Localised Routing, Distributed Execution: How LLMs Process Semantic Domains

Anonymous ACL submission

Abstract

Transformer language models excel across diverse semantic domains, yet how domain-specific processing is internally organised remains poorly understood. Mechanistic interpretability has established that attention layers route information and MLP layers store knowledge at the level of individual tokens and facts. We ask whether the same functional division holds at the coarser level of entire semantic domains. Using arXiv abstracts from six academic disciplines and operationalising domain identity through *domain concept vectors*, we show that both components carry comparable overall domain information, but in qualitatively distinct patterns: attention concentrates domain signal at specific layers while MLP distributes it monotonically across depth. Fine-tuning analysis shows MLP weight updates are $2.95\times$ more aligned with domain concept directions than attention, with the gate projection dominant across all adaptation metrics. Activation patching shows that patching attention at its peak layers causally affects domain shift more than non-peak layers. Together, these results establish that the attention/MLP functional division scales to semantic domain processing across multiple LLMs, with routing localised to identifiable attention layers, execution distributed across deep MLP computation.

1 Introduction

Large language models perform well across many domains — mathematics, physics, medicine, law. This raises a mechanistic question: how is domain-specific processing organised internally? At the level of individual tokens and facts, the answer is partially known. Attention heads route information across sequence positions (Elhage et al., 2021), with specialised heads identified as relation extractors and information movers (Yao et al., 2024). MLP layers function as key-value memories, storing associations between input patterns and output distributions (Geva et al., 2021), with factual recall causally localised to specific mid-layer MLP modules (Meng et al., 2022; Dai et al., 2022). Their roles are complementary: attention contextual patterns steer downstream MLP computation toward specific output directions (Neo et al., 2024; Xu and Chen, 2025).

However, a coarser question is yet to be addressed. How does this functional division hold at the level of *entire semantic domains*? Domain-level processing involves coherence across an entire text, having structural patterns and abstract conceptual organisation. Whether the attention/MLP division organises this kind of computation, and if so, whether it is localised or diffuse, is an open question. The token-level results can show that the division is an artifact of how gradients flow through different architectures, but not a principled organisational structure that scales with abstraction.

We use arXiv abstracts from six disciplines: computer science, electrical engineering, mathematics, physics, quantitative biology, and statistics (Clement et al., 2019). We operationalise domain identity through **domain concept vectors** (Park et al., 2025) (the mean-shifted activation direction for each domain at each layer) and use this single object consistently across all three experiments and three models.

Three experiments establish converging evidence. **Probing and separability** analysis shows that attention and MLP carry comparable overall domain information (linear probe accuracy 0.841 vs. 0.844), but in qualitatively different patterns. Attention is *spiky*: domain signal concentrates sharply at a small number of specific layers. MLP is *smooth*: domain signal increases monotonically across depth. The two components differ not in how much domain information they carry, but in how that information is distributed. **Fine-tuning adaptation** analysis shows that MLP weight updates during domain-specific fine-tuning are more

aligned with domain concept directions than attention updates, and are more domain-differentiated across domains. **Causal activation patching** provides the most direct evidence. Patching attention contributions along the domain concept direction at its peak layers produces much more domain shift in the final residual stream than patching non-peak layers. Downstream of the patch, MLP sublayers exclusively amplify the injected signal; downstream attention sublayers are passive, contributing near-zero incremental shift.

These results establish a unified account. The attention/MLP functional division is not confined to token-level processing, but scales to semantic domains. This finding is relevant to how we think about mechanistic interpretability at scale. Token-level analyses have demonstrated that transformers contain identifiable circuits operating on individual associations. Our results show that a higher-order version of the same organisational principle is at work when models process entire semantic domains. The findings remain consistent across Qwen2.5-7B, Pythia 2.8B and LLaMA 3.2-3B models. In summary, our contributions are:

- We introduce domain concept vectors as a unified mathematical object and use them consistently across three distinct experiments, connecting representational structure, adaptive behaviour, and causal control.
- We show that attention and MLP carry comparable overall domain information, but in qualitatively different layer-wise patterns: attention is localised and spiky, MLP is smooth and monotonically distributed.
- We provide causal evidence that attention hotspot layers control domain identity and that downstream MLP sublayers are the exclusive amplifiers of the routed signal.

2 Related Work

Attention vs MLP. The token-level division of labour between attention and MLPs is reviewed in the Introduction. All prior evidence operates at the level of individual tokens, facts, or circuits (Elhage et al., 2021; Geva et al., 2021, 2022; Meng et al., 2022; Dai et al., 2022; Yao et al., 2024; Neo et al., 2024; Xu and Chen, 2025). Whether this functional split extends to coarser semantic organisation, such as entire topic domains, has not been tested.

Concept Vectors and Geometric Structure. Our domain concept vectors rest on the linear representation hypothesis: concepts are encoded as directions in activation space (Mikolov et al., 2013), with causally separable concepts occupying orthogonal directions and cleanly separated categories forming a regular simplex (Park et al., 2023, 2025). Linear structure at the level of broad semantic features has been confirmed for sentiment (Tigges et al., 2023) and political ideology (Kim et al., 2025), both localised to middle layers, consistent with evidence that intermediate layers encode higher-level abstractions (Van Aken et al., 2019) and that activation geometry correlates with prediction difficulty (Viswanathan et al., 2025). Simplex-like structures have been linked to constrained Bayesian belief updating in attention heads (Piotrowski et al., 2025), and latent structure is extractable from activations via linear methods (Burns et al., 2022), though deep representations can diverge from human-interpretable similarity (Nguyen et al., 2015). Whether this geometric framework extends to domain-level organisation, and whether attention and MLP contributions differ, is what our experiments address.

Fine-Tuning Analysis. Pre-trained models lie on low-dimensional manifolds enabling effective adaptation via low-rank decomposition (Aghajanyan et al., 2021; Hu et al., 2022). Adaptation methods typically target attention despite evidence that selective MLP fine-tuning can be stronger (Chekalina et al., 2024), and weight analyses reveal distinct update patterns between LoRA and full fine-tuning (Liu et al., 2024). Fine-tuning induces structured, composable task vectors whose effectiveness depends on weight disentanglement (Ilharco et al., 2022; Ortiz-Jimenez et al., 2023), transferable across model versions within linearly connected basins (Lin et al., 2025). At the representation level, fine-tuning effects are highly localised (Jain et al., 2024; Yin et al., 2024), and an "imprint and shift" mechanism describes how token-level associations are written into MLP weights (Katz et al., 2024). However, all prior analyses operate at the token or fact level; whether MLP updates during domain-specific adaptation are differentiated across domains and aligned with domain concept directions has not been tested.

Activation Patching. Activation patching isolates functional computation inside transformers, from early causal tracing of facts to MLPs (Meng et al., 2022) and bias to attention heads (Vig et al., 2020), through finer-grained path patching (Goldowsky-Dill et al., 2023; Wang et al., 2022),

automated circuit discovery (Conmy et al., 2023), and gradient-based attribution patching (Syed et al., 2024). These build on causal abstraction (Geiger et al., 2022) and distributed alignment search (Geiger et al., 2024). However, replacing entire activation vectors mixes targeted signals with orthogonal information (Heimersheim and Nanda, 2024), and localisation often fails to predict downstream causal success (Hase et al., 2023; Todd et al., 2024). There is a need to track signal propagation.

To address these limitations, we patch along domain-concept directions, restrict the intervention to a single representational dimension, and track the domain-shift score at every intermediate layer to identify actual amplification.

3 Dataset and Concept Vectors

We study domain-level processing using arXiv abstracts from six academic disciplines: Computer Science (`cs`), Electrical Engineering (`eess`), Mathematics (`math`), Physics (`physics`), Quantitative Biology (`q-bio`), and Statistics (`stat`). We use 15,000 abstracts per domain (90,000 total), truncated to 384 tokens.

To make sure our analysis captures actual conceptual differences rather than superficial text patterns, all six datasets consist of formal academic abstracts that follow the same structure (problem, method, results). The important thing is whether the model's internal representations distinguish domains that *share* substantial vocabulary. As shown in Figure 1, the mean log-TF cosine similarity of word frequencies across all domains is 0.61, with the highest overlap between cs-eess (0.72) and stat-cs (0.71). Even the most distinctive domain, math, shares cosine similarity > 0.51 with every other domain. This high vocabulary overlap rules out trivial, vocabulary-based separation.

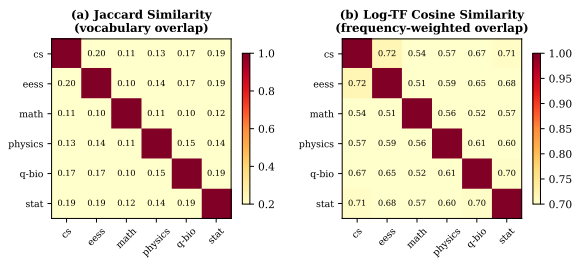

Figure 1: Vocabulary similarity between domains. **(a)** Jaccard similarity of unique token sets. **(b)** Log-TF cosine similarity of word frequency distributions.

Notation and Domain Concept Vectors. Let x^ℓ denote the residual stream activation at the input to layer ℓ. A standard transformer layer updates the residual stream via two additive components:

$$x^{\ell+1} = x^\ell + \mathrm{Attn}^\ell(x) + \mathrm{MLP}^\ell(x) \qquad (1)$$

For any component $c \in \{\mathrm{attn}, \mathrm{mlp}, \mathrm{resid}\}$, we define the **domain concept vector** for domain k at layer ℓ as the displacement from the global mean to the domain-specific mean:

$$\mathbf{d}_k^{c,\ell} = \mathbb{E}_{x\sim\mathcal{D}_k}[c^\ell(x)] - \mathbb{E}_x[c^\ell(x)] \qquad (2)$$

where $c^\ell(x)$ represents the output of that component for input x. This is the displacement from the global mean to the domain mean in the space of component c's activations at layer ℓ. It captures what is *specific* to domain k after removing what is common to all text. The concept vector is the simplest possible representation of domain identity in activation space. Whenever our analysis requires a purely directional axis, we denote the L_2-normalised unit vector with a hat: $\hat{\mathbf{d}}_k^{c,\ell} = \mathbf{d}_k^{c,\ell}/\|\mathbf{d}_k^{c,\ell}\|$. Because expectation is linear, these vectors satisfy an exact additive decomposition:

$$\mathbf{d}_k^{\ell+1} = \mathbf{d}_k^{\ell} + \mathbf{d}_k^{\mathrm{attn},\ell} + \mathbf{d}_k^{\mathrm{mlp},\ell} \qquad (3)$$

The domain signal in the residual stream at layer $\ell + 1$ is exactly the signal from the previous layer plus what attention added plus what MLP added. This identity connects all three of our experiments: we measure the same objects (concept vectors and their norms) through three different lenses.

Figure 2 shows t-SNE projections of residual stream activations at five layers. At layer 0, domains are partially mixed. By layer 22, the model has constructed well-separated clusters for each domain. Mathematics breaks away into distinct clusters the earliest. Whereas computer science and eess remain entangled with one another much deeper into the layers.

4 Probing and Separability

To characterise the division of labour between attention and MLP sublayers, we decompose each transformer layer's computation into its two additive contributions: the attention output $\mathrm{Attn}^\ell(x)$ and the MLP output $\mathrm{MLP}^\ell(x)$. We extract these contributions for 15,000 samples from each of six arXiv domains across all 28 layers of Qwen2.5-7B, using last-token pooling. To quantify how much domain-specific information these components carry at various depths, we evaluate them

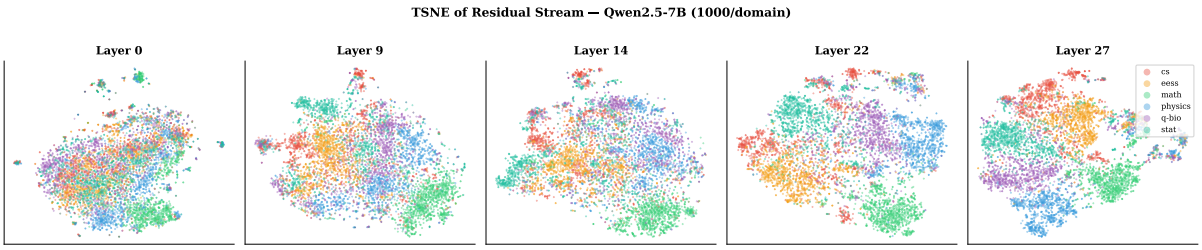

Figure 2: t-SNE projections of residual stream activations ($\mathbf{x}^\ell$) at selected layers, 1,000 samples per domain. Domain separation sharpens progressively with depth.

using two complementary methods. First, we use linear probing via a 6-way logistic regression to test whether the domain can be actively predicted from the activations. Second, we use the Fisher Discriminant Ratio (FDR) as a direct statistical metric(ratio of between-class to within-class scatter)as a measure of how cleanly the domain clusters naturally separate from one another, providing mathematical confirmation of our probing results.

Finding: Attention is spiky, MLP is smooth. Both analyses show a clear contrast in how attention and MLP handle domain information across the layers (Figure 3). The MLP contribution shows a smooth, increasing domain signal: probe accuracy rises steadily from 0.76 (layer 0) to 0.90 (layer 25), and FDR follows the same gradual ramp. The attention contribution, by contrast, is highly non-monotonic. Probe accuracy peaks early at layer 9 (0.89), drops to 0.77 at layer 17 (below even layer 0 levels), then spikes sharply to 0.91 at layer 22 before declining again. FDR confirms the identical pattern: attention exhibits sharp peaks at layers 2, 8-10, 19, and 22 separated by sharp drops. These two methods independently reproduce the same shape, establishing that attention's domain-relevant computation is concentrated at specific layers rather than distributed uniformly.

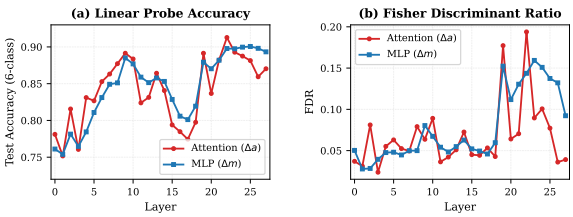

Figure 3: Domain classification from sublayer contributions across layers. **(a)** Linear probe accuracy (6-way logistic regression, 30% test split). **(b)** Fisher Discriminant Ratio (tr(S_B)/tr(S_W), higher = better separation).

Domain concept vectors reveal a second asymmetry (Figure 4). MLP concept vector norms grow rapidly with depth, reaching $\|\mathbf{d}_k\| \approx 80$ by

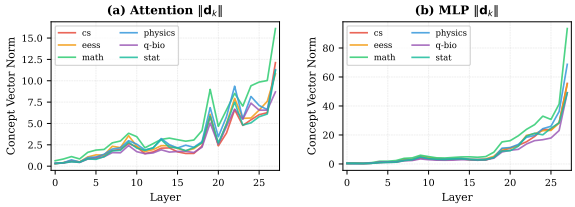

Figure 4: Domain concept vector norms $\|\mathbf{d}_k^\ell\|$ across layers for each of the six arXiv domains. **(a)** Attention contributions: norms remain small (0-16) with visible spikiness. **(b)** MLP contributions: norms grows rapidly (0-80).

layer 27, while attention concept norms remain smaller in magnitude ($\approx$ 16). Math is consistently the most separated domain in both components. Despite this large magnitude difference, unit-normalised FDR is virtually identical to raw FDR for both components (Figure 16, Appendix B), indicating that domain separation is driven by *direction*, not magnitude, in both sublayers. The magnitude difference reflects the overall scale of each sublayer's contribution rather than a qualitative difference in how domain information is encoded.

5 Fine-Tuning Analysis

To locate where the network permanently stores domain-specific concepts, we must look at how it learns. By tracking the exact weight modifications made during fine-tuning, we compare the adaptation patterns of both attention and MLP layers to see where the network prefers to write new domain-specific knowledge.

5.1 Fine-Tuning Setup

We fine-tune six independent models one for each arXiv domain starting from the exact same Qwen2.5-7B base checkpoint. We employ Low-Rank Adaptation LoRA (Hu et al., 2022) with rank $r = 16$ and $\alpha = 32$, targeting all query, key, value, output, gate, up, and down projections for finetuning over 3 epochs. As all six models diverge from

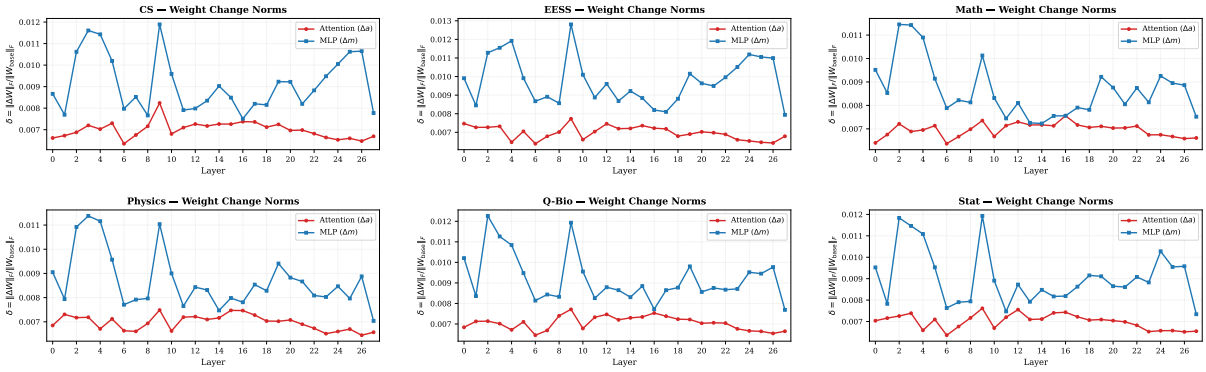

Figure 5: Relative update magnitude δ across all layers for each of the six domains. MLP updates consistently exceed attention updates across all layers and all domains.

an identical initialization, any structural differences in their weight updates are strictly domain-driven. Complete hyperparameters and training details are provided in Appendix C.4.

