

Features in Motion: Periodicity, Homeostasis, and Adaptivity in LLM Few-Shot Prompting

Ikhyun Cho Sewoong Lee Julia Hockenmaier

Siebel School of Computing and Data Science

The Grainger College of Engineering

University of Illinois Urbana-Champaign

{ihcho2, samuel127, juliahmr}@illinois.edu

Abstract

In this work, we characterize three novel representation-level dynamics shared across diverse large language models (LLMs) during few-shot prompting. Using sparse autoencoders (SAEs) as an analytical lens, we uncover the following phenomena: (1) *decimal periodicity effect*, in which feature activations exhibit a repeating modulo-10 pattern across exemplar positions; (2) *local feature homeostasis*, an internal balancing mechanism in which a feature’s activation in the current exemplar is inversely modulated by its activation in the preceding exemplar; and (3) *adaptive feature reorganization*, wherein specific features flexibly alter their semantic profiles in response to contextual changes. We show that these phenomena persist across multiple model families and tasks. Together, these findings offer fresh insights into the internal dynamics of in-context learning while highlighting crucial methodological caveats for future SAE-based interpretability studies.¹

1 Introduction

Representation engineering (RepE) has emerged as an increasingly important direction for understanding and controlling the internal computations of large language models (LLMs) (Zou et al., 2023; Li et al., 2023). While much existing work seeks to determine what individual representations encode (Bricken et al., 2023; Templeton et al., 2026), reliable interpretation also requires understanding how representations change dynamically as the model processes different instructions, demonstrations, and contextual constraints. These representation-level dynamics are crucial because they can reveal latent computations and prompt-processing regularities that are difficult to observe from model

outputs, yet may substantially influence the conclusions drawn from interpretability analyses and the effectiveness of representation-based interventions.

In this paper, we investigate such representation-level dynamics in the setting of few-shot prompting. Our goal is to identify latent patterns that have not yet been well documented, but nonetheless systematically shape model hidden states and recur across model families and scales. Understanding these dynamics is valuable not only for deepening our scientific understanding of LLM behavior, but also for revealing hidden biases and regularities that may affect practical interpretability studies.

To investigate these dynamics, we build on recent advances in mechanistic interpretability, using Sparse Autoencoders (SAEs) as a lens for analyzing internal LLM representations (Bricken et al., 2023; Templeton et al., 2026; Cho and Hockenmaier, 2025b). We focus on few-shot prompting, a canonical form of in-context learning, which has become a central paradigm for adapting LLMs to downstream tasks without parameter updates (Brown et al., 2020; Min et al., 2022; Cho et al., 2025c). Our analysis reveals three representation-level dynamics that consistently emerge as models process few-shot prompts:

1. **Decimal Periodicity Effect:** Models exhibit systematic periodic shifts in representation, tightly coupled to the modulo-10 position of few-shot exemplars. This behavior is driven primarily by exemplar position rather than semantic content, and can substantially confound SAE analyses by making a wide array of features highly sensitive to specific demonstration positions.
2. **Local Feature Homeostasis:** Models manifest a local balancing mechanism wherein the internal representation of a current exemplar is *inversely* modulated by the properties of the preceding exemplar.

¹Code available at https://github.com/ihcho2/Features_in_Motion.

3. **Adaptive Feature Reorganization:** We reveal that LLMs exhibit a systematic redistribution of features, where a wide range of features dynamically alter their activation profiles when the input context is selectively constrained.

We demonstrate that these patterns persist across a diverse suite of model families and scales, including Gemma 2 (2B, 9B), Qwen 3 (14B, 32B), and Llama 3.1 (8B), indicating that these representational dynamics are a shared property of modern LLMs. Overall, these findings highlight the rich latent structures underlying in-context learning, while raising important methodological caveats for future SAE-based research.

2 Preliminaries: Sparse Autoencoders

One of the most effective approaches for making the neural activations of large language models (LLMs) interpretable to humans is the sparse autoencoder (SAE) (Huben et al., 2024; Gao et al., 2025; Cho and Hockenmaier, 2026; Cho et al., 2025b). A major obstacle to understanding LLMs is that their internal representations are expressed as vectors that are not directly human-interpretable. Features disentangled by SAEs, however, are often regarded as analogous to a *microscope* that allows us to probe the internal mechanisms of large language models (Lindsey et al., 2025; Templeton et al., 2026).

Specifically, while individual neurons in LLMs tend to activate densely and participate in many unrelated computations, SAEs address this issue by learning representations in which only a small number of units activate for any given input. Thus, the intermediate neurons of an SAE, which activate sparsely, are referred to as *features*. Formally, let $\mathbf{x}_i^{(\ell)} \in \mathbb{R}^d$ denote the activation vector at layer ℓ of an LLM for the i -th data sample. A sparse autoencoder reconstructs the same activation:

$$\hat{\mathbf{x}}_i^{(\ell)} = W_{\text{dec}} \sigma(W_{\text{enc}} \mathbf{x}_i^{(\ell)}),$$

where W_{enc} and W_{dec} denote the encoder and decoder weight matrices, respectively. $\sigma(\cdot)$ is a non-linear function, and in this study, we use JumpReLU, as implemented in GemmaScope (Lieberum et al., 2024) and LlamaScope (He et al., 2024). In the loss function, the sparsity regularizer L_{sparsity} encourages only a small number of

features to be active for each input:

$$L_{\text{SAE}} = \frac{1}{N} \sum_{i=1}^N \left\| \mathbf{x}_i^{(\ell)} - \hat{\mathbf{x}}_i^{(\ell)} \right\|_2^2 + \lambda L_{\text{sparsity}}.$$

Prior work has identified neurons or SAE features that respond selectively to particular concepts, such as the Golden Gate Bridge or multilingual content (Templeton et al., 2026; Cho and Hockenmaier, 2025a; Deng et al., 2025). This line of research has also led to resources such as Neuronpedia (Lin and Bloom, 2023), which provide feature annotations—often generated using LLMs as judges—based on recurring activation patterns. While these resources are valuable, such annotations primarily characterize what a neuron or feature tends to correlate with, offering limited insight into how its activation changes dynamically as a model processes context. Our work addresses this gap by leveraging controlled few-shot prompting to identify consistent, recurring dynamics in LLM hidden states, thereby revealing a suite of shared internal representational behaviors.

3 Features in Motion: Hidden Dynamics of Few-Shot Prompting

Analysis Overview In this work, we identify three representation-level mechanisms that consistently recur across diverse models: (1) *decimal periodicity effect*, where models exhibit a repeating modulo-10 shift in their representations tied to exemplar position rather than semantic content; (2) *local feature homeostasis*, a balancing mechanism where the representation of a current exemplar is inversely modulated by the activations of the preceding exemplar; and (3) *adaptive feature reorganization*, where specific features significantly alter their activation profiles in response to constrained input contexts.

We conduct our experiments using three widely adopted LLM families: Gemma 2 (2B, 9B) (Lieberum et al., 2024), Qwen 3 (14B, 32B) (Yang et al., 2025), and Llama 3.1 (8B) (Grattafiori et al., 2024). These models were selected both for their strong open-weights performance and the availability of high-fidelity, pre-trained SAEs from the community. Unless otherwise specified, our analysis focuses on the SAEs trained on the middle layers of each model. For evaluation, we utilize three distinct prompt configurations: (1) text classification tasks using the AGNews (Zhang et al., 2015) dataset; (2) code generation and tool-use environments via

the Berkeley Function Calling Leaderboard (Patil et al., 2025); and (3) retrieval-augmented generation (RAG) question answering, utilizing the Natural Questions dataset (Kwiatkowski et al., 2019).

3.1 DECIMAL PERIODICITY EFFECT

3.1.1 Core Intuitions

Positional biases in language models have been extensively studied, most prominently through primacy and recency effects, where information appearing near the beginning or end of a prompt can disproportionately influence model behavior (Zhao et al., 2021; Lu et al., 2022; Liu et al., 2024). However, most prior work studies these effects through output-level analyses, such as accuracy and prediction bias. Comparatively less is known about how such positional biases manifest at the representation level. We therefore ask whether SAEs can provide a more fine-grained view into the model’s inherent positional structure.

We find that positional effects in LLMs are not limited to conventional primacy or recency patterns, but can also appear in a decimal-periodic form. In particular, we observe SAE features whose activations peak systematically at exemplar positions corresponding to multiples of ten, such as the 10th, 20th, 30th, 40th, and 50th exemplars. We refer to this phenomenon as the *decimal periodicity effect* (DPE). Importantly, this representational periodicity persists across a range of recent model families and model sizes, and appears in both base and instruction-tuned models, indicating a shared representational dynamic across recent LLMs.

3.1.2 Experimental Setup

We construct prompts using a standard format: a task instruction followed by a sequence of semantic blocks, such as few-shot exemplars in AGNews, tool definitions in BFCL, or retrieved documents in Natural Questions (Brown et al., 2020; Min et al., 2022). Example prompt templates are presented in Figure 6 of Appendix A. For each block, we probe the model at its boundary by extracting activations from the *final m tokens* of the block. These boundary-adjacent states are especially informative because, under causal attention, they can integrate the full preceding context within the semantic block while immediately preceding the transition to the next item. We set $m = 1$ for our primary analyses, and demonstrate in Appendix D that DPE persists robustly across alternative choices of m .

We construct 1,000 unique 50-shot prompts using exemplars randomly sampled from the training set. For each SAE feature, we compute its average activation over the designated boundary tokens to evaluate whether the resulting activation profile exhibits a periodic structure as a function of the block index. We quantify this structure by scanning a range of candidate periods and isolating the strongest, most stable periodic profiles (details in Appendix C). Given our large sample of randomized prompts ($N = 1,000$), persistent periodic activation profiles are unlikely to stem from semantic noise, indicating a genuine position-dependent artifact.

3.1.3 Results: Prevalence of Modulo-10 Features

Figure 1 summarizes our findings using Qwen3-14B Base and IT on AGNews. Each row in each heatmap corresponds to one selected SAE feature. Cell values are row-standardized mean activations: for feature f , the mean activation at block index k is centered and scaled by that feature’s mean and standard deviation across all block positions.

Representational Periodicity. We observe a pronounced form of representational periodicity within the model’s hidden states. As illustrated by the prominent vertical blue bands in the second and third columns of Figure 1, numerous SAE features exhibit sharp activation peaks at multiples of ten. Additionally, we frequently observe secondary harmonics emerging at multiples of five.

