble ablation experiments, while the bottom row shows corresponding ensemble replacement experiments. The horizontal dashed line shows the baseline performance. Dotted line at 25% is the chance of getting the correct answer under a uniform prior. **(a)** Ablating H^r_{AB} produces the largest decrease in top-3 accuracy around layers 15–17, whereas ablating H^r_{ABC} sharply decreases accuracy after layer 17. **(b)** Ablating H^r_{BC} or H^r_{CA} has little effect on performance across model depth. **(c)** Replacement experiments show a complementary transition where retaining H^r_{AB} preserves performance in the middle layers, while retaining H^r_{ABC} preserves performance in later layers. **(d)** Retaining H^r_{BC} or H^r_{CA} sharply decreases performance from layer 15 onward.

4 Causal Use of Interaction Terms

The previous section establishes that the interaction ensembles are highly organized by pairwise differences or by D, depending on the interaction ensemble and the layer. To test whether these ensembles are used in downstream computation, we intervene directly on the activation of the last token (Zhang & Nanda, 2024) using its ANOVA decomposition. In the first set of experiments in Section 4.1, we intervene on the last token at a particular layer to remove an interaction term. In Section 4.2, we instead remove all terms except for the mean and an interaction term. After each intervention, we let the model run without any further modification and score the performance.

4.1 Ensemble Ablations

We define *ensemble ablation* of an interaction ensemble H^r_S by subtracting the corresponding interaction vector $\boldsymbol{h}^r_S$ from the activation of the last token:

$$\widetilde{\boldsymbol{\phi}}^r(A,B,C) = \boldsymbol{\phi}^r(A,B,C) - \boldsymbol{h}^r_S, \tag{7}$$

for all $A, B, C \in \mathcal{X}$ and $S \in \{AB, BC, CA, ABC\}$.

Figure 3 shows the results of ensemble ablation experiments for every interaction term. We see that ablating H^r_{BC} or H^r_{CA} barely changes the top-3 accuracy relative to the baseline (Figure 3(b)). In contrast, ablating H^r_{AB} or H^r_{ABC} causes a drastic change in the model's performance, but at layers 15 and 18 respectively (Figure 3(a)).

4.2 Ensemble Replacement

In a complementary set of experiments, we perform *ensemble replacement* by retaining a single interaction ensemble together with the mean. Concretely, we replace the activation of the last token with

$$\widetilde{\boldsymbol{\phi}}^r(A,B,C) = \boldsymbol{\mu}^r + \boldsymbol{h}^r_S, \tag{8}$$

for all $A, B, C \in \mathcal{X}$ and $S \in \{AB, BC, CA, ABC\}$.

Figure 3(d) shows that retaining H^r_{BC} or H^r_{CA} alone induces a sharp decrease in top-3 accuracy around layers 14–15. However, Figure 3(c) shows that retaining H^r_{AB} preserves performance relative to the unmodified baseline until around layer 17, after which performance decreases sharply. Because we intervene only at the last-token position and at a particular layer, the model may recompute task-relevant information downstream of the intervention. We therefore also include a $\boldsymbol{\mu}^r$-only baseline to measure this recovery. Before layer 17, retaining H^r_{ABC} gives performance similar to the $\boldsymbol{\mu}^r$-only baseline. After layer 17, however, retaining H^r_{ABC} substantially improves performance over the $\boldsymbol{\mu}^r$-only baseline and even exceeds the unmodified baseline.

Together, ensemble ablation and replacement experiments show a transition in causal relevance from H^r_{AB} to H^r_{ABC} around layers 17–18. Between layers 15–17, ablating H^r_{AB} causes a sharp decrease in performance while retaining H^r_{AB} preserves it. The same aforementioned properties hold for H^r_{ABC} starting around layers 17–18. This transition occurs at the same depth where Section 3 finds that H^r_{ABC} becomes organized according to the answer D, the energy fraction of H^r_{AB} starts decreasing, and the energy fraction of H^r_{ABC} rises sharply. Although H^r_{BC} and H^r_{CA} are also strongly organized by their corresponding pairwise differences (Section 3), ablating them has little effect on the performance, and retaining them does not reproduce the layerwise replacement behavior of H^r_{AB}. Additional ablation and replacement experiments for different models and domains are in Appendices G and H, respectively. Appendix F further shows that ablating H^r_{AB} at layer 15 prevents the later emergence of H^r_{ABC}.

5 Steering

So far, we have shown that the ensembles H^r_{AB} and H^r_{ABC} are causally relevant, but this does not tell us if the model uses their organization according to γ or D. If the model indeed uses this organization, then changing $\gamma \to \gamma + \delta$ should shift the predicted answer by δ, that is, $D \to D + \delta$. In this section, we test this by steering along γ and D. There are multiple ways to steer using H^r_{AB}: shifting $B \to B + \delta$ such that $\boldsymbol{h}^r_{AB} \to \boldsymbol{h}^r_{A,B+\delta}$, shifting $A \to A - \delta$ such that $\boldsymbol{h}^r_{AB} \to \boldsymbol{h}^r_{A-\delta,B}$, or using an averaged vector that depends only on γ. Results using a combination of the first two strategies are reported in Appendix I. Here, we describe the last strategy. We define an averaged vector $\bar{\boldsymbol{h}}^r_\gamma$ by[4]

$$\bar{\boldsymbol{h}}^r_\gamma := \mathop{\mathbb{E}}_{B-A=\gamma} \boldsymbol{h}^r_{AB}. \tag{9}$$

To steer, we replace the activation $\boldsymbol{\phi}^r$ with $\widetilde{\boldsymbol{\phi}}^r$ according to

$$\widetilde{\boldsymbol{\phi}}^r(A,B,C) = \boldsymbol{\phi}^r(A,B,C) + \alpha\left(\bar{\boldsymbol{h}}^r_{\gamma+\delta} - \bar{\boldsymbol{h}}^r_\gamma\right). \tag{10}$$

Similarly, we define the steering vector $\bar{\boldsymbol{h}}_D$ for the ensemble H^r_{ABC} as

$$\bar{\boldsymbol{h}}^r_D := \mathop{\mathbb{E}}_{C+B-A=D} \boldsymbol{h}^r_{ABC} \tag{11}$$

and steer using Equation 10 after replacing $\bar{\boldsymbol{h}}^r_\gamma$ with $\bar{\boldsymbol{h}}^r_D$. The steering method is equivalent to the difference-of-means steering, analogous to that of Rimsky et al. (2024) and Subramani et al. (2022) (see Appendix B.2).

Let $D^* = D + \delta$ be the target answer. To measure the effect of steering, we compute the top-3 logit margin m_3, defined as the difference between the target logit z_{D^*} and the third-largest non-target logit z_3:

$$m_3 = z_{D^*} - z_3. \tag{12}$$

Thus $m_3 \geq 0$ iff the target logit is in top-3. Finally, we take the difference between steered and baseline m_3 to get

$$\Delta m_3 = m_3^{\text{steered}} - m_3^{\text{baseline}}. \tag{13}$$

In Figures 4(a–b), we report $\operatorname{median} \Delta m_3$ for steering using $\bar{\boldsymbol{h}}^r_\gamma$ and $\bar{\boldsymbol{h}}^r_D$ at $\alpha = 1$, and Figure 4(c) shows the top-3 accuracy after steering for $\delta = 4$. Steering using $\bar{\boldsymbol{h}}^r_\gamma$ becomes effective around layers 14–15, but around layers 17–18, its effect decreases and steering using $\bar{\boldsymbol{h}}^r_D$ becomes effective.

[4]The chosen steering vector $\bar{\boldsymbol{h}}^r_\gamma$ is closely related to the definition of the mode fraction in Equation 5. We show this correspondence in Appendix B.1.

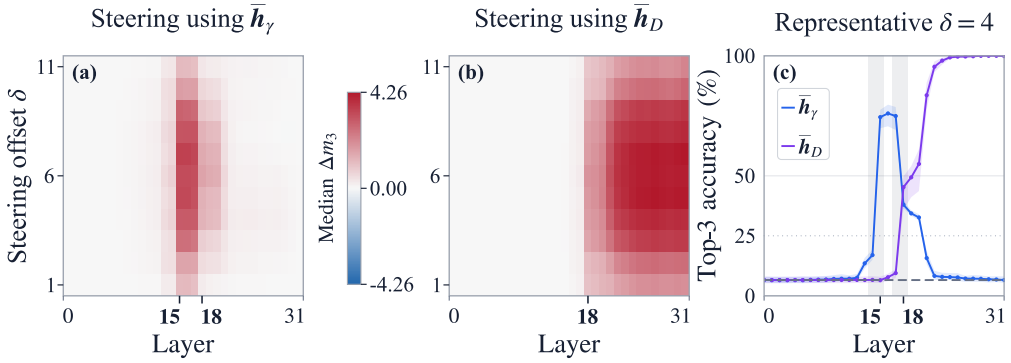

Figure 4: Steering using different vectors as a function of layer at $\alpha = 1$. The median is evaluated among all the prompts for each layer and δ. **(a)** We see that steering using $\bar{h}_\gamma$ is most effective between layers 15–17. **(b)** In contrast, steering using $\bar{h}_D$ becomes effective from layer 18 onward. **(c)** Top-3 accuracy for $\delta = 4$ after steering. The dashed line shows the unmodified model's top-3 accuracy with respect to the target answer $D + \delta$.

6 Cross-domain transfer

Our preceding experiments show that H^r_{AB} is causally relevant in the middle layers while H^r_{ABC} is relevant in the later layers. We now want to understand whether one interaction vector extracted from one context can substitute for the corresponding vector in another context performing the same underlying task. Prior work has shown that relational and task-level representations can be reused across contexts (Wang et al., 2024; Todd et al., 2024). We focus on the interaction ensemble H^r_{AB} here; the performance of transplanting H^r_{ABC} varies across model and domain combinations. Appendix J shows that H^r_{ABC} can be transplanted across domains when their subspaces are aligned. We will use *domain-1* to refer to the original set of prompts that we have discussed so far and we use *domain-2* for the following set of prompts:

```
Hi {name}, the {noun} is scheduled between {day} {A} o'clock and
{day} {B} o'clock. This is the same time as between {day}
{C} o'clock and {day}
```

where $A, B, C \in \mathcal{X}^{(2)} = \{1, \ldots, 12\}$ and (`name`, `noun`, `day`) are replicate variables (Appendix A.1). We then decompose the domain-2 ensemble according to Section 2.2. We transplant an interaction ensemble by replacing each domain-1 interaction vector with the corresponding domain-2 interaction vector indexed by the same values of the underlying cyclic variables. We consider

$$\widetilde{\boldsymbol{\phi}}^{r_1,(1)}(A,B,C) = \boldsymbol{\phi}^{r_1,(1)}(A,B,C) - \boldsymbol{h}^{r_1,(1)}_{AB} + \boldsymbol{h}^{r_2,(2)}_{AB}, \tag{14}$$

where superscripts $(1), (2)$ refer to the domain and r_1, r_2 refer to the corresponding replicate. Analogous to the ensemble replacement experiments in Section 4.2, we also consider

$$\widetilde{\boldsymbol{\phi}}^{r_1,(1)}(A,B,C) = \boldsymbol{\mu}^{r_1,(1)} + \boldsymbol{h}^{r_2,(2)}_{AB}. \tag{15}$$

In both cases, the model runs unmodified after the intervention and we score relative to domain-1. Figure 5(a) shows that replacing an ablated domain-1 interaction ensemble with the corresponding domain-2 ensemble restores the performance lost under ablation. Figure 5(b) shows that the domain-2 interaction ensemble also reproduces the layerwise trend observed in the replacement experiment in Section 4.2.

We next ask whether the organization according to γ also transfers across domains. Using the steering vector defined in Equation 9, we steer using the difference between the domain-1 vector for γ and the domain-2 vector for $\gamma + \delta$:

$$\widetilde{\boldsymbol{\phi}}^{r_1,(1)}(A,B,C) = \boldsymbol{\phi}^{r_1,(1)}(A,B,C) + \alpha \left(\bar{\boldsymbol{h}}^{r_2,(2)}_{\gamma+\delta} - \bar{\boldsymbol{h}}^{r_1,(1)}_{\gamma} \right). \tag{16}$$

If the organization by γ transfers between two domains, then the answer should shift accordingly. Figure 5(c) shows that steering vectors constructed across the two domains produce a steering pattern similar to that in Section 5.

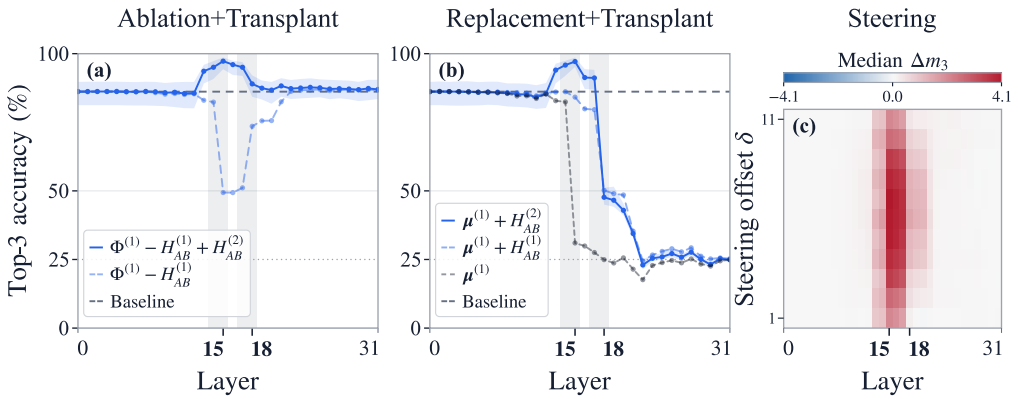

Figure 5: Shaded bands show the central 80% interval across 10 randomly sampled replicate pairs (r_1, r_2). **(a)** Transplanting the interaction ensemble $H_{AB}^{(2)}$ from domain-2 to domain-1 preserves performance. **(b)** Transplanting and retaining only the domain-1 mean $\boldsymbol{\mu}^{(1)}$ and domain-2 interaction ensemble $H_{AB}^{(2)}$ also preserves performance compared to the dashed blue within-domain replacement experiments. **(c)** Cross-domain steering using $\bar{\boldsymbol{h}}_{\gamma+\delta}^{r_2,(2)} - \bar{\boldsymbol{h}}_{\gamma}^{r_1,(1)}$ at $\alpha = 1$.

7 Discussion

We started by asking whether it is possible to separate variation due to individual concepts from variation arising through their joint dependence in an activation ensemble. We showed that using the functional ANOVA decomposition is a simple yet powerful technique to accomplish this task. The resulting interaction ensembles exhibit clear geometric structure, but not all geometrically organized ensembles are causally necessary for the task. For the ensembles that are causally relevant, their organization can be exploited to predictably steer the model.

We also found two results that we did not anticipate. First, retaining only the mean and the third-order interaction can outperform the unmodified model. Our interventions show that H_{ABC}^{r} is used by the LLM to output the correct answer, while other interaction ensembles either are unused or perform a different function. Therefore, it may happen that those other interaction ensembles interfere destructively in the computation in the last few layers. Second, transplanting H_{AB} between domains that perform the same underlying task recovers much of the performance lost under ablation. This suggests that the intermediate representation of the inferred relation is sufficiently compatible across domains to support downstream computation. Neither observation would have been accessible in the full activation ensemble without first isolating the joint dependence from the remaining components.

Our results also suggest a possible connection to the neuron-level analysis of Feucht et al. (2026). They identify a small set of MLP neurons at layer 18 that are associated with Fourier components of the arithmetic representation in Llama-3.1-8B. In our analysis, the same depth is where the three-way interaction H_{ABC}^{r} sharply increases in energy fraction, becomes organized by D, and becomes causally relevant. While we do not analyze individual neurons here, the interaction decomposition may provide a way to identify the layers in which such structure emerges and may also provide a model-agnostic starting point for neuron-level analyses.

Limitations

Scope. We have studied a controlled setting with three variables A, B, C that are drawn from the *same cyclic* concept whose composition has a *definitive* arithmetic answer. A natural extension is to allow A, B, C to vary over different, potentially non-cyclic concepts whose composition may not have a well-defined answer.

Exponential scaling. As we increase the number of variables p, the number of prompts required to furnish the ANOVA decomposition scales exponentially $\sim |\mathcal{X}|^p$.

AI USE STATEMENT

We have not used generative AI for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, supporting qualitative and thematic data analysis, interpreting results. The following does not apply to this work: generating synthetic data sets, assisting with translation, cleaning and reformatting datasets.

We have used generative AI (GPT-5.6 and GPT-6) to code, proofread the paper to improve readability, correct typos and grammatical errors. The AI-generated code was manually verified by the authors. The authors wrote the manuscript and reviewed every AI suggestion before incorporating/rejecting it.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Aryaman Arora, Dan Jurafsky, and Christopher Potts. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14638–14663, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.785. URL https://aclanthology.org/2024.acl-long.785/.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Cell-based representation of relational binding in language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47464–47524, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2194. URL https://aclanthology.org/2026.acl-long.2194/.

Xander Davies, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. Discovering variable binding circuitry with desiderata, 2023. URL https://arxiv.org/abs/2307.03637.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zb3b6oKO77.

Sheridan Feucht, Tal Haklay, Usha Bhalla, Daniel Wurgaft, Can Rager, Raphaël Sarfati, Jack Merullo, Thomas McGrath, Owen Lewis, Ekdeep Singh Lubana, Thomas Fel, and Atticus Geiger. Arithmetic in the wild: Llama uses base-10 addition to reason about cyclic concepts, 2026. URL https://arxiv.org/abs/2605.01148.

Hiroki Furuta, Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. Towards empirical interpretation of internal circuits and properties in grokked transformers on modular polynomials, 2024. URL https://arxiv.org/abs/2402.16726.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations, 2024. URL https://arxiv.org/abs/2303.02536.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Gemma Team. Gemma 3 technical report, 2025. URL https://goo.gle/Gemma3Report.

Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/linebreaks/index.html.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=w7LU2s14kE.

Wassily Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325, 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

Zhimin Hu, Lanhao Niu, and Sashank Varma. Language models represent and transform concepts with shared geometry. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=wg6Q5fPL1m.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition, 2025. URL https://arxiv.org/abs/2502.00873.

Dhruva Karkada, Daniel James Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetries in language statistics shape the geometry of model representations. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=XQeMPEkfdd.

Amit Arnold Levy and Mor Geva. Language models encode numbers using digit representations in base 10. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 385–395, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-190-2. doi: 10.18653/v1/2025.naacl-short.33. URL https://aclanthology.org/2025.naacl-short.33/.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024a. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Llama Team. Llama-3.2-3B. Hugging Face model card, 2024b. URL https://huggingface.co/meta-llama/Llama-3.2-3B.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple Word2Vec-style vector arithmetic. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5030–5047, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.281. URL https://aclanthology.org/2024.naacl-long.281/.