5.2 Weight Change Norms

If MLPs functions as a domain-specific knowledge storages, they should adapt more aggressively towards a new domain upon finetuning. To test this, we measure the **relative update magnitude** δ, defined as:

$$\delta_m^{(k),\ell} = \frac{\|\Delta W_m^{(k),\ell}\|_F}{\|W_{m,\text{base}}^{\ell}\|_F} \tag{4}$$

where $\Delta W_m^{(k),\ell} = (\alpha/r) \cdot B_m^{(k),\ell} A_m^{(k),\ell}$ is the effective LoRA adapter for domain k, layer ℓ, and module m[1]; $W_{m,\text{base}}^{\ell}$ denotes the corresponding frozen base weight.

MLP updates are on average $\mathbf{1.30}\times$ larger than attention (Figure 5) updates across all domains and all layers.

5.3 Domain Specificity of Updates

Larger MLP updates alone doesn't validate that these updates are domain-specific. It is therefore critical to test whether these updates are *differentiated* across domains. For each layer ℓ and module group, we vectorise the update matrices from all six domains and compute pairwise cosine similarity:

$$\text{sim}(k,k')^{\ell} = \frac{\langle \text{vec}(\Delta W_{\ell}^{(k)}),\ \text{vec}(\Delta W_{\ell}^{(k')}) \rangle}{\|\text{vec}(\Delta W_{\ell}^{(k)})\| \cdot \|\text{vec}(\Delta W_{\ell}^{(k')})\|} \tag{5}$$

This yields a 6×6 similarity matrix.

[1]We pool modules within each group (attention: `q/k/v/o_proj`; MLP: `gate/up/down_proj`) by averaging δ across constituent matrices.

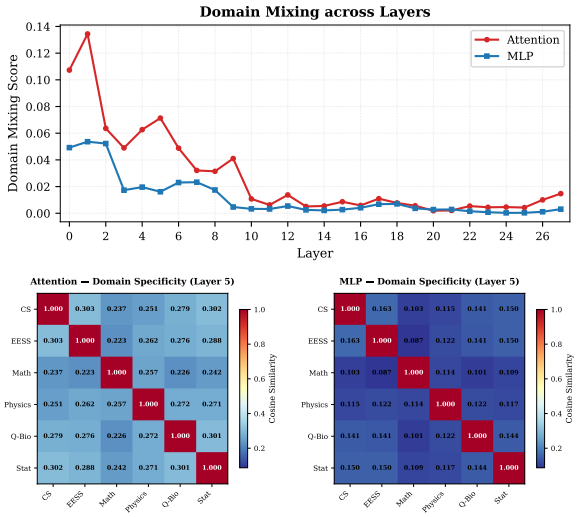

Figure 6: Domain specificity of weight updates across layers and domains. **Top**: Domain mixing score $\mathcal{M}^{\ell}$, lower corresponds to being more domain-specialized. **Bottom**: Pairwise cosine similarity heatmaps between domain update matrices at the most diverging attention-MLP layer. For other layers, see Appendix C.1.

For cleaner analysis, we define the **domain mixing score** as the mean squared error between S^{ℓ} and the identity matrix I (the ideal case where all update matrices are mutually orthogonal):

$$\mathcal{M}^{\ell} = \frac{1}{K(K-1)} \sum_{k \neq k'} \left(\text{sim}(k,k')^{\ell} - I\right)^2 \tag{6}$$

Averaged across all layers, the attention mixing score is $\mathbf{2.33}\times$ higher than that of MLP (Figure 6), indicating that attention updates are not merely smaller in magnitude but also more *generic* across different domains. In contrast, MLP updates are distinct and highly specialized to their domain.

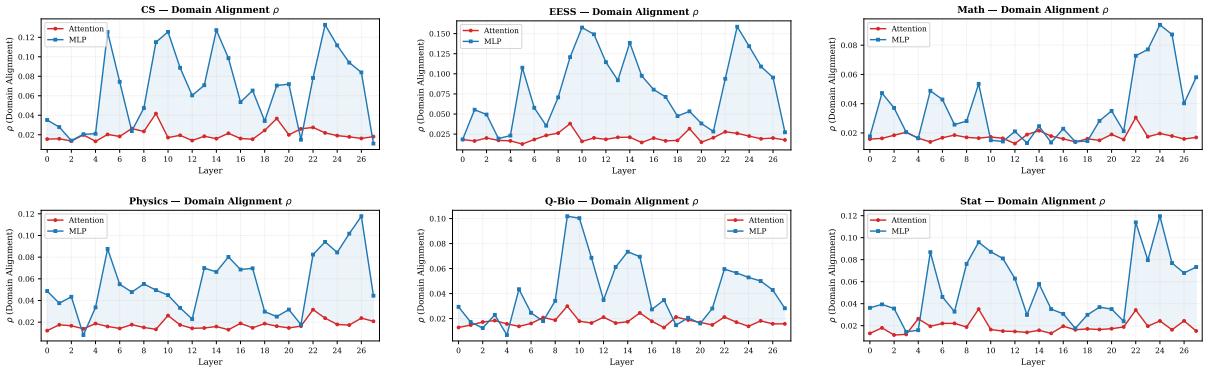

Figure 7: Domain alignment (Projection metric ρ) of weight updates across layers for all domains, measured by the projection metric ρ. MLP updates consistently show higher alignment than attention updates.

5.4 Domain Alignment

Having established that MLP updates are both larger and more domain-differentiated, we now evaluate whether these parameter changes are directionally aligned with the domain representations. To measure this, we project each update matrix onto its corresponding domain concept vector ($\hat{\mathbf{d}}_k^{\ell}$). We quantify alignment with **Projection magnitude** which measures the norm of projection of update matrix on its own domain's concept direction.

$$\rho_m^{(k),\ell} = \frac{\|\Delta W_m^{(k),\ell}\ \hat{\mathbf{d}}_k^{\mathrm{mlp},\ell}\|}{\|\Delta W_m^{(k),\ell}\|_F} \tag{7}$$

Averaged across all domains and all layers, MLP updates show $\mathbf{2.95}\times$ higher projection alignment than attention updates (Figure 7). This strong alignment confirms that the MLP weights are preferentially modified along the exact directions that represent each domain.

5.5 Module-Level Decomposition

After establishing MLP as the primary site for learning new domain knowledge, we look into their individual weight matrices, allowing us to pinpoint exactly which components drive these updates.

Table 1: Per-module summary (averaged across all domains and all layers).δ is relative update magnitude, ρ is projection magnitude and $\mathcal{M}$ is domain mixing score.

Module	$\delta(\uparrow)$	$\rho(\uparrow)$	$\mathcal{M}(\downarrow)$
q_proj	0.00685	0.0215	0.0107
k_proj	0.00658	0.0171	0.0145
v_proj	0.00765	0.0197	0.0902
o_proj	0.00744	0.0162	0.0431
gate_proj	**0.01230**	**0.0767**	**0.0036**
up_proj	0.00816	0.0338	0.0224
down_proj	0.00466	0.0297	0.0499

Table 1 suggests that the `gate_proj` matrix dominates on all three metrics: it has the highest relative update magnitude, the lowest domain mixing score (most specialised), and the highest alignment with domain concept vectors. This is interpretable from the SwiGLU (Shazeer, 2020) architecture:

$$\mathrm{MLP}(x) = W_{\mathrm{down}} \cdot \left[\sigma(W_{\mathrm{gate}} \cdot x) \odot (W_{\mathrm{up}} \cdot x)\right]$$

The gate projection W_{gate} acts as a **feature selector**, it controls which intermediate features are amplified and which are suppressed. By modifying W_{gate}, fine-tuning can selectively activate domain-relevant feature subspaces.

6 Causal Activation Patching

Neither probing nor fine-tuning establishes that the domain signal is *used* by downstream layers; we now test this causally by patching domain concept directions into individual sublayers and measuring the resulting shift in the final residual stream.

6.1 Experimental Setup

For each source text from domain A, we project out the source-domain component from the attention (or MLP) contribution at layer ℓ^* and inject a calibrated target-domain component along $\hat{\mathbf{d}}_B^{\mathrm{attn},\ell^*}$, with magnitude α matched to the natural projection of genuine domain-B activations (Appendix D.1). The resulting **final domain shift score** s^L measures the projection of the change in the final residual stream onto the normalised $A \to B$ concept difference. We evaluate across all 15 domain pairs ($\binom{6}{2}$, 200 samples each) under seven conditions: attention or MLP patching at hotspot and non-hotspot layers, full layer sweeps, and a random-direction specificity control (Table 18, Appendix D.2).

6.2 Results

Layer sweep. Attention patching shows a sharp, localised peak at L19 ($s^L = 25.56$), with a secondary peak at L22 (18.50), a 255× larger shift than the adjacent non-hotspot L17 (0.10) despite being only two layers apart (Figure 8). Outside these hotspots, attention patching produces near-zero effects. MLP patching shows a qualitatively different profile: negligible early-layer shifts rising sharply from L19 onward to $s^L = 74.32$ at L27, consistent with the monotonic late-layer execution role established by probing. All non-hotspot shifts have standard deviations exceeding their means, confirming they are indistinguishable from noise (full results in Table 19, Appendix D.3). For MLP, early hotspot patches yield near-zero effects while late patches surpass attention, indicating a crossover around L22 (Appendix D.4).

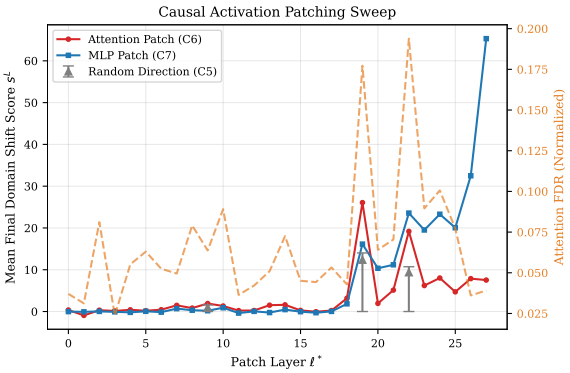

Figure 8: Mean final domain shift s^L as a function of patch layer ℓ^*. Attention peaks sharply at L19 ($s^L = 25.56$); MLP dominates in late layers (L27: $s^L = 74.32$). Grey band: random-direction control.

Specificity control. Random-direction interventions at hotspot layers produce approximately half the shift of targeted concept-direction patches (L19: 12.19 vs. 25.56; L22: 9.33 vs. 18.50), confirming that the domain concept vectors capture genuine causal directions, not generic perturbation effects.

6.3 Signal Propagation and Amplification

The mechanistic core of the experiment is the propagation trace (Figure 9). The L19 attention patch produces an immediate shift of $s^{19.5} = 8.77$, steadily amplified to $s^{28} = 25.56$, a 2.91× amplification. The decomposition is clean: at every downstream layer, attention contributes near-zero shift (functionally transparent), while every downstream MLP contributes a positive shift, accounting for the entirety of the amplification. This provides

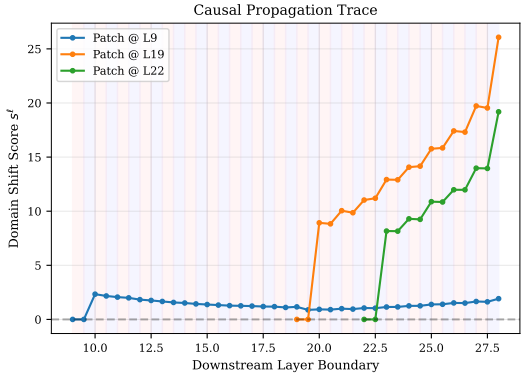

Figure 9: Propagation trace: domain shift s^ℓ at every sublayer boundary downstream of the L19 patch. Steady amplification from $s^{19.5} = 8.77$ to $s^{28} = 25.56$

the most direct evidence for the framework: attention at L19 *sets* the domain signal; downstream MLPs *execute* by incrementally writing domain-appropriate features into the residual stream.

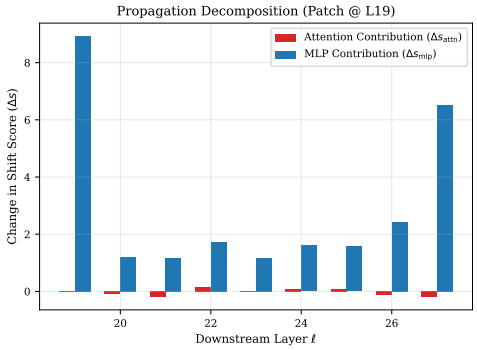

Figure 10: Sublayer decomposition: downstream attention contributes near-zero shift; every downstream MLP contributes a positive shift.

By contrast, the L9 patch shows partial *absorption*: the initial shift of $s^{9.5} = 2.93$ decreases through the mid-layer trough (L11–18) before stabilising at $s^{28} = 2.34$, confirming that mid-layer attention actively suppresses domain signal injected too early. The dose-response is monotonic across five injection magnitudes (Appendix D.8.1), and the L19 hotspot is consistent across all 15 domain pairs (Appendix D.6).

7 Generalization Across Model Families

The three experiments were conducted on Qwen2.5-7B as the primary model. We replicate these on LLaMA 3.2-3B and Pythia 2.8B, which differ from Qwen2.5-7B in architecture (SwiGLU vs. standard MLP), depth, and scale. We summarize cross-model results for each experiment below.

1. Probing: Attention universally exhibits localized, spiky domain signals, while MLP representa-

tions build smoothly across depth. **2. Fine-Tuning:** MLP weight updates consistently outpace attention in magnitude (1.09-1.40×) and domain alignment (up to 3.92×), driven entirely by the first MLP projection matrix. **3. Causal Patching:** The propagation mechanism replicates exactly; routed signals from attention hotspots are exclusively amplified by downstream MLPs, while downstream attention remains passive (See Appendix B ,C, D).

8 Discussion

The three experiments try to answer the central question from independent angles. Probing reveals the representational structure: attention concentrates domain signal at specific layers while MLP distributes it monotonically. Fine-tuning reveals the adaptive structure: when the model acquires domain-specific knowledge, it writes preferentially into MLP weights along the domain concept direction. Causal patching reveals the functional structure: injecting a domain concept direction into attention at its peak layer produces a directed, downstream shift that is exclusively amplified by MLP sublayers. These are not three separate findings, but three views of the same organisational principle, connected by a single mathematical object.

The most mechanistically specific finding is the mid-layer trough. Attention probe accuracy drops to 0.77 at layers 15–17, below even layer 0 levels. This is not an artifact of noisy estimation: causal patching confirms it functionally. A domain signal injected at L9 is partially absorbed through L11-18, arriving at the final layer weaker than at injection. By contrast, a signal injected at L19 is steadily amplified by every downstream MLP sublayer, with downstream attention contributing near-zero shift. We speculate that this establishes a three-phase temporal structure in domain processing. Early layers (0–10) perform initial domain recognition. Mid-layers (11-18) process domain-agnostic structure — likely syntactic, discourse-level, or genre-related computation shared across all six domains. Late layers (19-22) re-engage with domain identity, now with the benefit of the intermediate features built by the domain-agnostic phase. This structure falls directly out of the layer-wise profiles and is independently confirmed by causal intervention.

The fine-tuning results add a concrete mechanistic detail. In all three models, the first MLP projection (gate projection in Qwen and LLaMA, dense_h_to_4h in Pythia) dominates all adaptation metrics: largest update magnitude, lowest domain mixing, highest alignment with domain concept directions. This matrix controls which intermediate features are amplified in the MLP's nonlinear transformation. That domain adaptation preferentially modifies this gating function — and that the modification is aligned with domain concept directions — identifies a specific site where domain knowledge is physically written during learning.

One feature of our results deserves careful framing. Attention and MLP carry comparable *overall* domain information (probe accuracy 0.841 vs. 0.844). The finding is not that one component is more important than the other. It is that they organise domain information differently: attention concentrates it at a small number of layers, MLP distributes it across depth. These experiments characterize the distributional difference rather than the accuracy difference. The causal experiment sharpens this: it is not that attention carries *more* domain signal, but that the signal it carries at its peak layers is causally potent and triggers domain-consistent execution in downstream MLP.

These results extend the scope of the attention/MLP functional division established at the token level. Prior work showed that attention moves information and MLP transforms it for individual tokens and factual associations. We find the same division at the domain level, with the same character — localised routing, distributed execution — and with causal confirmation that routing at the right layer triggers execution downstream. The organisational seam between attention and MLP is not confined to token-scale processing. It structures computation at the level of semantic domains.

