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

Anonymous ACL submission

Abstract

In this work, we identify three representation-level dynamics shared across diverse large language models (LLMs) during the processing of few-shot prompts. These structural behaviors are not explicitly trained, yet they emerge reliably across model families and scales to systematically shape internal representations. Using Sparse Autoencoders (SAEs) as an analytical lens, we uncover three primary phenomena: (1) *a decimal periodicity effect,* where specific features exhibit a repeating modulo-10 activation pattern; (2) *local feature homeostasis*, a compensatory push–pull dynamic in which current representations are inversely modulated by prior exemplars; and (3) *adaptive feature redistribution,* wherein specific features flexibly shift their semantic profiles in response to contextual constraints. We validate these mechanisms across multiple model families and prompt formats. Together, they offer fresh insights into the internal dynamics of in-context learning and raise important caveats for future SAE-based interpretability studies.

1 Introduction

Modern large language models (LLMs) are trained with a seemingly simple objective: autoregressive next-token prediction (Brown et al., 2020). Yet a growing body of work has shown that this objective gives rise to remarkably rich internal computation and behavior. In particular, prior work has documented capabilities often described as "emergent" abilities—behaviors that are weak or absent in smaller models, but become pronounced as model scale increases (Wei et al., 2022). Nevertheless, the full range of such behaviors, and especially how to uncover them effectively, remains an open problem.

In this paper, we study this discovery problem. Our goal is to identify latent representational dynamics 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 prompt-processing 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., 2024). 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 (Radford et al., 2019; Brown et al., 2020; Min et al., 2022). 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.

3. **Adaptive Feature Redistribution:** 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 considerations for future SAE-based research.

2 Preliminaries: Sparse Autoencoders

One of the most effective approaches for making the neural activations of large language models interpretable to humans is the sparse autoencoder (SAE) (Sharkey and Beren, 2022). 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 (Team, 2024; Lindsey et al., 2025).

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:

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

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 (Lieberum et al., 2024), 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 multilingualism (Templeton et al., 2024), and has led to resources such as Neuronpedia (Lin, 2023), which provide annotations based on common next-token prediction patterns, often generated using LLMs as judges. While these resources are valuable, such annotations primarily capture *what* a neuron or feature tends to correlate with, and offer limited insight into *why* it becomes active during task execution. Our work addresses this gap by leveraging targeted prompt manipulations to study the dynamic behavior of LLM hidden states, revealing a suite of shared internal representational behaviors.

3 Intrinsic Representation Dynamics in Few-Shot Prompting

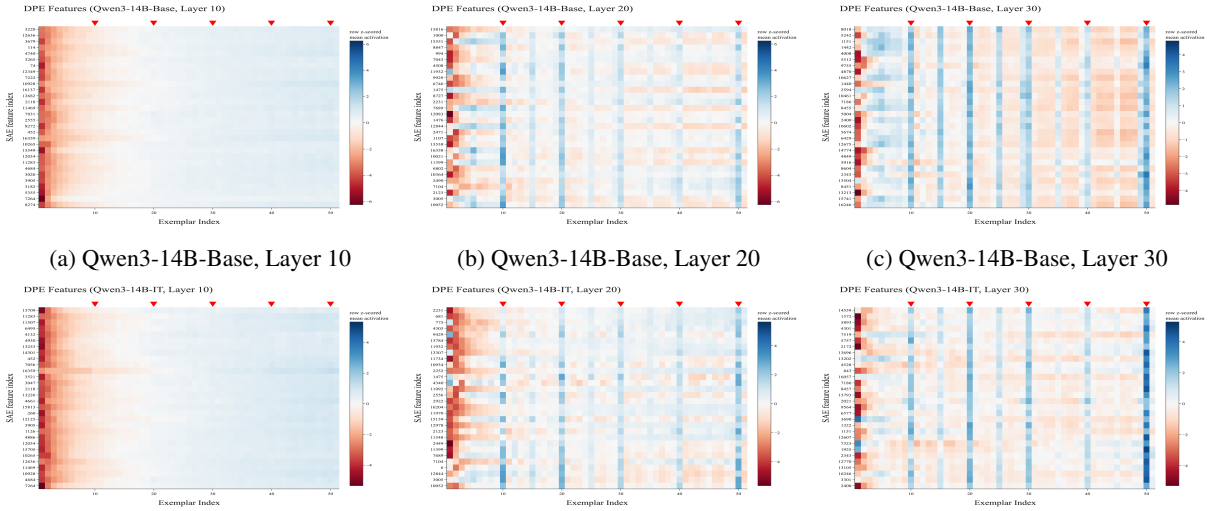
Overview In this work, we identify three representation-level mechanisms that consistently recur across diverse models: (1) *decimal periodicity effect*, where models exhibit a repeating modulo-10 shift in their representations tied to exemplar position rather than semantic content; (2) *local feature 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 redistribution*, 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) (Team, 2025), and Llama 3.1 (8B) (Dubey 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 Background and Intuition

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



(a) Qwen3-14B-Base, Layer 10 (b) Qwen3-14B-Base, Layer 20 (c) Qwen3-14B-Base, Layer 30

(d) Qwen3-14B-IT, Layer 10 (e) Qwen3-14B-IT, Layer 20 (f) Qwen3-14B-IT, Layer 30

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.

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 shown in Figures 5– 7. 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 C.1 that DPE persists robustly across alternative choices of m.

We construct 1,000 unique 50-shot prompts using exemplars randomly sampled from the AGNews 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. Because we utilize a sufficiently large sample size ($N = 1,000$) of randomized prompts, any remaining salient periodic activation profile is unlikely to be driven by semantic noise, establishing it as a genuine position-dependent artifact.

3.1.3 Results

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.

(1) **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.

(2) **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.

(3) **Training Resilience:** This pattern persists robustly across both base and instruction-tuned model variants. This invariance indicates that DPE is a foundational 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 C, we further verify that DPE persists across diverse tasks—specifically, tool calling and retrieval-augmented generation QA—and models, highlighting the broad generality of the DPE mechanism. Collectively, this consistency across different tasks and models demonstrates that DPE is a shared, highly systematic characteristic of modern 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 nearby non-modulo prompt length (e.g., a 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 positional artifacts rather than genuine task semantics. This serves as an important methodological caveat for future SAE-based studies of in-context learning.

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.

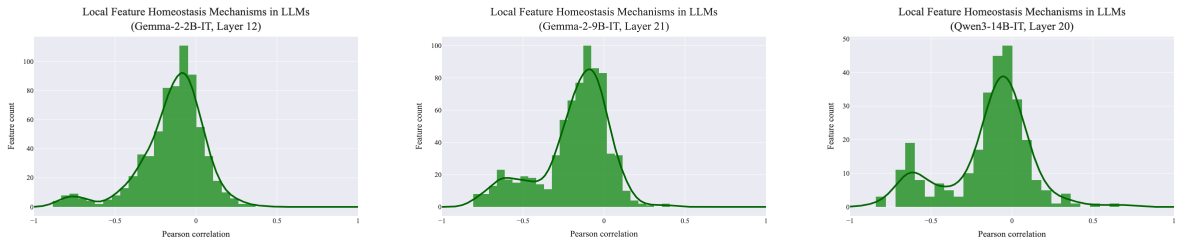
3.2 Local Feature Homeostasis

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

3.2.1 Background and Intuition

While it is well-established that few-shot exemplars improve downstream task performance (Brown et al., 2020; Min et al., 2022), the mechanisms 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. We demonstrate that this homeostatic, push-pull feature-level pattern is highly prevalent across model families. To our knowledge, this constitutes the first concrete feature-level evidence demonstrating inverse



(a) Gemma2-2B-IT (b) Gemma2-9B-IT (c) Qwen3-14B-IT

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.

cross-exemplar modulation in LLMs, which we term *Local Feature Homeostasis* (LFH).

3.2.2 Experimental Setup

To evaluate this effect, we curate a set of $M = 50$ target exemplars and, for each target, pair it with $K = 50$ distinct immediately preceding exemplars, yielding a total of $P = M \times K = 2{,}500$ prompt instances from the AGNews dataset. Each prompt is constructed in an 8-example format: the target exemplar occupies a fixed final position and the preceding exemplars are randomly sampled.

We first filter out features that activate too infrequently, using an average activation-frequency threshold of 5%. For the remaining features—typically around 100 to 200 features—we summarize, for each SAE feature f, its activation on the previous exemplar and on the target exemplar by averaging token-level activations over the corresponding exemplar span:

$$A^{(p)}_{f,\text{prev}} := \frac{1}{|E^{(p)}_{\text{prev}}|} \sum_{t \in E^{(p)}_{\text{prev}}} a^{(p)}_f(t),$$

$$A^{(p)}_{f,\text{tgt}} := \frac{1}{|E^{(p)}_{\text{tgt}}|} \sum_{t \in E^{(p)}_{\text{tgt}}} a^{(p)}_f(t), \qquad (1)$$

where $p \in \{1, \ldots, P\}$ indexes prompt instances, $a^{(p)}_f(t)$ denotes the token-level activation of feature f, and $E^{(p)}_{\text{prev}}$ and $E^{(p)}_{\text{tgt}}$ denote the token spans of the previous and target exemplars, respectively.