Mistral AI Team. Mistral-Small-24B-Base-2501. https://huggingface.co/mistralai/Mistral-Small-24B-Base-2501, 2025. Model card.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models, 2025. URL https://arxiv.org/abs/2505.18235.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Art B. Owen. *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

Lucas Prieto, Edward Stevinson, Melih Barsbey, Tolga Birdal, and Pedro A. M. Mediano. From data statistics to feature geometry: How correlations shape superposition. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7akSRQS5Xh.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Raphaël Sarfati, Eric Bigelow, Daniel Wurgaft, Siddharth Boppana, Jack Merullo, Atticus Geiger, Owen Lewis, Tom McGrath, and Ekdeep Singh Lubana. The shape of beliefs: Geometry, dynamics, and interventions along representation manifolds of language models' posteriors, 2026. URL https://arxiv.org/abs/2602.02315.

Social Security Administration. Top names over the last 100 years, n.d. URL https://www.ssa.gov/oact/babynames/decades/century.html. Accessed in 2026.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

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

Eric Todd, Jannik Brinkmann, Rohit Gandikota, and David Bau. In-context algebra. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 80695–80729, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/82aec8518602748540a42b783468c94d-Paper-Conference.pdf.

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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Zijian Wang, Britney Whyte, and Chang Xu. Locating and extracting relational concepts in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4818–4832, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.287. URL https://aclanthology.org/2024.findings-acl.287/.

Daniel Wurgaft, Can Rager, Matthew Kowal, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphaël Sarfati, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=pf3HKAGT2M.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. Pre-trained large language models use fourier features to compute addition. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=i4MutM2TZb.

A Models, Prompts, and Capabilities

In this work, we evaluate the robustness of our results across seven base models ranging from 3B to 24B parameters: Llama-3.2-3B, Llama-3.1-8B, Qwen-2.5-7B, Qwen-3-8B, Gemma-2-9B, Gemma-3-12B, Mistral-small-24B-Base-2501. Table 1 lists all the prompts for each model, while Table 4 shows their corresponding performances using top-1 and top-3 accuracy as metrics. Here, we report performance using all prompts, including cases with coincident variables, that is, when $A = B$, $B = C$, or $C = A$.

We use two conditions to fix the prompt template: **(a)** The model should be capable of answering the problem via the next token. **(b)** All variations over A, B, C must tokenize to the same width. Therefore we exclude evaluation of the hours domain for the Qwen, Gemma, and Mistral families since all of them use a single digit tokenizer.

A.1 List of variables

Tables 2 and 3 show the values of variables used for replicate variables and A, B, C respectively.

A.2 Capabilities

Figure 6 shows the top-1 and top-3 accuracy of each model indexed by γ over prompt templates in Table 1. Table 4 lists the aggregate top-1 and top-3 accuracy over replicates.

Table 1: Prompt templates used for each model and cyclic domain. A, B, C and the replicate variables vary as described in Tables 2 and 3. N/A indicates domains excluded because the cyclic values do not tokenize to a common width.

Model	Domain	Prompt template
Llama-3.2-3B	months	`Hi {name}, the {noun} is scheduled from {A} {dates} to {B} {dates}. This is the same as time from {C} {dates} to`
	hours	`Hi {name}, the {noun} is scheduled from {day} {A} o'clock to {day} {B} o'clock. This is the same time as from {day} {C} o'clock to {day}`
	weekdays	`Hi {name}, the {noun} is scheduled from {A} to {B}. This is the same as time from {C} to`
	music	`Hi {name}, on the musical scale, the interval from the note {A} to the note {B} is the same as the interval from the note {C} to the note`
Llama-3.1-8B	months	`Hi {name}, the {noun} is scheduled between {A} {dates} and {B} {dates}. This is the same time as between {C} {dates} and`

Continued on next page

Table 1 continued

Model	Domain	Prompt template
	hours	`Hi {name}, the {noun} is scheduled between {day} {A} o'clock and {day} {B} o'clock. This is the same time as between {day} {C} o'clock and {day}`
	weekdays	`Hi {name}, the {noun} is scheduled between {A} and {B}. This is the same as time between {C} and`
	music	`Hi {name}, on the musical scale, the interval from the note {A} to the note {B} is the same as the interval from the note {C} to the note`
Qwen-2.5-7B	months	`Hello {name}, the {noun} is scheduled from {A} {datesDouble} to {B} {datesDouble}. This is the same as duration from {C} {datesDouble} to`
	hours	N/A
	weekdays	`Hi {name}, the {noun} is scheduled from {A} to {B}. This is the same time as from {C} to`
	music	`Hi {name}, on the musical scale, the interval from the note {A} to the note {B} is the same as the interval from the note {C} to the note`
Qwen-3-8B	months	`Hello {name}, the {noun} is scheduled from {A} {datesDouble} to {B} {datesDouble}. This is the same time as from {C} {datesDouble} to`
	hours	N/A
	weekdays	`Hi {name}, the {noun} is scheduled from {A} to {B}. This is the same time as from {C} to`
	music	`Hi {name}, on the musical scale, the interval from the note {A} to the note {B} is the same as the interval from the note {C} to the note`
Gemma-2-9B	months	`Hello {name}, the {noun} is scheduled from {A} to {B}. This is the same as duration from {C} to`
	hours	N/A
	weekdays	`Hi {name}, the {noun} is scheduled from {A} to {B}. This is the same time as from {C} to`
	music	`Hi {name}, on the musical scale, the interval from the note {A} to the note {B} is the same as the interval from the note {C} to the note`
Gemma-3-12B	months	`Hello {name}, the {noun} is scheduled from {A} to {B}. This is the same time as from {C} to`

Continued on next page

Table 1 continued

Model	Domain	Prompt template
	hours	`N/A`
	weekdays	`Hi {name}, the {noun} is scheduled from {A} to {B}. This is the same time as from {C} to`
	music	`Hi {name}, on the musical scale, the interval from the note {A} to the note {B} is the same as the interval from the note {C} to the note`
Mistral-small-24B-Base-2501	months	`Hi {name}, the {noun} is scheduled from {A} to {B}. This is the same as time from {C} to`
	hours	`N/A`
	weekdays	`Hi {name}, the {noun} is scheduled between {A} and {B}. This is the same as time between {C} and`
	music	`Hi {name}, on the musical scale, the interval from the note {A} to the note {B} is the same as the interval from the note {C} to the note`

B THEORETICAL BACKGROUND

This appendix collects properties of the functional ANOVA decomposition used throughout the paper and establishes the identities underlying the mode fraction, the energy fraction and steering analyses. Let us revisit the functional ANOVA decomposition defined in the main text,

$$\boldsymbol{\phi}^r(A,B,C) = \boldsymbol{\mu}^r + \boldsymbol{h}^r_A + \boldsymbol{h}^r_B + \boldsymbol{h}^r_C + \boldsymbol{h}^r_{AB} + \boldsymbol{h}^r_{BC} + \boldsymbol{h}^r_{CA} + \boldsymbol{h}^r_{ABC},$$

where

$$\begin{aligned}
\boldsymbol{\mu}^r &:= \mathbb{E}_{ABC}(\boldsymbol{\phi}^r),\\
\boldsymbol{h}^r_A &:= \mathbb{E}_{BC}(\boldsymbol{\phi}^r) - \boldsymbol{\mu}^r,\\
\boldsymbol{h}^r_{AB} &:= \mathbb{E}_{C}(\boldsymbol{\phi}^r) - \boldsymbol{\mu}^r - \boldsymbol{h}^r_A - \boldsymbol{h}^r_B,\\
\boldsymbol{h}^r_{ABC} &:= \boldsymbol{\phi}^r - \boldsymbol{\mu}^r - \boldsymbol{h}^r_A - \boldsymbol{h}^r_B - \boldsymbol{h}^r_C - \boldsymbol{h}^r_{AB} - \boldsymbol{h}^r_{BC} - \boldsymbol{h}^r_{CA}.
\end{aligned}$$

To standardize notation, we use $\boldsymbol{h}^r_\emptyset = \boldsymbol{\mu}^r$. Using these definitions, it immediately follows that each non-constant decomposed term has zero mean with respect to each variable on which it depends.

Lemma 1 (Zero mean of ANOVA terms)**.** *For every nonempty $S \subseteq \{A,B,C\}$ and every $X \in S$,*

$$\mathbb{E}_X[\boldsymbol{h}^r_S] = \mathbf{0}. \tag{17}$$

Proof. Consider $\boldsymbol{h}^r_A$:

$$\begin{aligned}
\mathbb{E}_A(\boldsymbol{h}^r_A) &= \mathbb{E}_A\left(\mathbb{E}_{BC}(\boldsymbol{\phi}^r) - \boldsymbol{\mu}^r\right)\\
&= \mathbb{E}_{ABC}(\boldsymbol{\phi}^r) - \boldsymbol{\mu}^r\\
&= \boldsymbol{\mu}^r - \boldsymbol{\mu}^r\\
&= 0.
\end{aligned}$$

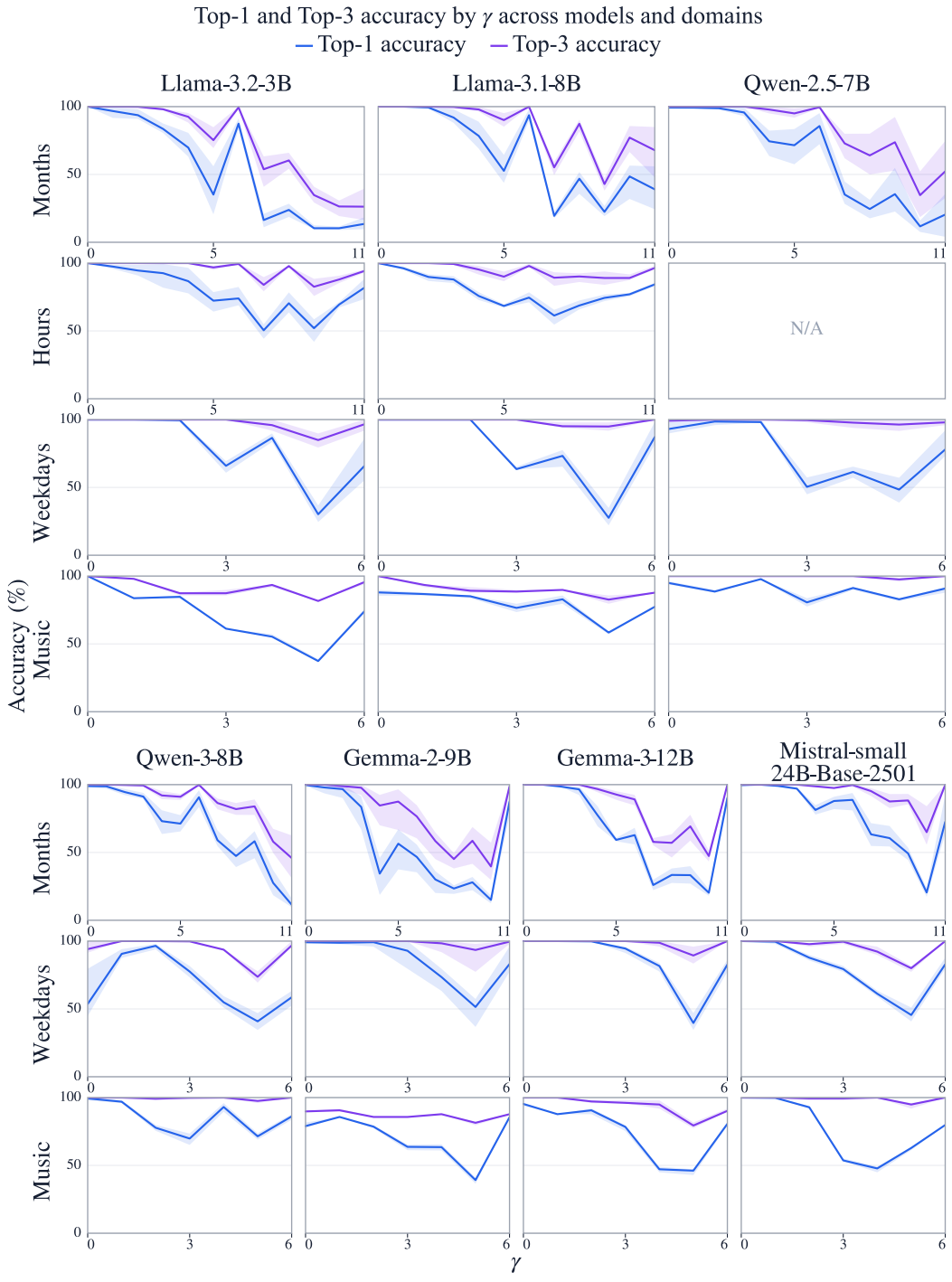

Figure 6: Top-1 and top-3 accuracy as a function of $\gamma = B - A$ across models and cyclic concepts. Columns correspond to models and rows to concepts. Curves show accuracy averaged across the 10 sampled replicates, with shaded bands showing the central 80% interval. Hours are omitted for Qwen, Gemma, and Mistral because the relevant numerical values do not tokenize to a common width. Top-3 accuracy is generally more stable across γ than top-1 accuracy.

Table 2: Replicate variables used in the experiments.

Variable	Values
`name`	Each name should tokenize to one token. There are 194 approved names for each Llama, Qwen, and Mistral model, and 200 for each Gemma model from the list of 200 most common names of the last century (Social Security Administration, n.d.).
`noun`	`conference`, `meeting`, `forum`, `summit`
`dates`	`1`,...,`28`
`datesDouble`	`10`,...,`28`
`day`	`Monday`, `Tuesday`, `Wednesday`, `Thursday`, `Friday`, `Saturday`, `Sunday`

Table 3: Values assigned to the cyclic variables A, B, and C in each domain. Months and hours have cycle length $N = 12$, while weekdays and musical notes have $N = 7$.

Domain	Values of A, B, and C
Months	`January`, `February`, `March`, `April`, `May`, `June`, `July`, `August`, `September`, `October`, `November`, `December`
Hours	`1`, `2`, `3`, `4`, `5`, `6`, `7`, `8`, `9`, `10`, `11`, `12`
Weekdays	`Monday`, `Tuesday`, `Wednesday`, `Thursday`, `Friday`, `Saturday`, `Sunday`
Music	`C`, `D`, `E`, `F`, `G`, `A`, `B`

Similarly, for $\boldsymbol{h}^r_{AB}$

$$\begin{aligned}\mathbb{E}_A(\boldsymbol{h}^r_{AB}) &= \mathbb{E}_A\left(\mathbb{E}_C(\boldsymbol{\phi}^r) - \boldsymbol{\mu}^r - \boldsymbol{h}^r_A - \boldsymbol{h}^r_B\right)\\ &= \mathbb{E}_{AC}(\boldsymbol{\phi}^r) - \boldsymbol{\mu}^r - \mathbb{E}_A(\boldsymbol{h}^r_A) - \boldsymbol{h}^r_B\\ &= (\boldsymbol{\mu}^r + \boldsymbol{h}^r_B) - \boldsymbol{\mu}^r - 0 - \boldsymbol{h}^r_B\\ &= 0.\end{aligned}$$

The same argument holds for $\mathbb{E}_B(\boldsymbol{h}_{AB}) = 0$. Finally, consider $\boldsymbol{h}^r_{ABC}$

$$\begin{aligned}\mathbb{E}_A(\boldsymbol{h}^r_{ABC}) &= \mathbb{E}_A\left(\boldsymbol{\phi}^r - \boldsymbol{\mu}^r - \boldsymbol{h}^r_A - \boldsymbol{h}^r_B - \boldsymbol{h}^r_C - \boldsymbol{h}^r_{AB} - \boldsymbol{h}^r_{BC} - \boldsymbol{h}^r_{CA}\right)\\ &= \mathbb{E}_A(\boldsymbol{\phi}^r) - \boldsymbol{\mu}^r - \mathbb{E}_A(\boldsymbol{h}^r_A) - \boldsymbol{h}^r_B - \boldsymbol{h}^r_C - \mathbb{E}_A(\boldsymbol{h}^r_{AB}) - \boldsymbol{h}^r_{BC} - \mathbb{E}_A(\boldsymbol{h}^r_{CA})\\ &= (\boldsymbol{\mu}^r + \boldsymbol{h}^r_B + \boldsymbol{h}^r_C + \boldsymbol{h}^r_{BC}) - \boldsymbol{\mu}^r - 0 - \boldsymbol{h}^r_B - \boldsymbol{h}^r_C - 0 - \boldsymbol{h}^r_{BC} - 0\\ &= 0.\end{aligned}$$

Similarly, $\mathbb{E}_B(\boldsymbol{h}^r_{ABC}) = \mathbb{E}_C(\boldsymbol{h}^r_{ABC}) = 0$. In general, it can be proved by induction. □

Definition 1 (Ensemble-averaged form)**.** *For $S, T \subseteq \{A, B, C\}$ and two decomposed ensembles H^r_S and H^r_T with vectors $\boldsymbol{h}^r_S \in H^r_S$ and $\boldsymbol{h}^r_T \in H^r_T$, define the ensemble-averaged form as*

$$\langle H^r_S, H^r_T\rangle_{\text{ens}} := \mathbb{E}_{ABC}\left[\langle \boldsymbol{h}^r_S, \boldsymbol{h}^r_T\rangle\right] \tag{18}$$

where the inner product on the RHS is the standard Euclidean inner product defined over $\mathbb{R}^{d_{model}}$.

Proposition 1 (Inner product)**.** *Under conditions of Definition 1, Equation 18 satisfies the properties of an inner product:*

1. ***Symmetry***. $\langle H^r_S, H^r_T\rangle_{\text{ens}} = \langle H^r_T, H^r_S\rangle_{\text{ens}}$.
2. ***Linearity***. $\langle \alpha H^r_S + \beta H^r_T, H^r_U\rangle_{\text{ens}} = \alpha\langle H^r_S, H^r_U\rangle_{\text{ens}} + \beta\langle H^r_T, H^r_U\rangle_{\text{ens}}$ *for any* $\alpha, \beta \in \mathbb{R}$.

Table 4: Top-1 and top-3 accuracy by model and domain. Values are mean ± standard deviation across 10 replicates, computed over all prompts, including coincident values of A, B, and C.