9 Conclusion

We asked whether the attention/MLP functional division, established at the token and fact level, scales to domain-level processing. Three experiments on six arXiv domains show that it does, but has a specific character: attention routing is localised to a small number of identifiable layers, while MLP execution is distributed across depth. Domain concept vectors serve as the unifying object connecting representational structure, adaptive behaviour, and causal control across all three experiments and three model families. These results establish that the organisational principle governing token-level computation in transformers extends to a qualitatively coarser level of abstraction.

Limitations

Dataset scope. All experiments use arXiv abstracts from six academic disciplines. While this controls for genre and surface form, it is a single text type. Whether the same patterns hold for more heterogeneous domain boundaries (e.g., fiction vs. technical writing, or cross-lingual domains) is untested.

Domain granularity. Our six domains are coarse, top-level arXiv categories. Finer-grained distinctions (e.g., cs.CL vs. cs.CV within computer science) may exhibit different layer-wise patterns. The concept vector framework applies at any granularity, but we have not verified whether the localised-routing, distributed-execution structure persists at sub-domain resolution.

Concept vector limitations. Domain concept vectors are first-order summaries: the direction from the global mean to the domain mean. They capture the dominant axis of domain variation but miss within-domain multimodality, higher-order geometric structure, and individual-sample variation. Domains whose internal structure is not well-described by a single centroid displacement may be poorly characterised by this formalism.

Probing caveats. Linear probing shows what information is decodable from activations, not what information is functionally used by downstream computation. We address this directly through causal patching, but the probing results alone should be interpreted as characterising the information landscape, not the computational mechanism.

Fine-tuning method. We use LoRA adaptation, not full fine-tuning. LoRA constrains the rank of weight updates, which may bias which modules appear most affected. The consistent gate-projection dominance across all three models (including Pythia, which uses a different MLP architecture) suggests this is not purely a LoRA artifact, but full fine-tuning comparisons would strengthen the claim.

Model selection. We test three decoder-only autoregressive models of moderate scale (2.8B–8B parameters). Encoder-only models, encoder-decoder architectures, and substantially larger models may exhibit different patterns. The consistency across three architecturally distinct models provides some evidence of generality, but broader verification is needed.

Pooling strategy. All main results use last-token pooling. Mean-pooled activations are collected but not reported as the primary analysis. If domain information is distributed across token positions rather than concentrated at the final position, last-token pooling may underestimate domain signal in certain layers or components.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 7319–7328.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Viktoriia A Chekalina, Anna Rudenko, Gleb Mezentsev, Aleksandr Mikhalev, Alexander Panchenko, and Ivan Oseledets. 2024. Sparsegrad: A selective method for efficient fine-tuning of mlp layers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14929–14939.

Colin B Clement, Matthew Bierbaum, Kevin P O'Keeffe, and Alexander A Alemi. 2019. On the use of arxiv as a dataset. *arXiv preprint arXiv:1905.00075*.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pages 7324–7338. PMLR.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pages 160–187. PMLR.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 30–45.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36:17643–17668.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert Dick, Hidenori Tanaka, Tim Rocktäschel, Edward Grefenstette, and David Krueger. 2024. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, volume 2024, pages 9944–10015.

Shahar Katz, Yonatan Belinkov, Mor Geva, and Lior Wolf. 2024. Backward lens: Projecting language model gradients into the vocabulary space. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2390–2422.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. *arXiv preprint arXiv:2503.02080*.

Pin-Jie Lin, Rishab Balasubramanian, Fengyuan Liu, Nikhil Kandpal, and Tu Vu. 2025. Efficient model development through fine-tuning transfer. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2617–2636.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Clement Neo, Shay B Cohen, and Fazl Barez. 2024. Interpreting context look-ups in transformers: Investigating attention-mlp interactions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16681–16697.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. 2023. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36:66727–66754.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2025. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, volume 2025, pages 76441–76463.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Mateusz Piotrowski, Paul M Riechers, Daniel Filan, and Adam S Shai. 2025. Constrained belief updates explain geometric structures in transformer representations. *arXiv preprint arXiv:2502.01954*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. 2024. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pages 17282–17333.

Betty Van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1823–1832.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. 2025. The geometry of tokens in internal representations of large language models. *arXiv preprint arXiv:2501.10573*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Ruichen Xu and Kexin Chen. 2025. Filtering with self-attention and storing with mlp: One-layer transformers can provably acquire and extract knowledge. *arXiv preprint arXiv:2508.00901*.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *Advances in neural information processing systems*, 37:118571–118602.

Fangcong Yin, Xi Ye, and Greg Durrett. 2024. Lofit: Localized fine-tuning on llm representations. *Advances in neural information processing systems*, 37:9474–9506.

Supplementary Materials

A Dataset Details

A.1 Domain Selection Rationale

We select six arXiv top-level categories that span a range of intellectual traditions:

The cs–stat pair is the intentionally hardest boundary: these domains share substantial methodological vocabulary ("regression," "estimator," "neural network") and overlap in subject matter (stat.ML vs cs.LG). If the model's domain processing mechanism works even for this pair, the finding is robust.

A.2 Data Collection and Preprocessing

The dataset was constructed using arXiv bulk metadata abstracts, which were filtered to include only papers with a single primary category to prevent cross-listing ambiguity. We randomly sampled 15,000 abstracts per domain, with the exception of the 'eess' domain, where all 14,560 available papers were utilized. The text was processed using the Qwen2.5 tokenizer, with sequences truncated or padded to a uniform length of 384 tokens. All selected samples were retained for the initial activation extraction phase, as train and test data splitting was deferred to the subsequent downstream probing experiments.

A.3 Vocabulary Overlap Analysis

Summary statistics:

- Unique tokens per domain: 84,835 (stat) to 123,235 (q-bio)
- Words appearing in *all six* domains: 11,979 (2.7% of total unique vocabulary)
- Mean pairwise Jaccard: 0.148
- Mean pairwise log-TF cosine: 0.613

The Jaccard similarity (10–20%) appears low because it counts unique types — rare technical terms inflate the denominator. The frequency-weighted cosine similarity (51–72%) is more informative: the *common* words that dominate actual text are shared across domains. This confirms that domain separation in the model's representations is not trivially explained by vocabulary differences.

A.4 Activation Extraction

For each input x, we decompose each layer's computation:

$$\mathbf{x}^{\ell+1} = \mathbf{x}^{\ell} + \underbrace{\text{Attn}(\mathbf{x}^{\ell})}_{\Delta\mathbf{a}^{\ell}} + \underbrace{\text{MLP}(\cdot)}_{\Delta\mathbf{m}^{\ell}} \tag{8}$$

using forward hooks that capture:

- $\Delta\mathbf{a}^{\ell}$: the output of the attention sublayer's output projection (`o_proj`), before addition to the residual stream.
- $\Delta\mathbf{m}^{\ell}$: the output of the MLP sublayer's down projection (`down_proj`), before addition to the residual stream.
- $\mathbf{x}^{\ell}$: the full residual stream after both sublayers.

We extract both last-token and mean-pooled representations; all main results use last-token pooling.

Storage: one HDF5 file per domain, each containing arrays of shape $(15000, 28, 3584)$ in float16. Total storage: $\sim$90 GB across all 6 domains.

A.5 t-SNE Projections

A.6 Definitions of Domain and Concept Vector

We define a "domain" as a stable area of scholarly inquiry with externally recognized boundaries maintained by the academic community. In this work, domains are operationalized using arXiv's top-level classification categories, providing a consistent framework navigated by researchers. We define a "concept vector" in the technical geometric sense introduced by Park et al. (2024), wherein categorical concepts correspond to simplices in the representation space of a language model.

Specifically, our domain concept vector $\mathbf{d}_k^{\ell}$ (Eq. 2) represents a distinct direction in the representation space corresponding to a categorical distinction. Although this formulation does not attempt to capture the complete semantic depth of a domain or represent an internal cognitive state, it provides an empirically measurable construct. The vector is mathematically well-defined, satisfies the decomposition identity detailed in Eq. 3, and serves as the fundamental entity across all presented experiments.

Domain	arXiv	Why included
Computer Science	`cs.*`	Empirical, systems-oriented; highest vocabulary overlap with stat
Elec. Engineering	`eess.*`	Engineering; shares methodology with physics and cs
Mathematics	`math.*`	Heavily symbolic; most distinctive vocabulary
Physics	`physics.*`	Close to math but experimentally grounded
Quant. Biology	`q-bio.*`	Biological systems; distinctive subject matter
Statistics	`stat.*`	Methodologically close to cs; intentional hard case

Table 2: Six arXiv domains and selection rationale.

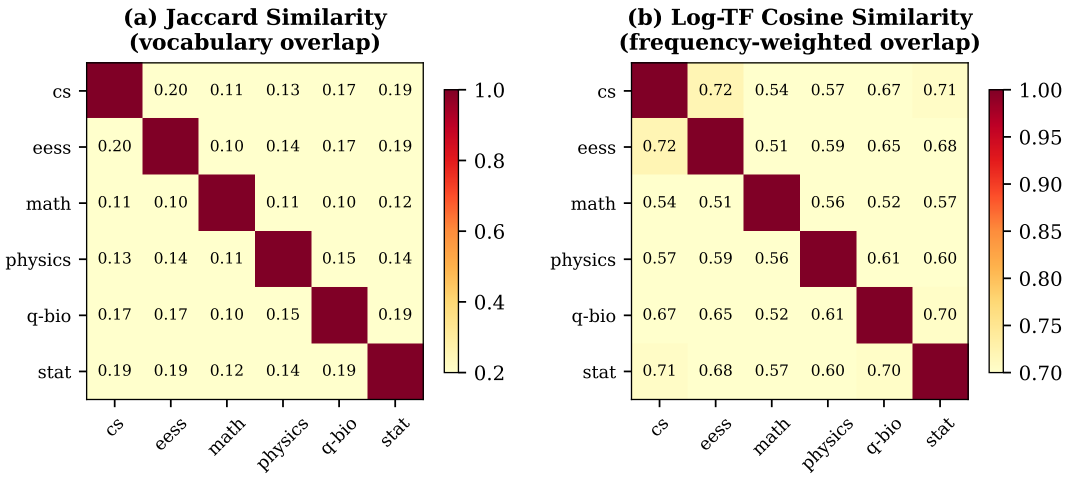

Figure 11: Full vocabulary overlap analysis. Left: Jaccard similarity of unique whitespace-tokenized words. Right: cosine similarity of log-TF weighted word frequency vectors. Key observations: (1) all pairs share cosine similarity > 0.51; (2) cs–eess (0.72) and stat–q-bio (0.70) are the most similar pairs; (3) math is the most lexically distinctive domain.

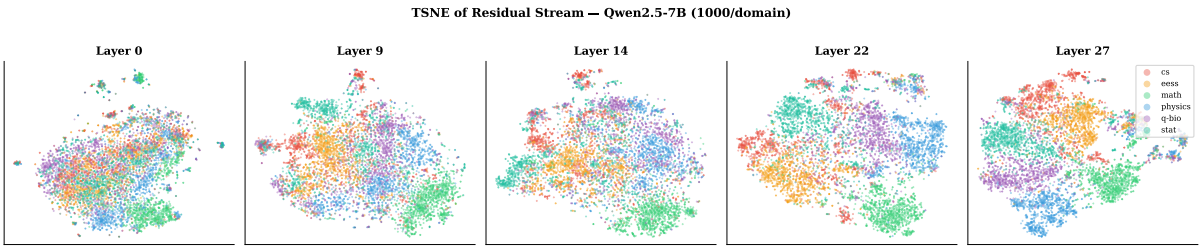

Figure 12: t-SNE of residual stream at layers 0, 9, 14, 22, 27 (1,000 samples/domain, perplexity=30). Progressive sharpening of domain clusters with depth.

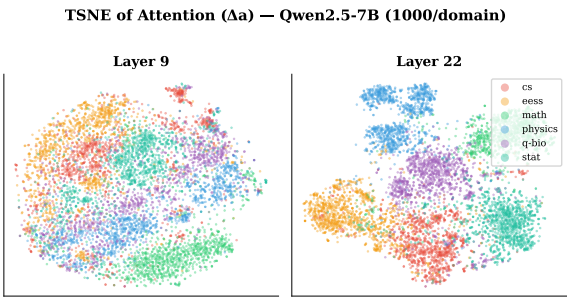

Figure 13: t-SNE of *attention contributions* (Δa) at layers 9 and 22 (the two probe accuracy peaks). At layer 9, domain structure is weak. At layer 22, attention contributions alone produce clear domain clusters. This is consistent with the spiky attention profile from the probing experiment.

Layer	Attn	MLP	Resid	Layer	Attn	MLP	Resid
0	.781	.761	.779	14	.840	.853	.866
1	.752	.754	.788	15	.794	.829	.861
2	.815	.781	.813	16	.784	.805	.856
3	.760	.764	.816	17	.774	.801	.852
4	.831	.784	.833	18	.797	.819	.849
5	.826	.810	.831	19	.891	.879	.889
6	.852	.831	.845	20	.836	.870	.894
7	.863	.849	.865	21	.882	.881	.900
8	.877	.851	.872	22	.912	.897	.911
9	.891	.885	.885	23	.892	.897	.918
10	.883	.876	.884	24	.887	.899	.914
11	.823	.858	.878	25	.881	.900	.916
12	.831	.851	.870	26	.859	.898	.915
13	.864	.857	.869	27	.870	.893	.913

Table 3: Qwen 2.5-7B: Per-layer linear probe accuracy for all three components.

B Probing and Seperability

This section presents the full results of the probing experiments (linear probe accuracy and Fisher Discriminant Ratio) for Qwen 2.5-7B, LLaMA 3.2-3B, and Pythia 2.8B.

B.1 Qwen 2.5-7B

Qwen's domain signal peaks late: residual stream reaches **0.918** accuracy at L23, attention peaks sharply at L22 (**0.913**), and MLP rises monotonically to **0.901** at L25. FDR confirms direction-based separation (raw $\approx$ normalised) with attention peaking at L22 (**0.194**).

Table 5: FDR summary (Qwen 2.5-7B). Raw and normalised values are nearly identical, showing that domain separation is direction-based. Attention peaks at **L22** (**0.194**).

Component	FDR Type	Peak Layer	Peak FDR	Mean FDR
Attn	Raw	**L22**	**0.179**	0.062
Attn	Normalised	L22	0.194	0.066
MLP	Raw	L23	0.164	0.075
MLP	Normalised	L23	0.159	0.078
Residual	Raw	L24	0.169	0.087
Residual	Normalised	L24	0.169	0.089

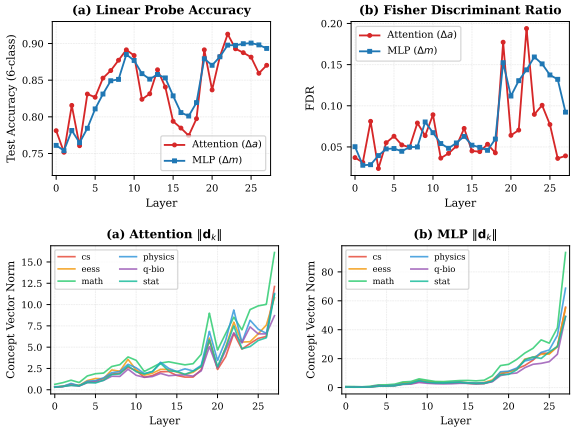

Figure 14: Qwen 2.5-7B: (Top) Domain classification accuracy and FDR. (Bottom) Concept vector norms.

Table 4: Linear probe accuracy summary (Qwen 2.5-7B). Residual stream peaks at **L23** with **0.918** accuracy, while Attention and MLP peak around **0.90-0.91**.

Component	Mean Acc.	Peak Layer	Peak Acc.
Attn	0.841	**L22**	**0.913**
MLP	0.844	**L25**	**0.901**
Residual	0.868	**L23**	**0.918**

B.2 LLaMA 3.2-3B

LLaMA shows a co-localised peak: both attention and MLP peak at **L20** (**0.902** and **0.893** respectively), while the residual stream peaks later at L25 (**0.910**). FDR is strongly attention-dominated, peaking at L20 (**0.275**), significantly higher than MLP.

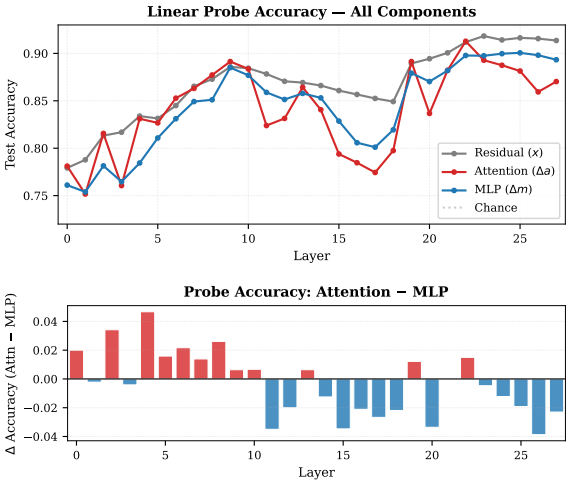

Figure 15: Qwen 2.5-7B: (Top) Probe accuracy all components. (Bottom) Attention minus MLP difference.