Depth-Dependent Prevalence. The abundance of these periodic features varies systematically as a function of network depth. The effect is highly pronounced in the mid-to-late layers, whereas it remains weak or nearly absent in the earlier ones.

Training Resilience. This pattern persists robustly across both base and instruction-tuned model variants. This invariance indicates that DPE is a mechanism crystallized during autoregressive pretraining that subsequent alignment techniques—such as supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF)—do not erase or override.

In Appendix D, we further demonstrate that DPE persists across diverse architectures and tasks—specifically, tool calling and retrieval-augmented generation QA—confirming that DPE is a shared, systematic characteristic of modern

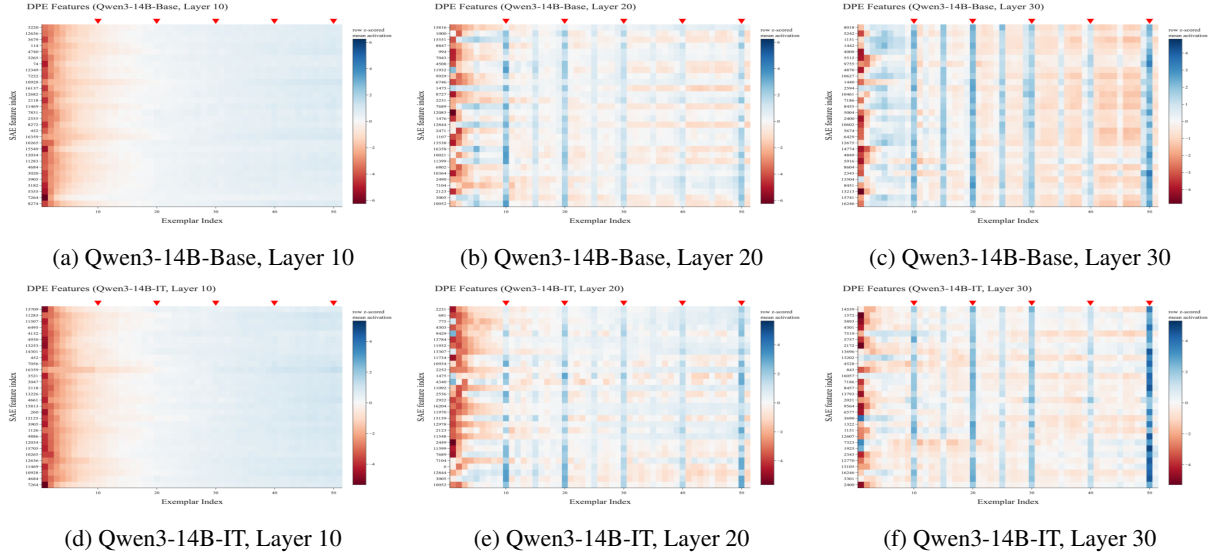


Figure 1: **The Decimal Periodicity Effect (DPE)**. Probing the hidden states at each exemplar boundary (the final token, $m = 1$, of exemplar k) reveals numerous features whose mean activations exhibit periodicities with respect to k . These features demonstrate a primary period of 10, with several displaying a secondary period of 5 (highlighted with red triangles). Across columns, the plots show that DPE is most heavily pronounced in mid-to-late layers; across rows, they demonstrate that the effect persists consistently across both base and instruction-tuned models.

autoregressive language models.

3.1.4 Practical Implications: Confounding Effects on Downstream SAE Analyses

We next show that the pervasiveness of DPE can confound standard feature-selection procedures in practical SAE analyses. Many SAE workflows prioritize features based on activation frequency or magnitude, treating frequently firing features as plausibly task-relevant (Kharlapenko et al., 2025; Deng et al., 2025). However, DPE can systematically distort this criterion. We show that some highly ranked features exhibit large activation-frequency gaps depending primarily on whether an exemplar appears at a modulo-10 position.

Table 1 lists the top-5 confounding features in Qwen3-14B-Base and Qwen3-14B-IT: for instance, feature Layer30-14512 activates on 88.33% of label-word positions at modulo-10 exemplars, but on only 25.82% of label-word positions at other exemplar positions. Consequently, an analyst studying a 10-shot prompt might identify this feature as highly task-relevant, whereas the same analysis on a prompt with a nearby non-modulo exemplar count (e.g., 12-shot) could suggest that the feature is much less relevant. This discrepancy arises despite the underlying task and prompt format remaining essentially unchanged. Thus, DPE introduces a substantial obstacle to robust feature interpretation: features may appear important because of

Model	SAE feature	Modulo-10 freq.	Other freq.	Gap
Qwen3-14B-Base	Layer30-14512	88.33%	25.82%	+62.51
Qwen3-14B-Base	Layer30-8457	71.07%	13.89%	+57.18
Qwen3-14B-Base	Layer30-8799	89.47%	35.04%	+54.42
Qwen3-14B-Base	Layer30-10919	59.67%	5.78%	+53.89
Qwen3-14B-Base	Layer30-10956	85.07%	33.38%	+51.68
Qwen3-14B-IT	Layer30-6966	98.73%	0.00%	+98.73
Qwen3-14B-IT	Layer30-375	97.93%	20.04%	+77.89
Qwen3-14B-IT	Layer30-6238	99.80%	27.73%	+72.07
Qwen3-14B-IT	Layer30-7096	58.20%	0.00%	+58.20
Qwen3-14B-IT	Layer30-8988	87.07%	34.32%	+52.75

Table 1: **Examples of DPE-confounded SAE features in Qwen3-14B-Base/IT**. Activation frequencies are measured at label-word positions for modulo-10 exemplar positions versus all other positions. Due to DPE, a wide range of features would appear highly active under a 10-shot-style prompt but substantially less active under nearby non-modulo prompt settings.

positional artifacts rather than genuine task semantics. This serves as an important methodological caveat for future SAE-based studies of in-context learning.

3.2 LOCAL FEATURE HOMEOSTASIS

We next introduce the second representational dynamic consistently observed across diverse LLMs: Local Feature Homeostasis (LFH).

3.2.1 Core Intuitions

While it is well-established that few-shot exemplars improve downstream task performance (Brown et al., 2020; Min et al., 2022), the mechanisms

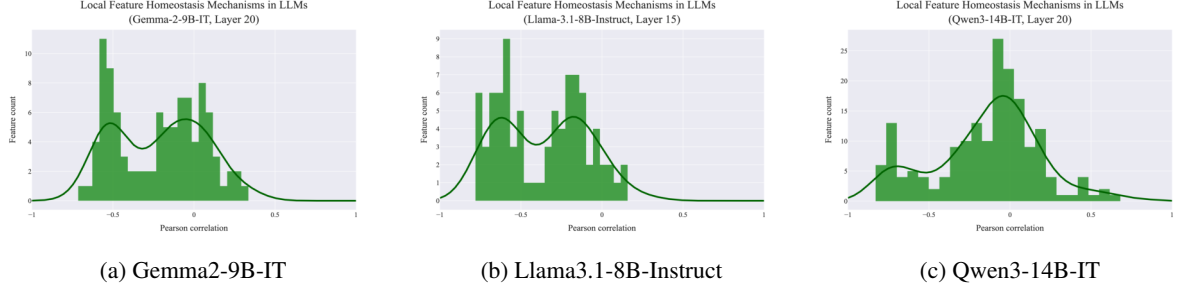


Figure 2: **Local Feature Homeostasis (LFH) Mechanism in LLMs.** Distribution of correlation values across all sufficiently active features. All models exhibit a bimodal-like distribution with a substantial left-shifted component, indicating a nontrivial population of features with inverse cross-exemplar modulation dynamics.

by which consecutive exemplars systematically interact at the representation level remain underexplored. We address this gap by examining the latent interactions between adjacent demonstrations.

Specifically, we identify an *inverse* cross-exemplar modulation mechanism: strong activation of a specific feature within a preceding exemplar systematically suppresses that same feature’s activation in the subsequent exemplar. Conversely, weak prior activation yields a compensatory, stronger subsequent activation. Figure 3 provides a representative example.

We demonstrate that this homeostatic, self-balancing feature-level pattern is highly prevalent across model families. To our knowledge, this provides the first feature-level evidence of inverse cross-exemplar modulation in LLMs, a mechanism we term *Local Feature Homeostasis* (LFH).

3.2.2 Experimental Setup

To evaluate this effect across a sufficiently deep context window, we adopt an 8-shot prompting regime. Specifically, we curate a set of $M = 50$ target exemplars. For each target exemplar, we construct $K = 50$ prompt variants by holding Examples 1–6 and the target Example 8 fixed while varying only the immediately preceding Example 7. This yields a total of $P = M \times K = 2,500$ prompt instances.

We first filter out features that fire too infrequently, using an average firing-frequency threshold of 5%. For each remaining feature f , we summarize its SAE activation on the preceding and target exemplars by averaging over the tokens within

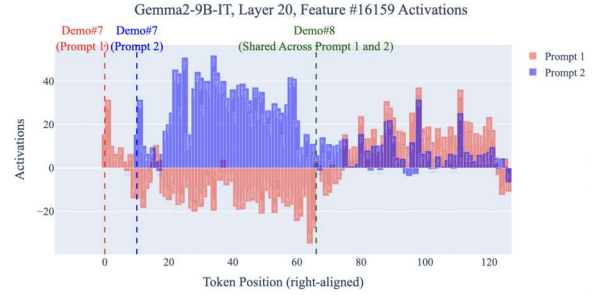


Figure 3: **Illustration of Local Feature Homeostasis (LFH).** Token-level activation trajectories are shown for a representative feature across the final two exemplars (Demos 7 and 8) of two 8-shot prompts. While Demo 7 differs between the two prompts, both share an identical target exemplar (Demo 8, right-aligned, onset marked in green). In Prompt 1 (red), the feature remains largely inactive during Demo 7, whereas in Prompt 2 (blue), it exhibits strong positive activation. Despite processing an identical target exemplar in Demo 8, the feature displays a flip: weaker prior activation elicits a stronger target response (red), while stronger prior activation leads to relative suppression (blue).

each exemplar:

$$A_{f,\text{prev}}^{(m,k)} := \frac{1}{|E_{\text{prev}}^{(m,k)}|} \sum_{t \in E_{\text{prev}}^{(m,k)}} z_f^{(m,k)}(t),$$

$$A_{f,\text{tgt}}^{(m,k)} := \frac{1}{|E_{\text{tgt}}^{(m,k)}|} \sum_{t \in E_{\text{tgt}}^{(m,k)}} z_f^{(m,k)}(t), \quad (1)$$

where $m \in \{1, \dots, M\}$ indexes the fixed target exemplars, $k \in \{1, \dots, K\}$ indexes the preceding-exemplar variants, and $z_f^{(m,k)}(t)$ denotes the token-level activation of feature f .