We then collect these values across all prompt instances into two vectors,

$$\mathbf{A}_{f,\text{prev}} = \left(A^{(1)}_{f,\text{prev}}, \ldots, A^{(P)}_{f,\text{prev}}\right),$$
$$\mathbf{A}_{f,\text{tgt}} = \left(A^{(1)}_{f,\text{tgt}}, \ldots, A^{(P)}_{f,\text{tgt}}\right),$$

and compute their Pearson correlation:

$$\rho_f := \text{corr}(\mathbf{A}_{f,\text{prev}}, \mathbf{A}_{f,\text{tgt}}).$$

Intuitively, if feature f were strictly expectation-independent, then varying the previous exemplar should have little 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 D). The observed pattern is striking: the distribution is approximately bimodal. One mode is centered near zero, which is broadly consistent with a null or weak-effect baseline, while a second, clearly left-shifted mode is concentrated in the $\rho_f < -0.5$ region.

This bimodal structure indicates that these inverse interactions are not rare exceptions, but rather a systematic component of the model's internal dynamics. In other words, LLMs exhibit a nontrivial class of features that display compensatory behavior: their responses are attenuated when the same feature was strongly activated in the immediately preceding exemplar, and amplified when the prior activation was weak. Moreover, the consistency of this pattern across all five evaluated models suggests that it reflects a shared representational dynamic rather than an isolated artifact. For an intuitive visual illustration of LFH, an instance-level

example is provided in Figure 16 (Appendix D).

3.2.4 Discussion and Practical Implications

(1) Geometric Consistency To better understand why LFH features exhibit strong source-to-target anti-correlations, we first examine their alignment with the model's native attention weights. Following the QK/OV circuit perspective of attention heads (Elhage et al., 2021), we use a weight-only diagnostic that asks whether an attention head's OV circuit maps a feature direction toward its negative counterpart. Concretely, for an SAE feature decoder direction d_f, we measure whether there exists an attention head h such that

$$W_O^h W_V^h d_f \approx -d_f.$$

Intuitively, this tests whether the head has the static geometric capacity to read feature f from the residual stream and write back a suppressive update along the same feature direction. This is precisely the kind of geometry predicted by LFH: if a feature is strongly expressed in the preceding exemplar, then the model may compensate by writing a negative update on the next exemplar.

Table 2 shows this weight-only diagnostic for the top-10 LFH features in `Gemma2-2B-IT`. For each feature, we report its ρ_f, the best-aligned attention head, the OV cosine, and the percentile of this OV cosine among all SAE features. Most of the top LFH features exhibit strongly negative OV cosines, often close to -1, indicating that the corresponding attention head maps the feature direction toward its negative direction. Moreover, seven of the ten features rank above the 95th percentile among all SAE features under this metric, with a median percentile near 99%. These results indicate that LFH features are not random directions. Instead, they are among the most strongly aligned with attention-head OV circuits possessing the precise geometry required to implement homeostasis-like behavior.

(2) Causal Validation. We next test whether the candidate attention heads identified above causally mediate the LFH effect. For a selected LFH feature and its corresponding candidate head, we intervene on the head's output contribution at the target-exemplar tokens and sweep the intervention strength $\alpha \in \{-5, -2, -1, 0, 1, 2, 5\}$. We use $\alpha = 0$ as the no-intervention baseline; $\alpha = -1$ ablates the head contribution, $\alpha > 0$ amplifies it, and $\alpha < -1$ reverses and amplifies the contribution. We then re-measure the source-to-target feature correlation under each intervention.

Feature	ρ_f	Best head	OV cosine	Percentile
L12-4758	-0.778	H2	-0.827	0.866
L12-13496	-0.753	H0	-0.336	0.041
L12-8745	-0.729	H2	-0.915	0.995
L12-14559	-0.724	H2	-0.875	0.957
L12-2245	-0.693	H2	-0.918	0.995
L12-12633	-0.626	H2	-0.951	0.996
L12-6457	-0.618	H2	-0.939	0.997
L12-5917	-0.601	H2	-0.919	0.982
L12-3607	-0.599	H2	-0.946	0.998
L12-14805	-0.589	H0	-0.758	0.758

Table 2: **Weight-only OV anti-write alignment of top LFH features in Gemma2-2B-IT.** For each top-ranked LFH feature, we identify the attention head whose OV circuit most strongly maps the SAE feature direction toward its negative counterpart. The OV cosine reports the cosine similarity between the value-side feature image and the output-side feature readout; more negative values indicate stronger anti-write geometry. Percentile is computed relative to all SAE features under the same best-head OV anti-write metric.

As shown in Figure 3, the attention-head intervention exhibits the predicted monotonic pattern: suppressing or ablating the candidate head weakens the LFH effect, whereas amplifying the head strengthens it, yielding a more negative source-to-target correlation. In contrast, analogous interventions on the MLP residual do not produce the same systematic strengthening of LFH. This contrast suggests that the dominant mediation pathway for LFH is attention-based rather than MLP-based.

Figure 3: **Attention heads (blue) causally mediate local feature homeostasis, whereas MLP (red) interventions fail to strengthen LFH.** Results are shown for top-ranked LFH features in `Gemma2-2B-IT`. Additional model results are provided in Figure 15.

(3) Confounding Effects on Downstream SAE Analyses. Just as DPE showed that SAE feature activations can be strongly shaped by demon-

stration 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 Redistribution

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 *adaptive redistribution:* 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 its activation on all other labels:

$$\Delta_l(f) := \bar{s}_{f,l} - \max_{c \neq l} \bar{s}_{f,c} > 0, \qquad (2)$$

where $\bar{s}_{f,c}$ denotes the average activation of feature f on examples from label c. As in Equation 1, example-level activation is computed by averaging activations over the tokens of that example.

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

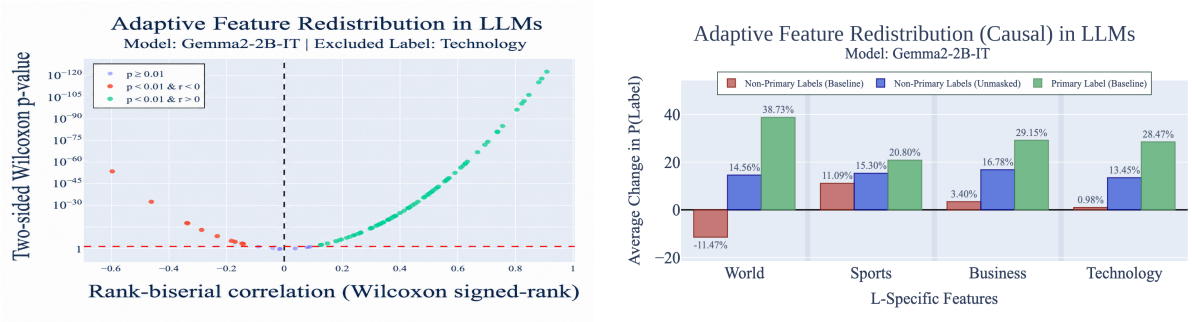
Figure 4a shows the result for `Gemma2-2B-IT`, using `Technology` as a representative excluded label. Results for other labels and models are provided in Appendix E. We observe a pronounced rightward and upward shift, indicating a substantial population of features whose activation on alternative labels increases once their primary semantic context is removed. We term this consistent, dynamic representational behavior, the *adaptive feature redistribution (AFR)*

3.3.4 Discussions and Practical Implications

(1) Causal Effects are Also Redistributed We next show that label-l-specific features do not merely change their activation distributions; their causal roles can also shift. To test this, we use feature steering. For each label l and each of the top 10 label-l-specific features ranked by Eq. 2, we add the corresponding feature direction to the hidden representations of the test-query tokens and measure the resulting change in the model's output distribution. We set the steering strength α to 10 times the feature's average activation across the dataset. Let

$$\mathrm{CE}^{u}_{f \to c}(\alpha)$$

denote the average causal effect of steering feature f with strength α on the model's probability of



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

(b) AFR affects not only feature activations, but also their causal effects.