Layer	Attn	MLP	Resid	Layer	Attn	MLP	Resid
0	.765	.753	.766	14	.878	.858	.867
1	.761	.727	.770	15	.884	.873	.872
2	.775	.752	.769	16	.882	.876	.879
3	.826	.790	.812	17	.871	.874	.881
4	.854	.827	.837	18	.868	.875	.882
5	.839	.811	.830	19	.891	.873	.884
6	.850	.817	.833	20	.902	.893	.897
7	.848	.811	.838	21	.868	.885	.897
8	.854	.832	.848	22	.871	.884	.897
9	.833	.824	.840	23	.899	.887	.901
10	.816	.801	.831	24	.878	.889	.907
11	.800	.805	.818	25	.890	.885	.910
12	.855	.826	.836	26	.879	.893	.909
13	.845	.824	.843	27	.886	.892	.905

Table 6: LLaMA 3.2-3B: Per-layer linear probe accuracy for all three components.

Table 7: Linear probe accuracy summary (LLaMA 3.2-3B). Residual stream peaks at **L25** (**0.910**), while both sublayers peak early at **L20**.

Component	Mean Acc.	Peak Layer	Peak Acc.
Attn	0.852	**L20**	**0.902**
MLP	0.841	**L20**	**0.893**
Residual	0.856	**L25**	**0.910**

Table 8: FDR summary (LLaMA 3.2-3B). Attention shows the highest FDR peaking at **L20** (**0.275**), significantly higher than MLP.

Component	FDR Type	Peak Layer	Peak FDR	Mean FDR
Attn	Raw	**L20**	**0.265**	0.088
Attn	Normalised	**L20**	**0.275**	0.094
MLP	Raw	L20	0.183	0.096
MLP	Normalised	L20	0.179	0.097
Residual	Raw	L20	0.181	0.102
Residual	Normalised	L20	0.182	0.103

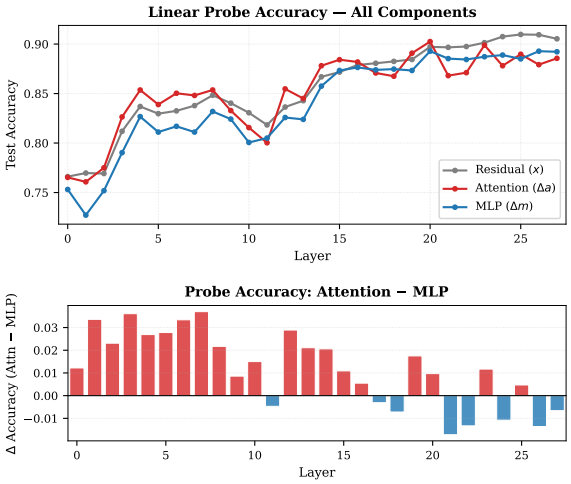

Figure 19: LLaMA 3.2-3B: (Top) Probe accuracy all components. (Bottom) Attention minus MLP difference.

B.3 Pythia 2.8B

Pythia's domain signal develops more gradually, with attention peaking mid-network (L20, **0.886**) and MLP peaking late (L27, **0.889**). FDR reveals a striking early–late split: attention FDR peaks at L19 (**0.224**) while MLP FDR peaks at L25 (**0.216**).

Layer	Attn	MLP	Resid	Layer	Attn	MLP	Resid
0	.803	.194	.798	16	.878	.841	.876
1	.818	.787	.810	17	.872	.858	.883
2	.842	.763	.819	18	.872	.870	.885
3	.843	.778	.841	19	.883	.863	.888
4	.835	.807	.847	20	.886	.869	.894
5	.840	.803	.843	21	.878	.870	.895
6	.836	.795	.837	22	.879	.870	.894
7	.837	.790	.832	23	.867	.885	.899
8	.823	.785	.830	24	.873	.879	.898
9	.812	.786	.828	25	.866	.888	.894
10	.806	.785	.826	26	.868	.887	.900
11	.799	.768	.823	27	.878	.888	.898
12	.819	.765	.824	28	.863	.886	.897
13	.802	.768	.814	29	.865	.888	.898
14	.871	.775	.858	30	.874	.888	.897
15	.861	.839	.866	31	.882	.882	.898

Table 9: Pythia 2.8B: Per-layer linear probe accuracy for all three components.

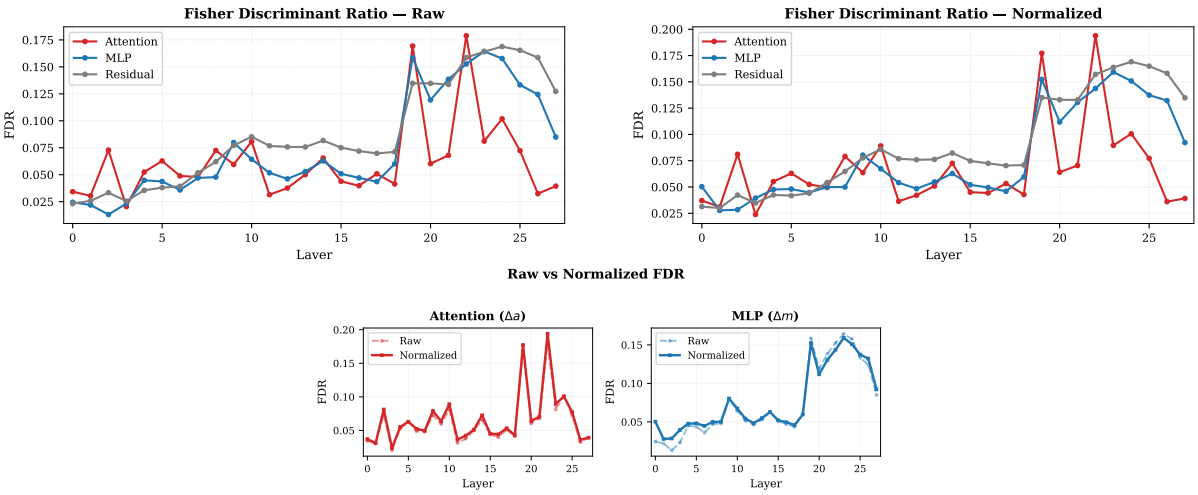

Figure 16: Qwen 2.5-7B: FDR Analysis.

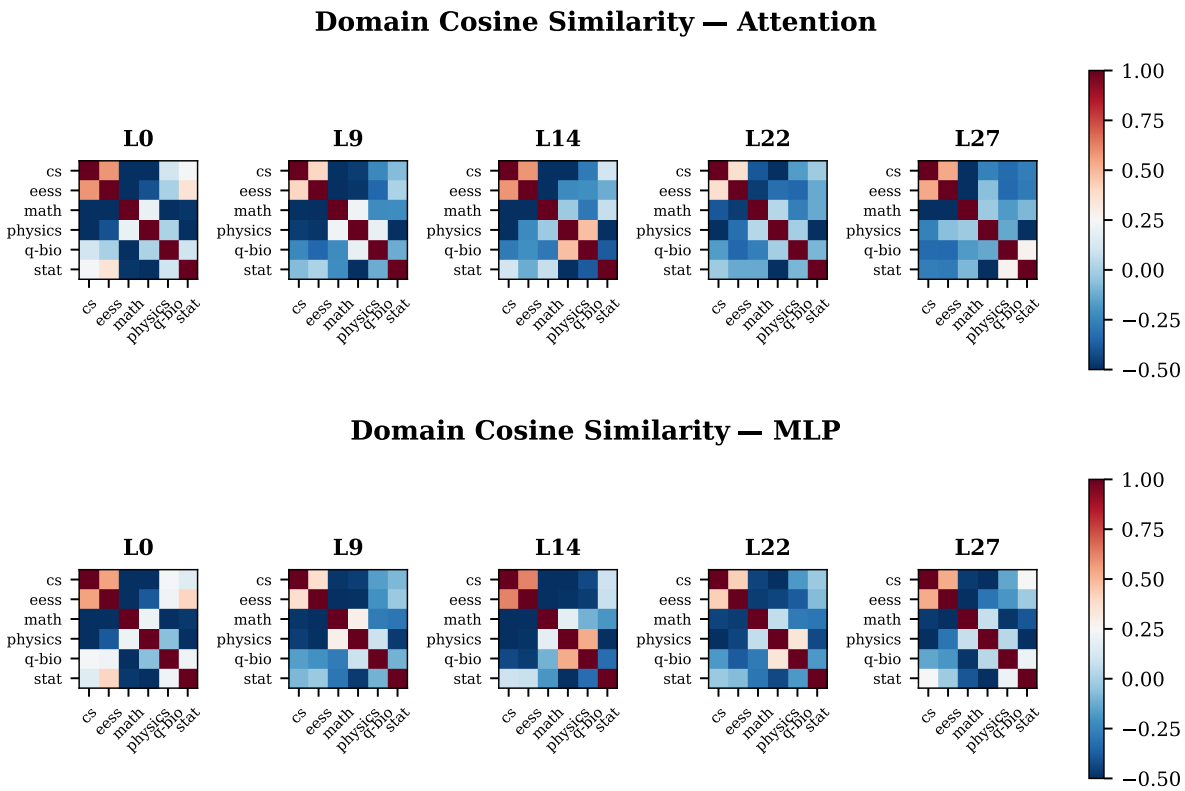

Figure 17: Qwen 2.5-7B: Cosine similarity heatmaps (Attention vs MLP).

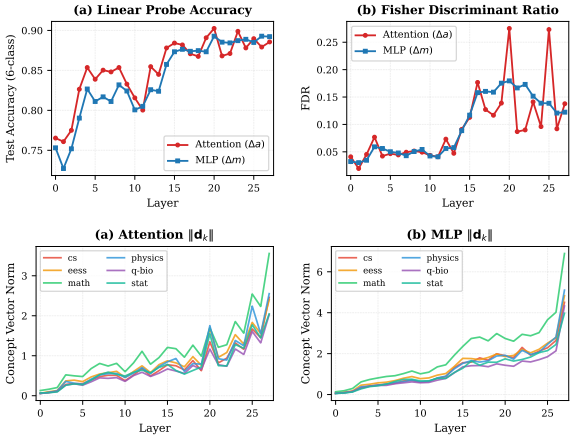

Figure 18: LLaMA 3.2-3B: (Top) Domain classification accuracy and FDR. (Bottom) Concept vector norms.

Table 10: Linear probe accuracy summary (Pythia 2.8B). Residual stream peaks at **L26** (**0.900**), with Attention peaking early at **L20** (**0.886**) and MLP peaking late at **L27** (**0.889**).

Component	Mean Acc.	Peak Layer	Peak Acc.
Attn	0.851	**L20**	**0.886**
MLP	0.813	**L27**	**0.889**
Residual	0.863	**L26**	**0.900**

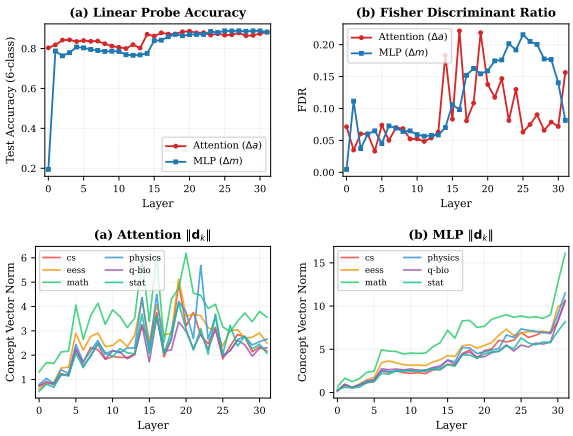

Figure 22: Pythia 2.8B: (Top) Domain classification accuracy and FDR. (Bottom) Concept vector norms.

Table 11: FDR summary (Pythia 2.8B). Attention peaks early at **L16/L19** (**0.224**), while MLP peaks late at **L25** (**0.216**).

Component	FDR Type	Peak Layer	Peak FDR	Mean FDR
Attn	Raw	**L19**	**0.224**	0.086
Attn	Normalised	**L16**	**0.222**	0.091
MLP	Raw	**L25**	**0.215**	0.111
MLP	Normalised	**L25**	**0.216**	0.115
Residual	Raw	L30	0.181	0.101
Residual	Normalised	L27	0.190	0.113

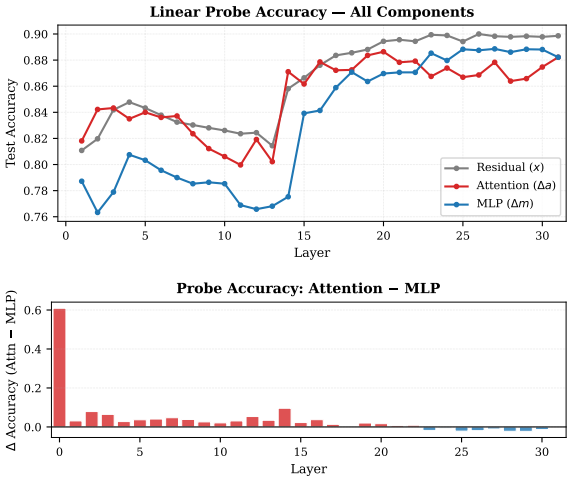

Figure 23: Pythia 2.8B: (Top) Probe accuracy all components (0th layer excluded for clarity). (Bottom) Attention minus MLP difference.

As claimed in section (section 4) that *"attention is spiky, MLP is smooth"*: attention carries domain signal at discrete layers while MLP increases monotonically. All three models validate this claim – Qwen (L9/L22 peaks), LLaMA (L20 peak), and Pythia (L20 peak) all show localised attention peaks, while MLP accuracy rises smoothly in all cases. The complementary finding that domain separation is direction-based (raw FDR $\approx$ normalised FDR) also holds universally across all three models.

B.4 Reproducibility

All probing and separability experiments were executed on an NVIDIA A6000 GPU (48GB VRAM) using bfloat16 precision across Qwen/Qwen2.5-7B ($d = 3584$), LLaMA 3.2-3B ($d = 3072$), and Pythia 2.8B ($d = 2560$). Model activations were extracted via last-token pooling from 5,000 samples per domain with a maximum sequence length of 512. The probing dataset consisted of 2,000 samples per domain (random seed 42), which were standardized using `StandardScaler` prior to a 70% training and 30% test stratified split. Classification was performed using `scikit-learn`'s multinomial logistic regression, optimized via the L-BFGS solver with a 1000-iteration maximum and an L2 penalty of $C = 1.0$. All probing and separability experiments were conducted in the following hardware and software environment:

C Fine-Tuning Domain Adaptation

This section details the results of LoRA fine-tuning for all three models.

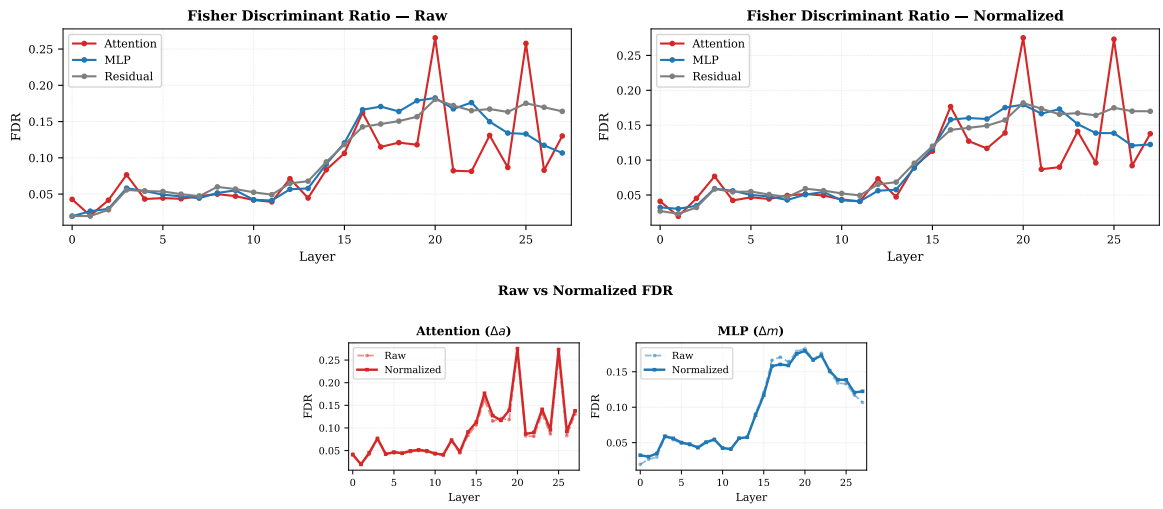

Figure 20: LLaMA 3.2-3B: FDR Analysis.