For each fixed target m , we compute the Pearson correlation across its K preceding-exemplar variants:

$$\rho_f^{(m)} := \text{corr}_k \left(A_{f,\text{prev}}^{(m,k)}, A_{f,\text{tgt}}^{(m,k)} \right).$$

We then define the overall LFH score by averaging these query-wise correlations:

$$\rho_f := \frac{1}{M} \sum_{m=1}^M \rho_f^{(m)}. \quad (2)$$

Intuitively, if feature f exhibits no cross-exemplar interaction, then varying the previous exemplar should exert no systematic effect on its response to the target exemplar. A more natural expectation might even assume a positive correlation: because the target exemplar is held constant across all K variants, one might expect strong activation on the previous exemplar to simply carry over. By contrast, a strongly negative ρ_f indicates the anti-relationship pattern: high activation on the previous exemplar is associated with relative *suppression* on the target exemplar, whereas weak prior activation is associated with relative *sensitization*. In this sense, a large negative ρ_f serves as a statistical signature of the LFH mechanism.

3.2.3 Results: A Substantial Population of Negatively Correlated Features

The resulting distribution of ρ_f is illustrated in Figure 2 (with results for additional models and tasks detailed in Appendix E). The observed pattern is striking: the distribution is markedly left-shifted and frequently bimodal, featuring one mode centered near zero (corresponding to a null or weak-effect baseline) and a second mode concentrated heavily in the strongly negative regime ($\rho_f < -0.5$).

This consistent bimodal distribution indicates that these inverse interactions are not rare exceptions, but a systematic property of internal model dynamics. Specifically, modern LLMs maintain a non-trivial population of features exhibiting homeostatic balancing behavior: activation is attenuated when the feature was strongly activated in the preceding exemplar, and amplified when prior activation was weak. The persistence of this pattern across all three model families confirms that it reflects a fundamental, shared representational dynamic rather than an architecture-specific artifact.

3.2.4 Discussion and Practical Implications

(1) Underlying Geometric Structure of LFH Features A natural follow-up question is: “Why do LFH features exhibit this inverse relationship?” We find that these features share a distinctive geometric structure. Crucially, LFH features are not

merely statistical anomalies; rather, their behavior is associated with a geometric signature embedded directly in the model’s native attention weights:

Geometric Signature of LFH Features

$$\text{LFH-OV}(f) := \min_h \cos(W_V^h d_f, (W_O^h)^\top e_f) \approx -1$$

To uncover this structure, we examine the alignment between SAE feature directions and attention heads through the QK/OV circuit framework (Elhage et al., 2021). Let d_f and e_f denote the SAE decoder and encoder directions of feature f , respectively. For each feature, we use a weight-driven geometric diagnostic that compares its value-space image, $W_V^h d_f$, with the head-space direction $(W_O^h)^\top e_f$, whose output projection would increase the feature’s activation. Their inner product satisfies

$$\langle W_V^h d_f, (W_O^h)^\top e_f \rangle = e_f^\top W_O^h W_V^h d_f.$$

Thus, a strongly negative cosine indicates that a residual-stream component along d_f is transformed by the head’s OV circuit into an update that projects negatively onto the feature’s encoder direction e_f , thereby suppressing its subsequent activation. This geometry provides a concrete architectural basis for the homeostasis-like behavior captured by LFH.

Table 2 reports this LFH-OV signature for the top-10 LFH features in Gemma2-2B-IT. For each feature, we identify the attention head that minimizes the signature and report the feature’s ρ_f , the corresponding signature value, and its percentile relative to all SAE features. The results reveal a strikingly consistent geometric pattern. Nine of the ten features have LFH-OV signatures between -0.959 and -0.866 , placing them above the 95th percentile among all SAE features. The median percentile is 0.998, indicating that the top LFH features exhibit some of the most extreme anti-write alignments in the entire SAE dictionary. These findings suggest that LFH features are not arbitrary residual-stream directions; rather, they are preferentially aligned with attention-head OV geometry well suited to implementing balancing, homeostasis-like updates.

(2) Causal Validation We next test whether the candidate attention heads identified above causally

Feature	ρ_f	LFH-OV signature	Percentile
L12-2245	-0.805	-0.900	0.994
L12-8745	-0.794	-0.895	0.994
L12-2992	-0.793	-0.946	0.999
L12-9632	-0.781	-0.937	0.999
L12-2863	-0.778	-0.959	1.000
L12-3117	-0.775	-0.866	0.986
L12-4758	-0.755	-0.661	0.802
L12-5160	-0.739	-0.957	1.000
L12-5185	-0.737	-0.954	1.000
L12-9059	-0.736	-0.914	0.997

Table 2: **Weight-driven LFH-OV alignment of the top LFH features in Gemma2-2B-IT.** For each feature f , we identify the attention head h that minimizes the LFH-OV signature, $\cos(W_V^h d_f, (W_O^h)^\top e_f)$, where d_f and e_f denote the feature’s SAE decoder and encoder directions, respectively. More negative values indicate stronger anti-write geometry: a residual-stream component along d_f induces an update with a negative projection onto e_f , thereby suppressing the feature’s subsequent preactivation. Percentiles are computed relative to all SAE features in the dictionary using each feature’s most negative head, with higher percentiles indicating more extreme anti-write alignment.

mediate the LFH effect. For the top-3 LFH features and their corresponding candidate head, we intervene on the head’s output contribution at the target-exemplar tokens and sweep the intervention strength over

$$\alpha \in \{-5, -2, -1, 0, 1, 2, 5\}.$$

The intervention scales the original head output by $(1 + \alpha)$: $\alpha = 0$ is the unmodified baseline, $\alpha = -1$ fully ablates the head, $\alpha > 0$ amplifies its contribution, and $\alpha < -1$ reverses and amplifies it. For each intervention strength, we then re-measure the correlation between the feature’s activation on the source exemplar and its activation on the target.

As shown in Figure 4, intervening on the targeted attention head produces the predicted monotonic trend: attenuating or ablating the head weakens the LFH correlation, whereas amplifying its output accentuates the effect, yielding a progressively stronger negative correlation. By contrast, analogous interventions on MLP outputs produce negligible shifts (additional model evaluations are provided in Appendix E).

This provides direct evidence that attention heads causally implement LFH rather than passively tracking it. Together, these findings establish that LFH is primarily an attention-mediated mechanism rather than an MLP-driven phenomenon,

confirming that it represents a highly structured internal dynamic in LLMs rather than a statistical artifact.

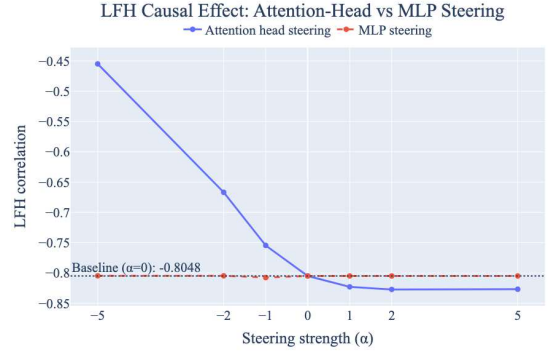


Figure 4: **Attention heads (blue) causally mediate LFH, whereas MLP interventions (red) remain virtually flat and exert negligible effects.** Results are shown for the top-ranked LFH feature L12-2245 in Gemma2-2B-IT.

(3) Confounding Effects on Downstream SAE Analyses. Just as DPE showed that SAE feature activations can be strongly shaped by demonstration position, LFH reveals a second source of prompt-induced confounding: the feature content of the immediately preceding exemplar. Since a substantial fraction of SAE features exhibit negative ρ_f , their activations on the test-query segment can be systematically shifted by the final demonstration in the prompt. In particular, if a fixed few-shot prefix elicits unusually high activation of an LFH-sensitive feature in its last demonstration, the same feature may be suppressed in the subsequent test-query representation; conversely, unusually weak activation in the final demonstration often induces compensatory overactivation on the query. An SAE analysis conducted under only a single prefix may therefore mistake a local homeostatic artifact for a genuine task- or example-level property of the query representation.

This suggests an important methodological precaution for downstream SAE studies: query-side feature activations should be interpreted in light of the immediately preceding context. In practice, this can be done by evaluating features across multiple prompt prefixes, varying the final demonstration, or explicitly measuring the relationship between final-demonstration and query-side activations. Accounting for LFH in this way can substantially stabilize feature interpretation and reduce the risk of attributing prompt-induced activation shifts to

task semantics.

3.3 ADAPTIVE FEATURE REORGANIZATION

3.3.1 Core Intuitions

What happens to a model’s internal representations when a structural constraint is introduced? We find that models contain a nontrivial set of features that exhibit flexible reorganization: when a prompt artificially restricts the available label space—for instance, by omitting one class from a standard N -class classification task—features strongly aligned with the missing class do not simply become inactive. Instead, their activation patterns are redirected toward the remaining viable labels, dynamically reconfiguring their semantic profiles to accommodate the newly imposed contextual boundaries.

3.3.2 Experimental Setup

To test this representational adaptability, we compare feature activations under the original classification prompt and a controlled prompt in which a target label l is removed. We define a feature f as label-specific to l if its average activation on label- l examples exceeds that on every other label:

$$\Delta_l(f) := \bar{s}_{f,l} - \max_{c \neq l} \bar{s}_{f,c} > 0, \quad (3)$$

where $\bar{s}_{f,c}$ denotes the mean activation of feature f across examples with class label c . Following our DPE experimental setup, we evaluate feature activations at the final prompt token. We retain the top 500 features ranked by $\Delta_l(f)$, subject to a minimum firing frequency threshold of 5%, thereby focusing on features that are both frequently active and task-relevant.