Figure 4: **Illustration of Adaptive Feature Redistribution 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.

predicting label c, where $u \in \{\text{orig}, \text{ctrl}\}$ indicates the original or label-exclusion prompt setting. For a feature originally specific to label l, Figure 4b compares:

$$\underbrace{\max_{c \neq l} \mathrm{CE}^{\text{orig}}_{f \to c}(\alpha)}_{\text{strongest non-primary effect in the original setting}},$$

$$\underbrace{\max_{c \neq l} \mathrm{CE}^{\text{ctrl}}_{f \to c}(\alpha)}_{\text{strongest non-primary effect under label exclusion}},$$

$$\underbrace{\mathrm{CE}^{\text{orig}}_{f \to l}(\alpha)}_{\text{original primary-label effect}}.$$

If AFR manifests causally, the strongest non-primary effect should increase under label exclusion:

$$\max_{c \neq l} \mathrm{CE}^{\text{ctrl}}_{f \to c}(\alpha) \;>\; \max_{c \neq l} \mathrm{CE}^{\text{orig}}_{f \to c}(\alpha).$$

The original primary-label effect serves as a reference for the feature's baseline causal role in the unconstrained setting.

As shown in Figure 4b, with additional models in Figure 22 of Appendix E, the strongest non-primary causal effect is consistently larger under label exclusion than in the original setting, while remaining below the feature's original primary-label effect. This pattern indicates *partial causal role transfer:* when the primary label is unavailable, the feature is not silenced, but is re-engaged in support of the remaining labels.

(2) Implications for Mechanistic Interpretability. AFR dynamics highlight the contextual flexibility of model representations and SAE features, challenging the view of features as static, binary semantic detectors. While that perspective may describe certain features, our results show that a nontrivial subset redistributes both activation and causal influence under varying contextual constraints. Consequently, interpreting a feature solely through its top-activating examples or via steering in a single prompt setting can overlook roles that emerge only under alternative conditions. A faithful account of a feature's causal role therefore requires evaluating interventions across diverse prompts and constraint regimes rather than in a single fixed context.

4 Conclusion

We have shown that large language models develop systematic internal dynamics during few-shot prompt processing that are not explicitly trained for, yet emerge consistently across model families and scales. By using sparse autoencoders as an analytical lens, we make these latent dynamics directly observable and characterize three recurring phenomena: the decimal periodicity effect (DPE), local feature homeostasis (LFH), and adaptive feature redistribution (AFR).

Together, these findings demonstrate the value of sparse autoencoders for uncovering hidden but structured representational dynamics in LLMs. At the same time, they also highlight important methodological caveats for SAE-based interpretability: feature activations can be shaped not only by task semantics, but also by prompt position, local exemplar history, and contextual constraints.

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 different 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, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

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. 2025. Revisiting in-context learning inference circuit in large language models. In *International Conference on Learning Representations*, volume 2025, pages 13081–13117.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

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*.

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.

Tom Lieberum, Senthooran 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. *arXiv preprint arXiv:2408.05147*.

Johnny Lin. 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 Jermyn, 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 Parikh, 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*.

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*.

Sewon Min, Xinxi Lyu, Junifer Arihashi, 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 (EMNLP)*, 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*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Lee Sharkey and Dan Braun Beren. 2022. [interim research report] taking features out of superposition with sparse autoencoders.

Alibaba Cloud Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Language Model Interpretability Team. 2024. Gemma scope: Helping the safety community shed light on the inner workings of language models.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

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. *arXiv preprint arXiv:2305.14160*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems (NIPS)*, pages 649–657.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*.

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 Figures 5, 6, and 7.

Prompt used for local feature homeostasis (LFH) For studying LFH, we fix all the components except the last few-shot exemplar. An example prompt set is shown in Figures 8 and 9, where only example 7 is different.

B Details on the Hypothesis Tests Used in the Paper

For each feature, we formed paired activation differences $d_i = a_i^{controlled} - a_i^{original}$ 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

$$r_{rb} = \frac{2W_+}{R} - 1 \tag{3}$$

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 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.

AGNews Few-Shot Prompt (excerpt)

```
<bos><start_of_turn>user
Pretend that you are an expert in news topic classification. For a given news article, you have to
assess the topic, determining whether it is world, sports, business, or technology.

Example 1:
News article:
Bush: Kerry Proposals Will Hurt Economy WASHINGTON - President Bush renewed his criticism of
rival John Kerry's domestic proposals Friday as his campaign prepared a new television and warning
that liberals in Congress would join Kerry in harming job growth and the economy by raising taxes on
small businesses.  "To make sure the economy continues to grow, we'll be wise about how we spend
the money - the people's money...
Topic:
World

Example 2:
News article:
Deutsche rides Germany #39;s tepid recovery FRANKFURT Deutsche Bank said Friday that third quarter
net income rose a better-than-expected 18 percent from a year earlier, to 680 million, helped by the
sale of property in New York that was damaged by the Sept.
Topic:
Business

Example 3:
News article:
Oil Scales New High as Norway Strike Hits (Reuters) Reuters - Oil scaled a new high above #36;55
a barrel\Monday as a threat by Norwegian ship owners to halt production\from the world's third
biggest exporter stoked fears of a\winter supply crunch.
Topic:
Business

Example 4:
News article:
Irish invite for De Lima The Brazilian runner jostled by an Irish-born ex-priest in the Olympic marathon
has been invited to Ireland by the country #39;s Sports Minister.
Topic:
Sports
```

Figure 5: **AGNews Prompt Template.**

C Extended Decimal Periodicity Effect (DPE) Results

C.1 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., 2025). As illustrated in Figure 13, we find that the identical modulo-10 periodicity persists within the representations of these label words (Figure 13). This confirms that the effects of DPE are spread more broadly across tokens in the prompt, not merely at formatting boundaries.

C.2 DPE Results on Additional Models

Figures 10, 11, and 12 provide DPE results on Gemma2-2B, Gemma2-9B, and Qwen3-32B, respectively. Llama3.1-8B-Base results are provided in Figure 13.

D Additional Results for Local Feature Homeostasis (LFH)

In this section, we provide extended empirical results to further characterize the dynamics and ubiquity of Local Feature Homeostasis (LFH):

- **Instance-Level Case Study:** Figure 16 presents a qualitative visualization of the LFH mechanism in `Gemma2-9B-IT`. It illustrates an "activation flip" phenomenon by comparing two prompts that terminate in an identical test query. In Prompt 2, a strong feature activation immediately preceding the test query heavily suppresses the response on the query itself. Conversely, Prompt 1 lacks this prior stimulus and consequently exhibits a robust response

on the exact same query, demonstrating how prior context inversely regulates subsequent activation magnitudes.

- **Generality Across Tasks:** Figure 17 evaluates LFH across two alternative task paradigms—tool calling (BFCL) and question answering (NQ)—confirming that the phenomenon generalizes broadly across diverse contexts.

- **Prevalence in Base Models:** Figure 18 provides LFH evaluation results on base architectures, verifying that these homeostatic dynamics emerge natively during pre-training rather than as an artifact of instruction tuning.

E Detailed Results on Adaptive Feature Redistribution (AFR) in LLMs

We present a comprehensive analysis of AFR mechanism in Figures 20 and 21. 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 models we evaluate.

Figure 22 summarizes the causal role reassignment associated with the AFR effect in `Gemma2-2B-IT` and `Llama3.1-8B-Instruct`. 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 19.

BFCL Prompt (excerpt)

```
You are a helpful assistant with access to a set of tools.
Given the user's request, choose the most appropriate tool
and provide the tool call.

Available tools:

Tool 1:
   name: get_weather
   arguments:
     - city: string
     - date: string
   Description: Get the weather forecast for a city on a given date.

Tool 2:
   name: calculator
   arguments:
     - expression: string
   Description: Evaluate a mathematical expression.

Tool 3:
   name: search_flights
   arguments:
     - origin: string
     - destination: string
     - date: string
   Description: Search for available flights.

Tool 4:
   name: restaurant_search
   arguments:
     - city: string
     - cuisine: string
   Description: Find restaurants by location and cuisine.

User request:
"What will the weather be like in Boston tomorrow?"
```

Illustrative excerpt showing 4 tools rather than the full prompt.

Figure 6: **BFCL Prompt Template.**

Natural Questions Prompt (excerpt)

```
You are a careful retrieval-augmented question answering
assistant. You are given a question and an ordered list of
retrieved Wikipedia documents.

Question:
does a kite have 2 pairs of parallel sides

Retrieved documents:

Document 1:
 - title: The Princess Bride (film)
 - text: Gene Siskel and Roger Ebert gave the film a "two
   thumbs up" rating on their television program. Ebert also
   wrote a favorable print review in the Chicago Sun-Times ...

Document 2:
 - title: Tri-Cities, Washington
 - text: From 2005 to 2016, there was also a professional
   indoor football team called the Tri-Cities Fever. The team
   came to the Tri-Cities in 2005 as an expansion team ...

Document 3:
 - title: Flood control
 - text: Floods are caused by many factors or combinations of
   factors, including prolonged heavy rainfall, accelerated
   snowmelt, severe winds over water, unusual high tides ...

Document 4:
 - title: Kite (geometry)
 - text: In Euclidean geometry, a kite is a quadrilateral whose
   four sides can be grouped into two pairs of equal-length
   adjacent sides. Unlike a parallelogram, a kite does not
   generally have two pairs of parallel sides ..."
```

*Illustrative excerpt showing **4** retrieved documents rather than the full prompt.*

Figure 7: **Natural Questions Prompt Template.**

<bos><start_of_turn>user
Pretend that you are an expert in news topic classification. For a given news article, you have to assess the topic, determining whether it is world, sports, business, or technology.