Model	Performance metric	Months	Hours	Weekdays	Music
Llama-3.2-3B	Top-1 accuracy	$53.4 \pm 2.2\%$	$78.5 \pm 3.2\%$	$78.3 \pm 2.5\%$	$70.9 \pm 0.3\%$
	Top-3 accuracy	$72.2 \pm 2.8\%$	$95.2 \pm 1.1\%$	$96.8 \pm 0.8\%$	$91.9 \pm 0.5\%$
Llama-3.1-8B	Top-1 accuracy	$66.0 \pm 3.9\%$	$79.9 \pm 1.8\%$	$78.8 \pm 1.1\%$	$79.3 \pm 0.7\%$
	Top-3 accuracy	$84.8 \pm 2.6\%$	$94.7 \pm 1.5\%$	$98.6 \pm 0.3\%$	$90.2 \pm 0.4\%$
Qwen-2.5-7B	Top-1 accuracy	$62.7 \pm 7.0\%$	N/A	$75.4 \pm 2.4\%$	$89.5 \pm 0.6\%$
	Top-3 accuracy	$82.5 \pm 6.8\%$	N/A	$98.7 \pm 0.8\%$	$99.7 \pm 0.2\%$
Qwen-3-8B	Top-1 accuracy	$68.5 \pm 4.5\%$	N/A	$67.5 \pm 3.4\%$	$84.9 \pm 1.7\%$
	Top-3 accuracy	$86.5 \pm 2.6\%$	N/A	$94.0 \pm 1.0\%$	$99.5 \pm 0.3\%$
Gemma-2-9B	Top-1 accuracy	$58.3 \pm 5.9\%$	N/A	$85.4 \pm 2.6\%$	$70.7 \pm 0.6\%$
	Top-3 accuracy	$78.8 \pm 6.2\%$	N/A	$98.8 \pm 1.7\%$	$87.0 \pm 0.3\%$
Gemma-3-12B	Top-1 accuracy	$66.5 \pm 1.8\%$	N/A	$85.5 \pm 1.4\%$	$75.1 \pm 1.3\%$
	Top-3 accuracy	$84.1 \pm 1.7\%$	N/A	$98.3 \pm 1.1\%$	$94.0 \pm 0.8\%$
Mistral-small-24B-Base-2501	Top-1 accuracy	$76.7 \pm 2.7\%$	N/A	$79.5 \pm 0.7\%$	$76.7 \pm 0.7\%$
	Top-3 accuracy	$94.3 \pm 1.6\%$	N/A	$95.7 \pm 0.7\%$	$99.1 \pm 0.3\%$

3. ***Positive definiteness***. $\langle H_S^r, H_S^r \rangle_{\text{ens}} \geq 0$ *and the equality holds iff* $H_S^r = 0$.

Proof. **Symmetry.** By symmetry of the Euclidean inner product,

$$\langle H_S^r, H_T^r \rangle_{\text{ens}} = \mathbb{E}_{ABC} \left[\langle \boldsymbol{h}_S^r, \boldsymbol{h}_T^r \rangle \right] = \mathbb{E}_{ABC} \left[\langle \boldsymbol{h}_T^r, \boldsymbol{h}_S^r \rangle \right] = \langle H_T^r, H_S^r \rangle_{\text{ens}} .$$

Linearity. For any $\alpha, \beta \in \mathbb{R}$ and ensembles H_S^r, H_T^r, H_U^r, linearity of the Euclidean inner product and of expectation gives

$$\begin{aligned} \langle \alpha H_S^r + \beta H_T^r, H_U^r \rangle_{\text{ens}} &= \mathbb{E}_{ABC} \left[\langle \alpha \boldsymbol{h}_S^r + \beta \boldsymbol{h}_T^r, \boldsymbol{h}_U^r \rangle \right] \\ &= \alpha \langle H_S^r, H_U^r \rangle_{\text{ens}} + \beta \langle H_T^r, H_U^r \rangle_{\text{ens}} . \end{aligned}$$

Positive definiteness. For every ensemble H_S^r,

$$\langle H_S^r, H_S^r \rangle_{\text{ens}} = \mathbb{E}_{ABC} \left[\| \boldsymbol{h}_S^r \|_2^2 \right] \geq 0.$$

A nonnegative random variable has expectation zero if and only if it is zero. Hence

$$\langle H_S^r, H_S^r \rangle_{\text{ens}} = 0 \quad \Longleftrightarrow \quad \boldsymbol{h}_S^r = 0.$$

This is equivalent to $H_S^r = 0$ in the ensemble space.

Thus the form satisfies all three inner product axioms. □

Proposition 2 (Orthogonality of the ANOVA terms)**.** *Under the full ensemble, any two distinct terms in the decomposition are orthogonal under the inner product in Definition 1:*

$$\langle H_S^r, H_T^r \rangle_{\text{ens}} = 0, \qquad S \neq T. \tag{19}$$

Here, $S, T \subseteq \{A, B, C\}$

Proof. Let $S \neq T$. Without loss of generality, choose a variable $X \in S \setminus T$. Since $\boldsymbol{h}_T^r$ does not depend on X, we may average over X first:

$$\begin{aligned} \langle H_S^r, H_T^r \rangle_{\text{ens}} &= \mathbb{E}_{\{ABC\} \setminus \{X\}} \left[\langle \mathbb{E}_X \left[\boldsymbol{h}_S^r \right], \boldsymbol{h}_T^r \rangle \right] \\ &= 0, \end{aligned}$$

where the final equality follows from Lemma 1. Hence all distinct terms are pairwise orthogonal. □

Corollary 1 (Energy decomposition)**.** *Under the conditions of Proposition 2,*

$$\mathbb{E}_{ABC}\left\|\boldsymbol{\phi}^r(A,B,C)-\boldsymbol{h}^r_{\emptyset}\right\|_2^2=\sum_{\emptyset\neq S\subseteq\{A,B,C\}}\mathbb{E}_S\left\|\boldsymbol{h}^r_S\right\|_2^2. \tag{20}$$

Proof. By Equation 2,

$$\boldsymbol{\phi}^r-\boldsymbol{h}^r_{\emptyset}=\sum_{\emptyset\neq S\subseteq\{A,B,C\}}\boldsymbol{h}^r_S.$$

Expanding the squared norm and applying Proposition 2 eliminates all cross terms:

$$\begin{aligned}\mathbb{E}_{ABC}\left\|\boldsymbol{\phi}^r-\boldsymbol{h}^r_{\emptyset}\right\|_2^2&=\sum_{\emptyset\neq S\subseteq\{A,B,C\}}\mathbb{E}_{ABC}\left\|\boldsymbol{h}^r_S\right\|_2^2\\&=\sum_{\emptyset\neq S\subseteq\{A,B,C\}}\mathbb{E}_S\left\|\boldsymbol{h}^r_S\right\|_2^2,\end{aligned}$$

where the second equality follows because $\boldsymbol{h}^r_S$ depends only on the variables contained in S. □

B.1 Mode fraction and cluster analysis

In Section 3 of the main text, we defined mode fraction using the DFT. It can instead be defined directly without invoking the DFT. Here, we show the equivalence. Figure 1 suggests that the interaction ensembles form clusters indexed by γ or D.

Definition 2 (Cluster vectors)**.** *For H^r_{AB}, we define the centroid of the cluster indexed by γ as*

$$\bar{\boldsymbol{h}}^r_{\gamma}:=\mathbb{E}[\boldsymbol{h}^r_{AB}\,|\,B-A=\gamma]\,. \tag{21}$$

Similarly, for H^r_{ABC}, we define the centroid of the cluster indexed by D as

$$\bar{\boldsymbol{h}}^r_{D}:=\mathbb{E}[\boldsymbol{h}^r_{ABC}\,|\,C+B-A=D]\,. \tag{22}$$

Note that these are the same vectors used in the steering experiments in Section 5. We now define the corresponding projection operators for each ensemble.

Definition 3 (Projection operator)**.** *Let P_γ be the projection operator that operates on H^r_{AB} to give the corresponding cluster vector:*

$$P_\gamma H^r_{AB}:=\bar{\boldsymbol{h}}^r_{\gamma} \tag{23}$$

Similarly, we define P_D as

$$P_D H^r_{ABC}:=\bar{\boldsymbol{h}}^r_{D} \tag{24}$$

Thus, $P_\gamma H^r_{AB}$ is constant over pairs with the same value of γ, while $P_D H^r_{ABC}$ is constant over triples with the same value of D. Now we define cluster fraction using these projection operators and show that it is equivalent to the definition of mode fraction in Section 3.

Definition 4 (Cluster fraction)**.** *Let $c(H^r)$ be the cluster fraction defined by*

$$c(H^r_{AB}):=\frac{\mathbb{E}_\gamma\left\|P_\gamma H^r_{AB}\right\|_2^2}{\mathbb{E}_{AB}\left\|\boldsymbol{h}^r_{AB}\right\|_2^2}=\frac{\mathbb{E}_\gamma\left\|\bar{\boldsymbol{h}}^r_{\gamma}\right\|_2^2}{\mathbb{E}_{AB}\left\|\boldsymbol{h}^r_{AB}\right\|_2^2} \tag{25}$$

and

$$c(H^r_{ABC}):=\frac{\mathbb{E}_D\left\|P_D H^r_{ABC}\right\|_2^2}{\mathbb{E}_{ABC}\left\|\boldsymbol{h}^r_{ABC}\right\|_2^2}=\frac{\mathbb{E}_D\left\|\bar{\boldsymbol{h}}^r_{D}\right\|_2^2}{\mathbb{E}_{ABC}\left\|\boldsymbol{h}^r_{ABC}\right\|_2^2} \tag{26}$$

Proposition 3 (Equivalence between cluster fraction and mode fraction)**.** *If $c(H^r)$ is the cluster fraction as in Definition 4 and $m(H^r)$ is the mode fraction as defined in Section 3, then*

$$c(H^r_S)=m(H^r_S), \tag{27}$$

for $S\in\{AB,BC,CA,ABC\}$.

Proof. We suppress the superscript r for readability.

First consider H_{AB}. Let

$$\boldsymbol{p}_{AB} := P_\gamma H_{AB} = \bar{\boldsymbol{h}}_\gamma.$$

Here, $\gamma = B - A$, and hence $B = A + \gamma$. The discrete Fourier transform (DFT) of $\boldsymbol{p}$ is

$$\begin{aligned}
\widehat{\boldsymbol{p}}_{k_A k_B} &= \sum_{A,B} \boldsymbol{p}_{AB} e^{-2\pi i (k_A A + k_B B)/N} \\
&= \sum_{A,\gamma} \bar{\boldsymbol{h}}_\gamma e^{-2\pi i [(k_A + k_B)A + k_B \gamma]/N} \\
&= \left(\sum_A e^{-2\pi i (k_A + k_B) A/N} \right) \left(\sum_\gamma \bar{\boldsymbol{h}}_\gamma e^{-2\pi i k_B \gamma/N} \right).
\end{aligned}$$

Here, N is the length of the cycle; $N = 12$ for months and hours, while $N = 7$ for weekdays and musical notes. The sum over A vanishes unless $k_A + k_B = 0$. Therefore,

$$\widehat{\boldsymbol{p}}_{k_A k_B} = \mathbf{0} \qquad \text{whenever } k_A \neq -k_B.$$

For frequencies $(k_A, k_B) = (-k, k)$,

$$\widehat{\boldsymbol{p}}_{-k,k} = N \sum_\gamma \bar{\boldsymbol{h}}_\gamma e^{-2\pi i k \gamma/N} \tag{28}$$

$$= \sum_{A,\gamma} \boldsymbol{h}_{A,A+\gamma} e^{-2\pi i k \gamma/N} \tag{29}$$

$$= \sum_{A,B} \boldsymbol{h}_{AB} e^{-2\pi i k (B-A)/N} \tag{30}$$

$$= \widehat{\boldsymbol{h}}_{-k,k}, \tag{31}$$

where the second equality uses

$$\bar{\boldsymbol{h}}_\gamma = \frac{1}{N} \sum_A \boldsymbol{h}_{A,A+\gamma}.$$

Thus, the Fourier transform of $P_\gamma H_{AB}$ agrees with that of H_{AB} on the modes $k_A = -k_B$ and vanishes on all other modes.

Consider $c(H_{AB})$:

$$\begin{aligned}
c(H_{AB}) &= \frac{\mathbb{E}_{AB} \left\| \boldsymbol{p}_{AB} \right\|_2^2}{\mathbb{E}_{AB} \left\| \boldsymbol{h}_{AB} \right\|_2^2} \\
&= \frac{\sum_{k_A = -k_B} \left\| \widehat{\boldsymbol{p}}_{k_A k_B} \right\|_2^2}{\sum_{k_A, k_B} \left\| \widehat{\boldsymbol{h}}_{k_A k_B} \right\|_2^2} \\
&= \frac{\sum_{k_A = -k_B} \left\| \widehat{\boldsymbol{h}}_{k_A k_B} \right\|_2^2}{\sum_{k_A, k_B} \left\| \widehat{\boldsymbol{h}}_{k_A k_B} \right\|_2^2} \\
&= m(H_{AB}),
\end{aligned}$$

where we get the second equality using Parseval's identity and the third equality using Equation 31.

A similar proof holds for $c(H_{ABC}) = m(H_{ABC})$. □

B.2 Steering vector

Finally, we show the correspondence between the choice of steering vector as in Section 5 and the conditional mean of the activation ensemble.

Proposition 4. *If $\bar{\boldsymbol{h}}^r_\gamma$ is the vector according to Definition 2, then*

$$\bar{\boldsymbol{h}}^r_\gamma = \mathbb{E}_{ABC}\left[\boldsymbol{\phi}^r(A,B,C) \mid B - A = \gamma\right] - \boldsymbol{\mu}^r. \tag{32}$$

Similarly,

$$\bar{\boldsymbol{h}}^r_D = \mathbb{E}_{ABC}\left[\boldsymbol{\phi}^r(A,B,C) \mid C + B - A = D\right] - \boldsymbol{\mu}^r. \tag{33}$$

Proof. To see this, start with the conditional average over $\boldsymbol{\phi}^r$:

$$\begin{aligned}\mathbb{E}[\boldsymbol{\phi}^r \mid B - A = \gamma] = \boldsymbol{\mu}^r &+ \mathbb{E}[\boldsymbol{h}^r_A \mid B - A = \gamma] + \mathbb{E}[\boldsymbol{h}^r_B \mid B - A = \gamma] \\ &+ \mathbb{E}[\boldsymbol{h}^r_C \mid B - A = \gamma] + \mathbb{E}[\boldsymbol{h}^r_{AB} \mid B - A = \gamma] \\ &+ \mathbb{E}[\boldsymbol{h}^r_{BC} \mid B - A = \gamma] + \mathbb{E}[\boldsymbol{h}^r_{CA} \mid B - A = \gamma] \\ &+ \mathbb{E}[\boldsymbol{h}^r_{ABC} \mid B - A = \gamma]\,.\end{aligned}$$

Using Lemma 1, we get

$$\mathbb{E}[\boldsymbol{h}^r_A \mid B - A = \gamma] = \mathbb{E}[\boldsymbol{h}^r_B \mid B - A = \gamma] = \mathbb{E}[\boldsymbol{h}^r_C \mid B - A = \gamma] = \mathbf{0},$$

and

$$\mathbb{E}[\boldsymbol{h}^r_{BC} \mid B - A = \gamma] = \mathbb{E}[\boldsymbol{h}^r_{CA} \mid B - A = \gamma] = \mathbb{E}[\boldsymbol{h}^r_{ABC} \mid B - A = \gamma] = \mathbf{0}.$$

Hence,

$$\mathbb{E}[\boldsymbol{\phi}^r \mid B - A = \gamma] = \boldsymbol{\mu}^r + \mathbb{E}[\boldsymbol{h}^r_{AB} \mid B - A = \gamma]\,.$$

By definition,

$$\bar{\boldsymbol{h}}^r_\gamma = \mathbb{E}[\boldsymbol{h}^r_{AB} \mid B - A = \gamma]\,,$$

and therefore

$$\bar{\boldsymbol{h}}^r_\gamma = \mathbb{E}[\boldsymbol{\phi}^r \mid B - A = \gamma] - \boldsymbol{\mu}^r.$$

A similar proof holds for $\bar{\boldsymbol{h}}^r_D$. □

Therefore, we get the following identities using Proposition 4:

$$\bar{\boldsymbol{h}}^r_{\gamma+\delta} - \bar{\boldsymbol{h}}^r_\gamma = \mathbb{E}[\boldsymbol{\phi}^r \mid B - A = \gamma + \delta] - \mathbb{E}[\boldsymbol{\phi}^r \mid B - A = \gamma]\,, \tag{34}$$

$$\bar{\boldsymbol{h}}^r_{D+\delta} - \bar{\boldsymbol{h}}^r_D = \mathbb{E}[\boldsymbol{\phi}^r \mid C + B - A = D + \delta] - \mathbb{E}[\boldsymbol{\phi}^r \mid C + B - A = D]\,. \tag{35}$$

Thus, the steering using $\bar{\boldsymbol{h}}^r_{\gamma+\delta} - \bar{\boldsymbol{h}}^r_\gamma$ is equivalent to the difference-of-means steering.

B.3 Coincident variables

We always compute the interaction ensembles using the full set of prompts, including cases with $A = B$, $B = C$, or $C = A$. Table 5 reports the mean norm of the interaction vectors grouped by $\gamma = B - A$, $\gamma' = C - B$, and $\gamma'' = A - C$ for a representative ensemble using the months prompt template at layer 15 of Llama-3.1-8B. The vectors corresponding to $\gamma = 0,\ \gamma' = 0,\ \gamma'' = 0$ have substantially larger norms than the rest of the ensemble and can therefore bias quantitative results toward coincident-variable cases. We exclude such prompts when reporting the results in the main text. Unless stated otherwise, we repeat our analyses without this exclusion in the following appendices. This also serves as a sanity check that our conclusions are not driven by the coincident-variable cases.