We then remove label l from the few-shot context by constructing prompts whose exemplars are drawn only from the remaining labels. Next, for each label l , we measure how the average activation of label- l -specific features changes on the remaining labels after label l is excluded from the prompt. We visualize these shifts using volcano-style plots: the x-axis shows the matched-pairs rank-biserial correlation, the y-axis shows the two-sided Wilcoxon p-value on an inverted log scale. Positive x-values indicate predominantly positive paired activation differences, while negative values indicate the reverse. This provides a nonparametric summary of both directional effect size and evidence for systematic paired shifts; further details are provided in Appendix B.

3.3.3 Results: Abundance of AFR Features

Figure 5a shows the result for Gemma2-2B-IT, using World as a representative excluded label. Results for other labels and models are provided in Appendix F. We observe a pronounced rightward and upward shift, indicating a substantial population of features whose activation on alternative labels increases once their primary semantic context is removed. We term this dynamic representational behavior, the *adaptive feature reorganization* (AFR). The consistency of this pattern across labels and models suggests that AFR is a shared representational dynamic across model families and scales.

3.3.4 Discussions and Practical Implications

(1) Causal Effects Are Also Reorganized We next test whether features exhibiting strong activation-level AFR also reorganize their causal effects. For each label l , we retain features whose original mean activation is highest on l and whose firing frequency exceeds the specified threshold (we use 5% as the default threshold). We then rank these features by the matched-pairs rank-biserial effect of their controlled-minus-original activation shift on the retained labels and select the top 10 causal features, denoted $\mathcal{F}_l$.

For each selected feature f , we steer the model by adding a fixed multiple of its SAE decoder direction at the selected test-query tokens:

$$h_t \leftarrow h_t + \alpha W_f^{\text{dec}}.$$

We quantify the resulting change in target label probability across 400 original prompts (100 examples per label) and 300 matched prompts in which label l is removed. Specifically, let

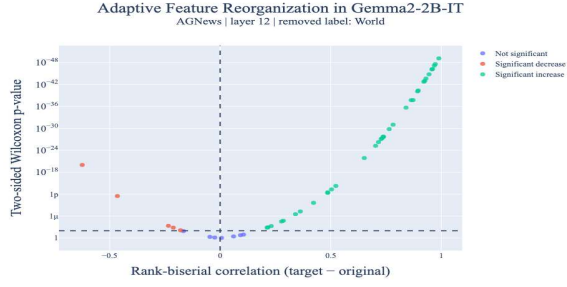
$$\text{CE}_{f \rightarrow c}^u(\alpha) = \frac{1}{|\mathcal{D}_l^u|} \sum_{x \in \mathcal{D}_l^u} [P_{\alpha, f}^u(c | x) - P_0^u(c | x)]$$

where $u \in \{\text{orig}, \text{ctrl}(-l)\}$. Figure 5b reports

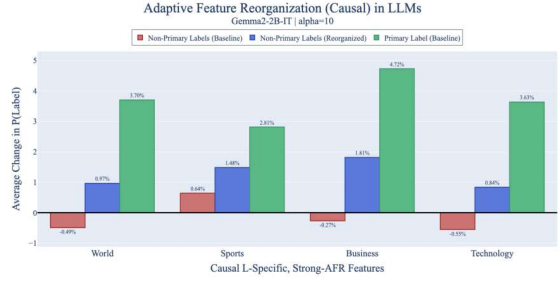
$$\begin{aligned} B_{\text{red}}(l) &= \frac{1}{|\mathcal{F}_l|} \sum_{f \in \mathcal{F}_l} \max_{c \neq l} \text{CE}_{f \rightarrow c}^{\text{orig}}(\alpha), \\ B_{\text{blue}}(l) &= \frac{1}{|\mathcal{F}_l|} \sum_{f \in \mathcal{F}_l} \max_{c \neq l} \text{CE}_{f \rightarrow c}^{\text{ctrl}(-l)}(\alpha), \\ B_{\text{green}}(l) &= \frac{1}{|\mathcal{F}_l|} \sum_{f \in \mathcal{F}_l} \text{CE}_{f \rightarrow l}^{\text{orig}}(\alpha). \end{aligned}$$

Thus, causal reorganization predicts

$$B_{\text{red}}(l) < B_{\text{blue}}(l) < B_{\text{green}}(l).$$



(a) Excluding a feature’s primary label often increases its activations toward the remaining labels.



(b) AFR features dynamically reorganize their causal effect. We observe a consistent ordering (red < blue < green).

Figure 5: Illustration of Adaptive Feature Reorganization in LLMs. (a) Volcano plots show a pronounced positive skew, revealing a substantial subset of features that acquire stronger activation toward alternative labels when their primary semantic trigger is removed. (b) Steering results for Gemma2-2B-IT show partial causal role reassignment. For each label l , we compare the strongest non-primary causal effect in the original setting (red), the strongest non-primary causal effect under label exclusion (blue), and the original primary-label causal effect (green). The typical ordering red < blue < green indicates partial causal role transfer under contextual constraint.

In other words, when the primary label is removed from the task, part of the feature’s causal influence becomes redirected toward the remaining labels.

As shown in Figure 5b (with additional model results in Figure 15 of Appendix F), this ordering holds across all four labels and evaluated models. These results support *partial causal reorganization*: causal features that exhibit strong AFR at the activation level also redirect part of their causal influence toward the remaining labels when their original primary label becomes unavailable.

(2) Implications for Mechanistic Interpretability

AFR dynamics highlight the contextual flexibility of internal model representations and SAE features, challenging the standard view of features as static, binary semantic detectors (e.g., a “dog” feature or “Golden Gate Bridge” feature (Bricken et al., 2023)). While a static perspective may adequately describe certain features, our findings demonstrate that a non-trivial population dynamically reorganizes both its activation profile and causal influence under varying contextual constraints. Consequently, interpreting a feature solely through its top-activating dataset examples or single-prompt steering risks overlooking critical roles that emerge under alternative conditions. A faithful characterization of a feature’s causal role therefore necessitates evaluating interventions across diverse prompts and constraint regimes rather than within a single, fixed context.

4 Conclusion

In this work, we uncover three previously uncharacterized representation-level dynamics that consistently emerge across diverse model families and scales during few-shot prompt processing. By leveraging Sparse Autoencoders (SAEs) as an analytical lens, we render these latent behaviors directly observable and formalize three recurring phenomena: Decimal Periodicity Effect (DPE), Local Feature Homeostasis (LFH), and Adaptive Feature Reorganization (AFR).

Together, these findings demonstrate the power of SAEs in revealing structured representational dynamics while exposing critical methodological implications for feature-based interpretability. Specifically, feature activations are not governed solely by static task semantics, but are dynamically modulated by prompt position, local exemplar history, and contextual constraints. Accounting for DPE, LFH, and AFR is therefore essential for developing more stable, reliable, and semantically faithful interpretations of model internals.

5 Limitations

Our study relies on publicly available, high-quality sparse autoencoders (SAEs), which inherently limits the breadth of our analysis. For instance, prominent models such as Qwen3 only have high-quality SAEs available at specific layer intervals (e.g., 1/4, 2/4, and 3/4 layer depths), which constrains our ability to track continuous layer-by-layer dynamics across the entire network. Additionally, because these open-source SAEs are trained using differ-

ent methodologies, hyperparameter configurations, and target sparsities across different model families, making direct quantitative comparisons between models remains a challenge.

References

- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2(5):6.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Hakaze Cho, Mariko Kato, Yoshihiro Sakai, and Naoya Inoue. 2025a. Revisiting in-context learning inference circuit in large language models. In *International Conference on Learning Representations*, volume 2025, pages 13081–13117.
- Ikhyun Cho and Julia Hockenmaier. 2025a. Analyzing multilingualism in large language models with sparse autoencoders. In *Second Conference on Language Modeling*.
- Ikhyun Cho and Julia Hockenmaier. 2025b. Toward efficient sparse autoencoder-guided steering for improved in-context learning in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28961–28973.
- Ikhyun Cho and Julia Hockenmaier. 2026. Understanding primacy effects in large language models with sparse autoencoders. In *Third Conference on Language Modeling*.
- Ikhyun Cho, Gaeul Kwon, and Julia Hockenmaier. 2024. Tutor-ICL: Guiding large language models for improved in-context learning performance. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9496–9506.
- Ikhyun Cho, Gaeul Kwon, and Julia Hockenmaier. 2025b. On the versatility of sparse autoencoders for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19531–19538.
- Ikhyun Cho, Changyeon Park, and Julia Hockenmaier. 2025c. The power of bullet lists: A simple yet effective prompting approach to enhancing spatial reasoning in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3047–3057.
- Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.
- Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pages 26721–26754.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. 2024. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*.
- Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pages 7827–7845.
- Dmitrii Kharlapenko, Stepan Shabalin, Fazl Barez, Arthur Conmy, and Neel Nanda. 2025. Scaling sparse feature circuit finding for in-context learning. *arXiv preprint arXiv:2504.13756*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530.
- Tom Lieberum, Senthoooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300.

Johnny Lin and Joseph Bloom. 2023. Neuronpedia: Interactive reference and tooling for analyzing neural networks. *Software available from neuronpedia.org*.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermy, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. [On the biology of a large language model](#). *Transformer Circuits Thread*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.

Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 11048–11064.

Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E Gonzalez. 2025. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. 2026. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *arXiv preprint arXiv:2605.29358*.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855.

An Yang, Anpeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Prompt Examples

For our evaluation, we utilize three distinct prompt configurations: (1) text classification tasks using the AGNews (Zhang et al., 2015) dataset; (2) code generation and tool-use environments via the Berkeley Function Calling Leaderboard (Patil et al., 2025); and (3) retrieval-augmented generation (RAG) question answering, utilizing the Natural Questions dataset (Kwiatkowski et al., 2019). An example prompt template for each dataset is shown in Figure 6.

B Details on the Hypothesis Tests Used in the Paper

For each feature, we formed paired activation differences $d_i = a_i^{\text{controlled}} - a_i^{\text{original}}$ across aligned examples and applied the Wilcoxon signed-rank procedure to these differences. We then computed the positive-rank sum W_+ and obtained an approximate two-sided p-value from the asymptotic normal z-statistic associated with the Wilcoxon signed-rank test. Under this test, the null hypothesis is that the distribution of paired differences is symmetric about zero, and the two-sided alternative is that it is not. Because the procedure is rank-based, it does not require the paired differences to be normally distributed, although the p-values reported here rely on the asymptotic normal approximation rather than exact enumeration.