Example 1:
News article:
Bush: Kerry Proposals Will Hurt Economy WASHINGTON - President Bush renewed his criticism of rival John Kerry's domestic proposals Friday as his campaign prepared a new television ad warning that liberals in Congress would join Kerry in harming job growth and the economy by raising taxes on small businesses. "To make sure the economy continues to grow, we'll be wise about how we spend the money - the people's money...
Topic:
World

Example 2:
News article:
Deutsche rides Germany #39;s tepid recovery FRANKFURT Deutsche Bank said Friday that third quarter net income rose a better-than-expected 18 percent from a year earlier, to 680 million, helped by the sale of property in New York that was damaged by the Sept.
Topic:
Business

Example 3:
News article:
Oil Scales New High as Norway Strike Hits (Reuters) Reuters - Oil scaled a new high above #36;55 a barrel\Monday as a threat by Norwegian ship owners to halt production\from the world's third biggest exporter stoked fears of a\winter supply crunch.
Topic:
Business

Example 4:
News article:
Irish invite for De Lima The Brazilian runner jostled by an Irish-born ex-priest in the Olympic marathon has been invited to Ireland by the country #39;s Sports Minister.
Topic:
Sports

Example 5:
News article:
Pak. scribes on visit to J amp;K Wagah (Amritsar), Oct. 4. (PTI): A 15-member delegation of Pakistani journalists on Sunday crossed over to India on their way to perhaps a first ever visit to Jammu and Kashmir by a team of scribes from that country.
Topic:
World

Example 6:
News article:
Space shuttle delayed again The space shuttle will not fly before May 2005, according to NASA officials. This pushes the shuttle #39;s return-to-flight schedule back by two months, and postpones a vital servicing mission to the International Space Station.
Topic:
Technology

Example 7:
News article:
Teacher: On tape, prince says he cheated READING, England -- A teacher #39;s tape recording of Prince Harry purportedly admitting he cheated on an art exam is insufficient to warrant a renewed investigation into the incident, an examination board said Friday.
Topic:
World

Example 8:
News article:
Polish Hostage Freed in Iraq Already in Warsaw WARSAW (Reuters) - A Polish woman kidnapped in Iraq last month has been freed and flown to Poland and said she was treated well, raising hopes for other foreign hostages.
Topic:
World

Figure 8: **Prompt used in local feature homeostasis (LFH).**

<bos><start_of_turn>user
Pretend that you are an expert in news topic classification. For a given news article, you have to assess the topic, determining whether it is world, sports, business, or technology.

Example 1:
News article:
Bush: Kerry Proposals Will Hurt Economy WASHINGTON - President Bush renewed his criticism of rival John Kerry's domestic proposals Friday as his campaign prepared a new television ad warning that liberals in Congress would join Kerry in harming job growth and the economy by raising taxes on small businesses. "To make sure the economy continues to grow, we'll be wise about how we spend the money - the people's money...
Topic:
World

Example 2:
News article:
Deutsche rides Germany #39;s tepid recovery FRANKFURT Deutsche Bank said Friday that third quarter net income rose a better-than-expected 18 percent from a year earlier, to 680 million, helped by the sale of property in New York that was damaged by the Sept.
Topic:
Business

Example 3:
News article:
Oil Scales New High as Norway Strike Hits (Reuters) Reuters - Oil scaled a new high above #36;55 a barrel\Monday as a threat by Norwegian ship owners to halt production\from the world's third biggest exporter stoked fears of a\winter supply crunch.
Topic:
Business

Example 4:
News article:
Irish invite for De Lima The Brazilian runner jostled by an Irish-born ex-priest in the Olympic marathon has been invited to Ireland by the country #39;s Sports Minister.
Topic:
Sports

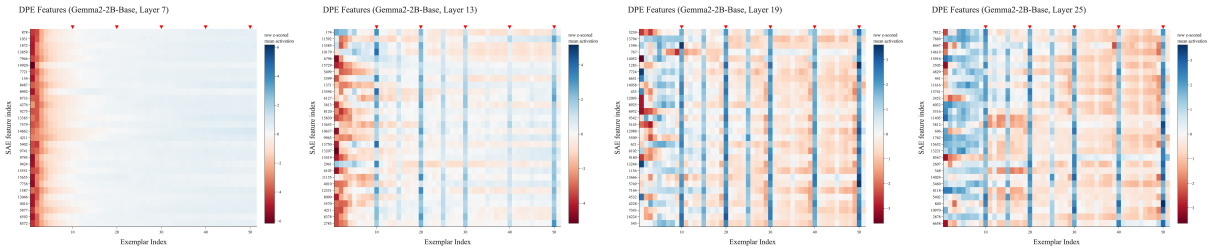
Example 5:
News article:
Pak. scribes on visit to J amp;K Wagah (Amritsar), Oct. 4. (PTI): A 15-member delegation of Pakistani journalists on Sunday crossed over to India on their way to perhaps a first ever visit to Jammu and Kashmir by a team of scribes from that country.
Topic:
World

Example 6:
News article:
Space shuttle delayed again The space shuttle will not fly before May 2005, according to NASA officials. This pushes the shuttle #39;s return-to-flight schedule back by two months, and postpones a vital servicing mission to the International Space Station.
Topic:
Technology

Example 7:
News article:
US to Question China About Bush Comment WASHINGTON - The State Department will ask the Chinese government about comments by a former senior Chinese official who accused President Bush of trying to quot;rule over the whole world.
Topic:
World

Example 8:
News article:
Polish Hostage Freed in Iraq Already in Warsaw WARSAW (Reuters) - A Polish woman kidnapped in Iraq last month has been freed and flown to Poland and said she was treated well, raising hopes for other foreign hostages.
Topic:
World

Figure 9: **Prompt used in local feature homeostasis (LFH) (2).**

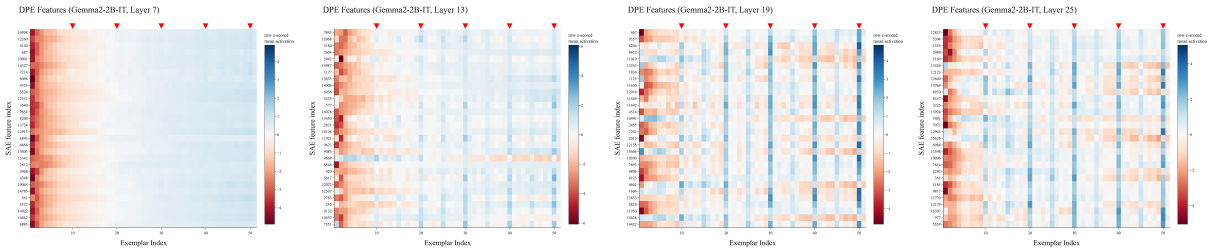


(a) Gemma2-2B-Base, Layer 7

(b) Gemma2-2B-Base, Layer 13

(c) Gemma2-2B-Base, Layer 19

(d) Gemma2-2B-Base, Layer 25



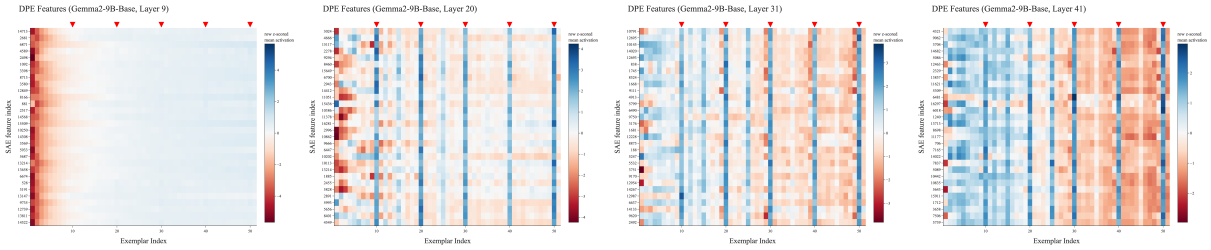
(e) Gemma2-2B-IT, Layer 7

(f) Gemma2-2B-IT, Layer 13

(g) Gemma2-2B-IT, Layer 19

(h) Gemma2-2B-IT, Layer 25

Figure 10: **DPE Results on Gemma2-2B-Base and Gemma2-2B-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.