C Mode fraction

The main text reports the layerwise organization of the interaction ensembles for Llama-3.1-8B on the months domain. Here, we test whether the same pattern generalizes across models and cyclic domains. Recall the definition of mode fraction for H^r_{AB} as defined in the main text:

$$m(H^r_{AB}) := \frac{\sum_{k_A=-k_B} \left\|\widehat{\boldsymbol{h}}^r_{k_A k_B}\right\|_2^2}{\sum_{k_A,k_B} \left\|\widehat{\boldsymbol{h}}^r_{k_A k_B}\right\|_2^2}. \tag{36}$$

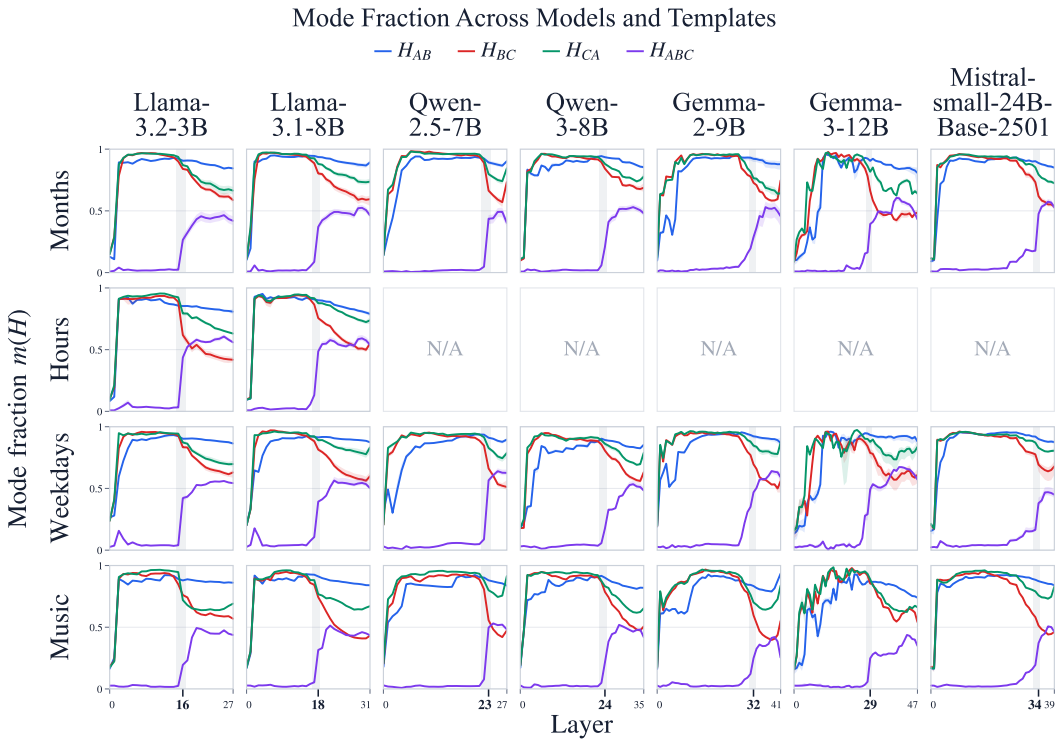

Figure 7: Mode fraction of the interaction ensembles across models and cyclic domains. Columns correspond to models and rows to domains; shaded bands show the central 80% interval across 10 replicates. The vertical gray band marks the transition in causal relevance from H^r_{AB} to H^r_{ABC} identified by the ablation and replacement experiments. Across all models and prompt templates, the pairwise interaction ensembles are strongly organized by their corresponding pairwise differences before H^r_{ABC} becomes organized by D. The transition in H^r_{ABC} is more distributed across depth in the Gemma models.

Table 5: Mean norm of each interaction vector categorized by $\gamma = B - A$, $\gamma' = C - B$, and $\gamma'' = A - C$, respectively, for a single replicate of the months prompt template at layer 15 of Llama-3.1-8B. The interaction vectors associated with zero difference have substantially larger norms than the remaining classes. This motivates the exclusion of coincident-variable prompts from the quantitative analyses in the main text.

Ensemble	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
H_{AB}	7.445	5.875	3.899	3.568	3.144	2.551	2.391	2.307	2.595	2.700	3.188	4.023
H_{BC}	4.343	1.921	0.918	0.961	0.971	1.021	0.853	0.976	1.022	1.378	1.377	1.575
H_{CA}	6.448	3.013	1.311	1.457	1.478	1.624	1.476	1.487	1.263	1.220	1.589	3.189

Similarly, we define the corresponding mode fractions for H^r_{BC} and H^r_{CA} using modes satisfying $k_B + k_C = 0$ and $k_C + k_A = 0$ respectively. For H^r_{ABC}, the mode fraction is defined using modes $(k_A, k_B, k_C) = (-k, k, k)$.

Figure 7 shows the mode fractions for all interaction ensembles using the prompts and models as detailed in Table 1. Across all models and prompt templates, the second-order interaction ensembles become strongly organized by their corresponding pairwise differences before the three-way interaction becomes organized by D. The increase in $m(H^r_{ABC})$ occurs near the same relative depth within a particular model across domains. The transition is sharp for most models but is distributed across several layers in the Gemma models. Notice that we report Llama-3.1-8B on the months domain again: we include prompts with coincident variables and we see that the mode fraction is inflated as a result of their inclusion.

D Energy Fraction

Recall Corollary 1 from Appendix B:

$$\mathbb{E}_{A,B,C} \left\| \boldsymbol{\phi}^r(A, B, C) - \boldsymbol{\mu}^r \right\|_2^2 = \sum_S \mathbb{E}_S \left\| \boldsymbol{h}^r_S \right\|_2^2, \tag{37}$$

where S ranges over non-empty subsets of $\{A, B, C\}$. This motivates us to define the *energy fraction* of each ensemble as

$$e(H^r_S) = \frac{\mathbb{E}_S\Big(\left\| \boldsymbol{h}^r_S \right\|_2^2 \Big)}{\mathbb{E}_{ABC}\Big(\left\| \boldsymbol{\phi}^r(A, B, C) - \boldsymbol{\mu}^r \right\|_2^2 \Big)}, \qquad \text{with} \qquad \sum_S e(H^r_S) = 1. \tag{38}$$

We next test whether the layerwise redistribution of energy fraction observed in the main text generalizes across models and domains. Figure 8 shows the energy fraction for each interaction ensemble. Across all models and prompt templates, the energy fraction of H^r_{AB} decreases near the same depth at which the energy fraction of H^r_{ABC} increases. As with mode fraction, this transition is spread over several layers in the Gemma models.

E Controls for Llama-3.1-8B

The strong organization of the pairwise interaction ensembles by their corresponding pairwise differences does not by itself imply that this structure is task-specific. We therefore test whether similar organization arises when the cyclic tokens co-occur without the relational computation used in the main task. We design three control prompts as listed in Table 6. For the *non-cyclic* control, we construct a template similar to the Llama-3.1-8B months template but use $A, B, C \in \{$berry, apple, orange, banana, peach, pear, fig, mango, lemon, lime, cherry, plum$\}$ instead of the usual calendar months. The *off-by-1* control retains A, B, C as calendar months, but the answer depends only on C; A and B therefore co-occur in the prompt without contributing to the answer. The *list* control also retains calendar months but removes the arithmetic relation entirely; A, B, C are presented as items in a list and the model is asked to return B.

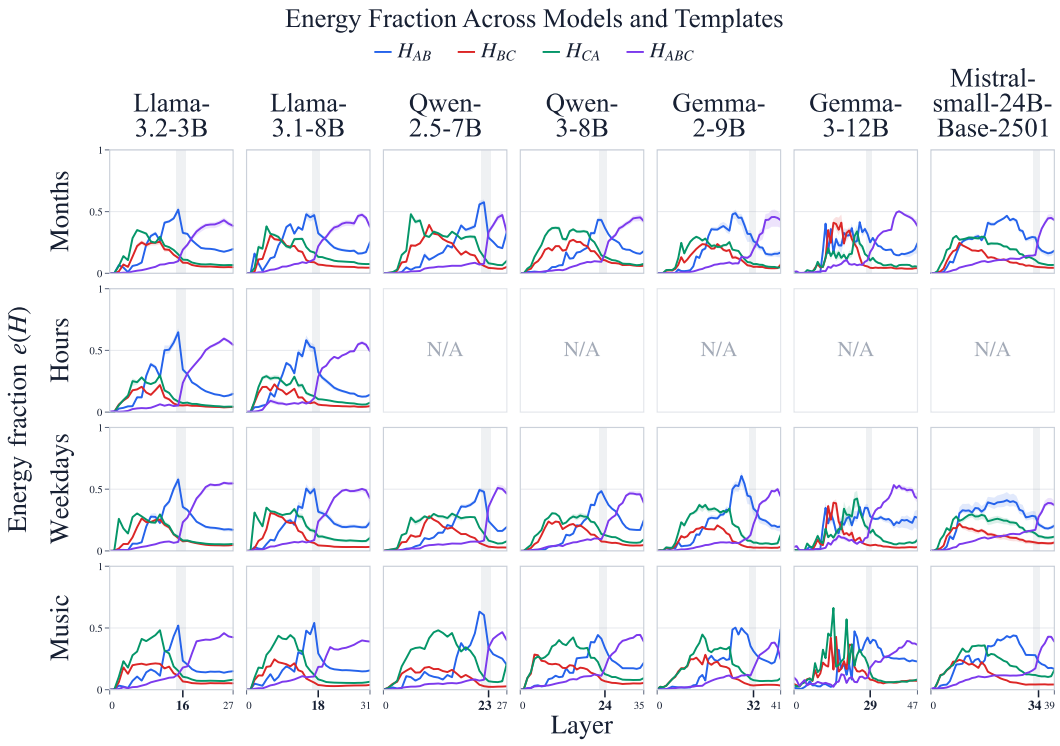

Figure 8: Energy fraction of the interaction ensembles across models and cyclic domains. Columns correspond to models and rows to domains; shaded bands show the central 80% interval across 10 replicates. The vertical gray band marks the transition in causal relevance identified by the ablation and replacement experiments. Across all models and prompt templates, the energy fraction of H^r_{AB} decreases near the same depth at which the energy fraction of H^r_{ABC} increases.

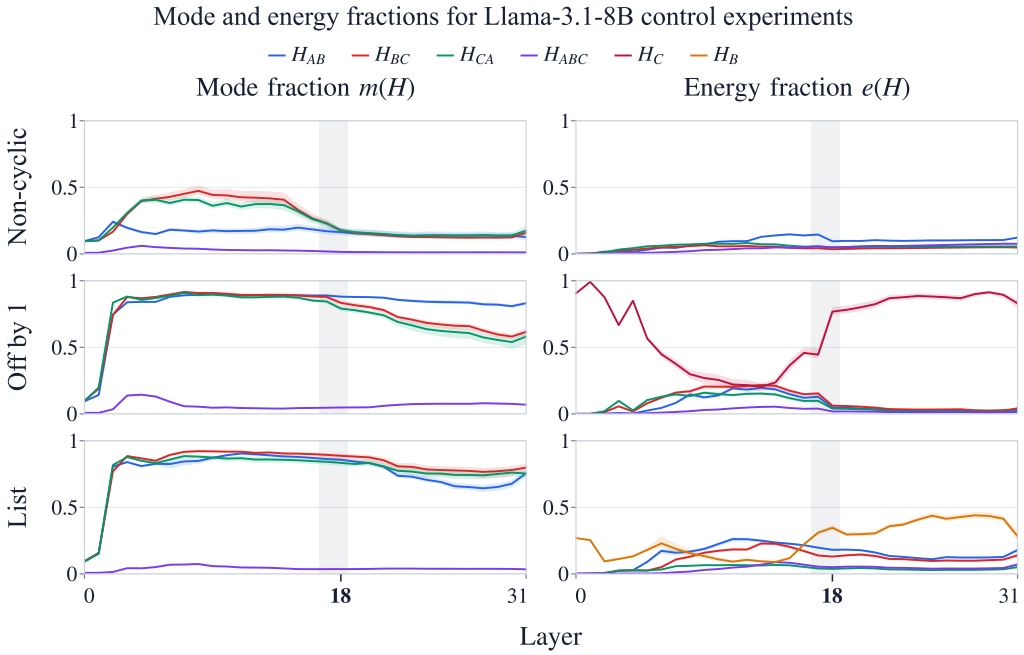

Figure 9: Mode and energy fractions for three control prompt templates using Llama-3.1-8B. The *non-cyclic* control replaces months with an unordered vocabulary, while the *off-by-1* and *list* controls retain calendar months but remove the relational computation of the main task. In the latter two controls, the pairwise interaction ensembles remain strongly organized by their corresponding pairwise differences despite having a small energy fraction. In contrast, H^r_{ABC} remains negligible in both mode and energy fraction across all controls. All results reported here exclude vectors corresponding to the coincident-variable cases.

Table 6: Control prompt templates for Llama-3.1-8B.

Control	Prompt template
Non-cyclic	`Hello {name}, the {noun} is scheduled between {A} {dates} and {B} {dates}. This is the same time as between {C} {dates} and`
Off by 1	`Hello {name}, the {noun} is scheduled between {A} {dates} and {B} {dates}. The month that comes after {C} {dates} is`
List	`Hello {name}, the talks are on {A} {dates}, on {B} {dates}, and on {C} {dates}. The second talk is on`

In Figure 9, we have computed the mode and energy fractions for the control prompt templates. In all three controls, the mode and energy fractions of H^r_{ABC} remain close to zero. In contrast, when calendar months co-occur in the *off-by-1* and *list* controls, H^r_{AB}, H^r_{BC}, and H^r_{CA} can still have large mode fractions, but retain a small energy fraction. This suggests that pairwise-difference geometry can arise from co-occurrence of cyclic tokens alone, whereas the D-organized interaction ensemble H^r_{ABC} requires the task structure.

F Causal handoff

Section 4 shows that H^r_{AB} and H^r_{ABC} are causally relevant at different depths, but this does not establish whether the later H^r_{ABC} representation depends on the earlier H^r_{AB} representation. To test this, we ablate H^r_{AB} at layer 15 of Llama-3.1-8B on the months domain, allow the model to run normally thereafter, and recompute the ANOVA decomposition at every subsequent layer. We then compare the mode and energy fractions of the resulting H^r_{ABC} with those of the unmodified model.

Figure 10 shows the comparison of mode and energy fractions between the ablated and the unmodified ensemble for H^r_{ABC}. In the unmodified case, both the mode fraction and energy fraction of H^r_{ABC} rise sharply around layer 18. After ablating H^r_{AB} at layer 15, neither transition occurs at that depth. The mode fraction recovers slightly in later layers, but the energy fraction remains substantially reduced. Thus, the emergence of the later H^r_{ABC} representation depends on the earlier H^r_{AB} representation.

G Ensemble ablation

We next test whether the layerwise ablation pattern from the main text generalizes across models and cyclic concepts. We conduct the ensemble ablation experiments for all prompts and models in Table 1. Recall that we intervene on the last token position at a particular layer using

$$\widetilde{\boldsymbol{\phi}}^r(A,B,C) = \boldsymbol{\phi}^r(A,B,C) - \boldsymbol{h}^r_S, \tag{39}$$

for all $A, B, C \in \mathcal{X}$ and $S \in \{AB, BC, CA, ABC\}$.

Figures 12 and 13 report the results of the ensemble ablation experiments using top-1 and top-3 accuracy as metrics, respectively. We observe that across every model and domain, the layerwise behavior of interaction ensembles under ablation remains the same. Ablating H^r_{AB} always affects the accuracy at an earlier network depth than ablating H^r_{ABC}. However, ablating H^r_{BC} or H^r_{CA} has little to no effect on performance. Ablating H^r_{ABC} at later layers decreases the performance dramatically, with performance dropping to near chance when it is ablated near the final layers.

H Ensemble replacement

We repeat the ensemble replacement experiments across all models and prompt templates using the following intervention

$$\widetilde{\boldsymbol{\phi}}^r(A,B,C) = \boldsymbol{\mu}^r + \boldsymbol{h}^r_S, \tag{40}$$

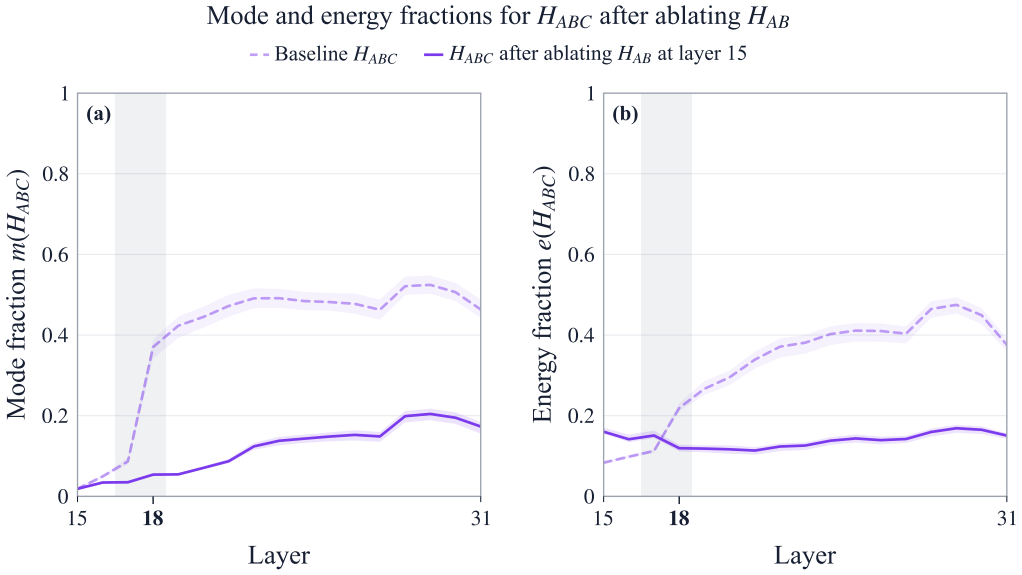

Figure 10: Downstream H^r_{ABC} after ablating H^r_{AB} at layer 15. Solid curves show the mode and energy fractions of the recomputed H^r_{ABC} after the intervention; dashed curves show the corresponding quantities in the unmodified model. The vertical gray band marks layer 18, where H^r_{ABC} normally becomes organized by D and increases sharply in energy fraction. **(a)** After ablating H^r_{AB}, the mode fraction does not undergo its normal increase at layer 18, although some organization recovers later but never regains the unmodified baseline. **(b)** The energy fraction likewise fails to increase at layer 18 and remains substantially below that of the unmodified ensemble. These results show that the normal emergence of H^r_{ABC} depends on the earlier H^r_{AB} representation.

for all $A, B, C \in \mathcal{X}$ and $S \in \{AB, BC, CA, ABC\}$.

Figures 14 and 15 show the corresponding results using top-1 and top-3 accuracy respectively. Retaining H^r_{BC} or H^r_{CA} together with the mean performs similarly to the μ^r-only baseline. Retaining H^r_{AB} preserves performance in intermediate layers but becomes insufficient later in depth, while retaining H^r_{ABC} becomes effective at later layers and can exceed the unmodified baseline.

Together with Appendix G, these results show that the transition from H^r_{AB} to H^r_{ABC} is reproduced by both ablation and replacement interventions across models and prompt templates.

I STEERING

We report steering results across all models and prompt templates using the averaged vectors introduced in Section 5. For $S \in \{\gamma, D\}$, we intervene according to

$$\widetilde{\boldsymbol{\phi}}^r(A, B, C) = \boldsymbol{\phi}^r(A, B, C) + \alpha \left(\bar{\boldsymbol{h}}^r_{S+\delta} - \bar{\boldsymbol{h}}^r_S \right). \tag{41}$$

To compare steering effects across models, domains, and layers, we report the normalized median Δm_3. We normalize by

$$\Delta m_3^{\max} = \max_{S,\delta,r,\ell} \left| \Delta m_3^{r,\ell}(\bar{\boldsymbol{h}}^r_{S+\delta} - \bar{\boldsymbol{h}}^r_S) \right|. \tag{42}$$

Figure 16 shows the resulting steering profiles at $\alpha = 1$. Steering with $\bar{\boldsymbol{h}}^r_\gamma$ is effective in intermediate layers, whereas steering with $\bar{\boldsymbol{h}}^r_D$ becomes effective at later layers. These regions coincide with the layers where H^r_{AB} and H^r_{ABC} are found to be causally relevant in Appendices G and H.