To summarize effect direction and size, we report the matched-pairs rank-biserial correlation

$$r_{rb} = \frac{2W_+}{R} - 1 \quad (4)$$

where $R = \frac{n_{eff}(n_{eff}+1)}{2}$ is the total sum of ranks over the n_{eff} nonzero paired differences. Values near +1 indicate that positive controlled-minus-original differences dominate the ranked comparisons, values near -1 indicate dominance of negative differences, and values near 0 indicate little directional dominance. In the volcano-style

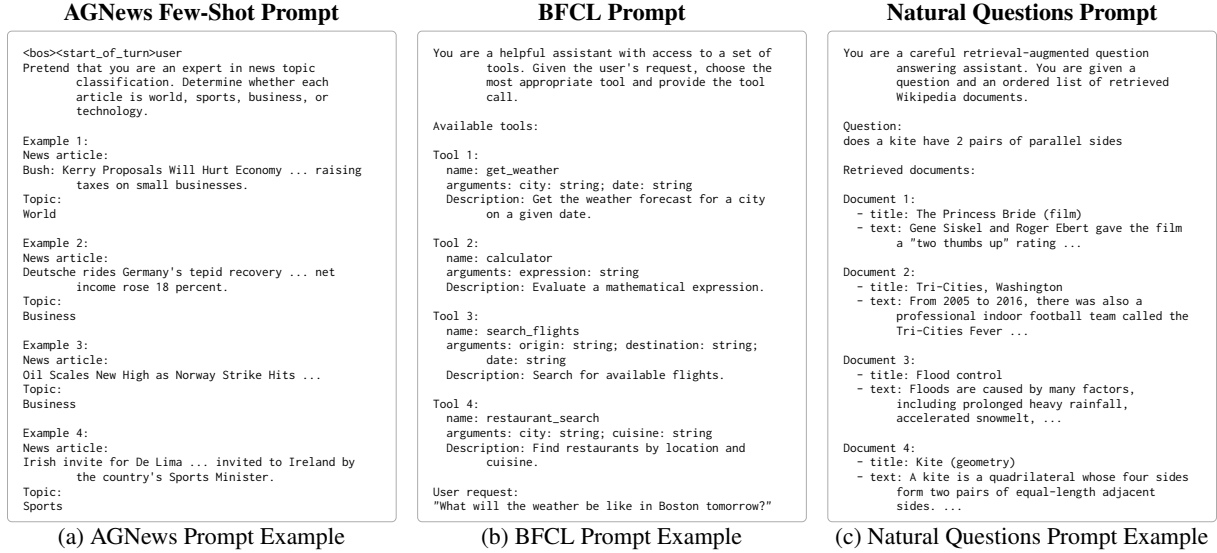


Figure 6: **Prompt templates used in our experiments.** Each panel shows an illustrative four-item excerpt rather than the full prompt.

plots, the x-axis shows r_{rb} and the y-axis shows $-\log_{10}$ of the corresponding p-value, so points farther from zero horizontally and higher vertically indicate stronger directional effects and stronger evidence for a systematic paired shift.

C Measuring the Decimal Periodicity Effect

Each prompt contains 50 few-shot demonstrations followed by a final labeled target example. For a prompt p , model layer ℓ , SAE feature f , and completed-example count $j \in \{1, \dots, 50\}$, we measure the feature at token positions defined relative to the beginning of Example $j + 1$. Let $\mathcal{T}_{p,j}$ denote these selected positions. We define

$$s_{p,j,f}^{(\ell)} := \frac{1}{|\mathcal{T}_{p,j}|} \sum_{t \in \mathcal{T}_{p,j}} z_{p,t,f}^{(\ell)}, \quad (5)$$

where $z_{p,t,f}^{(\ell)}$ is the activation of feature f at token t . In our default configuration, $\mathcal{T}_{p,j}$ contains the last token of each exemplar block.

We prespecify the modulo-10 target set

$$\mathcal{M} = \{10, 20, 30, 40, 50\},$$

which corresponds to the states immediately before Examples 11, 21, 31, 41, and 51. We also examined other modulo- k patterns, but modulo-10 was the only consistently salient one. For each feature, we pool its values at these positions across prompts and compare them with its values at all remaining

completed-example counts. Let

$$\mu_{f,\mathcal{M}}^{(\ell)} = \frac{1}{P|\mathcal{M}|} \sum_{p=1}^P \sum_{j \in \mathcal{M}} s_{p,j,f}^{(\ell)},$$

and define $\mu_{f,\neg\mathcal{M}}^{(\ell)}$ analogously over $j \notin \mathcal{M}$. We quantify the feature-level effect using

$$d_f^{(\ell)} = \frac{\mu_{f,\mathcal{M}}^{(\ell)} - \mu_{f,\neg\mathcal{M}}^{(\ell)}}{\sigma_{f,\text{pooled}}^{(\ell)}}, \quad (6)$$

where $\sigma_{f,\text{pooled}}^{(\ell)}$ is the pooled standard deviation of the two groups. Positive values indicate features that are more strongly expressed at multiples of 10, whereas negative values indicate periodic troughs.

To obtain a layer-level measure while reducing feature-selection bias, we deterministically divide the prompts into two halves. Within each split S , we first compute the mean positional profile

$$\mu_{f,j}^{(\ell,S)} = \frac{1}{|S|} \sum_{p \in S} s_{p,j,f}^{(\ell)}$$

and standardize it across completed-example counts:

$$\tilde{\mu}_{f,j}^{(\ell,S)} = \frac{\mu_{f,j}^{(\ell,S)} - \text{mean}_{j'} \mu_{f,j'}^{(\ell,S)}}{\text{std}_{j'} \mu_{f,j'}^{(\ell,S)}}.$$

The split-specific periodicity score is then

$$q_f^{(\ell,S)} = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \tilde{\mu}_{f,j}^{(\ell,S)} - \frac{1}{|\neg\mathcal{M}|} \sum_{j \notin \mathcal{M}} \tilde{\mu}_{f,j}^{(\ell,S)}. \quad (7)$$

We select the top- K features according to q_f on one split and evaluate their mean score on the other; we then reverse the split roles and average the two results. This cross-validated top- K contrast measures how strongly each layer contains a reproducible population of features whose activation profiles peak at the prespecified modulo-10 positions.

D Extended Decimal Periodicity Effect Results

Pervasiveness Across Diverse Token Positions

A critical question is whether the DPE effect is simply an artifact occurring exclusively at structural boundary tokens (e.g., immediately preceding an “Example” marker). If DPE is localized strictly to these boundaries, it could be reduced to a minor formatting artifact. However, our analysis demonstrates that DPE manifests across a broader range of token positions. To demonstrate this, we investigate the “label words” in in-context learning (ICL) classification—tokens that have garnered significant attention as the primary drivers of ICL task execution (Wang et al., 2023; Cho et al., 2025a, 2024). As illustrated in Figure 7, we find that the identical modulo-10 periodicity persists within the representations of these label words. This confirms that the effects of DPE are spread more broadly across tokens in the prompt, not merely at formatting boundaries.

DPE Results on Additional Models

Figure 9 provides DPE results on Gemma2-2B, Gemma2-9B, and Qwen3-32B, respectively. Llama3.1-8B-Base results are provided in Figure 7.

E Additional Results for Local Feature Homeostasis

In this section, we present extended empirical results to further characterize the dynamics and ubiquity of Local Feature Homeostasis (LFH). Figure 11 evaluates LFH across two distinct task paradigms—tool calling (BFCL) and open-domain question answering (NQ)—confirming that the phenomenon generalizes broadly across diverse context regimes.

Figure 10 presents extended causal experiments on LFH via attention head steering across Gemma2-9B-IT, Llama3.1-8B-Instruct, and Qwen3-14B-IT, confirming that attention heads consistently mediate the LFH mechanism.

F Detailed Results on Adaptive Feature Reorganization (AFR) in LLMs

We present a comprehensive analysis of AFR mechanism in Figures 13 and 14. Our results yield several key observations. As shown in the volcano-style plots, label-specific features exhibit a consistent positive shift in activation on remaining labels when their primary trigger label is excluded from the input context. This pattern indicates that, in the absence of a primary semantic trigger, the affected features do not become dormant; instead, the model activates the latent—normally unused—functional roles to process the remaining information. Importantly, this pattern is consistent across all four labels and three model families we evaluate.

Figure 15 summarizes the causal role reassignment associated with the AFR effect in Gemma2-2B-IT, Gemma2-9B-IT, Llama3.1-8B-Instruct, and Qwen3-14B-IT. We observe the same qualitative pattern as in the main text. Taken together, these consistent results suggest that AFR behavior is a shared representational dynamic in LLMs.

To evaluate AFR on alternative tasks, we utilize the BFCL dataset. Because function definitions lack explicit category labels, we partition them into two distinct groups—Math and Non-Math—to serve as proxy labels. The remaining experimental procedures remain unchanged. Using Gemma2-9B-IT and Qwen3-14B-IT as representative models, the results are summarized in Figure 12.

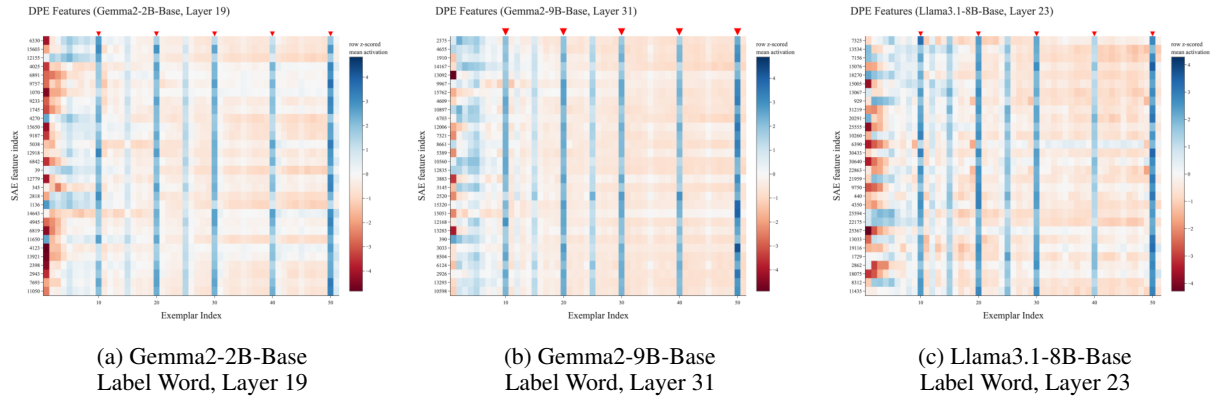


Figure 7: **Persistence of DPE across varied token positions.** Evaluated on Gemma2-2B-Base, Gemma2-9B-Base, and Llama3.1-8B-Base, the decimal periodicity effect is not isolated to the block-final token but extends to alternative positions within the sequence. Illustrated figures measure activations at the corresponding label words, demonstrating that this positional periodicity remains robust to the specific probe placement.