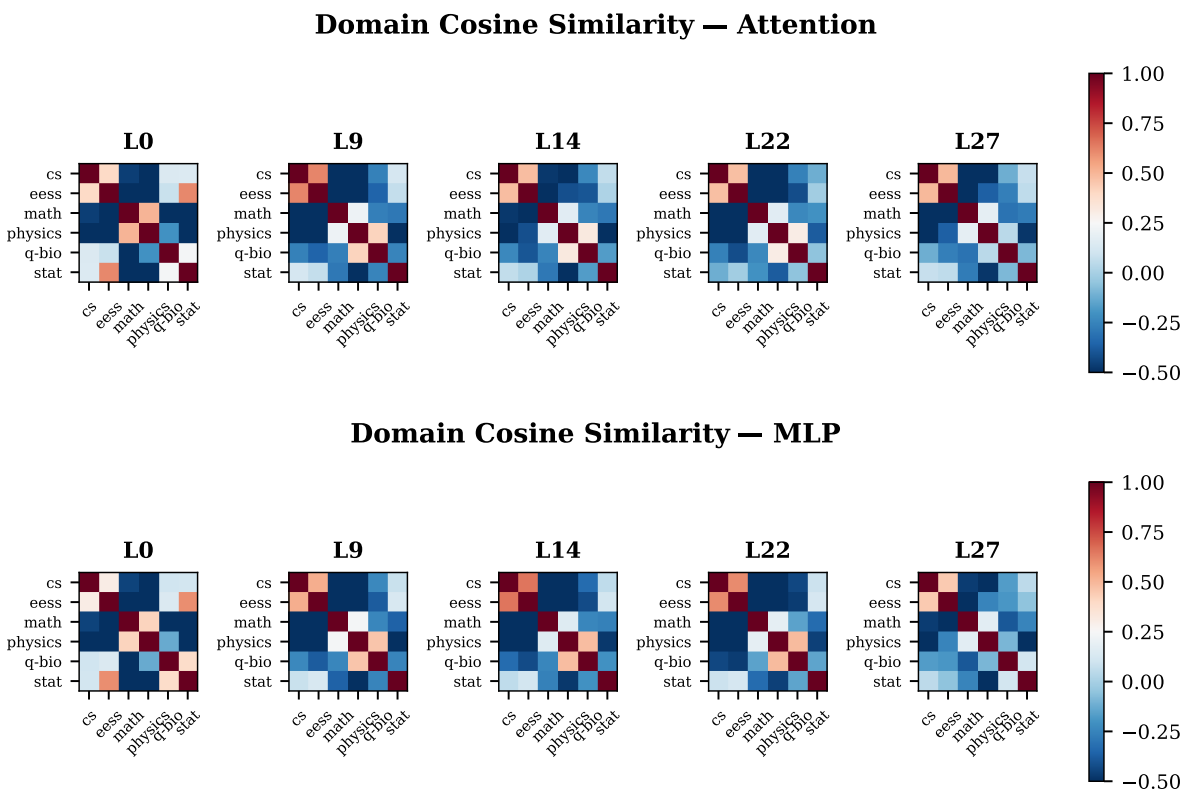

Figure 21: LLaMA 3.2-3B: Cosine similarity heatmaps (Attention vs MLP).

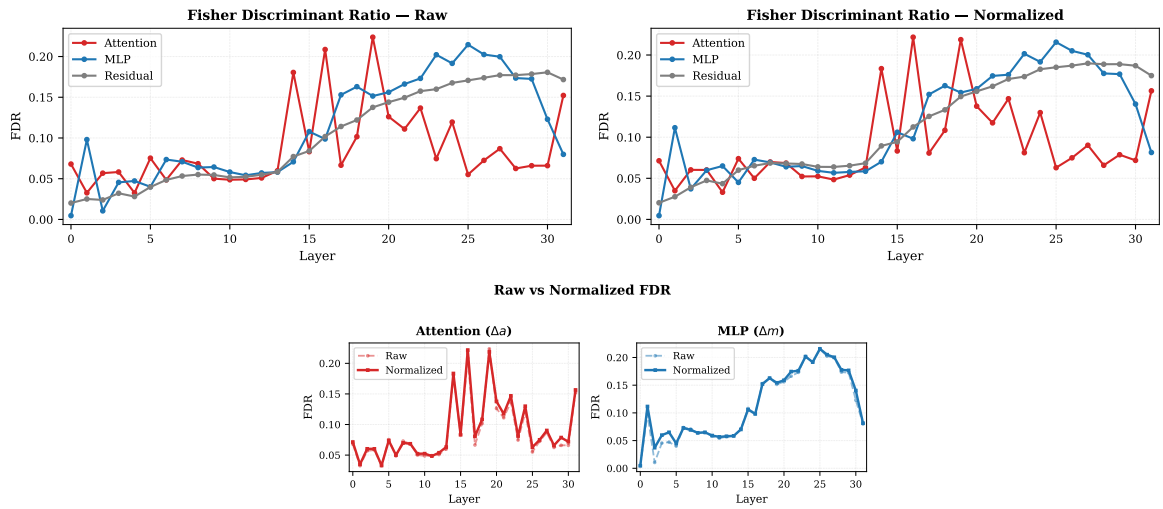

Figure 24: Pythia 2.8B: FDR Analysis.

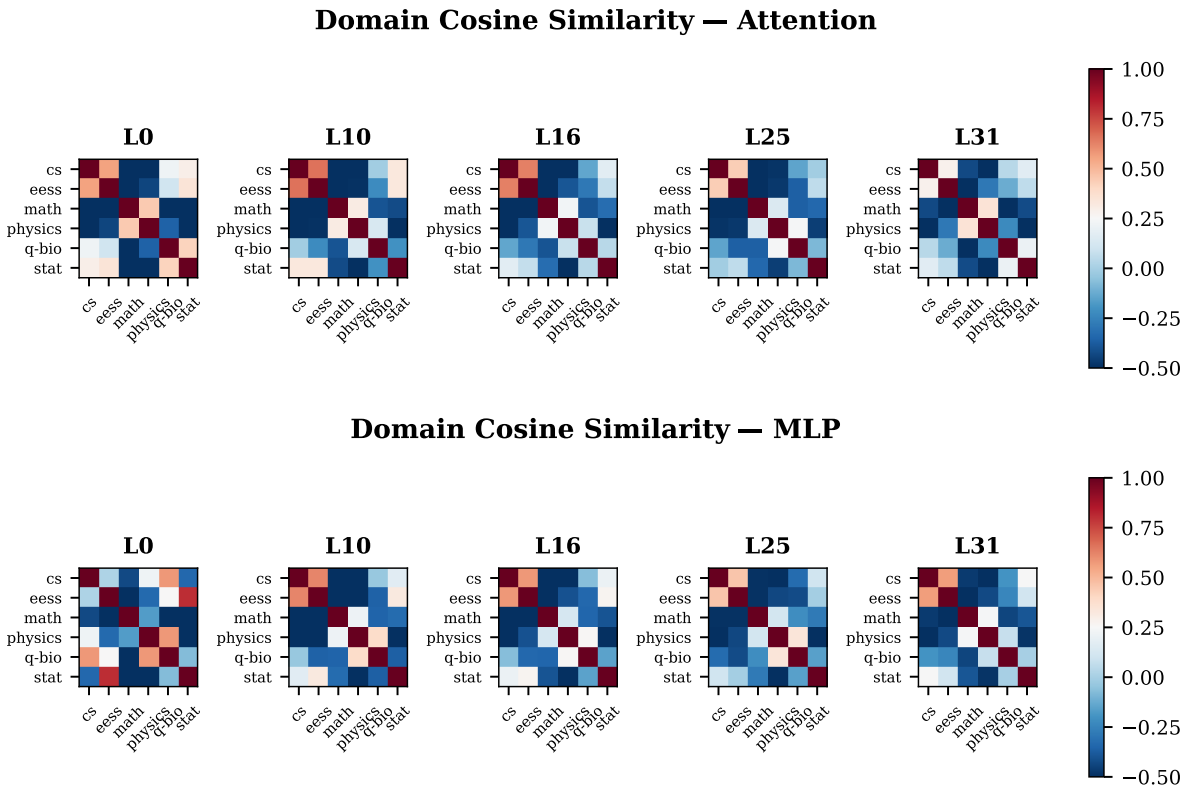

Figure 25: Pythia 2.8B: Cosine similarity heatmaps (Attention vs MLP).

Table 12: Qwen 2.5-7B: Mean relative update magnitude $\bar{\delta}$ across layers.

Domain	δ_{attn}	δ_{mlp}	Ratio
CS	0.00700	0.00913	1.30
EESS	0.00698	0.00976	1.40
Math	0.00696	0.00864	1.24
Physics	0.00697	0.00870	1.25
Q-Bio	0.00705	0.00920	1.31
Stat	0.00700	0.00905	1.29

Table 13: Qwen 2.5-7B: Per-module summary.

Module	δ	ρ	$\mathcal{M}$
q_proj	0.00685	0.0215	0.0107
k_proj	0.00658	0.0171	0.0145
v_proj	0.00765	0.0197	0.0902
o_proj	0.00744	0.0162	0.0431
gate_proj	**0.01230**	**0.0767**	**0.0036**
up_proj	0.00816	0.0338	0.0224
down_proj	0.00466	0.0297	0.0499

C.1 Qwen 2.5-7B

MLP updates are consistently **1.24–1.40**$\times$ larger than attention updates across all six domains ($\delta_{\text{mlp}} \approx 0.009$ vs. $\delta_{\text{attn}} \approx 0.007$). The `gate_proj` module dominates with the highest weight change ($\delta =$ **0.01230**) and alignment ($\rho =$ **0.0767**).

C.2 LLaMA 3.2-3B

LLaMA shows a similar MLP-dominant adaptation pattern (MLP/Attn ratio: **1.16–1.33**$\times$), but with lower absolute magnitudes ($\delta_{\text{mlp}} \approx 0.007$ vs. 0.009 for Qwen). The `gate_proj` remains the top MLP module ($\delta =$ **0.00842**, $\rho =$ **0.0318**), though domain specificity is weaker overall.

C.3 Pythia 2.8B

Pythia (GPT-NeoX architecture) confirms the MLP adaptation pattern with ratios of **1.09–1.28**$\times$. The `dense_h_to_4h` module (analogous to `gate_proj`) shows the highest $\delta =$ **0.00988** and $\rho =$ **0.0291**.

As we claimed in (section 5), (1) MLP weight updates are larger than attention updates, (2) MLP updates are more domain-differentiated (lower cross-domain similarity), and (3) MLP updates are more

Table 14: LLaMA 3.2-3B: Mean relative update magnitude $\bar{\delta}$ across layers.

Domain	δ_{attn}	δ_{mlp}	Ratio
CS	0.00578	0.00731	1.26
EESS	0.00581	0.00773	1.33
Math	0.00567	0.00672	1.19
Physics	0.00568	0.00661	1.16
Q-Bio	0.00582	0.00712	1.22
Stat	0.00579	0.00731	1.26

Table 15: LLaMA 3.2-3B: Per-module summary.

Module	δ	ρ	$\mathcal{M}$
q_proj	0.00517	0.0218	0.0017
k_proj	0.00423	0.0198	0.0011
v_proj	0.00850	0.0225	0.0025
o_proj	0.00787	0.0177	0.0010
gate_proj	**0.00842**	**0.0318**	**0.0009**
up_proj	0.00713	0.0199	0.0009
down_proj	0.00483	0.0218	0.0010

Table 16: Pythia 2.8B: Mean relative update magnitude $\bar{\delta}$ across layers.

Domain	δ_{attn}	δ_{mlp}	Ratio
CS	0.00647	0.00826	1.28
EESS	0.00646	0.00825	1.28
Math	0.00646	0.00728	1.13
Physics	0.00639	0.00696	1.09
Q-Bio	0.00641	0.00707	1.10
Stat	0.00651	0.00741	1.14

Table 17: Pythia 2.8B: Per-module summary.

Module	δ	ρ	$\mathcal{M}$
query_key_value	0.00610	0.0282	0.0019
dense	0.00819	0.0192	0.0011
dense_h_to_4h	**0.00988**	**0.0291**	**0.0011**
dense_4h_to_h	0.00347	0.0250	0.0030

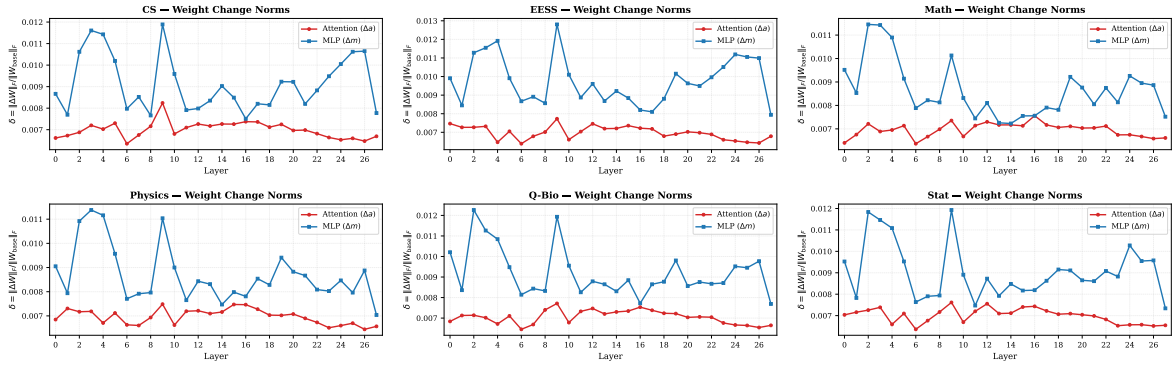

Figure 26: Qwen 2.5-7B: Relative update magnitude δ across layers for all six domains.

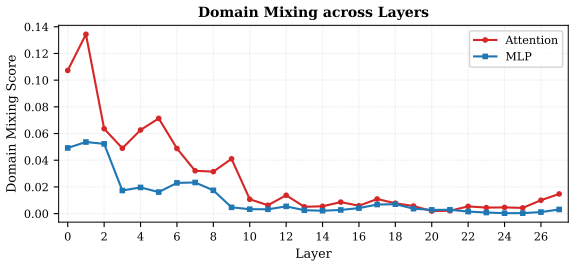

Figure 27: Qwen 2.5-7B: Domain mixing score across layers.

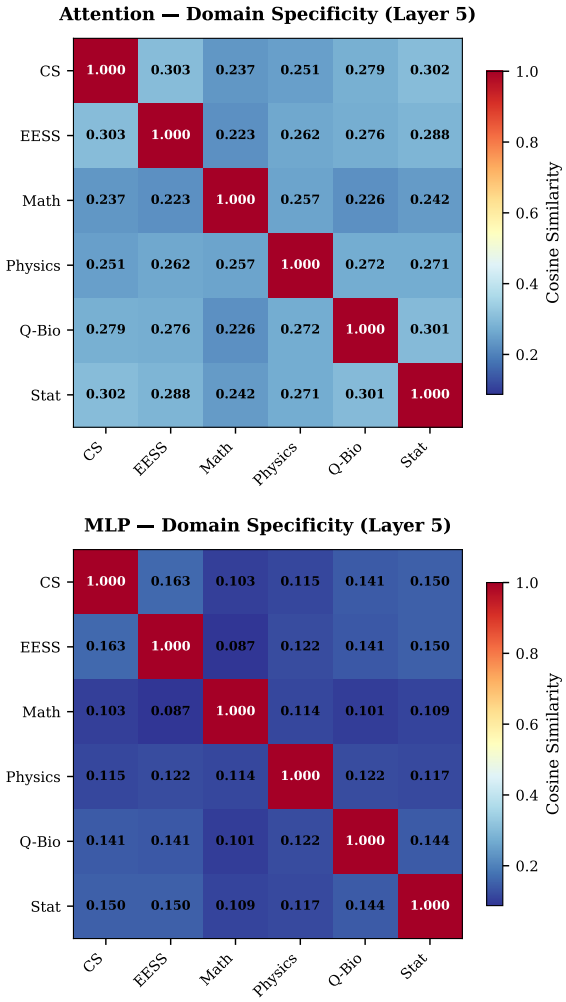

Figure 28: Qwen 2.5-7B: Cross-domain similarity heatmaps.

aligned with domain concept vectors. All three findings are **validated** across all three models. In all three models, the gating module (`gate_proj` in Qwen/LLaMA, `dense_h_to_4h` in Pythia) consistently shows the highest weight change magnitude and domain alignment, supporting the `gate_proj` anomaly as a universal architectural phenomenon across SwiGLU and standard MLP designs.

C.4 Reproducibility

All domain-specific adaptation experiments were conducted on an NVIDIA A6000 GPU (48GB VRAM) utilizing bfloat16 precision across Qwen/Qwen2.5-7B, LLaMA 3.2-3B, and Pythia 2.8B. Low-Rank Adaptation (LoRA) was applied to all linear modules, encompassing attention (`q_proj`, `k_proj`, `v_proj`, `o_proj`) and MLP (`gate_proj`, `up_proj`, `down_proj`, or equivalent Pythia modules), configured with a rank $r = 16$, a scaling factor $\alpha = 32$, and a dropout of 0.1. The training procedure operated with a learning rate of 1×10^{-4}, an effective batch size of 32, and a maximum sequence length of 256, running for 3 epochs ($\sim$282 optimizer steps per domain). Overall, the total computation time was approximately 1.5 hours per model, corresponding to $\sim$15–20 minutes per domain.