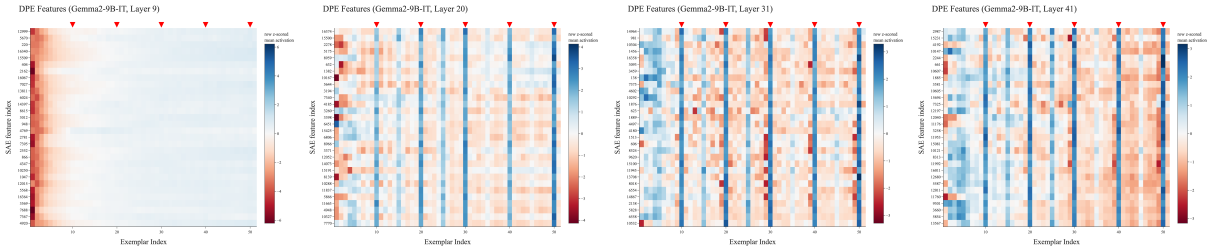


(a) Gemma2-9B-Base, Layer 9

(b) Gemma2-9B-Base, Layer 20

(c) Gemma2-9B-Base, Layer 31

(d) Gemma2-9B-Base, Layer 41



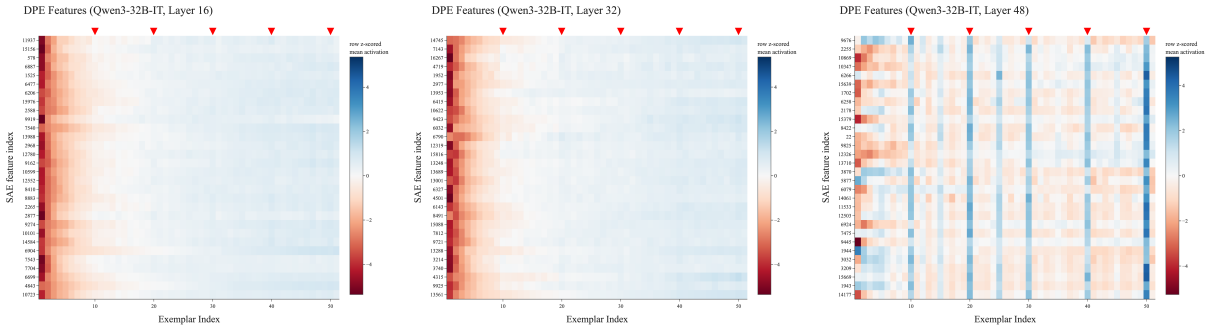
(e) Gemma2-9B-IT, Layer 9

(f) Gemma2-9B-IT, Layer 20

(g) Gemma2-9B-IT, Layer 31

(h) Gemma2-9B-IT, Layer 41

Figure 11: **DPE Results on Gemma2-9B-Base and Gemma2-9B-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.

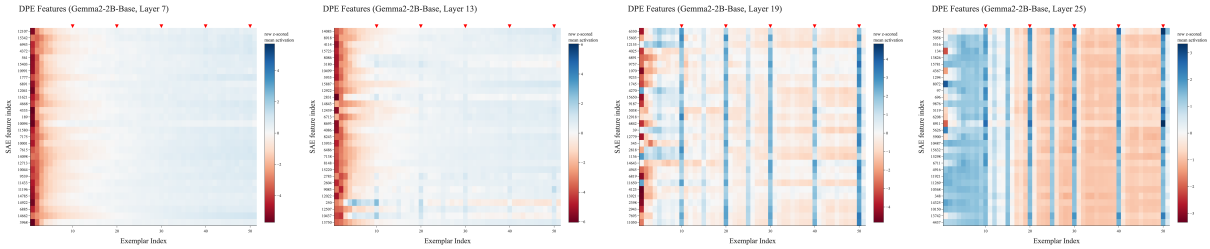


(a) Qwen3-32B-IT, Layer 16

(b) Qwen3-32B-IT, Layer 32

(c) Qwen3-32B-IT, Layer 48

Figure 12: **DPE Results on 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.

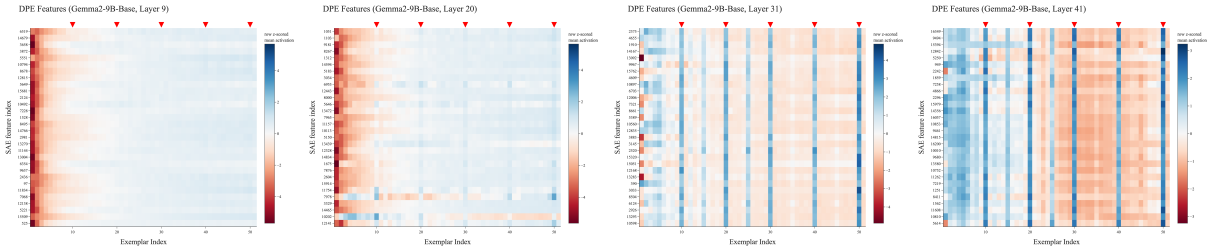


(a) Gemma2-2B-Base Label Word, Layer 7

(b) Gemma2-2B-Base Label Word, Layer 13

(c) Gemma2-2B-Base Label Word, Layer 19

(d) Gemma2-2B-Base Label Word, Layer 25

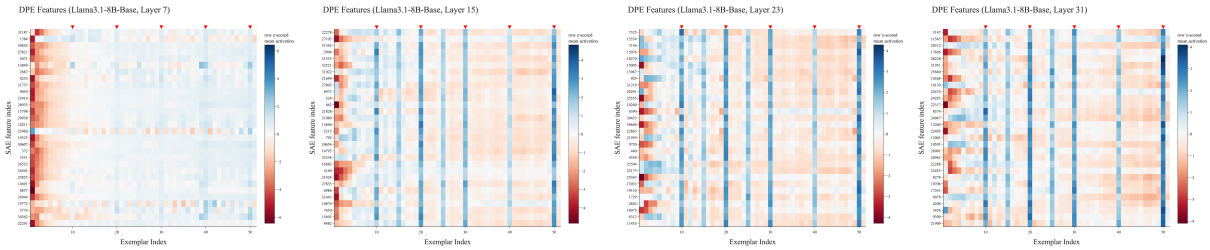


(e) Gemma2-9B-Base Label Word, Layer 9

(f) Gemma2-9B-Base Label Word, Layer 20

(g) Gemma2-9B-Base Label Word, Layer 31

(h) Gemma2-9B-Base Label Word, Layer 41



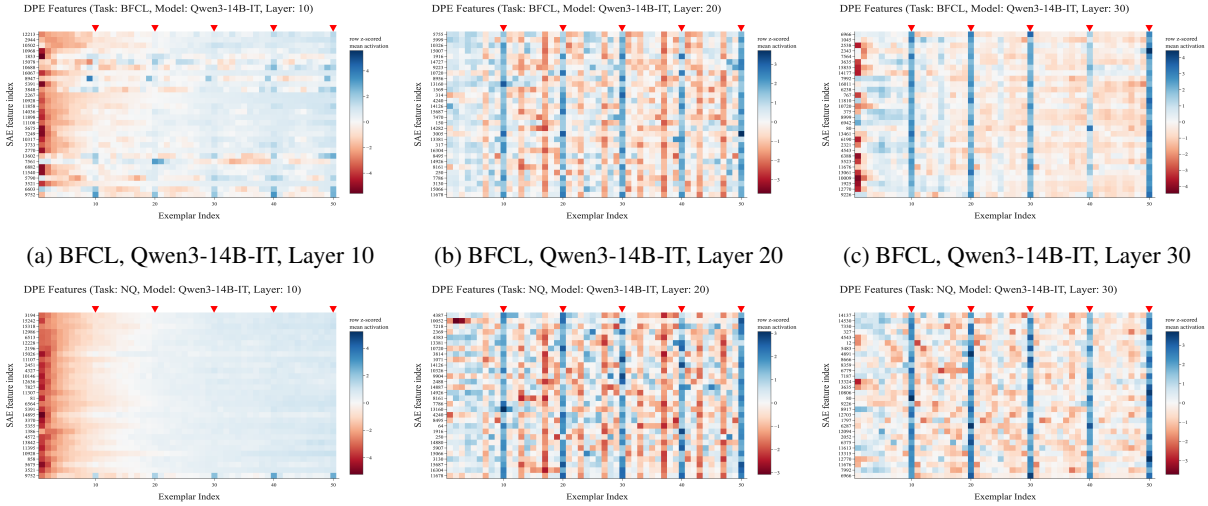
(i) Llama3.1-8B-Base Label Word, Layer 7

(j) Llama3.1-8B-Base Label Word, Layer 15

(k) Llama3.1-8B-Base Label Word, Layer 23

(l) Llama3.1-8B-Base Label Word, Layer 31

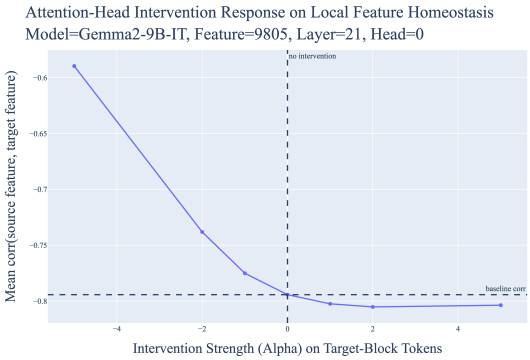
Figure 13: **Persistence of DPE across varied token positions.** Evaluated on Gemma2-2B-Base, Gemma2-9B-Base, and Llama3.1-8B-Base, the decimal periodicity effect is not isolated to the block-final token but extends to alternative positions within the sequence. Illustrated figures measure activations at the corresponding label words, demonstrating that this positional periodicity remains robust to the specific probe placement.