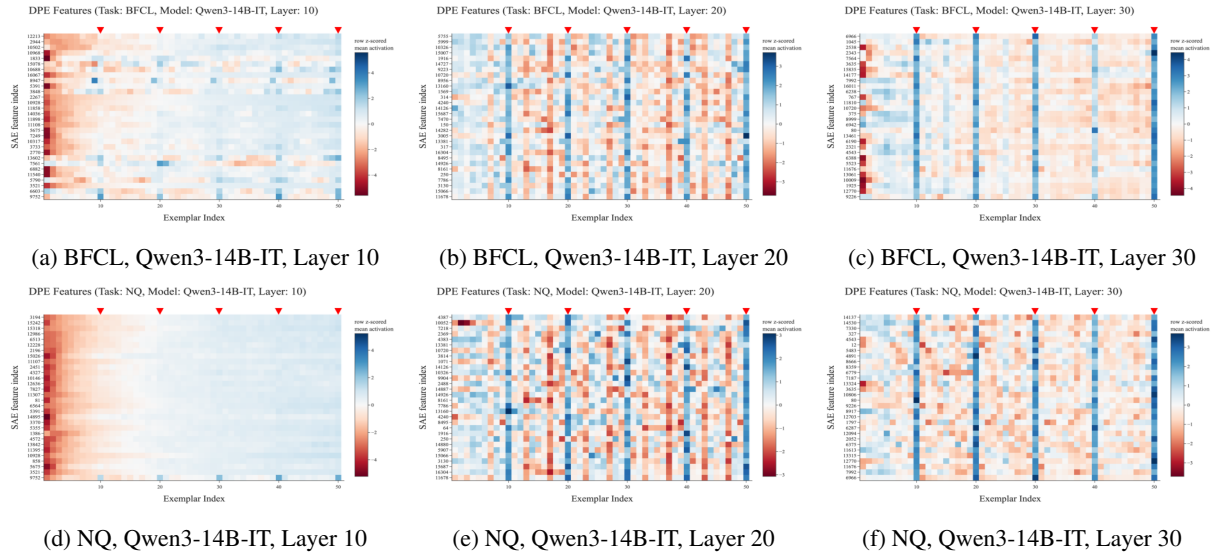


Figure 8: **The Decimal Periodicity Effect (DPE) on Tool-Calling (BFCL) and Retrieval Augmented Generation (RAG) Question Answering (Natural Questions).**

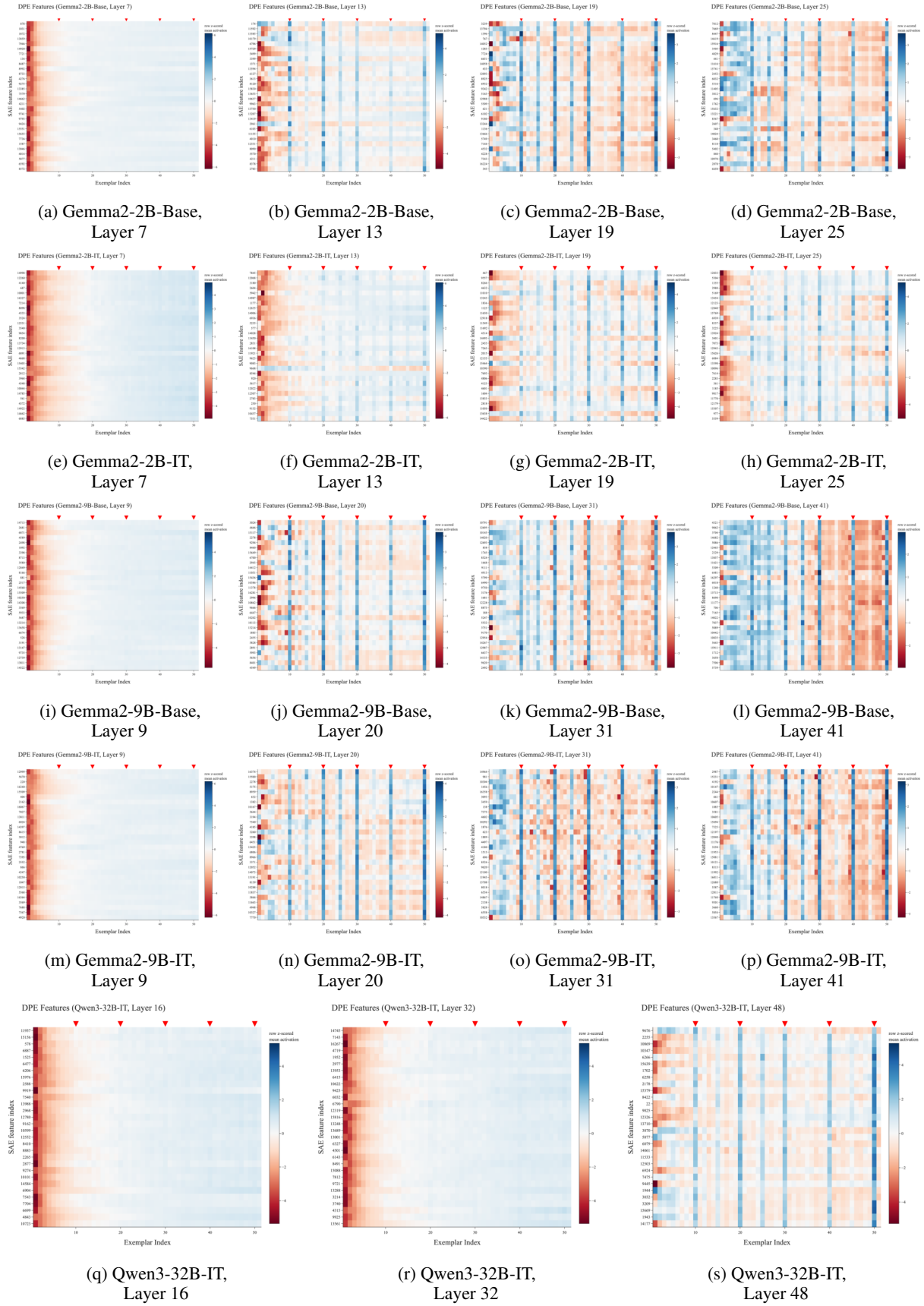
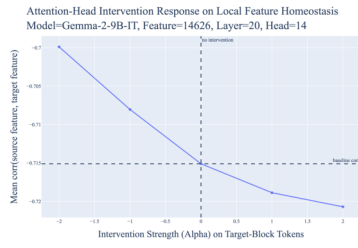
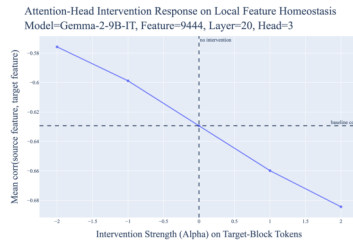


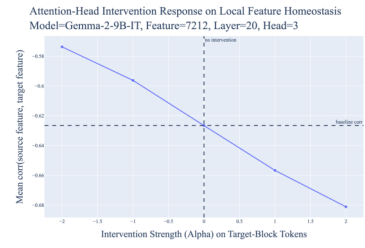
Figure 9: **DPE Results on Gemma2-2B-Base/IT, Gemma2-9B-Base/IT, and Qwen3-32B-IT.** Each row in each heatmap corresponds to one selected SAE feature. Cell values are row-standardized mean activations: for feature f , the mean activation at block index k is centered and scaled by that feature’s mean and standard deviation across all block counts.