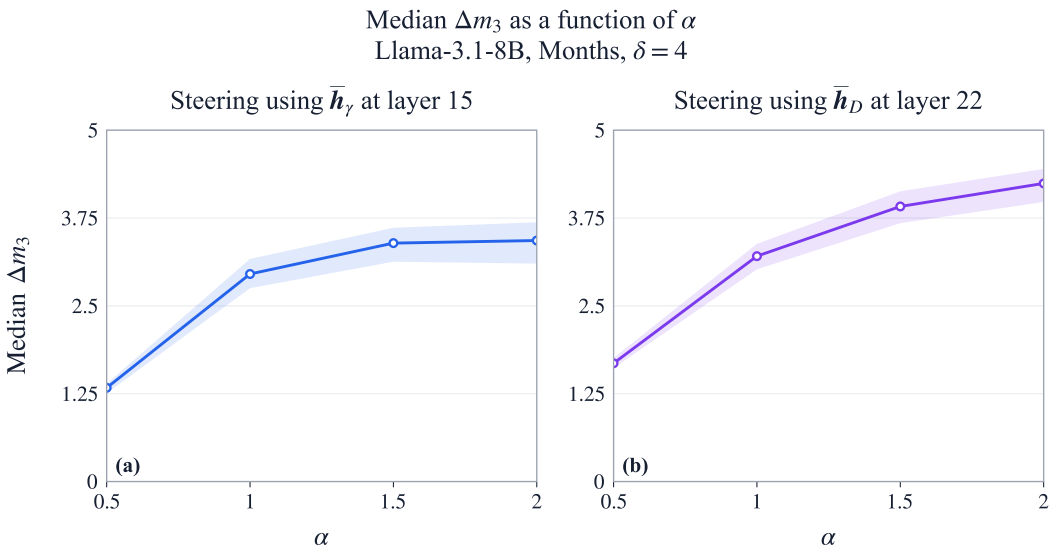

Figure 11: Median Δm_3 as a function of steering strength α for the averaged steering vectors. **(a)** Steering with $\bar{\boldsymbol{h}}^r_\gamma$ at layer 15. **(b)** Steering with $\bar{\boldsymbol{h}}^r_D$ at layer 22. Both panels use the months prompt template with Llama-3.1-8B.

I.1 Sensitivity to steering strength

The experiments in the main text use $\alpha = 1$. To test whether the observed steering effects depend strongly on this choice, Figure 11 shows the median Δm_3 as a function of α at representative layers for $\bar{\boldsymbol{h}}^r_\gamma$ and $\bar{\boldsymbol{h}}^r_D$. In both cases, the steering effect varies smoothly with α.

I.2 Symmetrized pointwise steering

The averaged vectors above depend only on the relation label γ or D. As a complementary intervention, we construct prompt-specific counterfactual interaction vectors by averaging over the coordinate changes that produce the same shift in the corresponding relation.

For H^r_{AB}, both $B \to B + \delta$ and $A \to A - \delta$ change $\gamma \to \gamma + \delta$ and therefore $D \to D + \delta$. We define

$$\widetilde{\boldsymbol{h}}^r_{AB|\delta} = \frac{1}{2}\left(\boldsymbol{h}^r_{A,B+\delta} + \boldsymbol{h}^r_{A-\delta,B}\right). \tag{43}$$

Similarly, each of $A \to A - \delta$, $B \to B + \delta$, and $C \to C + \delta$ changes D to $D + \delta$, so we define

$$\widetilde{\boldsymbol{h}}^r_{ABC|\delta} = \frac{1}{3}\left(\boldsymbol{h}^r_{A-\delta,B,C} + \boldsymbol{h}^r_{A,B+\delta,C} + \boldsymbol{h}^r_{A,B,C+\delta}\right). \tag{44}$$

When $\delta = 0$, these expressions reduce to the original interaction vectors. We steer using

$$\widetilde{\boldsymbol{\phi}}^r(A,B,C) = \boldsymbol{\phi}^r(A,B,C) + \alpha\left(\widetilde{\boldsymbol{h}}^r_{S|\delta} - \boldsymbol{h}^r_S\right), \tag{45}$$

where $S \in \{AB, ABC\}$.

Figure 17 shows the normalized median Δm_3 obtained using this intervention. The same layerwise pattern appears as with the averaged steering vectors: steering using $\widetilde{\boldsymbol{h}}^r_{AB|\delta}$ is effective in intermediate layers, whereas steering using $\widetilde{\boldsymbol{h}}^r_{ABC|\delta}$ becomes effective at later layers.

J Cross-domain transfer

We report the cross-domain transfer experiments across models and domains. Figures 18 and 19 show top-1 and top-3 accuracy after transplanting H^r_{AB} from one domain into another following

ablation of the corresponding domain-1 interaction. Figures 20 and 21 show the corresponding replacement experiments, in which the domain-1 mean is retained together with the domain-2 interaction. Figure 22 shows cross-domain steering using $\bar{\boldsymbol{h}}_\gamma$, including $\delta = 0$ as a control.

Direct transplantation of H^r_{AB} generally recovers part of the performance lost under ablation, although the degree of recovery depends on the source and target domains. In particular, transfer between the months and hours domains is asymmetric for the Llama models: months→hours decreases performance relative to the unmodified model, whereas hours→months improves the performance (Figures 18 and 19). In contrast, direct transfer of H^r_{ABC} is substantially less consistent across domains (Figures 23 and 24).

J.1 Alignment of interaction subspaces

The inconsistent direct transfer of H^r_{ABC} may arise because different domains use different residual-stream subspaces for the same underlying interaction structure. We therefore align the interaction subspaces before transplantation. We perform the alignment in a $q = N - 1$ dimensional subspace and divide the replicates into a training set (70%), denoted by r_t, and a held-out test set (30%), denoted by r_e.

For domain $c \in \{1, 2\}$, let

$$W^{(c)}_{AB} = \left[\boldsymbol{h}^{r_t,(c)}_{AB}\right] \in \mathbb{R}^{tN^2 \times d_{\text{model}}}, \tag{46}$$

and similarly,

$$W^{(c)}_{ABC} = \left[\boldsymbol{h}^{r_t,(c)}_{ABC}\right] \in \mathbb{R}^{tN^3 \times d_{\text{model}}}. \tag{47}$$

Both $W^{(c)}_{AB}$ and $W^{(c)}_{ABC}$ have zero mean by Lemma 1.

We describe the alignment procedure for W_{AB}; the same procedure is applied to W_{ABC}. Consider transplanting the domain-2 interaction into domain-1. For each domain, we first compute

$$W^{(c)}_{AB} = U^{(c)}\Sigma^{(c)}\left(V^{(c)}\right)^T, \tag{48}$$

where the columns of $V^{(c)}$ are the right singular vectors. Let $V^{(c)}_q \in \mathbb{R}^{d_{\text{model}} \times q}$ contain the leading $q = N - 1$ right singular vectors. We project the interaction ensembles into these subspaces:

$$Z^{(c)}_{AB} = W^{(c)}_{AB} V^{(c)}_q \in \mathbb{R}^{tN^2 \times q}. \tag{49}$$

We then align the domain-2 coordinates to domain-1 by solving the orthogonal Procrustes problem

$$R^*_{AB} = \underset{R^T R = \boldsymbol{I}_q}{\arg\min} \left\| Z^{(2)}_{AB} R - Z^{(1)}_{AB} \right\|^2_F, \tag{50}$$

where $R^*_{AB} \in \mathbb{R}^{q \times q}$ is the orthogonal transformation that best aligns the two coordinate systems.

Transforming a domain-2 interaction vector into the domain-1 residual-stream basis then consists of projection onto the domain-2 subspace, alignment in the q-dimensional coordinate system, and reconstruction in the domain-1 basis. We therefore define

$$Q = V^{(2)}_q R^*_{AB} \left(V^{(1)}_q\right)^T \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}. \tag{51}$$

We evaluate the learned alignment only on held-out replicates. For the aligned ablation-transplant experiment, we use

$$\widetilde{\boldsymbol{\phi}}^{r_e,(1)}(A, B, C) = \boldsymbol{\phi}^{r_e,(1)}(A, B, C) - \boldsymbol{h}^{r_e,(1)}_{AB} + \boldsymbol{h}^{r_e,(2)}_{AB} Q. \tag{52}$$

For the corresponding aligned replacement experiment, we use

$$\widetilde{\boldsymbol{\phi}}^{r_e,(1)}(A, B, C) = \boldsymbol{\mu}^{r_e,(1)} + \boldsymbol{h}^{r_e,(2)}_{AB} Q. \tag{53}$$

Figures 25 and 26 show the aligned ablation-transplant results for both H^r_{AB} and H^r_{ABC}, while Figures 27 and 28 show the corresponding replacement experiments. Alignment improves cross-domain transfer for both interaction ensembles and produces a particularly large improvement for H^r_{ABC}, whose direct transfer is otherwise inconsistent across domains. This suggests that part of the failure of direct H^r_{ABC} transplantation can be explained by differences in the residual-stream subspaces used by different domains.

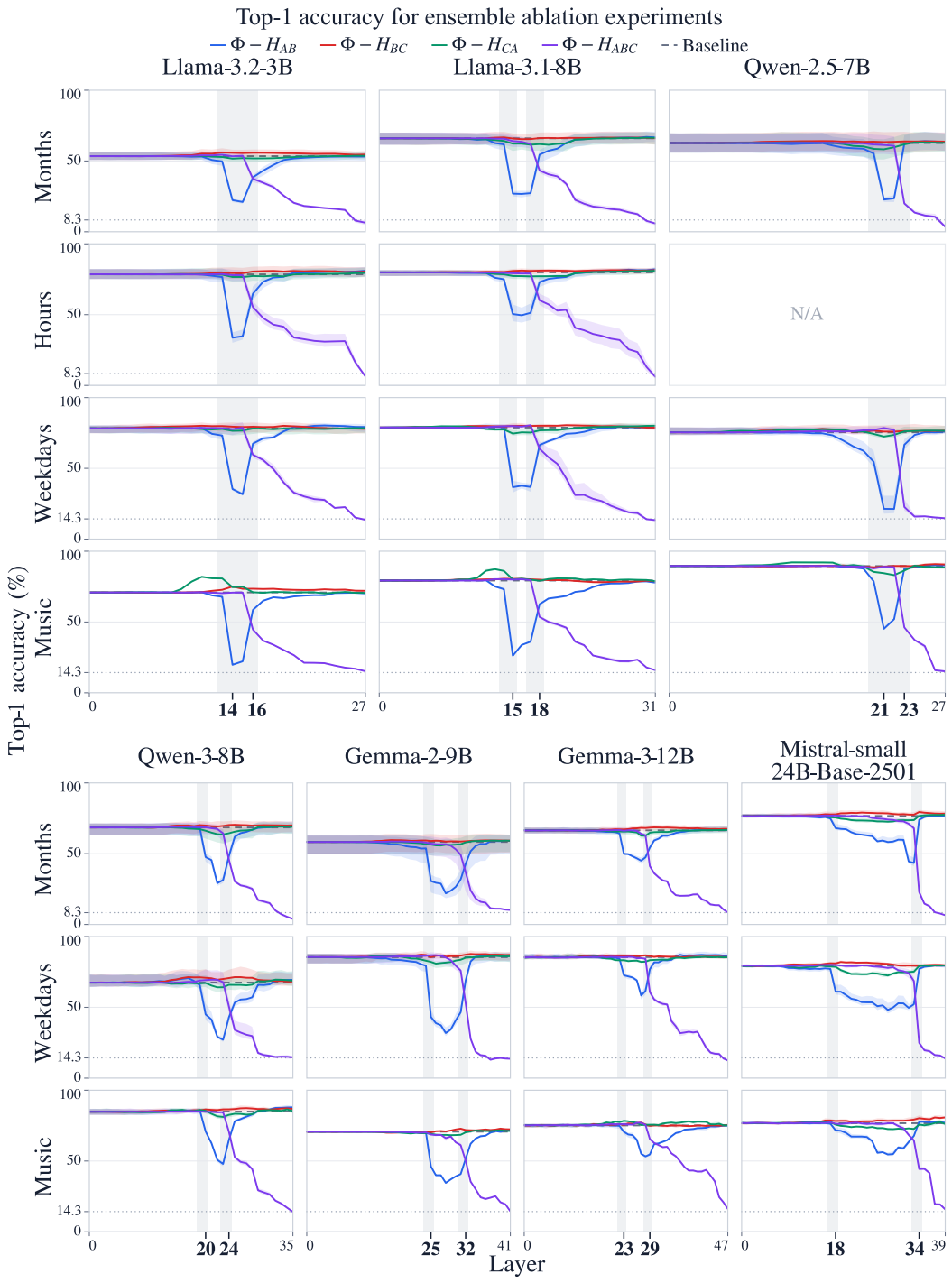

Figure 12: Top-1 accuracy after ablating an interaction ensemble at the last-token position, shown as a function of layer. The vertical shaded bands mark the model-specific intermediate layers where ablating H^r_{AB} produces its largest drop in performance. Across models and prompts, ablating H^r_{BC} or H^r_{CA} has little effect on accuracy, whereas ablating H^r_{AB} or H^r_{ABC} substantially degrades performance. The effect of ablating H^r_{AB} occurs earlier in depth than the effect of ablating H^r_{ABC}.

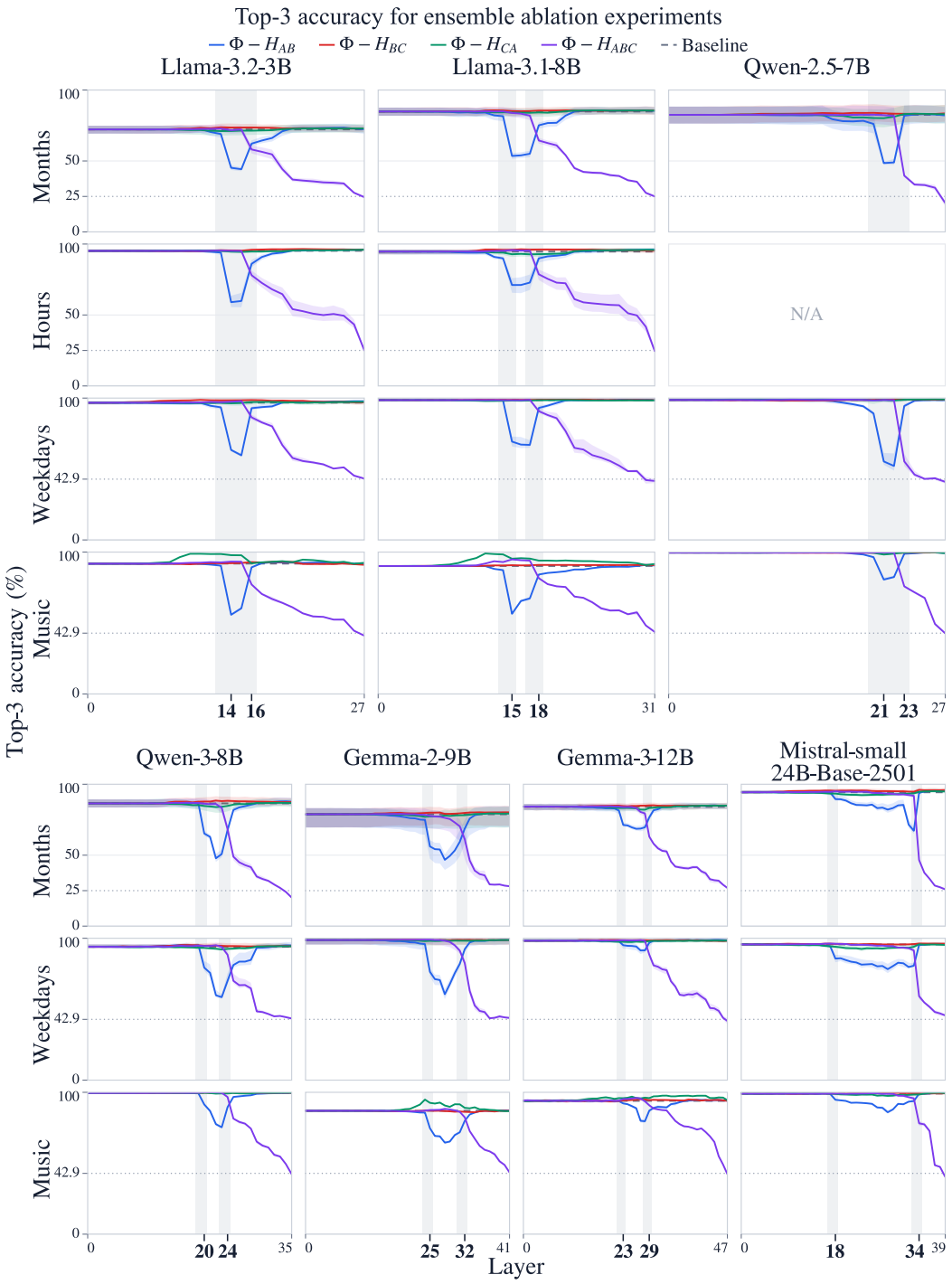

Figure 13: Top-3 accuracy after ablating each interaction ensemble at the last-token position, shown as a function of layer. Across models and prompt templates, ablating H^r_{BC} or H^r_{CA} has little effect on performance, whereas ablating H^r_{AB} or H^r_{ABC} substantially decreases accuracy. The effect of ablating H^r_{AB} consistently occurs earlier in depth than the effect of ablating H^r_{ABC}.