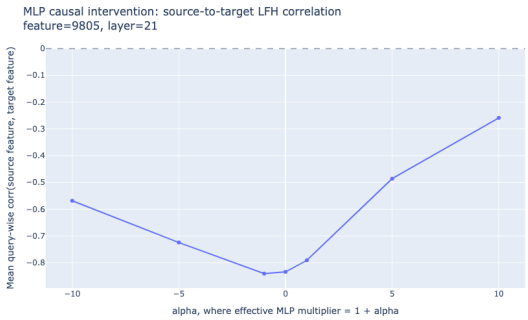
(a) BFCL, Qwen3-14B-IT, Layer 10 (b) BFCL, Qwen3-14B-IT, Layer 20 (c) BFCL, Qwen3-14B-IT, Layer 30

(d) NQ, Qwen3-14B-IT, Layer 10 (e) NQ, Qwen3-14B-IT, Layer 20 (f) NQ, Qwen3-14B-IT, Layer 30

Figure 14: **The Decimal Periodicity Effect (DPE) on Tool-Calling (BFCL) and Retrieval Augmented Generation (RAG) Question Answering (NaturalQuestions).**



(a) Attention Head can both weaken and strengthen the LFH effect.



(b) MLP cannot effectively strengthen the LFH effect.

Figure 15: **Additional Causal Results on LFH, using Gemma2-9B-IT.** Likewise, intervening on attention heads can both weaken and strengthen the LFH effect in expected ways, while MLP cannot.

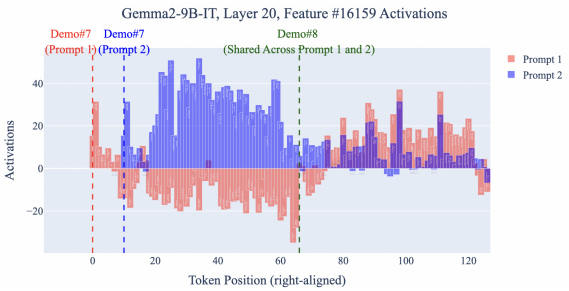
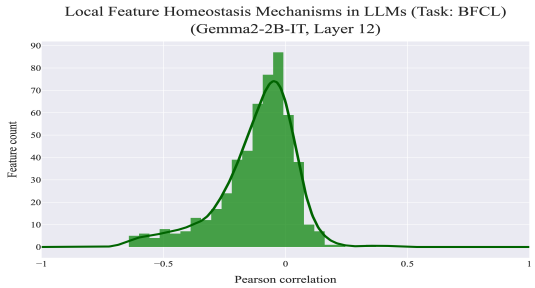
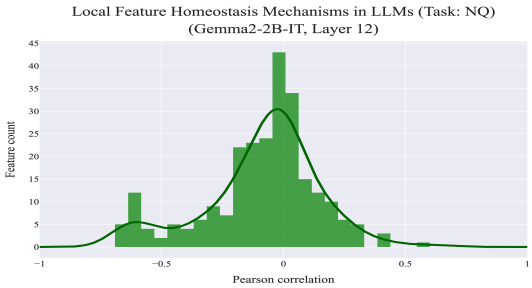


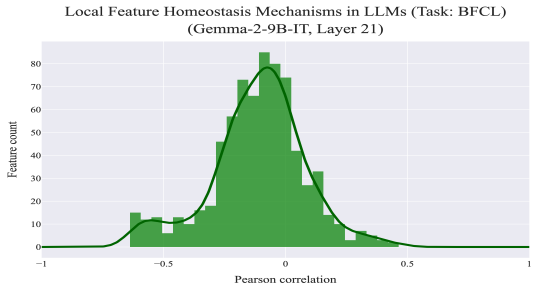
Figure 16: **The figure illustrates what we term *local feature homeostasis* behavior.** We show the token-level activation trajectories of a representative feature under two 8-shot prompts, isolating the final two exemplars (Demos 7 and 8). While the preceding exemplar (Demo 7) differs between the two prompts, both share an identical target exemplar (Demo 8), shown on the right with its onset marked in green (tokens are right-aligned). In Prompt 1 (red), the feature remains largely inactive during Demo 7, whereas in Prompt 2 (blue), it exhibits strong positive activation. Despite the target exemplar being exactly the same, the feature demonstrates inverse dynamics: weaker prior activation leads to a stronger response on the target, while stronger prior activation elicits a weaker response. The corresponding prompts are provided in Figures 8 and 9 in Appendix A.