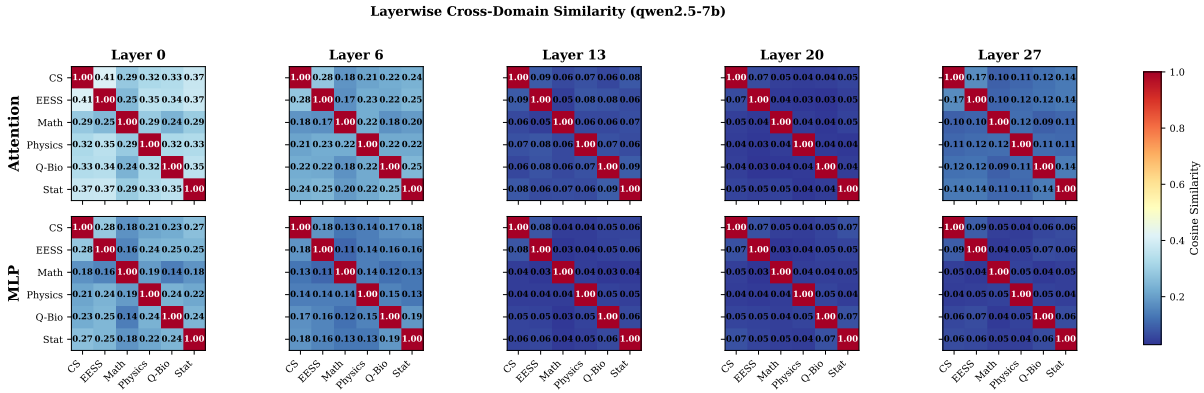

Figure 29: Qwen 2.5-7B: Layer-wise cross-domain similarity heatmaps (Attention vs. MLP). Domain specificity strongly manifests in the MLP layers across the depth of the network.

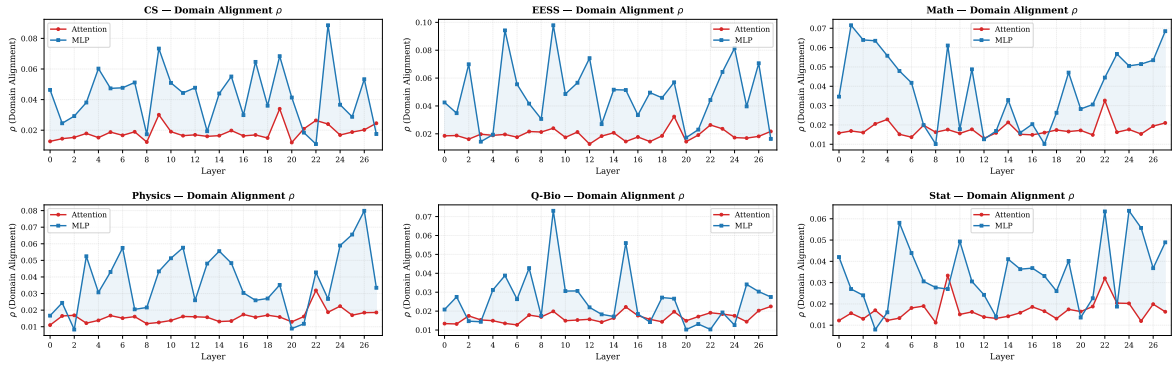

Figure 30: Qwen 2.5-7B: Domain alignment (projection metric ρ) across layers.

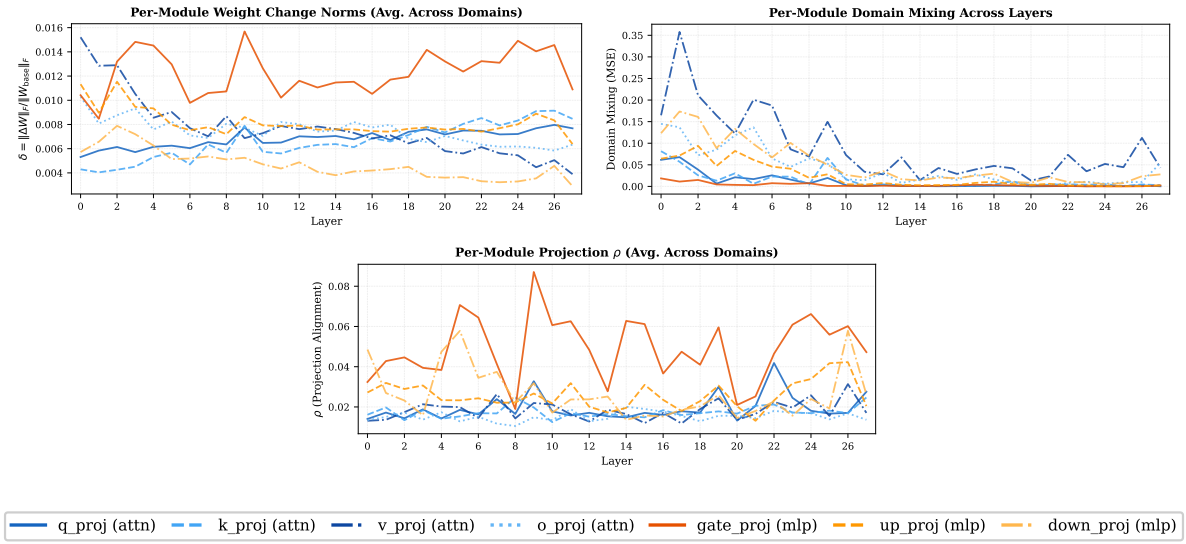

Figure 31: Qwen 2.5-7B: Module-wise metrics.

Figure 32: LLaMA 3.2-3B: Relative update magnitude δ across layers.

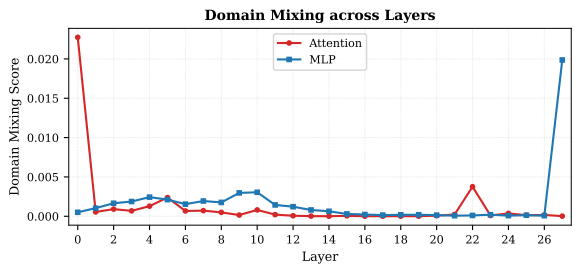

Figure 33: LLaMA 3.2-3B: Domain mixing score across layers.

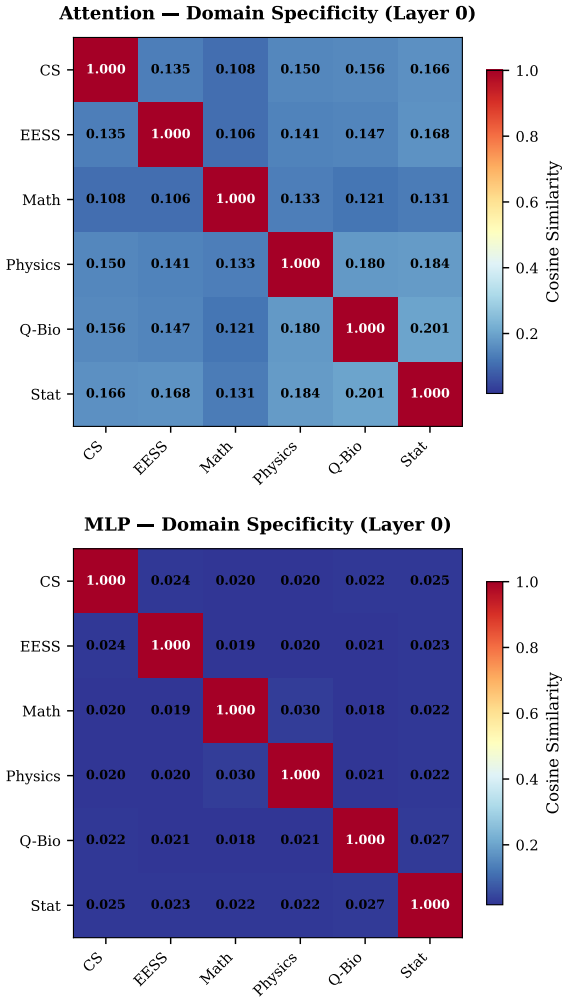

Figure 34: LLaMA 3.2-3B: Cross-domain similarity heatmaps.

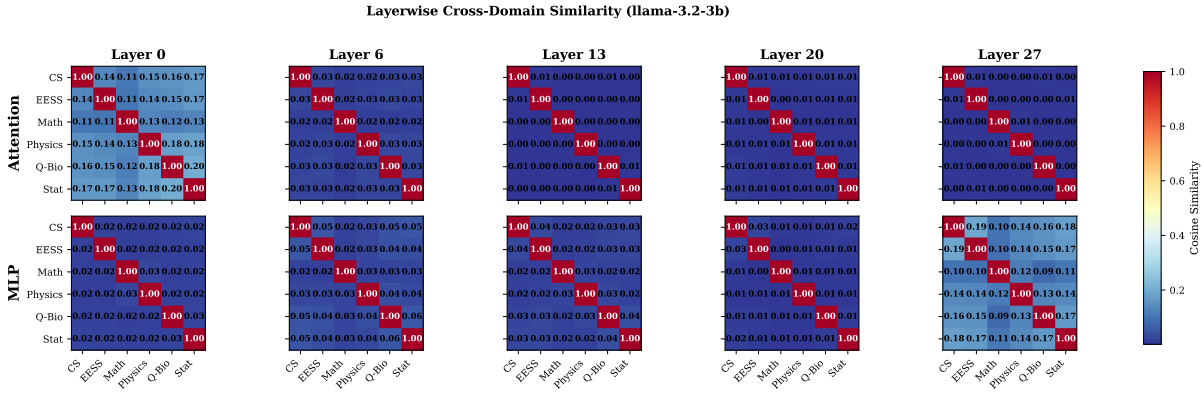

Figure 35: LLaMA 3.2-3B: Layer-wise cross-domain similarity heatmaps (Attention vs. MLP). Domain specificity is much weaker than Qwen and shows less contrast between Attention and MLP.

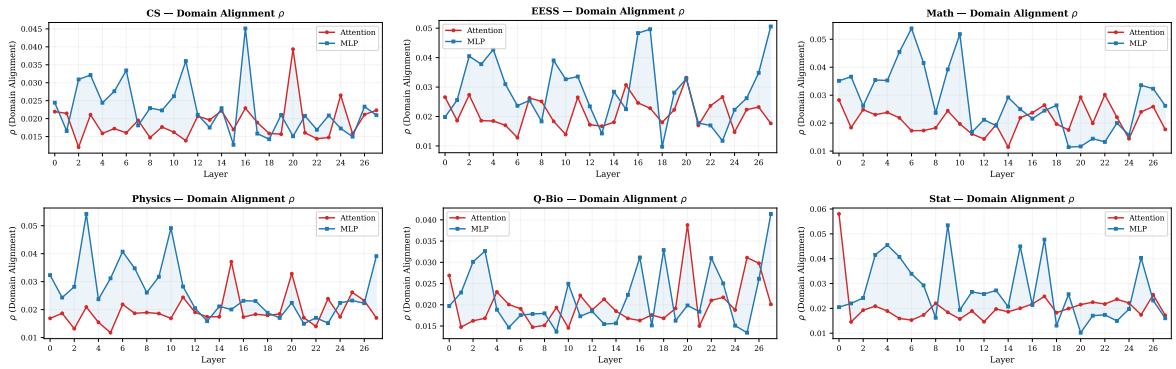

Figure 36: LLaMA 3.2-3B: Domain alignment (projection metric ρ) across layers.

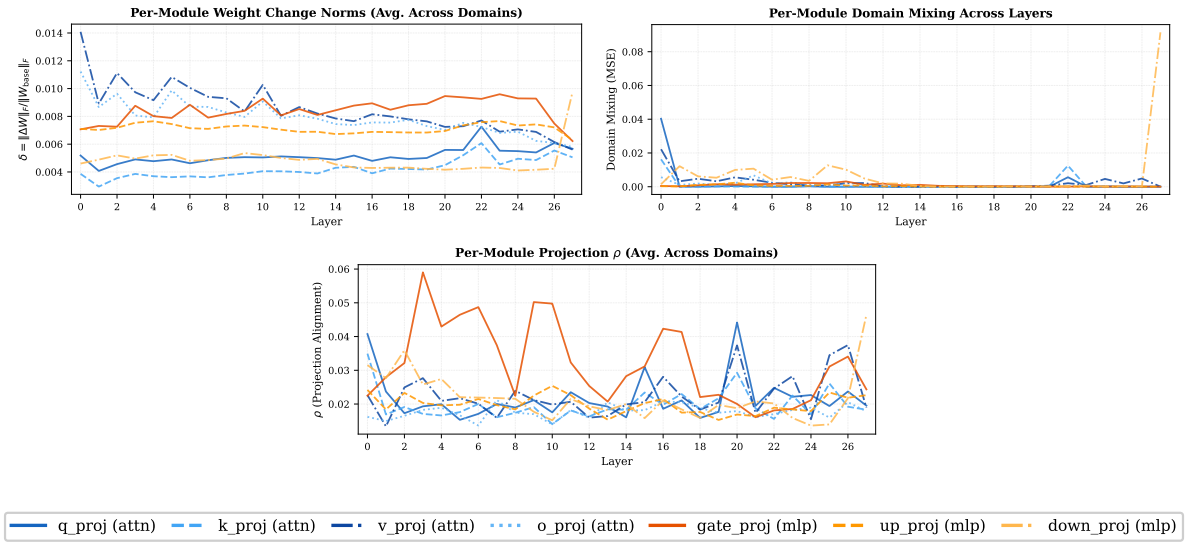

Figure 37: LLaMA 3.2-3B: Module-wise metrics.

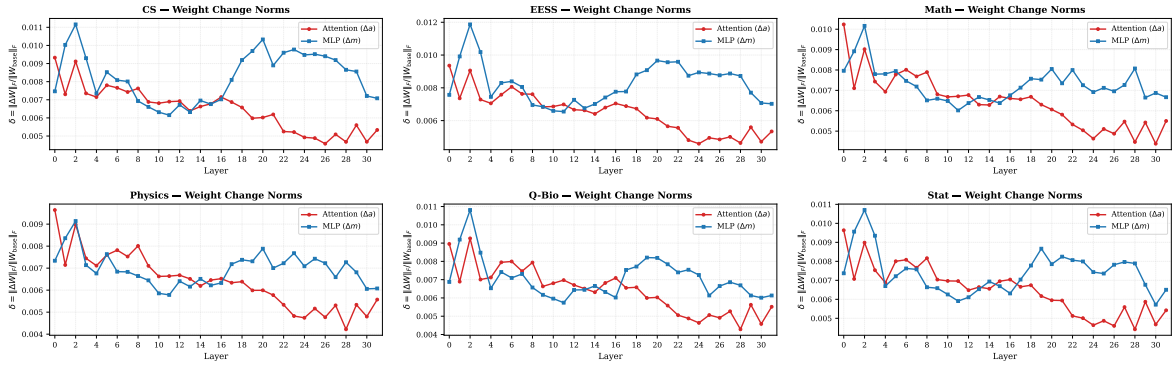

Figure 38: Pythia 2.8B: Relative update magnitude δ across layers.

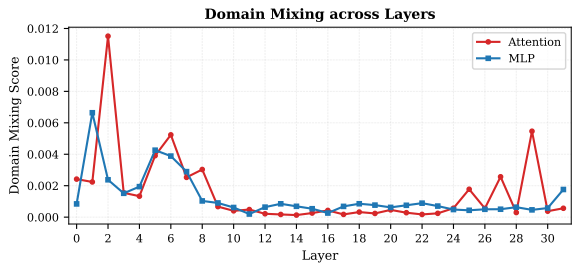

Figure 39: Pythia 2.8B: Domain mixing score across layers.

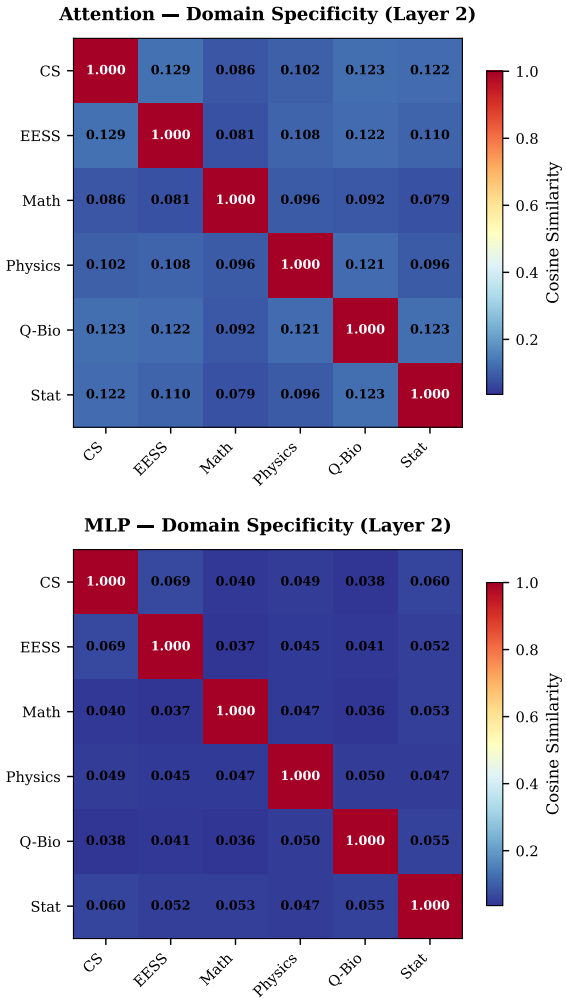

Figure 40: Pythia 2.8B: Cross-domain similarity heatmaps.