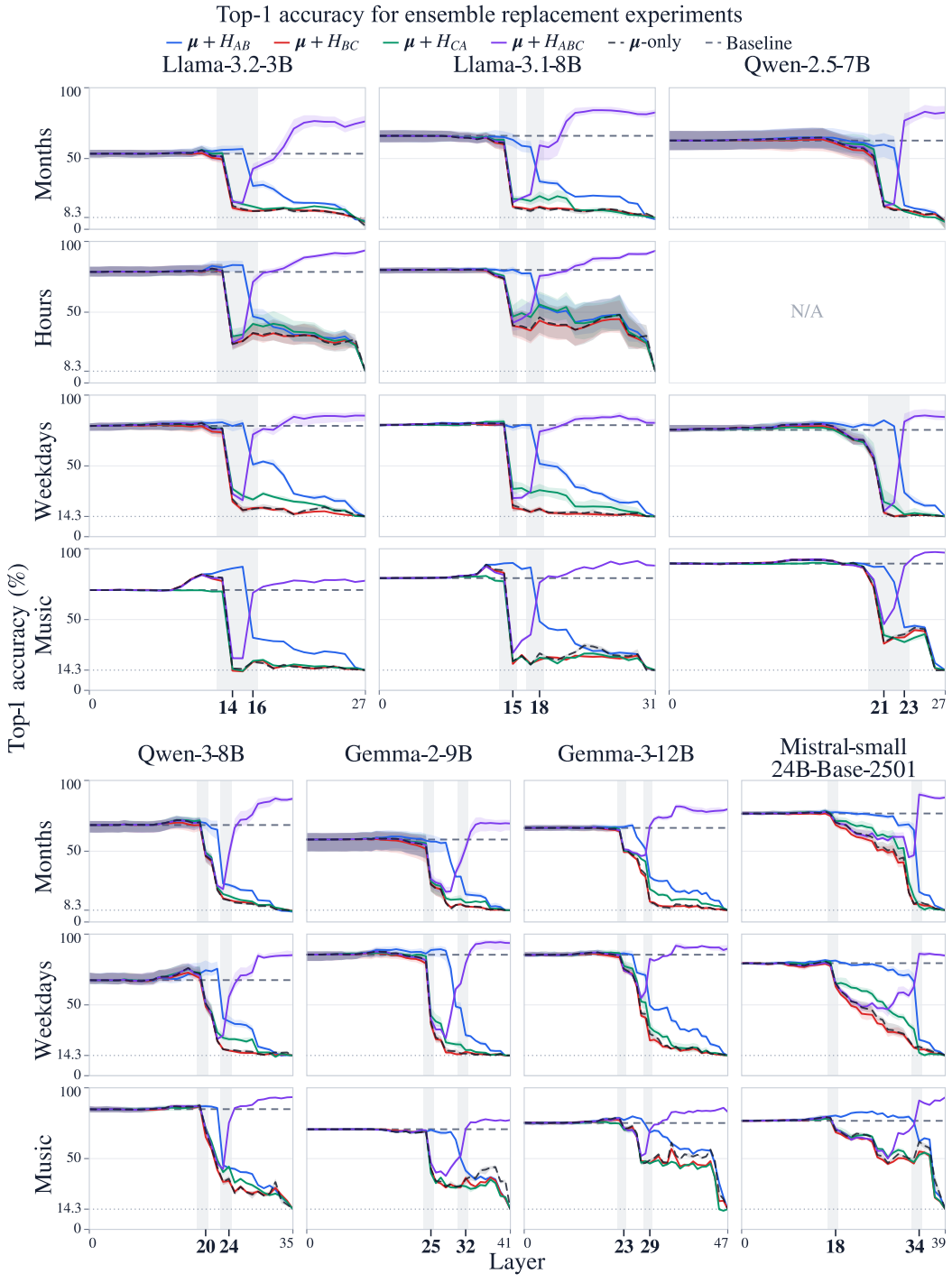

Figure 14: Top-1 accuracy after replacing the activation of the last token with only the mean $\boldsymbol{\mu}^r$ and an interaction ensemble, shown as a function of layer. The vertical shaded bands mark the intermediate layers where ablating H^r_{AB} produced the largest drop in performance, as shown in Figures 12 and 13. Replacing with H^r_{BC}, H^r_{CA}, or H^r_{ABC} substantially degrades performance at the first vertical band, whereas replacing with H^r_{AB} does not affect performance until the second vertical band. At later layers, retaining H^r_{ABC} preserves performance and can exceed the unmodified baseline.

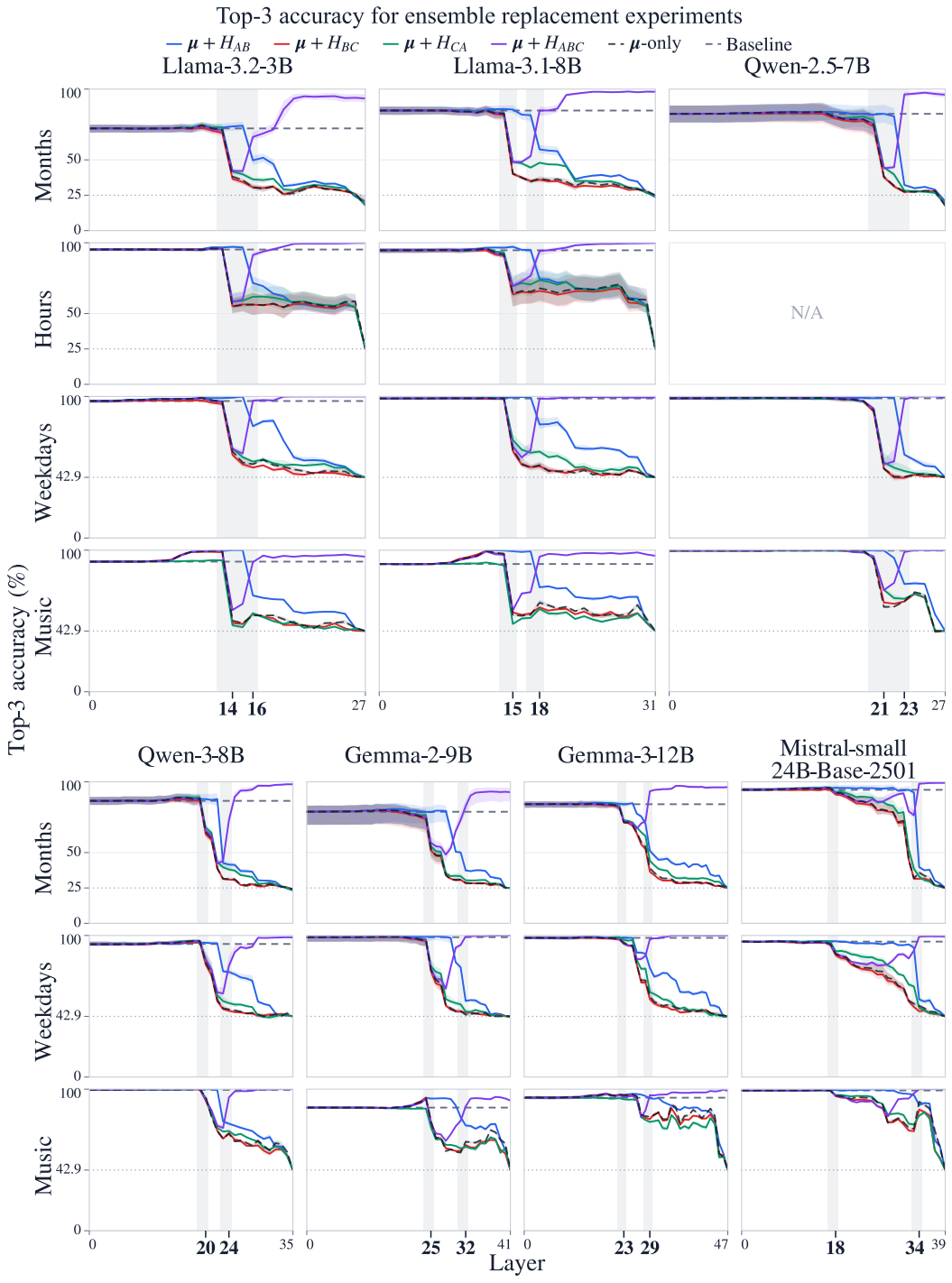

Figure 15: Top-3 accuracy after replacing the last-token activation with the ensemble mean $\boldsymbol{\mu}^r$ and one interaction term. Across models and prompt templates, retaining H^r_{AB} preserves performance in intermediate layers but becomes insufficient later in depth, while retaining H^r_{ABC} becomes effective at later layers and can outperform the unmodified baseline. Retaining H^r_{BC} or H^r_{CA} generally degrades the performance beginning from the first vertical band.

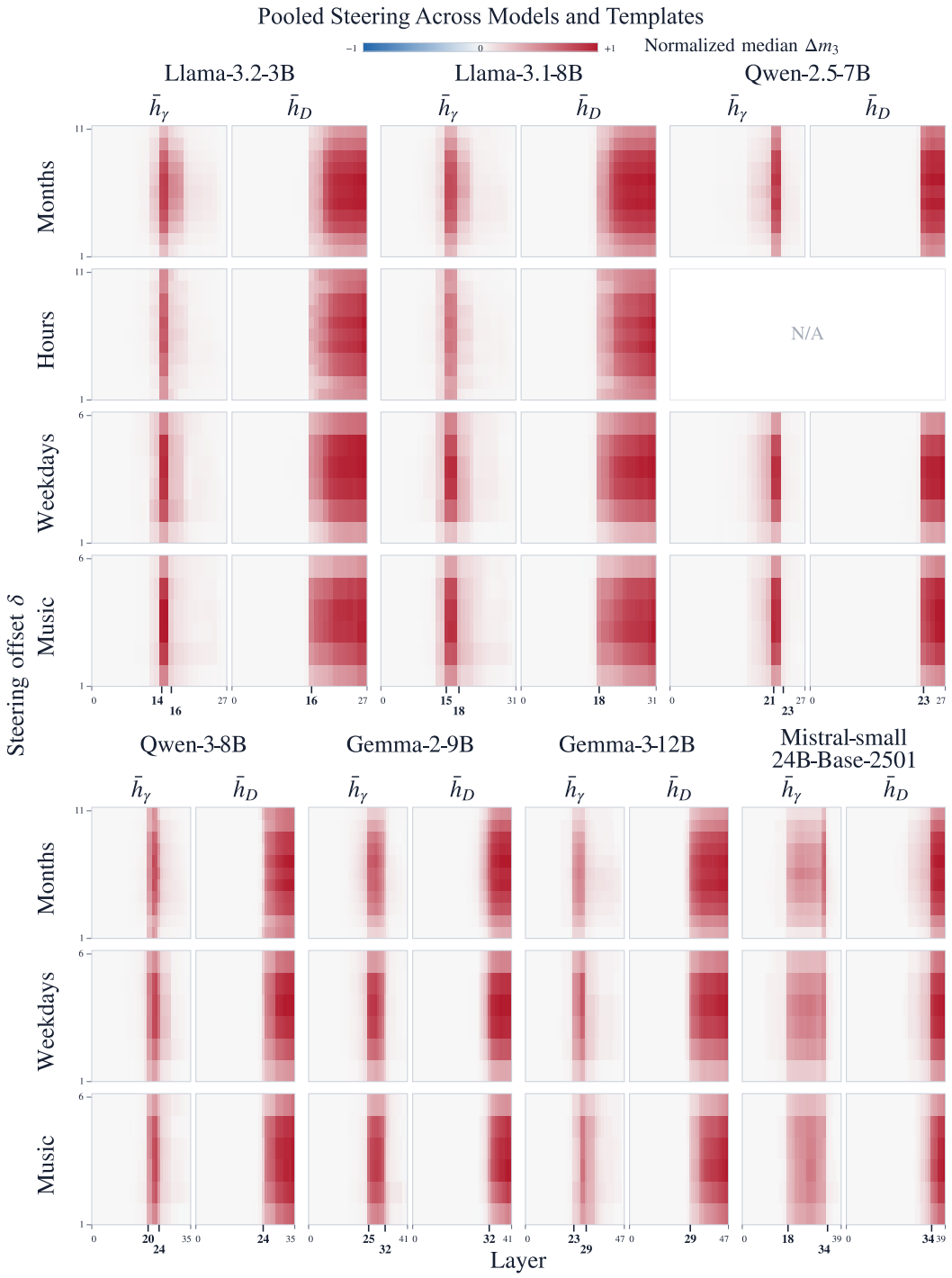

Figure 16: Heatmap of steering using $\bar{\boldsymbol{h}}_\gamma$ and $\bar{\boldsymbol{h}}_D$ across every model and prompt. Color indicates normalized median Δm_3. Steering with $\bar{\boldsymbol{h}}_\gamma$ is most effective in intermediate layers, with boldface layer labels marking the layers where H^r_{AB} was found to be causally relevant. At later layers, steering with $\bar{\boldsymbol{h}}_D$ becomes effective.

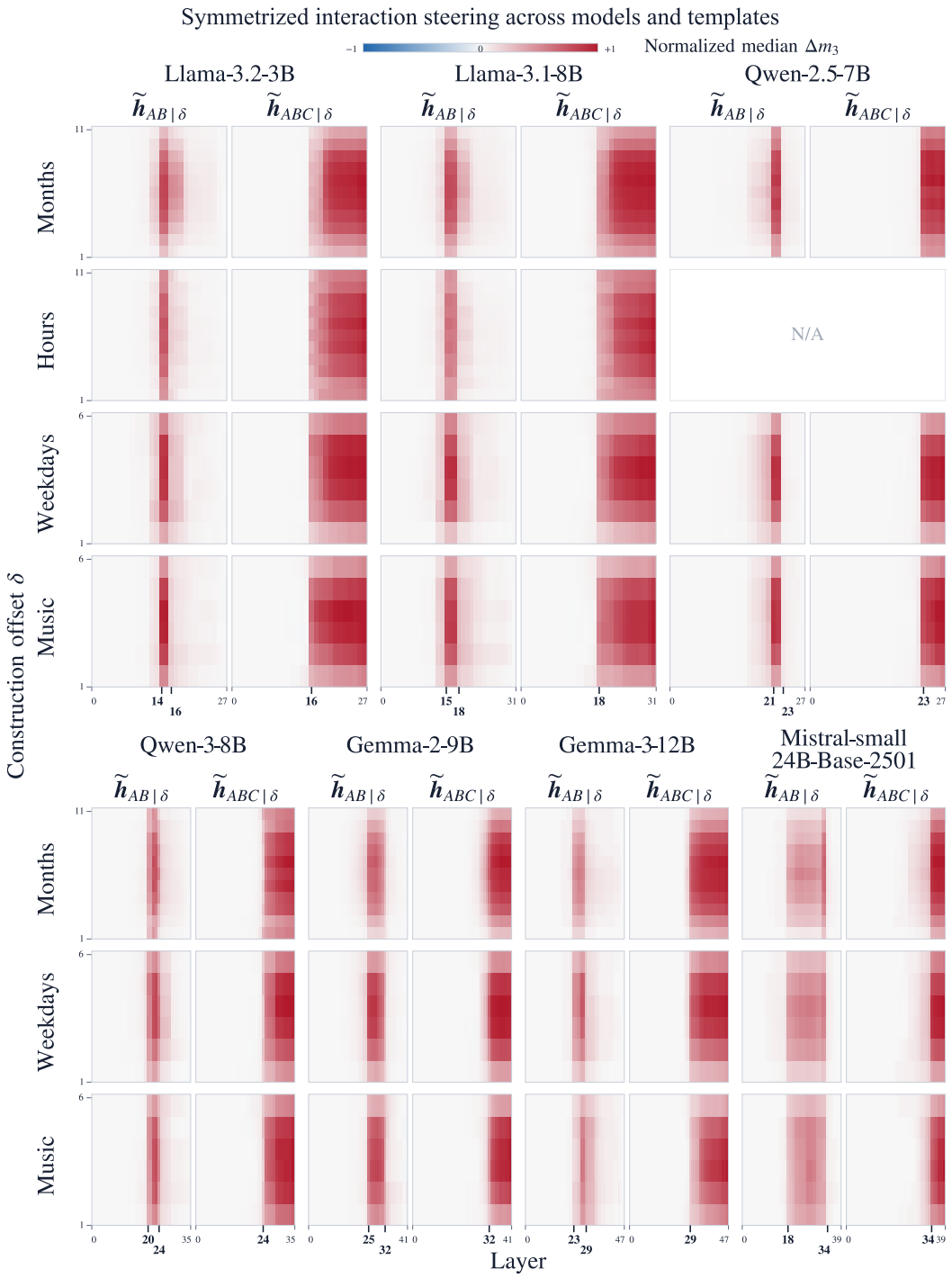

Figure 17: Heatmap of steering using $\widetilde{h}_{AB|\delta}$ and $\widetilde{h}_{ABC|\delta}$ across every model and prompt. Color indicates normalized median Δm_3. As with the steering results in Figure 16, steering using $\widetilde{h}_{AB|\delta}$ is effective in intermediate layers, while steering using $\widetilde{h}_{ABC|\delta}$ becomes effective at later layers.

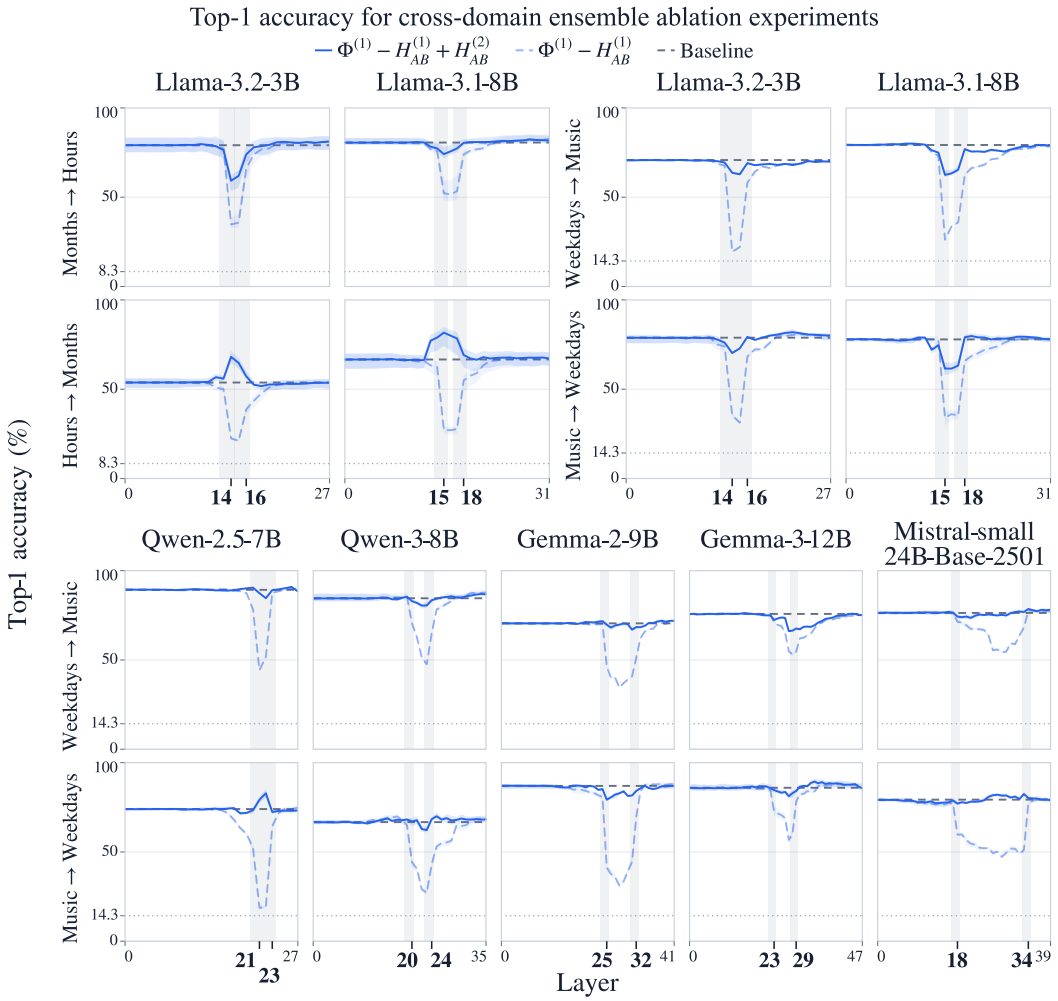

Figure 18: Top-1 accuracy for cross-domain ablation transplant of H^{r}_{AB}. Solid curves show performance after replacing an ablated domain-1 interaction vector with the corresponding domain-2 vector; dashed blue curves show ablation without transplantation and dashed black curves show the unmodified baseline. Successful transfer is indicated by recovery from the dashed blue ablation curve toward the baseline. Direct transplantation generally recovers performance, although the amount of recovery depends on the model and cross-domain pair.