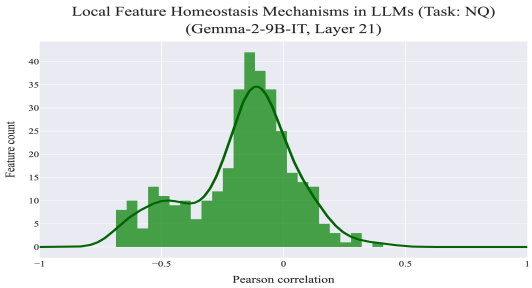
(a) Gemma2-2B-IT, BFCL



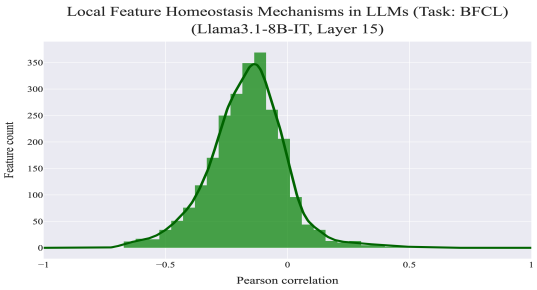
(b) Gemma2-2B-IT, NQ



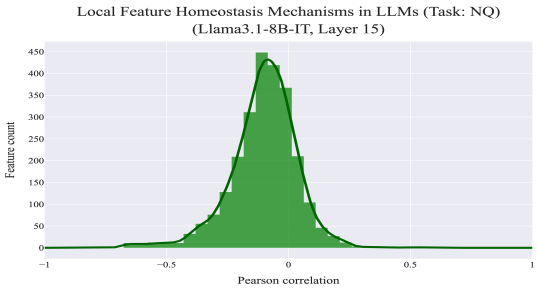
(c) Gemma2-9B-IT, BFCL



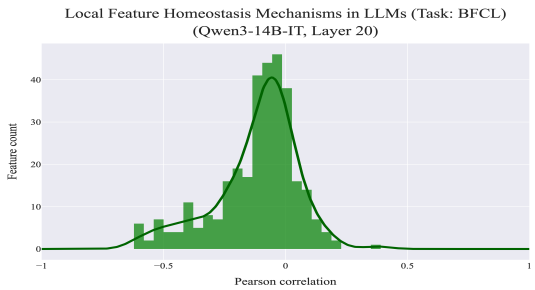
(d) Gemma2-9B-IT, NQ



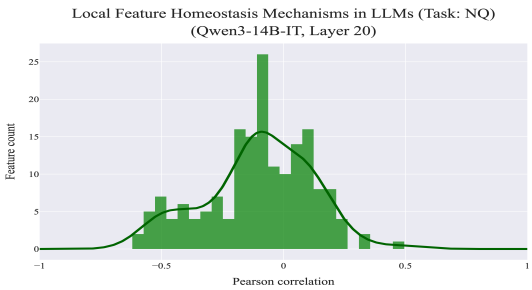
(e) Llama3.1-8B-IT, BFCL



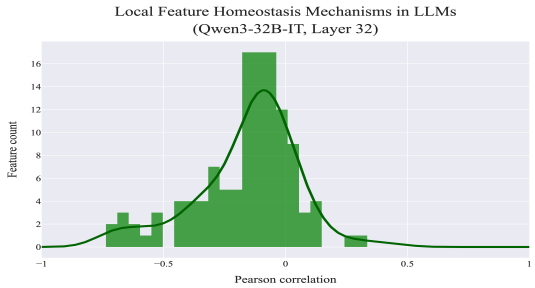
(f) Llama3.1-8B-IT, NQ



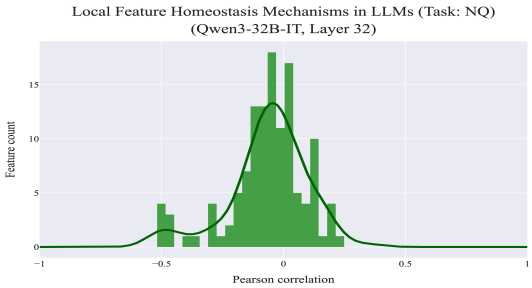
(g) Qwen3-14B-IT, BFCL



(h) Qwen3-14B-IT, NQ

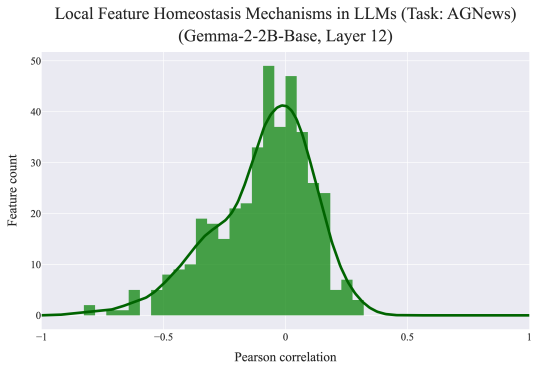


(i) Qwen3-32B-IT, BFCL

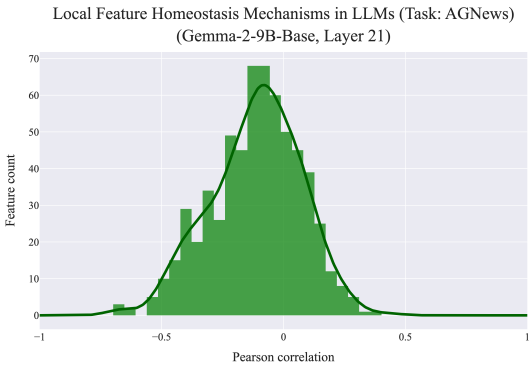


(j) Qwen3-32B-IT, NQ

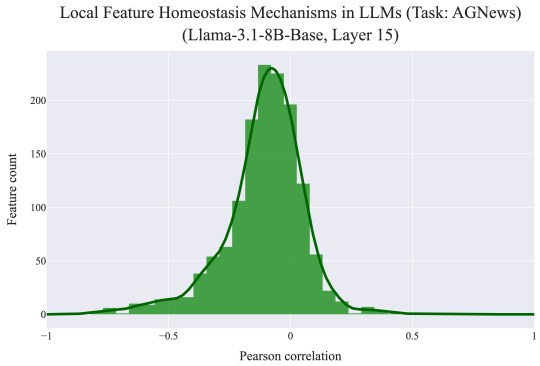
Figure 17: **Local Feature Homeostasis (LFH) across BFCL and NQ.** We use the instruction-tuned models as representatives.



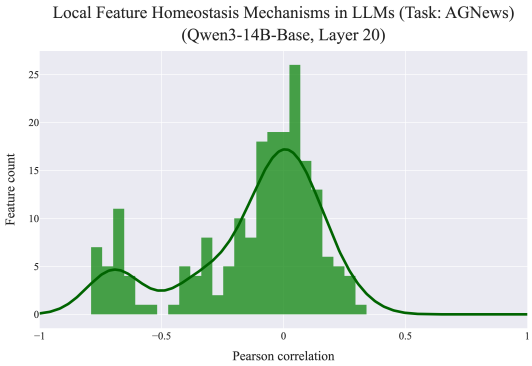
(a) Gemma2-2B-Base, AGNews



(b) Gemma2-9B-Base, AGNews

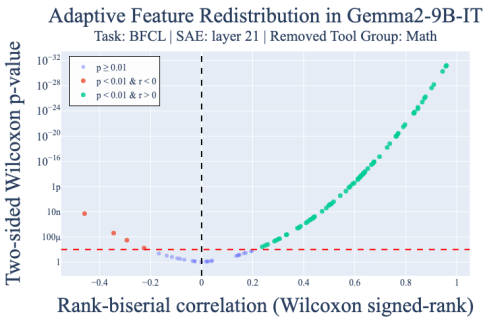


(c) Llama3.1-8B-Base, AGNews

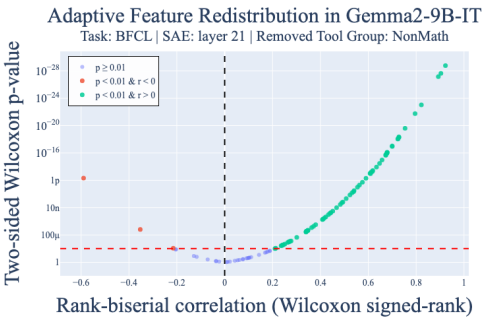


(d) Qwen3-14B-Base, AGNews

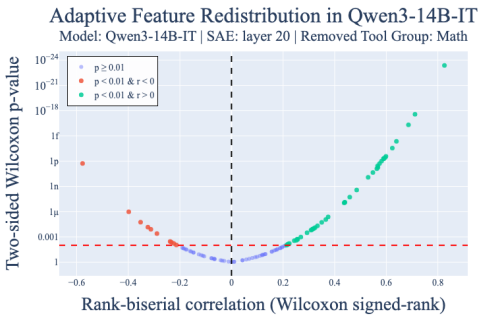
Figure 18: **Local Feature Homeostasis (LFH) across base models.** Results are evaluated on the AGNews dataset using Gemma2-2B-Base, Gemma2-9B-Base, Llama3.1-8B-Base, and Qwen3-14B-Base.



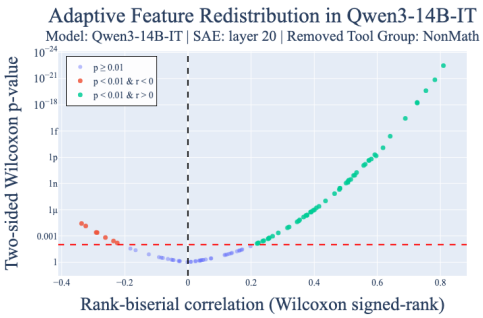
(a) `Gemma2-9B-IT, L21, Removing "Math"`



(b) `Gemma2-9B-IT, L21, Removing "Non-Math"`

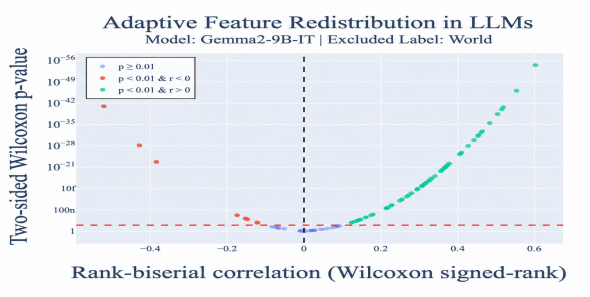


(c) `Qwen3-14B-IT, L20, Removing "Math"`

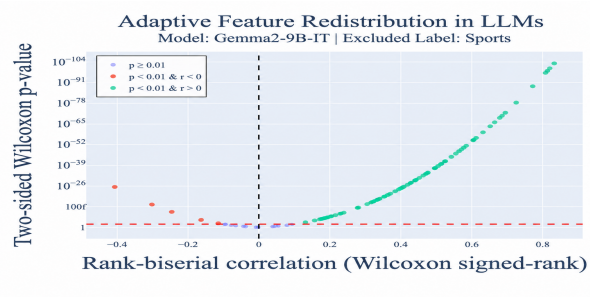


(d) `Qwen3-14B-IT, L20, Removing "Non-Math"`

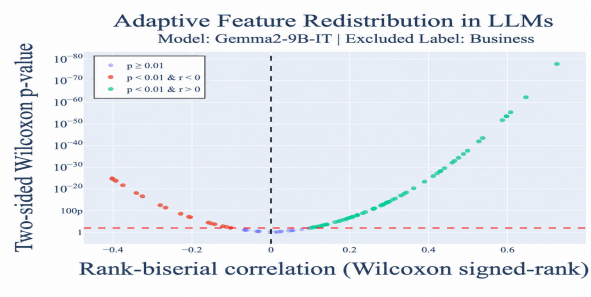
Figure 19: **AFR effect across the BFCL task `Gemma2-9B-IT` and `Qwen3-14B-IT`.** Consistent patterns are observed.