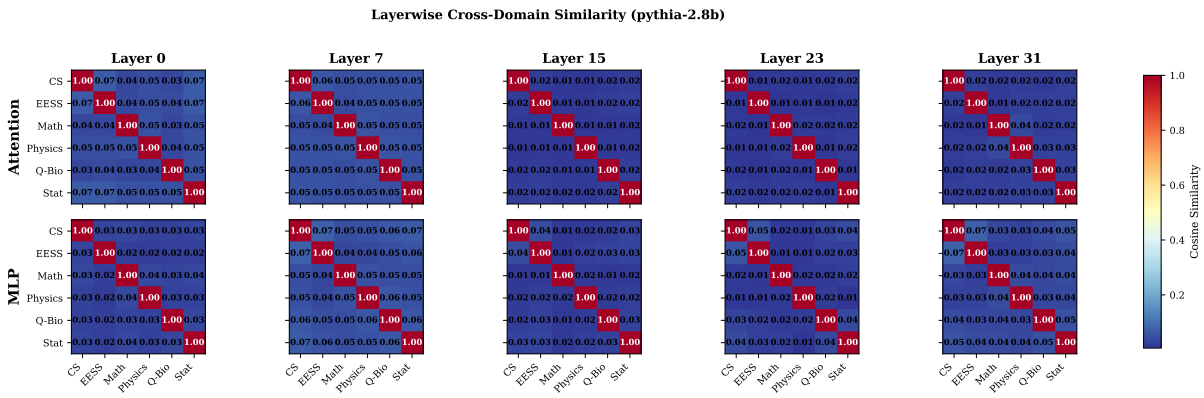

Figure 41: Pythia 2.8B: Layer-wise cross-domain similarity heatmaps (Attention vs. MLP). Cross-domain similarity remains high across all layers, indicating little domain-specific specialization.

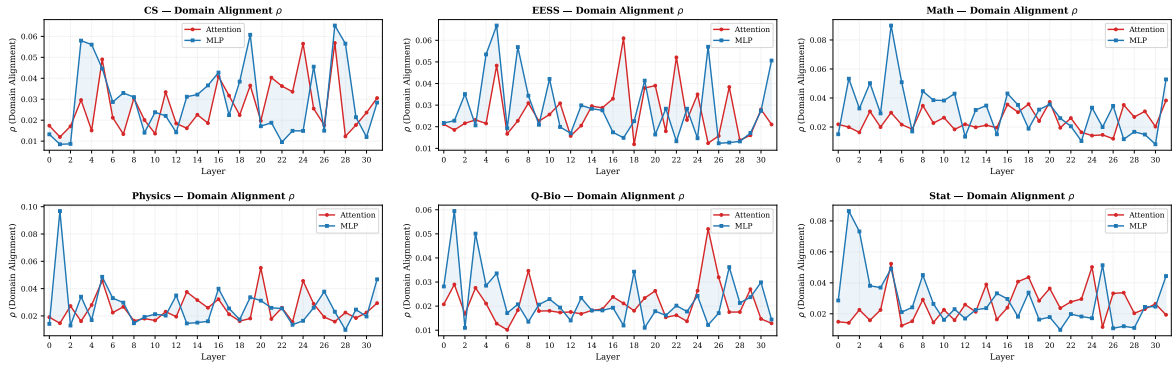

Figure 42: Pythia 2.8B: Domain alignment (projection metric ρ) across layers.

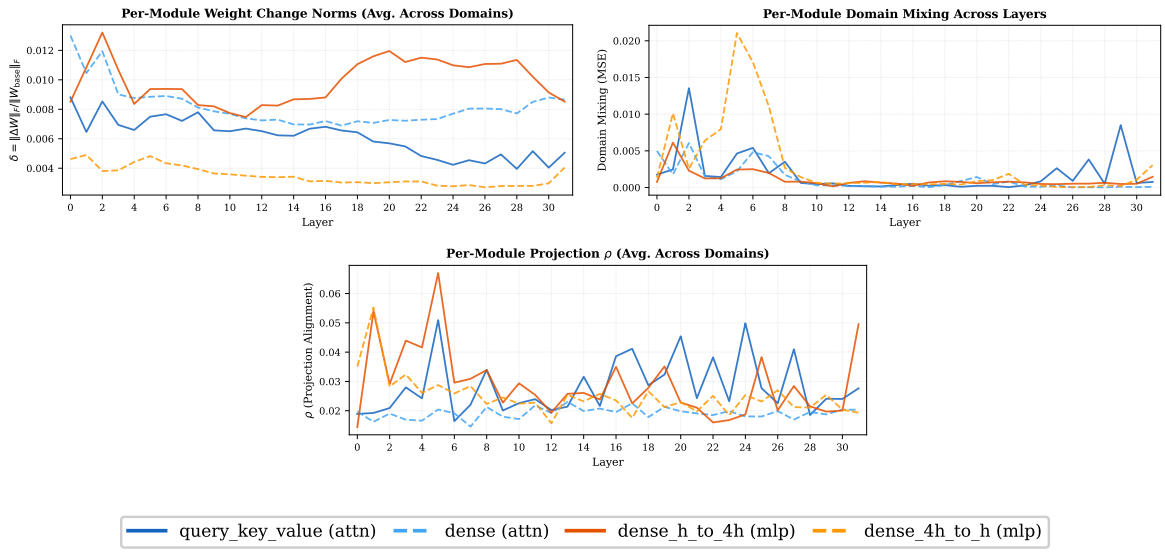

Figure 43: Pythia 2.8B: Module-wise metrics.

D Causal Activation Patching

This appendix provides the full patching formulation (Section 6 describes the intervention conceptually), formal metric definitions, experimental conditions, detailed results tables, and reproducibility information.

D.1 Formal Definitions

Patching intervention. For a source text $x_A \in \mathcal{D}_A$ and target domain B, we replace the attention contribution at layer ℓ^* by:

$$\begin{aligned}\widetilde{\mathrm{Attn}}^{\ell^*}(x_A) &= \mathrm{Attn}^{\ell^*}(x_A) \\ &\quad - \langle \mathrm{Attn}^{\ell^*}(x_A),\, \hat{\mathbf{d}}_A^{\mathrm{attn},\ell^*} \rangle \hat{\mathbf{d}}_A^{\mathrm{attn},\ell^*} \\ &\quad + \alpha \cdot \hat{\mathbf{d}}_B^{\mathrm{attn},\ell^*}\end{aligned} \tag{9}$$

where $\hat{\mathbf{d}}_k^{\mathrm{attn},\ell}$ is the unit domain concept vector for domain k's attention contributions at layer ℓ. The first two terms project out the source-domain component; the third injects a calibrated target-domain component. The same formulation applies when patching MLP contributions.

Calibration of injection magnitude. The injection magnitude α is calibrated on a held-out validation split (200 samples per domain) to prevent circular contamination from the training split used to compute concept vectors:

$$\alpha_B^\ell = \mathbb{E}_{x \sim \mathcal{D}_B}\left[\langle \mathrm{Attn}^\ell(x),\, \hat{\mathbf{d}}_B^{\mathrm{attn},\ell} \rangle\right] \tag{10}$$

This ensures that the injected component has the same expected projection magnitude as genuine domain-B text, making the perturbation naturalistic rather than arbitrary.

Perturbation magnitude check. To verify that interventions remain within the natural activation regime, we compute the perturbation ratio:

$$r = \frac{\|\widetilde{\mathrm{Attn}}^{\ell^*} - \mathrm{Attn}^{\ell^*}\|}{\sqrt{\mathrm{tr}(S_W^{\ell^*})/N}} \tag{11}$$

where the denominator is the average within-class standard deviation of attention contributions at layer ℓ^*. Values of $r \in [0.1, 1.0]$ indicate that the perturbation lies within the natural spread of activations.

Final domain shift score. The primary outcome metric measures how far the patched residual stream has moved toward domain B relative to domain A:

$$s^L(\ell^*) = \left\langle \widetilde{x}^L - x^L(x_A),\, \frac{\mathbf{d}_B^L - \mathbf{d}_A^L}{\|\mathbf{d}_B^L - \mathbf{d}_A^L\|} \right\rangle \tag{12}$$

where x^L is the final-layer residual stream and $\mathbf{d}_k^L$ are the layer-L domain concept vectors. Positive values indicate a shift toward domain B.

Propagation trace and sublayer decomposition. To track how the intervention propagates through the network, we compute the shift score at every sublayer boundary downstream of ℓ^*:

$$s^\ell = \left\langle \widetilde{x}^\ell - x^\ell(x_A),\, \frac{\mathbf{d}_B^\ell - \mathbf{d}_A^\ell}{\|\mathbf{d}_B^\ell - \mathbf{d}_A^\ell\|} \right\rangle \tag{13}$$

We evaluate this at both integer boundaries (x^ℓ, layer input) and half-integer boundaries ($x^{\ell+0.5} = x^\ell + \mathrm{Attn}^\ell(x)$, post-attention), yielding a sublayer-resolved decomposition:

$$\Delta s_{\mathrm{attn}}^\ell = s^{\ell+0.5} - s^\ell \tag{14}$$

$$\Delta s_{\mathrm{mlp}}^\ell = s^{\ell+1} - s^{\ell+0.5} \tag{15}$$

D.2 Experimental Conditions

Table 18 summarises the seven experimental conditions. Hotspot layers (L9, L19, L22) were identified by the FDR peaks in the probing experiment (Section 4). Trough layers (L4, L15, L17) serve as non-hotspot controls.

Condition C5 draws 10 random unit vectors $\hat{\mathbf{r}}$ per domain pair and patches along those directions instead of the domain concept vector, using the same calibrated α, to test whether the causal effect is specific to domain concept directions or is a generic consequence of any perturbation.

Table 18: Experimental conditions for causal activation patching.

Cond.	Component	Layers	Purpose
C1	Attention	9, 19, 22	Hotspot intervention
C2	Attention	4, 15, 17	Non-hotspot control
C3	MLP	9, 19, 22	Hotspot intervention (MLP)
C4	MLP	4, 15, 17	Non-hotspot control (MLP)
C5	Random dir.	9, 19, 22	Specificity control
C6	Attention	0–27	Full layer sweep
C7	MLP	0–27	Full layer sweep

D.3 Complete Focal Results

Table 19 provides the full statistical detail for the hotspot vs. non-hotspot comparison (C1–C4), including standard deviations and domain classifier probability shifts ΔP_{tgt}.

Table 19: Mean final domain shift score s^L and linear probe probability shift ΔP_{target} for focal conditions (averaged over 15 domain pairs). C1/C3 patch at hotspot layers; C2/C4 patch at non-hotspot layers.

Cond.	Comp.	Layer	$\bar{s}^L$	$\sigma(s^L)$	ΔP_{tgt}
C1	Attn	L9	2.34	2.18	4.4×10^{-5}
C1	Attn	L19	25.56	15.63	5.3×10^{-4}
C1	Attn	L22	18.50	5.62	2.6×10^{-4}
C2	Attn	L4	0.52	0.47	1.1×10^{-5}
C2	Attn	L15	0.44	1.11	3.0×10^{-6}
C2	Attn	L17	0.10	1.66	2.0×10^{-6}
C3	MLP	L9	0.21	1.01	8.0×10^{-6}
C3	MLP	L19	17.35	6.53	2.8×10^{-4}
C3	MLP	L22	25.85	15.94	4.7×10^{-4}
C4	MLP	L4	−0.12	0.76	2.0×10^{-6}
C4	MLP	L15	0.07	1.17	-2.0×10^{-6}
C4	MLP	L17	−0.13	1.36	3.0×10^{-6}

Several additional patterns emerge from the complete data. At L9, attention carries a measurable causal signal (2.34) but MLP yields essentially zero (0.21), indicating that at this depth, attention has begun domain routing while MLP has not yet engaged in domain-specific execution. By L22, the MLP (25.85) has surpassed the corresponding attention effect (18.50), marking the crossover to MLP-dominated causal influence.

Specificity control. Table 20 compares targeted concept-direction interventions against random-direction controls (C5). The non-zero random control shifts are expected: any perturbation at a causally influential layer will produce some downstream effect, but random directions project only partially onto the domain axis. The $\sim 2\times$ ratio indicates that roughly half of the targeted intervention's effect is attributable to the *specific direction* of the concept vector, with the remainder reflecting the generic causal influence of the layer.

Table 20: Targeted concept-direction intervention vs. random-direction control (C5). Values are mean s^L across 15 domain pairs (10 random directions per pair).

Layer	Targeted (C1)	Random (C5)	Ratio
L9	2.34	1.00	2.34×
L19	25.56	12.19	2.10×
L22	18.50	9.33	1.98×

D.4 Attention–MLP Crossover Analysis

The crossover between the attention and MLP causal effect curves (Figure 8) is quantitatively informative. At L19, attention patching ($s^L = 25.56$) exceeds MLP patching (17.35) by $1.47\times$. By L22, MLP patching (25.85) surpasses attention (18.50) by $1.40\times$. By L27, MLP dominance is overwhelming (74.32 vs. 7.86, a $9.5\times$ ratio). This crossover provides direct causal evidence for the division of labour: attention routes in mid-layers; MLPs execute in late layers.

D.5 Dose-Response Details

To verify that the intervention operates through a graded, linear mechanism rather than a binary threshold, we sweep the injection magnitude through five levels: $\alpha \in \{0.25\alpha_0,\ 0.5\alpha_0,\ \alpha_0,\ 2\alpha_0,\ 4\alpha_0\}$, where α_0 is the calibrated value from Equation 10.

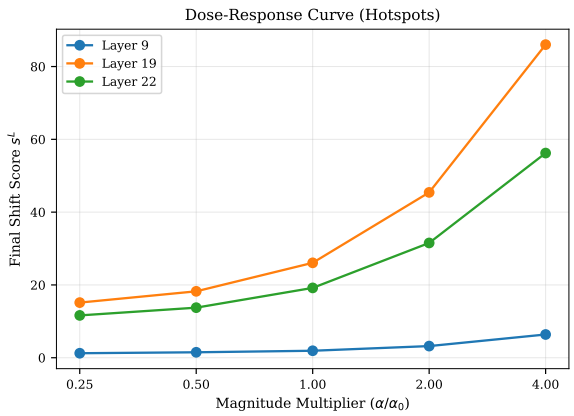

Figure 44: Dose-response curve: final domain shift s^L as a function of injection magnitude multiplier (α/α_0) for hotspot layers L9, L19, and L22 (log-scale x-axis). All three layers show monotonic, approximately linear dose-response relationships.

Table 21: Dose-response values: mean s^L at each magnitude multiplier (averaged over 15 pairs).

Layer	0.25×	0.5×	1.0×	2.0×	4.0×
L9	1.49	1.57	2.34	4.61	8.14
L19	15.01	17.95	25.56	44.68	83.63
L22	11.38	13.58	18.50	29.98	53.41

The dose-response is monotonic and approximately linear at all three hotspot layers (Figure 44, Table 21). Doubling the injection magnitude roughly doubles the final shift (L19: $25.56 \rightarrow 44.68$, ratio 1.75). Quadrupling produces a $\sim 3.3\times$ increase. The sub-linear scaling at high magnitudes suggests mild saturation, consistent with the bounded geometry of the residual stream. Critically, there is no threshold effect: even at $0.25\alpha_0$, the intervention produces a measurable shift, confirming that the concept directions function as continuous steering vectors rather than binary switches.

D.6 Per-Pair Consistency

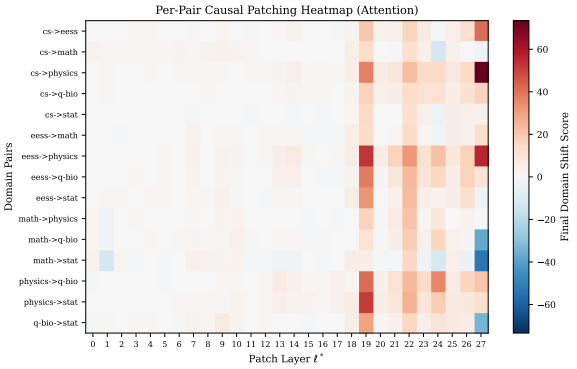

Figure 45: Per-pair causal patching heatmap (attention, C6). Each row is one of the 15 domain pairs; each column is a patch layer. Colour intensity encodes the final domain shift s^L. The L19 column shows uniformly strong shifts across all pairs.

The heatmap (Figure 45) confirms that the L19 attention hotspot is consistent across all 15 domain pairs. No pair shows a qualitatively different pattern: every row exhibits a strong L19 column and a secondary L22 column. The strongest individual shifts occur for pairs involving `physics` (e.g., `eess`→`physics`: $s^L = 54.93$), while same-superfield pairs (e.g., `cs`→`stat`: $s^L = 13.72$) show weaker but still clearly positive shifts. This universality rules out the possibility that the hotspot reflects a quirk of particular domain pairs and instead points to a general-purpose architectural mechanism for domain routing.