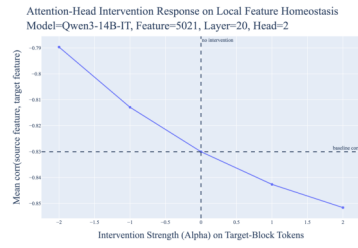
(a) Gemma2-9B-IT, Feature 14626



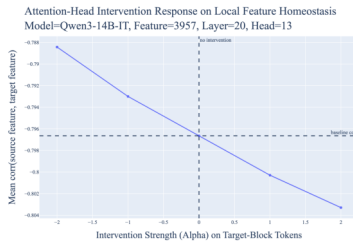
(b) Gemma2-9B-IT, Feature 9444



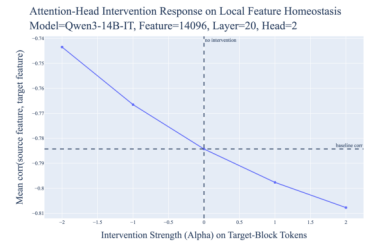
(c) Gemma2-9B-IT, Feature 7212



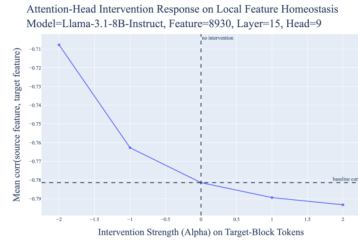
(d) Qwen3-14B-IT, Feature 5021



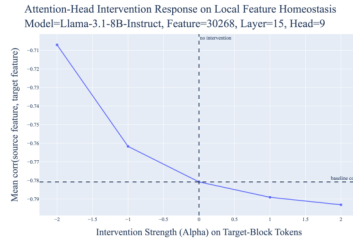
(e) Qwen3-14B-IT, Feature 3957



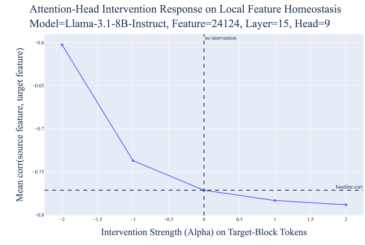
(f) Qwen3-14B-IT, Feature 14096



(g) Llama3.1-8B-Instruct, Feature 8930

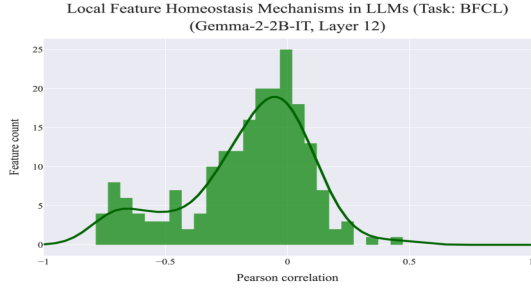


(h) Llama3.1-8B-Instruct, Feature 30268

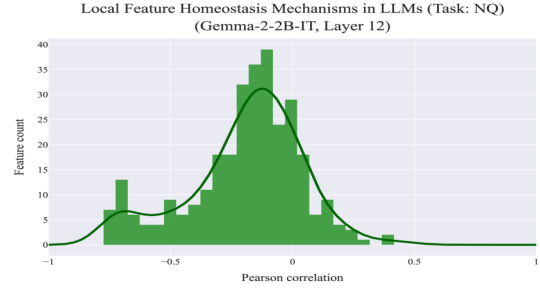


(i) Llama3.1-8B-Instruct, Feature 24124

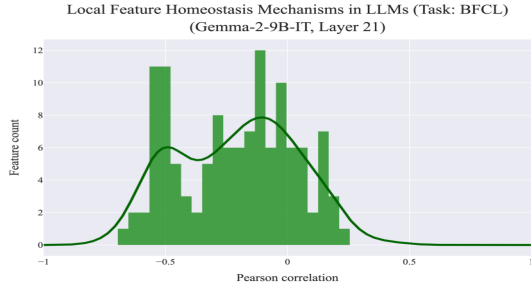
Figure 10: Extended causal evaluation of LFH. Interventions on the top three LFH features across Gemma2-9B-IT, Llama3.1-8B-Instruct, and Qwen3-14B-IT confirm that specific attention heads directly mediate the homeostatic mechanism.



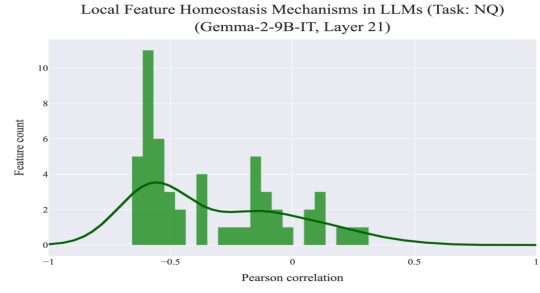
(a) Gemma2-2B-IT, BFCL



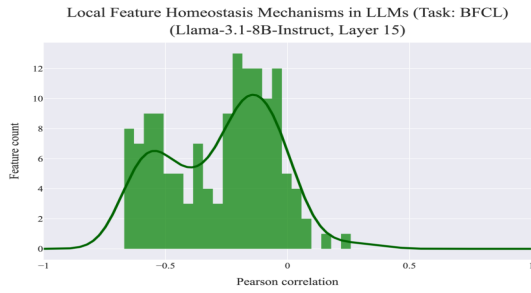
(b) Gemma2-2B-IT, NQ



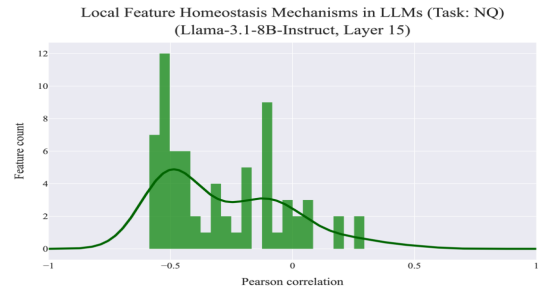
(c) Gemma2-9B-IT, BFCL



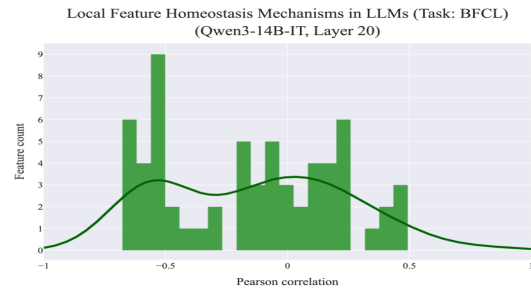
(d) Gemma2-9B-IT, NQ



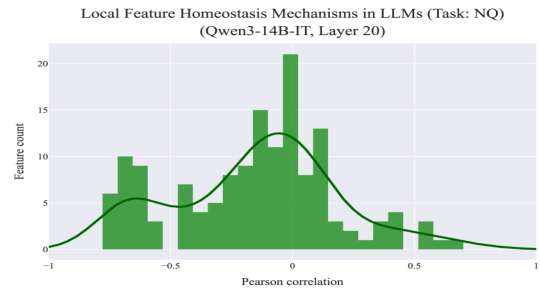
(e) Llama3.1-8B-Instruct, BFCL



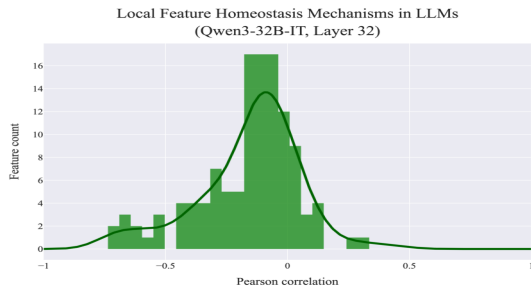
(f) Llama3.1-8B-Instruct, NQ



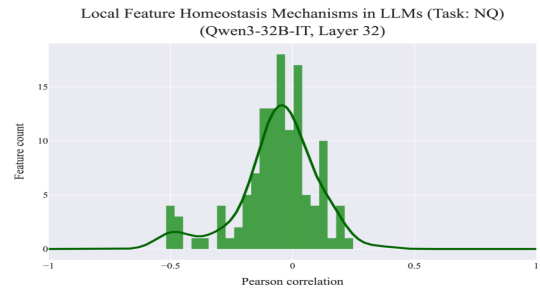
(g) Qwen3-14B-IT, BFCL



(h) Qwen3-14B-IT, NQ

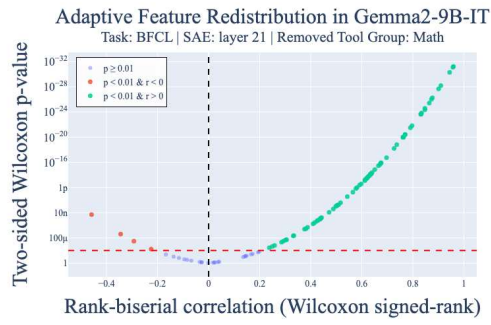


(i) Qwen3-32B-IT, BFCL

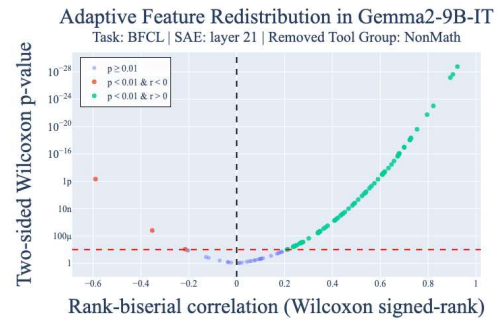


(j) Qwen3-32B-IT, NQ

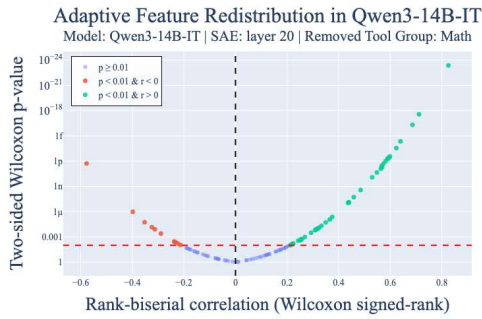
Figure 11: **Local Feature Homeostasis (LFH) across BFCL and NQ.** We use the instruction-tuned models as representatives.



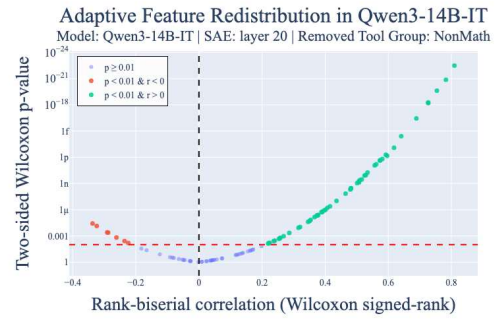
(a) Gemma2-9B-IT, L21, Removing “Math”



(b) Gemma2-9B-IT, L21, Removing “Non-Math”

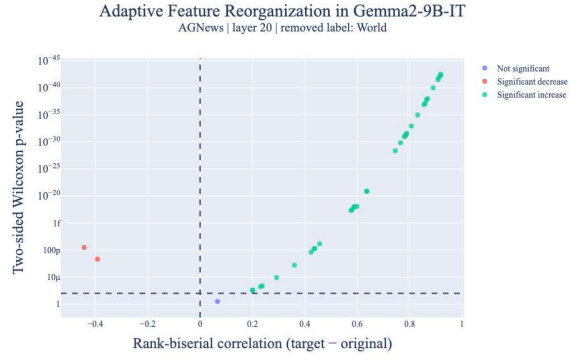


(c) Qwen3-14B-IT, L20, Removing “Math”

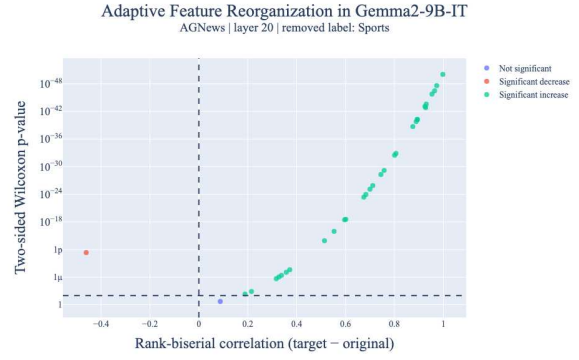


(d) Qwen3-14B-IT, L20, Removing “Non-Math”

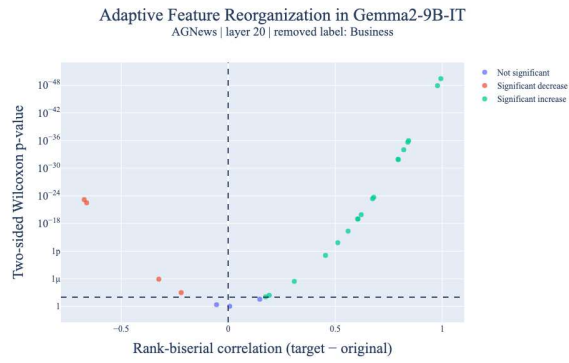
Figure 12: AFR effect on the BFCL task for Gemma2-9B-IT and Qwen3-14B-IT. Consistent patterns are observed.