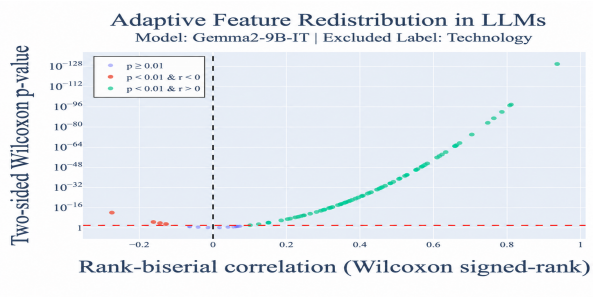
(a) `Gemma2-9B-IT, L20, Removing "World"`



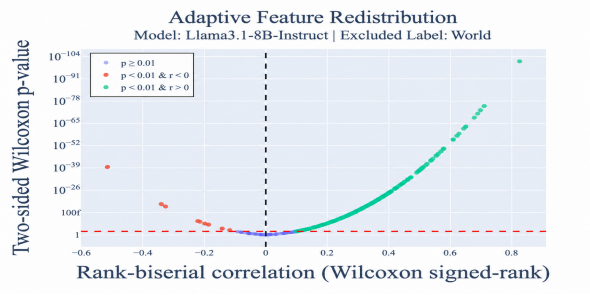
(b) `Gemma2-9B-IT, L20, Removing "Sports"`



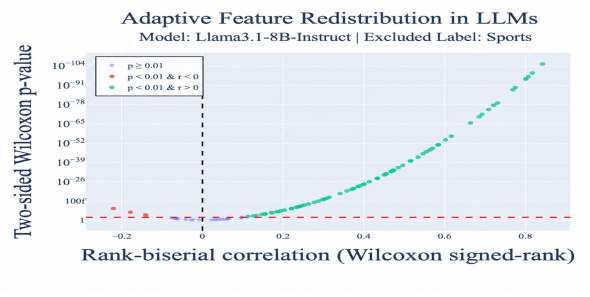
(c) `Gemma2-9B-IT, L20, Removing "Business"`



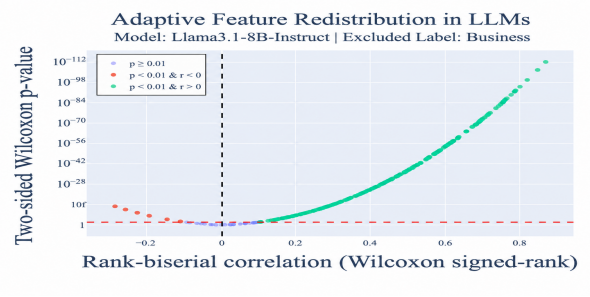
(d) `Gemma2-9B-IT, L20, Removing "Technology"`



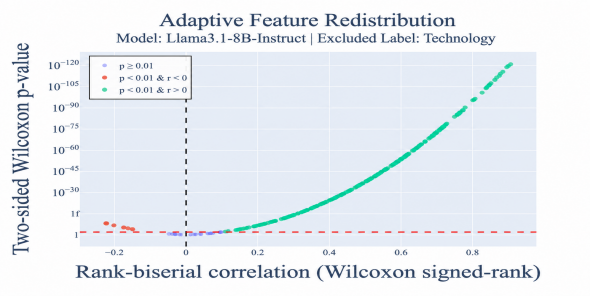
(e) `Llama3.1-8B-IT, L15, Removing "World"`



(f) `Llama3.1-8B-Instruct, L15, Removing "Sports"`

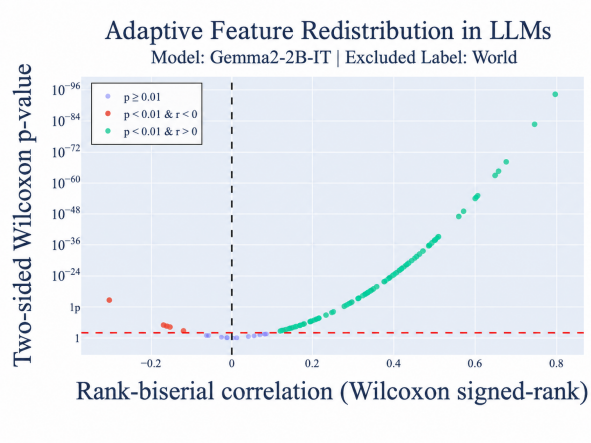


(g) `Llama3.1-8B-IT, L15, Removing "Business"`

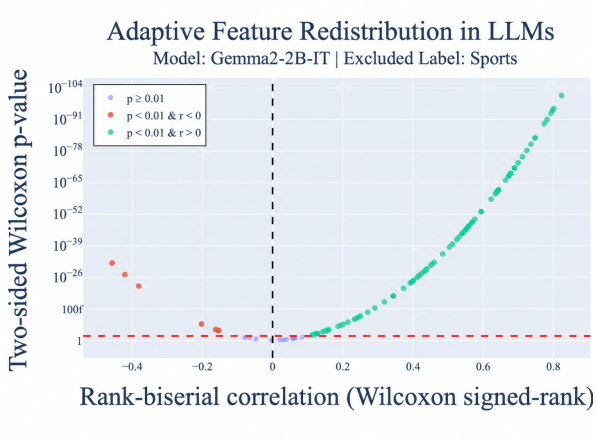


(h) `Llama3.1-8B-Instruct, L15, Removing "Technology"`

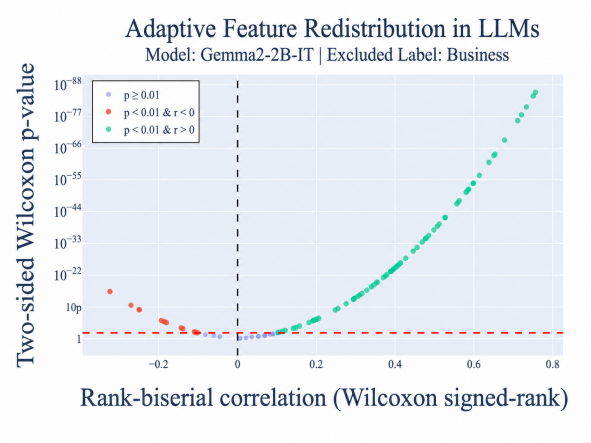
Figure 20: **Adaptive Feature Redistribution in `Gemma2-9B-IT` and `Llama3.1-8B-Instruct`.** Consistent patterns are observed across model families.



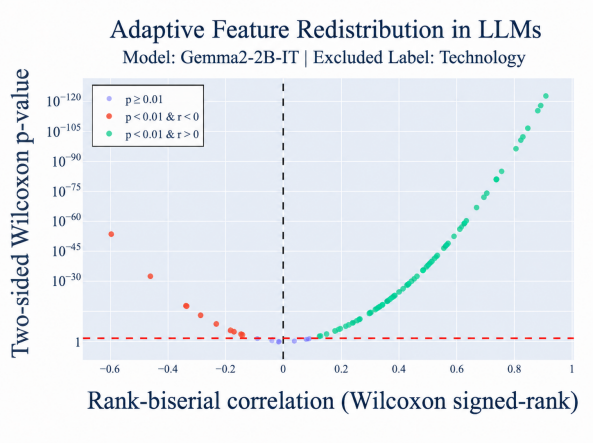
(a) `Gemma2-2B-IT, L12, Removing "World"`



(b) `Gemma2-2B-IT, L12, Removing "Sports"`

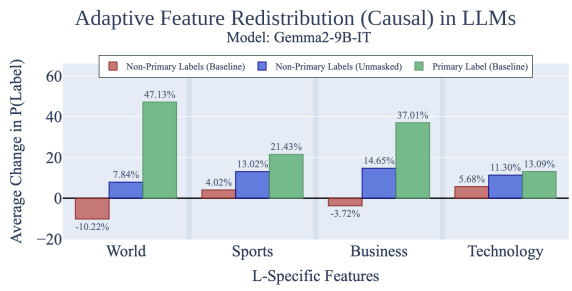


(c) `Gemma2-2B-IT, L12, Removing "Business"`

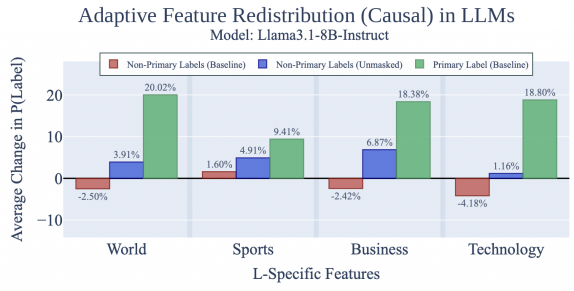


(d) `Gemma2-2B-IT, L12, Removing "Technology"`

Figure 21: **Adaptive Feature Redistribution in `Gemma2-2B-IT`.** Consistent patterns are observed across different labels.



(a) `Gemma2-9B-IT`



(b) `Llama3.1-8B-Instruct`

Figure 22: **Causal Intervention Results on `Gemma2-9B-IT` and `Llama3.1-8B-Instruct`.** Due to computational constraints, we restrict this analysis to models with fewer than 10 billion parameters.