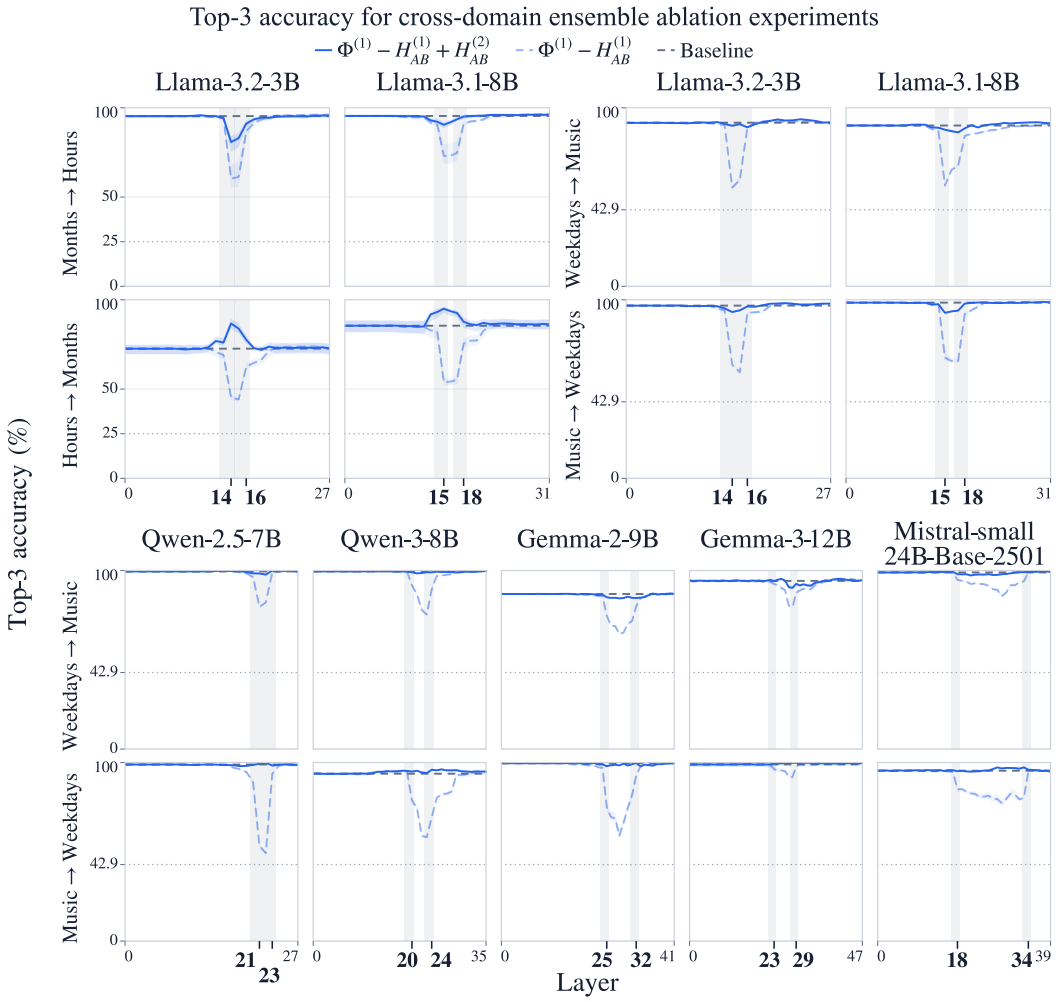

Figure 19: Top-3 accuracy for direct cross-domain ablation transplant of H^r_{AB}. As in Figure 18, transplantation generally recovers part of the performance lost under ablation, with variation across model and cross-domain pairs.

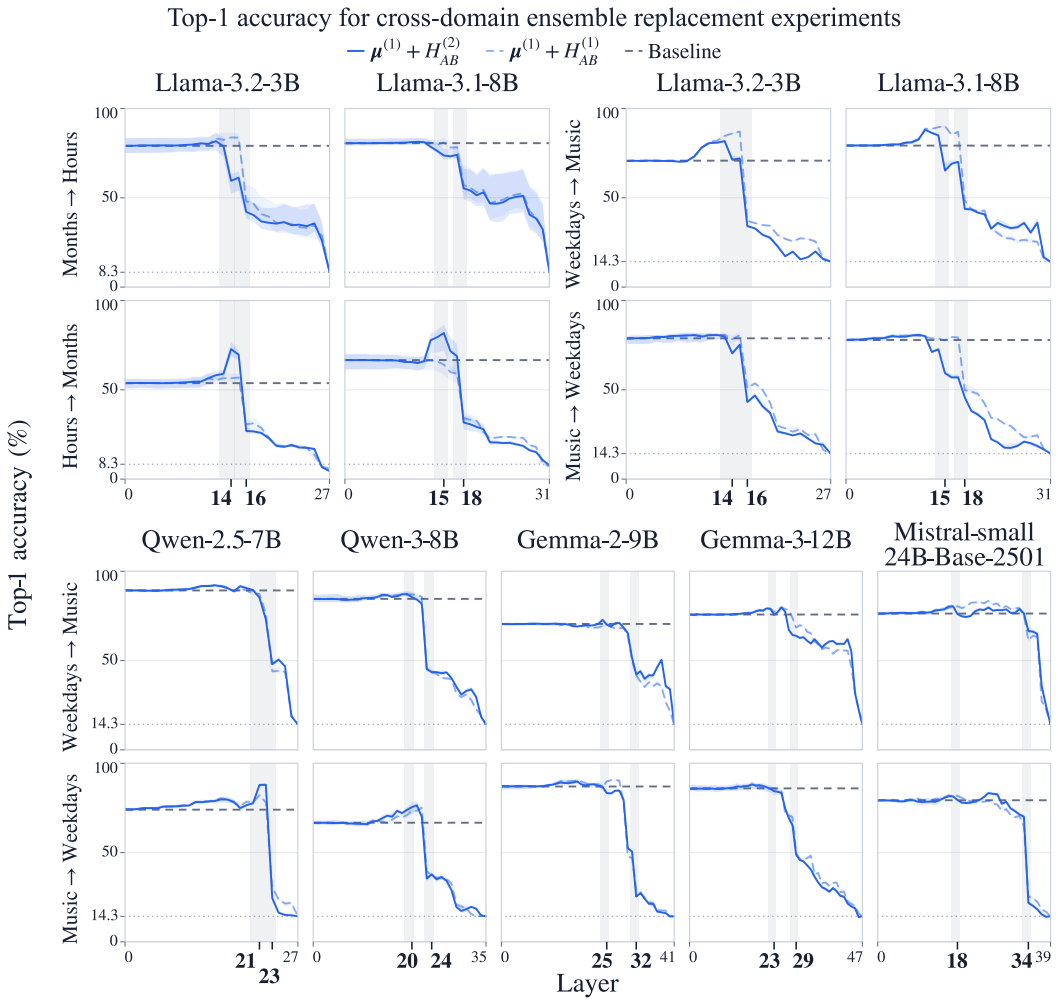

Figure 20: Top-1 accuracy for cross-domain replacement using H_{AB}^{τ}. Solid curves retain the domain-1 mean together with the domain-2 interaction vector; dashed blue curves show the corresponding within-domain replacement experiment. Solid and dashed trajectories follow each other, which indicates that the domain-2 interaction can substitute for the domain-1 interaction in downstream computation.

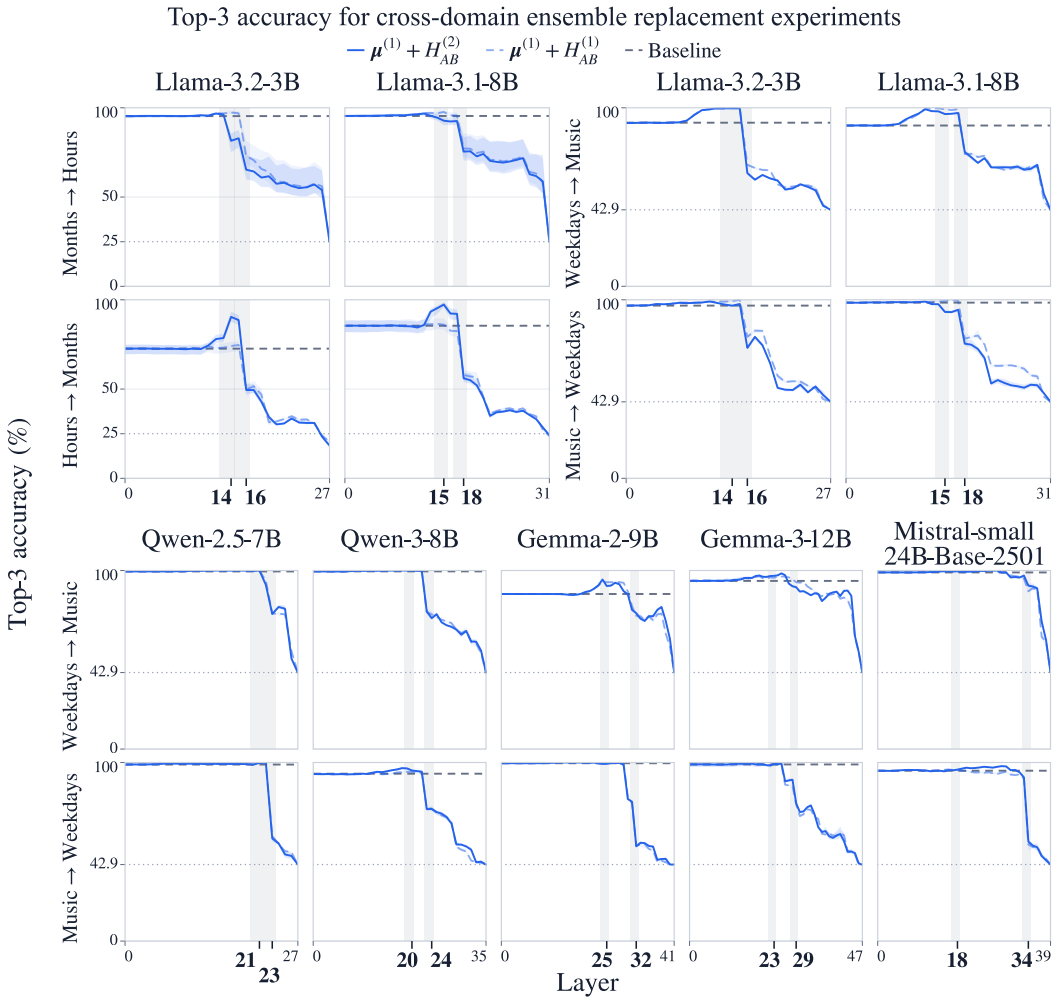

Figure 21: Top-3 accuracy for direct cross-domain replacement using H_{AB}^{τ}. Across all models and cross-domain pairs, the transplanted domain-2 interaction reproduces the layerwise behavior of the corresponding within-domain replacement.

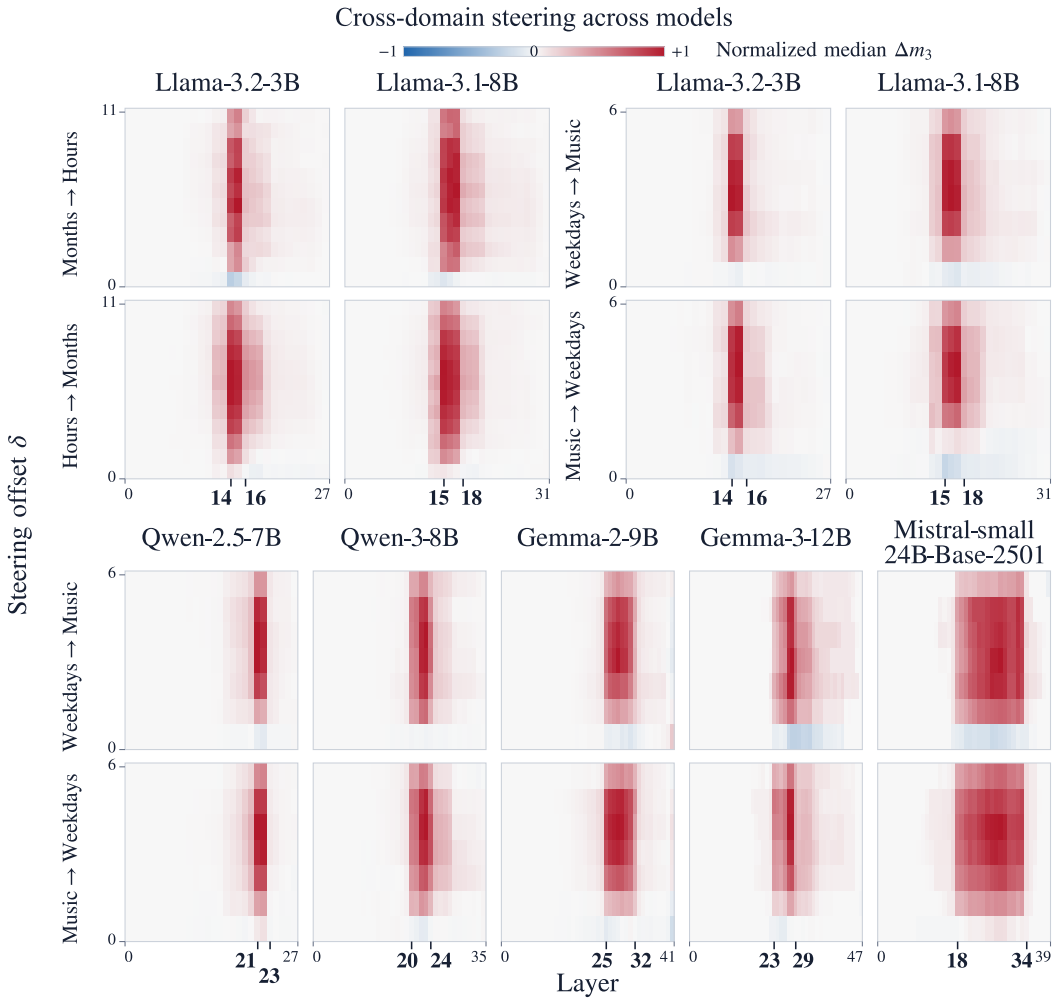

Figure 22: Cross-domain steering using vectors extracted from domain-2. Color indicates normalized median Δm_3. Steering is strongest in the same intermediate layers where H^r_{AB} is causally relevant. The $\delta = 0$ case provides a non-trivial control: replacing the domain-1 relation vector with the corresponding domain-2 vector should preserve the original answer if the relation representation transfers across domains.

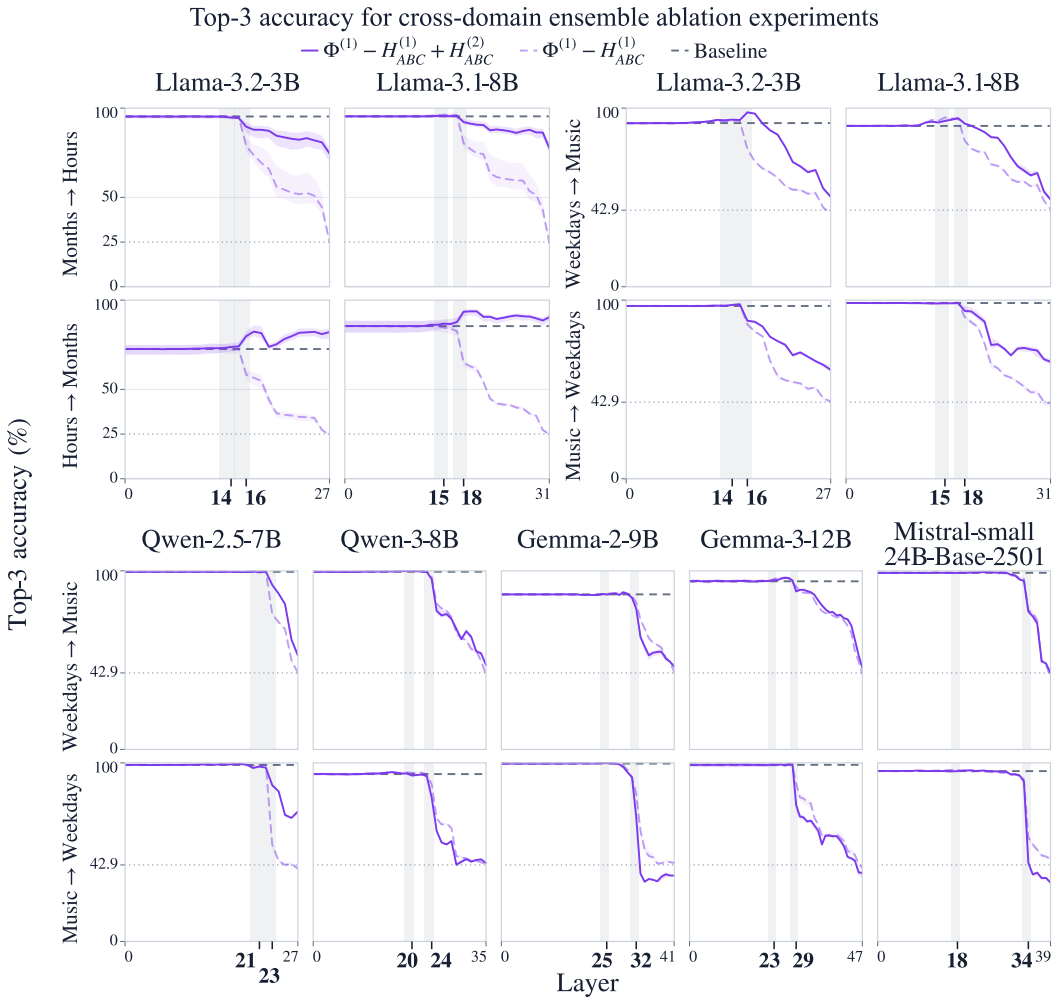

Figure 23: Top-3 accuracy for cross-domain ablation transplant of H^r_{ABC}. Unlike H^r_{AB}, direct transfer of the three-way interaction is not consistently successful across domains. Recovery in performance is observed for the months↔hours transfers in the Llama models, but most other model–domain pairs show little improvement over ablation.