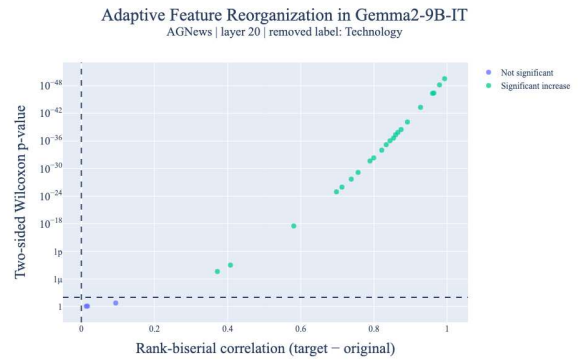
(a) Gemma2-9B-IT, L20, Removing “World”



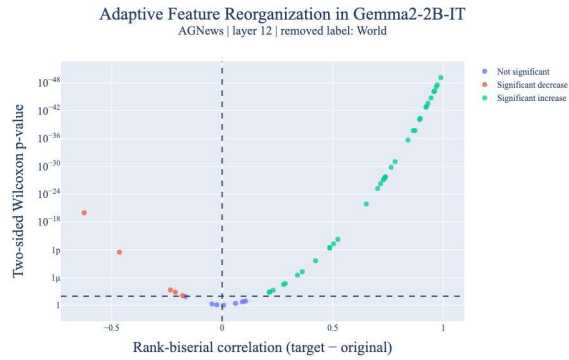
(b) Gemma2-9B-IT, L20, Removing “Sports”



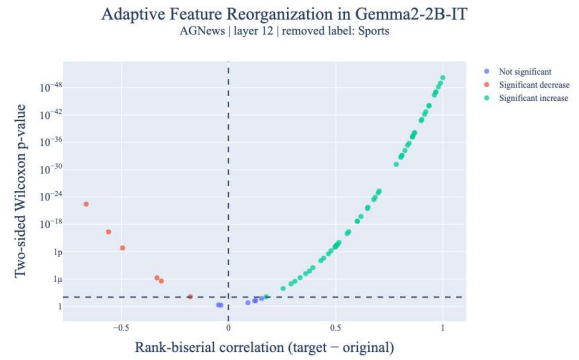
(c) Gemma2-9B-IT, L20, Removing “Business”



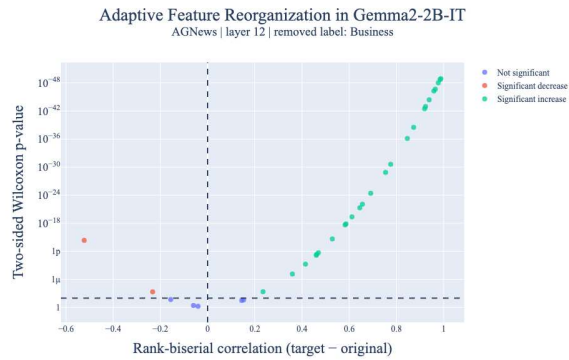
(d) Gemma2-9B-IT, L20, Removing “Technology”



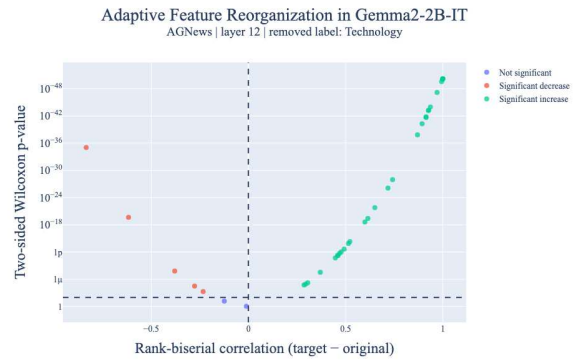
(e) Gemma2-2B-IT, L12, Removing “World”



(f) Gemma2-2B-IT, L12, Removing “Sports”

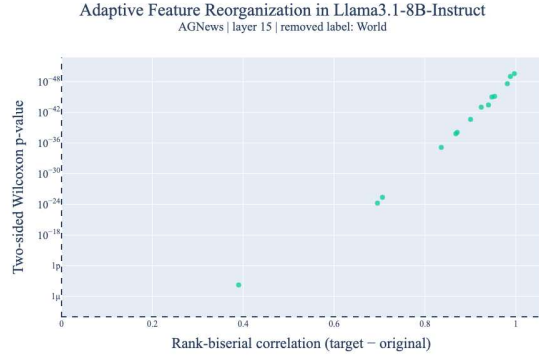


(g) Gemma2-2B-IT, L12, Removing “Business”

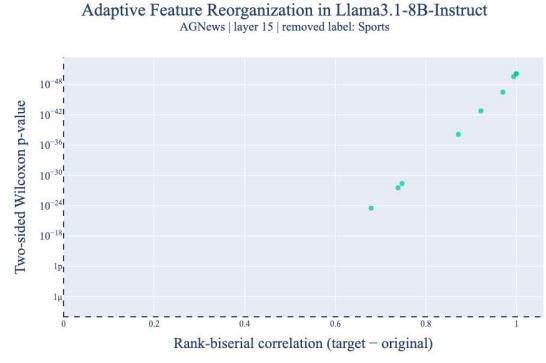


(h) Gemma2-2B-IT, L12, Removing “Technology”

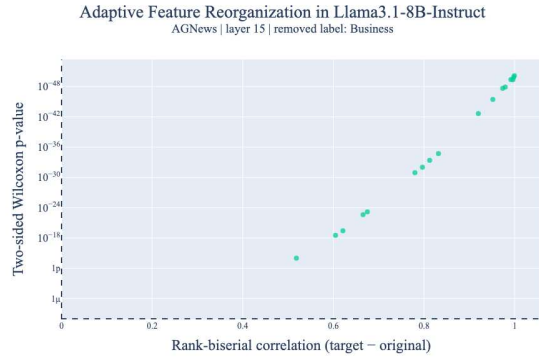
Figure 13: **Adaptive Feature Reorganization in Gemma2-2B-IT and Gemma2-9B-IT.** Consistent patterns are observed across model families.



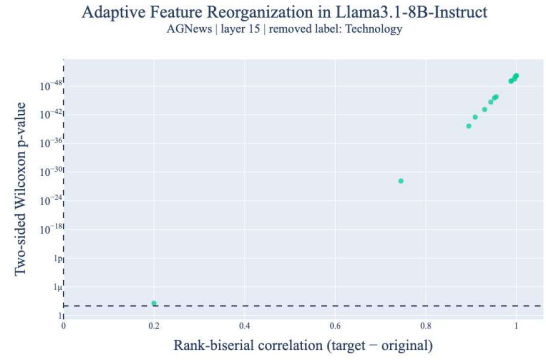
(a) Llama3.1-8B-Instruct, L15, Removing “World”



(b) Llama3.1-8B-Instruct, L15, Removing “Sports”

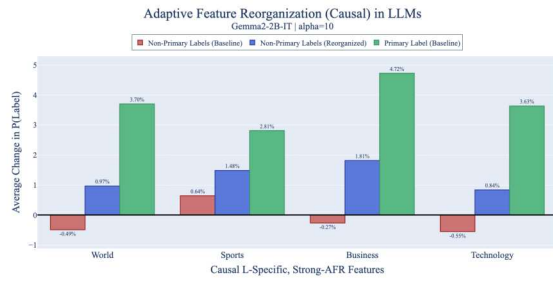


(c) Llama3.1-8B-Instruct, L15, Removing “Business”

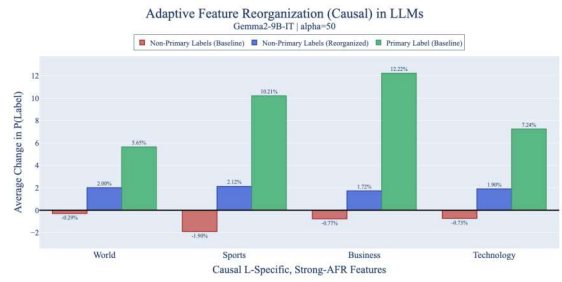


(d) Llama3.1-8B-Instruct, L15, Removing “Technology”

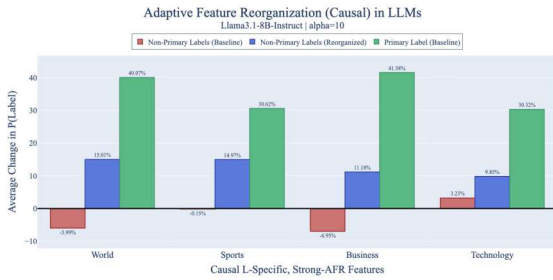
Figure 14: **Adaptive Feature Reorganization in Llama3.1-8B-Instruct.** Consistent patterns are observed across different labels.



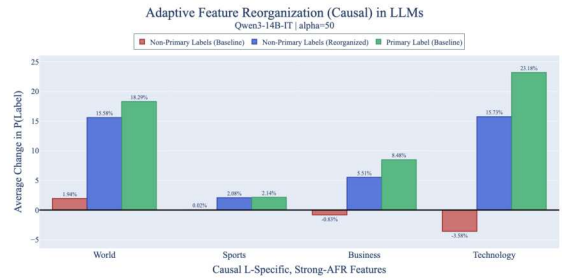
(a) Gemma2-2B-IT



(b) Gemma2-9B-IT



(c) Llama3.1-8B-Instruct



(d) Qwen3-14B-IT

Figure 15: **AFR Causal Intervention Results on Gemma2-2B-IT, Gemma2-9B-IT, Llama3.1-8B-Instruct, and Qwen3-14B-IT.** The red < blue < green ordering is consistently satisfied.