D.7 LLaMA 3.2-3B

LLaMA's causal profile is weaker but structurally consistent: attention peaks at L25 ($\bar{s}^L$ = **2.38**), MLP also peaks at L25 ($\bar{s}^L$ = **1.63**). Targeted patching exceeds random by **2.07**× at the hotspot, and dose-response grows monotonically ($1.46 \rightarrow 6.01$).

Table 22: LLaMA 3.2-3B: Experimental conditions. Hotspot layers (L16, L20, L25) were identified by Experiment 1's FDR peaks. Trough layers (L4, L11, L15) serve as non-hotspot controls.

Cond.	Component	Layers	Purpose
C1	Attention	16, 20, 25	Hotspot intervention
C2	Attention	4, 11, 15	Non-hotspot control
C3	MLP	16, 20, 25	Hotspot intervention (MLP)
C4	MLP	4, 11, 15	Non-hotspot control (MLP)
C5	Random direction	16, 20, 25	Specificity control
C6	Attention	0–27	Full layer sweep
C7	MLP	0–27	Full layer sweep

Table 23: LLaMA 3.2-3B: Mean final domain shift score s^L. Attention interventions (C1) peak at **L25** ($\bar{s}^L$ = **2.38**), while MLP interventions (C3) peak at **L25** ($\bar{s}^L$ = **1.63**).

Cond.	Component	Layer	$\bar{s}^L$	$\sigma(s^L)$	ΔP_{tgt}
C1	Attention	L16	1.32	0.46	0.001
C1	Attention	L20	2.26	0.57	0.007
C1	Attention	**L25**	**2.38**	0.47	0.006
C2	Attention	L4	0.03	0.03	0.000
C2	Attention	L11	0.13	0.14	0.000
C2	Attention	L15	1.50	0.66	0.001
C3	MLP	L16	0.92	0.53	0.000
C3	MLP	L20	1.19	0.41	0.000
C3	MLP	**L25**	**1.63**	0.53	0.001
C4	MLP	L4	0.01	0.01	0.000
C4	MLP	L11	-0.04	0.08	0.000
C4	MLP	L15	0.71	0.47	0.000

Table 24: LLaMA 3.2-3B: Targeted vs. random-direction control. Targeted interventions yield strongly higher shifts, peaking at **2.07**× (L25).

Layer	Targeted (C1)	Random (C5)	Ratio
L16	1.32	0.73	1.82×
L20	2.26	1.24	1.82×
L25	**2.38**	**1.15**	**2.07**×

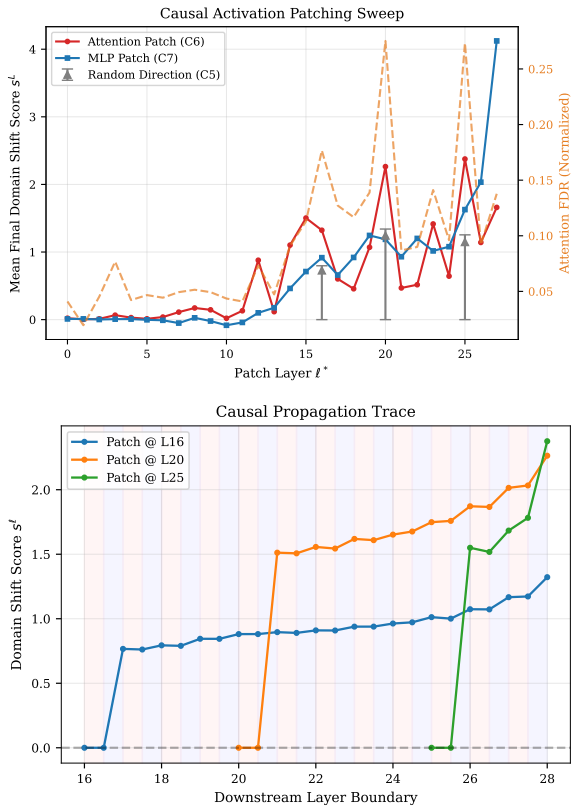

Figure 46: LLaMA 3.2-3B: (Top) Causal effect curve. (Bottom) Propagation trace.

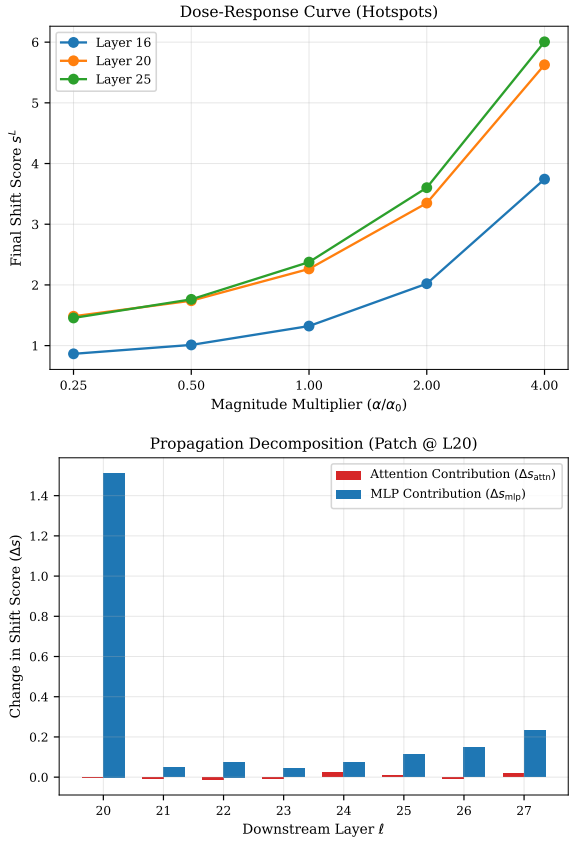

Figure 47: LLaMA 3.2-3B: (Top) Dose-response curve. (Bottom) Propagation decomposition.

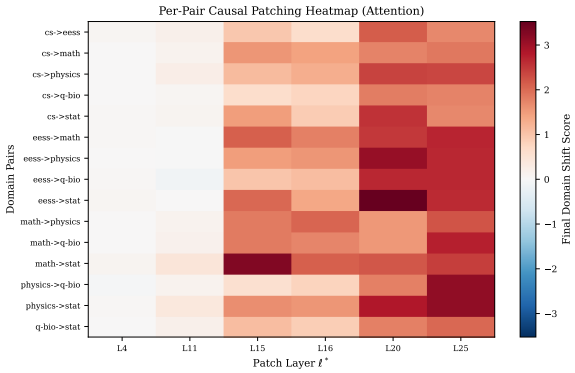

Figure 48: LLaMA 3.2-3B: Per-pair causal patching heatmap.

Table 25: LLaMA 3.2-3B: Dose-response values. Effect sizes grow monotonically, with **L25** exhibiting the most explosive growth to **6.01**.

Layer	**0.25×**	**0.5×**	**1.0×**	**2.0×**	**4.0×**
L16	0.86	1.01	1.32	2.02	3.74
L20	1.48	1.74	2.26	3.35	5.63
L25	**1.46**	**1.76**	**2.38**	**3.60**	**6.01**

D.8 Pythia 2.8B

Pythia exhibits the most symmetric causal profile: attention peaks early at L14 ($\bar{s}^L$ = **5.81**), MLP peaks late at L22 ($\bar{s}^L$ = **2.14**). Targeted/random ratio is **1.78**× and dose-response ranges from 3.85 → 15.92 at L14.

Table 26: LLaMA 3.2-3B: Summary of causal findings. Attention routing is localized at **L25** ($s^L = \mathbf{2.38}$), and MLPs amplify with $\Delta s_{\text{mlp}} > \mathbf{0}$.

Finding	Evidence	Key Value
Attention routing is localised	C1 vs. C2: L25 vs. L4	2.38 vs. 0.03
MLP execution is late-layer	C7 sweep: L25	$s^L = 1.63$
Concept directions are causal	C1 vs. C5: targeted vs. random	2.07× ratio
MLPs amplify, attention is passive	Propagation decomposition (L25)	$\Delta s_{\text{mlp}} > 0$; $\Delta s_{\text{attn}} \approx 0$
Effect is graded, not binary	Dose-response at L25	$1.46 \rightarrow 6.01$
Hotspot is universal	Per-pair heatmap at L25	All 15 pairs positive

Table 27: Pythia 2.8B: Experimental conditions. Hotspot layers (L14, L19, L22) were identified by Experiment 1's FDR peaks. Trough layers (L4, L11, L17) serve as non-hotspot controls.

Cond.	Component	Layers	Purpose
C1	Attention	14, 19, 22	Hotspot intervention
C2	Attention	4, 11, 17	Non-hotspot control
C3	MLP	14, 19, 22	Hotspot intervention (MLP)
C4	MLP	4, 11, 17	Non-hotspot control (MLP)
C5	Random direction	14, 19, 22	Specificity control
C6	Attention	0–31	Full layer sweep
C7	MLP	0–31	Full layer sweep

Table 28: Pythia 2.8B: Mean final domain shift score s^L. Attention interventions (C1) peak at **L14** ($\bar{s}^L = \mathbf{5.81}$), while MLP interventions (C3) peak at **L22** ($\bar{s}^L = \mathbf{2.14}$).

Cond.	Component	Layer	$\bar{s}^L$	$\sigma(s^L)$	ΔP_{tgt}
C1	Attention	**L14**	**5.81**	2.02	0.003
C1	Attention	L19	5.55	3.21	0.011
C1	Attention	L22	4.95	3.08	0.003
C2	Attention	L4	0.13	0.13	0.000
C2	Attention	L11	1.02	0.82	0.000
C2	Attention	L17	3.08	1.91	0.002
C3	MLP	L14	-0.46	0.73	−0.000
C3	MLP	L19	0.64	0.47	0.000
C3	MLP	**L22**	**2.14**	1.31	0.000
C4	MLP	L4	0.04	0.08	0.000
C4	MLP	L11	-0.38	0.21	0.000
C4	MLP	L17	2.32	1.16	0.000

Table 29: Pythia 2.8B: Targeted vs. random-direction control. Targeted interventions yield strongly higher shifts, peaking at **1.78**× (L14).

Layer	Targeted (C1)	Random (C5)	Ratio
L14	**5.81**	**3.26**	**1.78**×
L19	5.55	3.48	1.60×
L22	4.95	2.48	2.00×

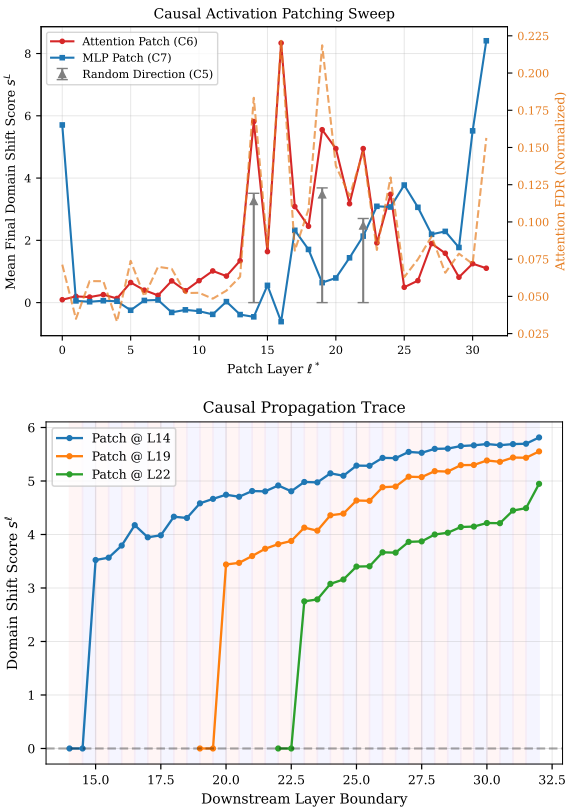

Figure 49: Pythia 2.8B: (Top) Causal effect curve. (Bottom) Propagation trace.

D.8.1 Dose-Response Details

To verify that the intervention operates through a graded, linear mechanism rather than a binary threshold, we sweep the injection magnitude through five levels: $\alpha \in \{0.25\alpha_0,\ 0.5\alpha_0,\ \alpha_0,\ 2\alpha_0,\ 4\alpha_0\}$, where α_0 is the calibrated value from Equation 10.

Table 30: Pythia 2.8B: Dose-response values. Effect sizes grow monotonically, with **L14** exhibiting the most explosive growth to **15.92**.

Layer	0.25×	0.5×	1.0×	2.0×	4.0×
L14	**3.85**	**4.46**	**5.81**	**8.95**	**15.92**
L19	3.95	4.46	5.55	7.87	12.97
L22	3.07	3.69	4.95	7.57	13.00

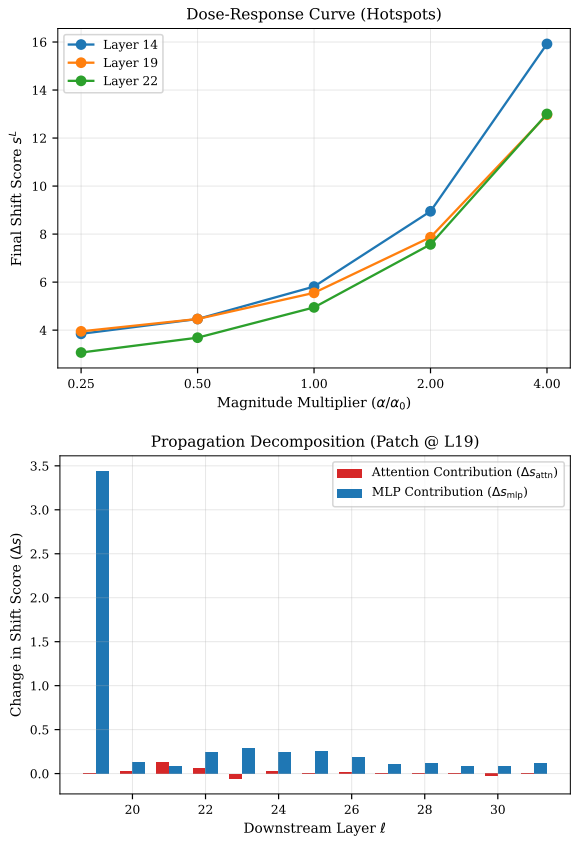

Figure 50: Pythia 2.8B: (Top) Dose-response curve. (Bottom) Propagation decomposition.

D.8.2 Per-Pair Consistency

In section (section 6), we have claimed six findings: (1) attention routing is localised, (2) MLP execution is late-layer, (3) concept directions are causal, (4) MLPs amplify while attention is passive, (5) the effect is graded, and (6) the hotspot

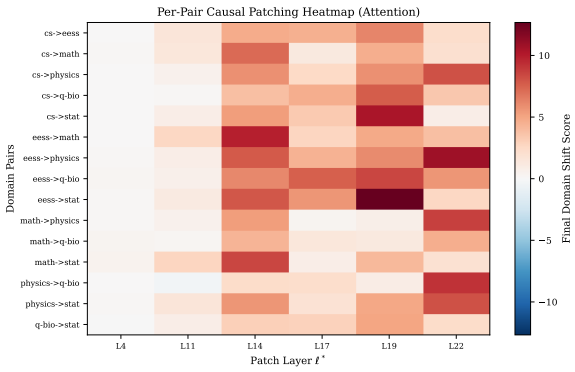

Figure 51: Pythia 2.8B: Per-pair causal patching heatmap.

is universal across domain pairs. All six findings are **validated** across all three models. The absolute magnitudes differ (Qwen $\gg$ Pythia $>$ LLaMA), but the qualitative pattern – localised attention hotspots, monotonic MLP sweep, direction-specific effects, dose-dependent growth holds universally.

D.9 Reproducibility

All causal activation patching experiments were executed on an NVIDIA A6000 GPU (48GB VRAM) using the Qwen/Qwen2.5-7B model in bfloat16 precision (28 transformer layers, hidden dimension $d = 3584$). A held-out validation split of 200 samples per domain (random seed 100) provided the 200 patching samples per source domain. Random control directions (C5) were established using 10 random unit vectors per domain pair at each hotspot layer, requiring a total computation time of $\sim$8 hours.

Table 31: Pythia 2.8B: Summary of causal findings. Attention routing is localized at **L14** ($s^L = \mathbf{5.81}$), and MLPs amplify with $\Delta s_{\text{mlp}} > \mathbf{0}$.

Finding	**Evidence**	**Key Value**
Attention routing is localised	C1 vs. C2: L14 vs. L4	5.81 vs. 0.13
MLP execution is late-layer	C7 sweep: L22	$s^L = 2.14$
Concept directions are causal	C1 vs. C5: targeted vs. random	$1.78\times$ ratio
MLPs amplify, attention is passive	Propagation decomposition (L14)	$\Delta s_{\text{mlp}} > 0$; $\Delta s_{\text{attn}} \approx 0$
Effect is graded, not binary	Dose-response at L14	$3.85 \rightarrow 15.92$
Hotspot is universal	Per-pair heatmap at L14	All 15 pairs positive