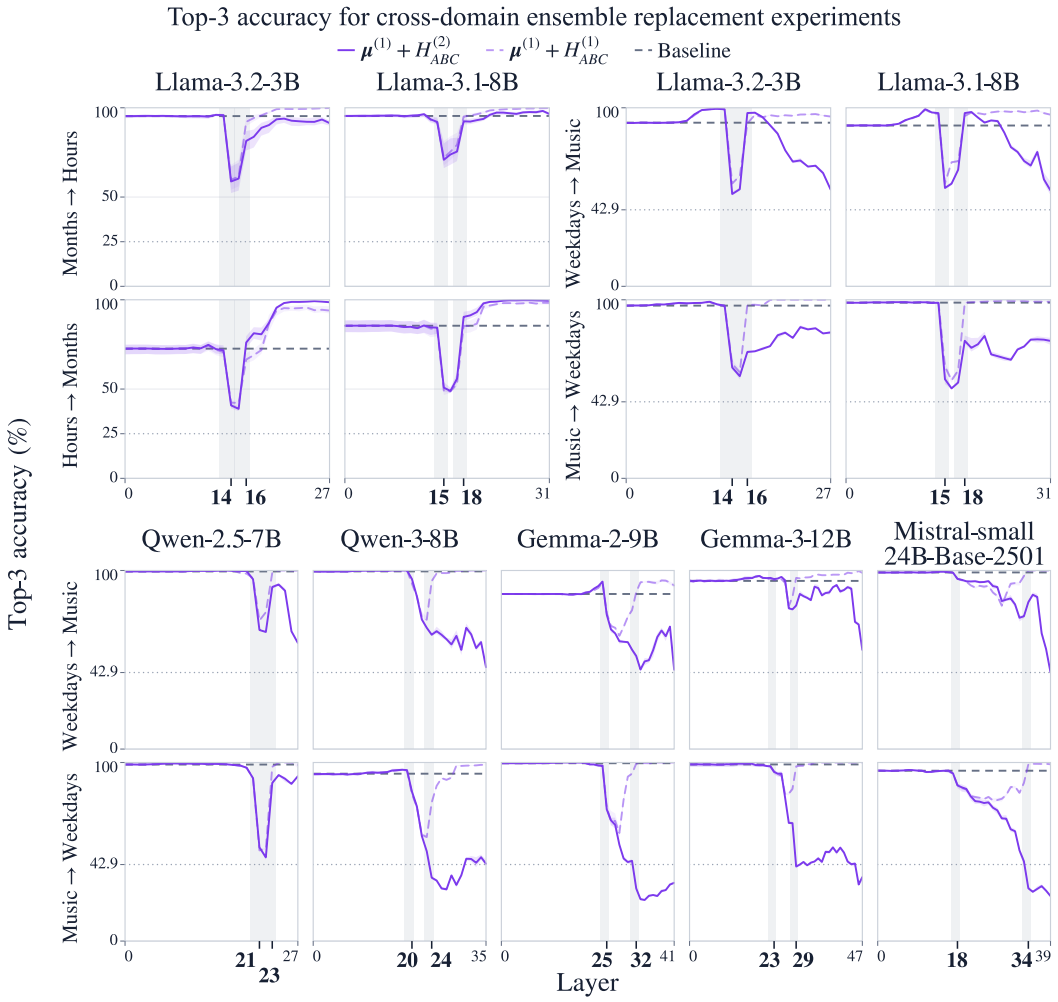

Figure 24: Top-3 accuracy for cross-domain replacement using H^r_{ABC}. Consistent with Figure 23, the ability of a domain-2 H^r_{ABC} to substitute for its domain-1 counterpart depends strongly on the model and domain pair.

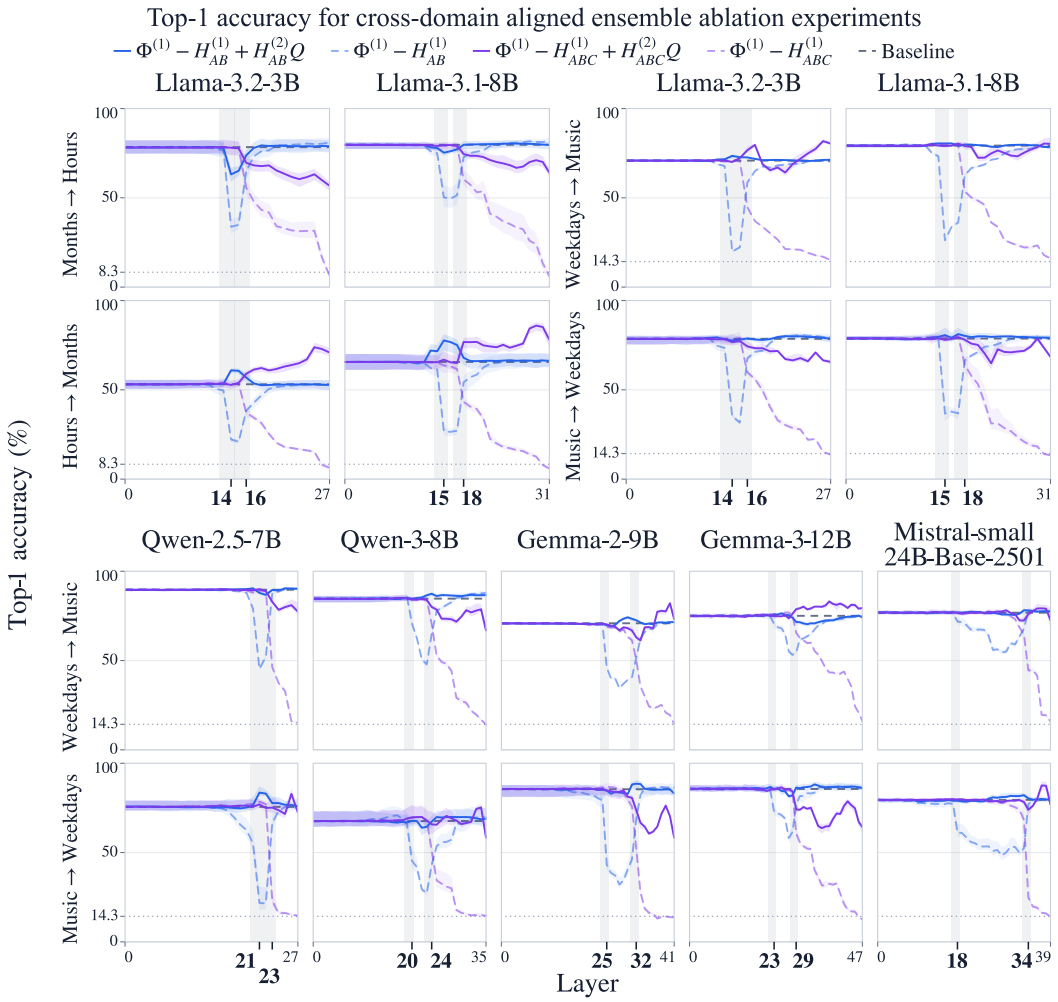

Figure 25: Top-1 accuracy for cross-domain ablation transplant after aligning the interaction subspaces using training replicates and evaluating on held-out replicates. Solid curves show aligned domain-2 interaction vectors transplanted into domain 1; dashed curves show the corresponding ablation and unmodified baselines. Alignment improves transfer for both H^r_{AB} and H^r_{ABC}, although the degree of recovery varies across models and domain pairs.

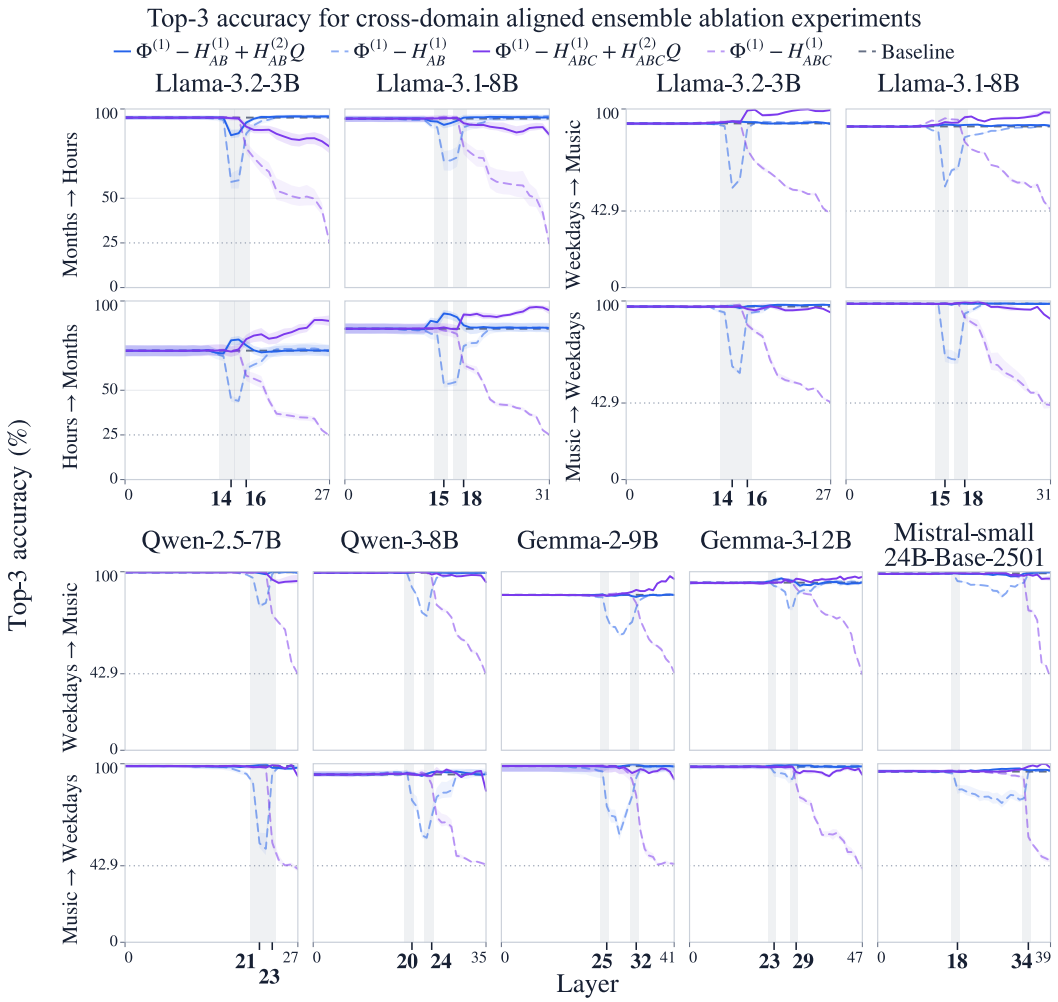

Figure 26: Top-3 accuracy for cross-domain ablation transplant after interaction-subspace alignment. Alignment improves recovery over the ablated baseline for both H^r_{AB} and H^r_{ABC}, with the largest qualitative change occurring for H^r_{ABC}, whose direct transfer was often weak.

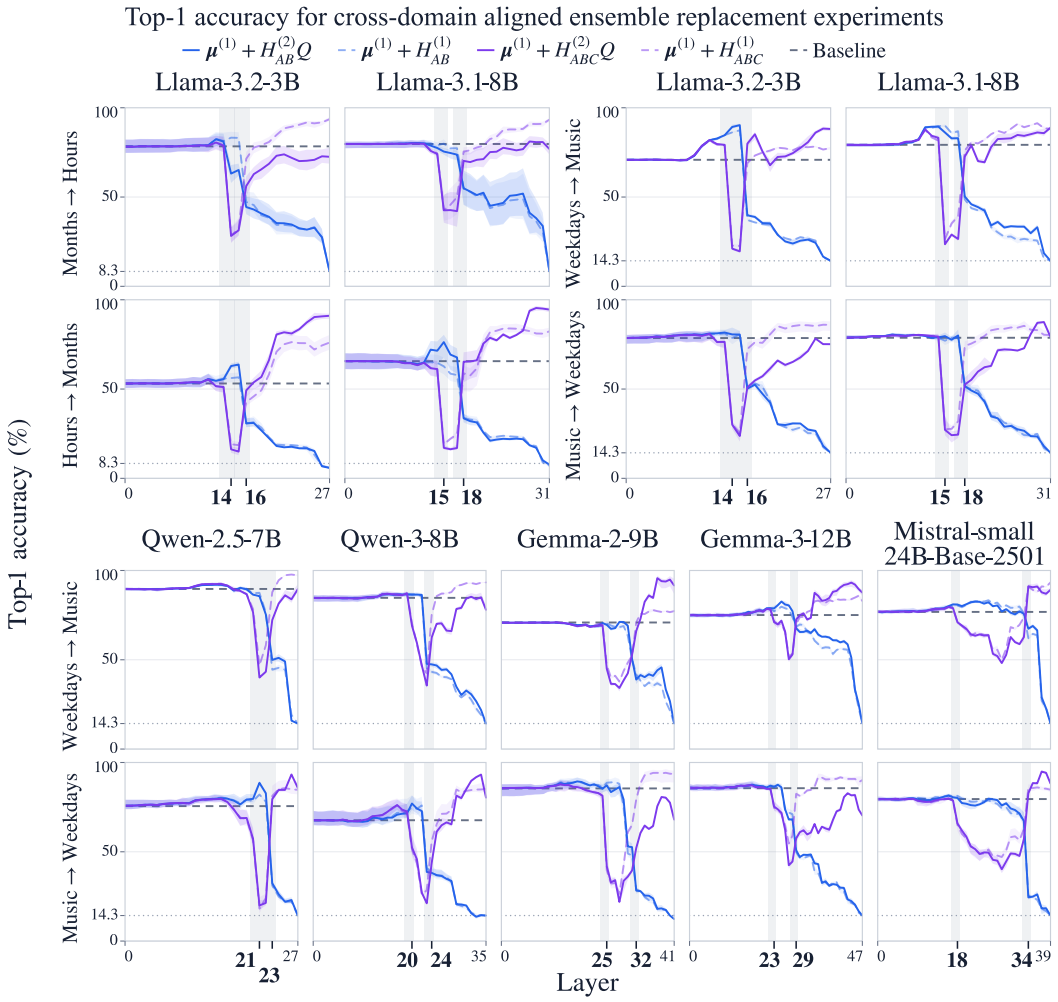

Figure 27: Top-1 accuracy for cross-domain replacement after aligning the interaction subspaces. Solid curves retain the domain-1 mean together with an aligned domain-2 interaction vector; dashed curves show the corresponding within-domain replacement. Alignment allows the transplanted interaction to reproduce the within-domain replacement trajectory more closely, particularly for H^{r}_{ABC}.

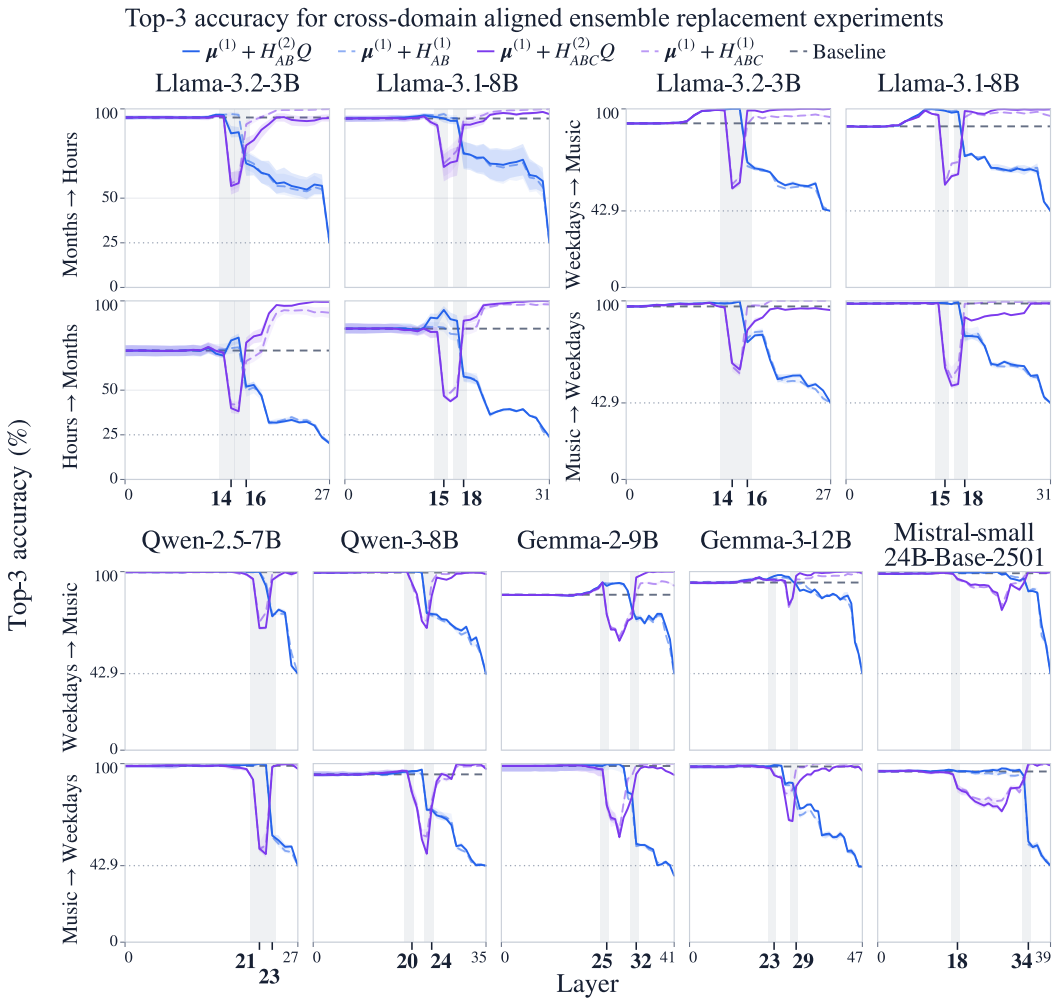

Figure 28: Top-3 accuracy for cross-domain replacement after interaction-subspace alignment. As in the top-1 results in Figure 27, alignment improves agreement between cross-domain and within-domain replacement, especially for H^{r}_{ABC}, although the degree of recovery depends on the model and domain.