# From Behavior to Geometry: A Causal and Geometric Analysis of LoRA-Based Domain Adaptation

**Yizhe Wang[1,2], Liu He[1], Zhenhua Ling[1,2]**
[1]Interdisciplinary Research Center for Linguistic Sciences,
[2]National Engineering Research Center of Speech and Language Information Processing,
University of Science and Technology of China
No.96, JinZhai Road, Baohe District, Hefei, Anhui, P.R.China
{wangyz, zhling}@ustc.edu.cn, heliummn@mail.ustc.edu.cn

**Abstract**

Parameter-efficient fine-tuning with Low-Rank Adaptation (LoRA) often improves a large language model's in-domain performance at the cost of cross-domain generalization. We investigate the mechanistic basis for this trade-off, asking whether LoRA creates new discriminative directions in representation space (emergence) or merely reshapes pre-existing ones. Using a Word Sense Disambiguation testbed, we couple controlled behavioral evaluation with causal localization and geometric diagnostics. We find LoRA learns new, spatially localized discriminative directions in the middle layers of the network, focused at token positions critical for the task. This "subspace extension" account explains why LoRA-tuned models excel on in-domain data but struggle to transfer. As a proof of concept, we introduce a mechanistically informed LoRA configuration that concentrates capacity in the identified layers, promotes rank diversity, and applies light answer-token calibration. Without increasing training budget, it yields consistent improvements in both in- and cross-domain settings, demonstrating that mechanistic insight can guide more efficient adaptation.

**Keywords:** LoRA, Mechanistic interpretability, Domain adaptation, Domain generalization, Word sense disambiguation

## 1. Introduction

Low-Rank Adaptation (LoRA) (Hu et al., 2022) has become a standard technique for adapting large language models (LLMs) to specialized domains, offering an efficient alternative to full fine-tuning (Huang, 2025; Wang et al., 2025a; Christophe et al., 2024). While effective at boosting in-domain performance, LoRA-tuned models often fail to generalize across domains (Afzal et al., 2025; Xu et al., 2025). This trade-off between specialization and transferability presents a critical barrier to deploying models in real-world, multifaceted environments.

To understand and mitigate this poor transfer, we must first uncover the mechanisms of LoRA-based adaptation. However, existing mechanistic analyses offer limited insight. Most studies focus on scenarios where base models already perform well (Radford et al., 2019) or are fine-tuned on general data (Prakash et al., 2024; Wang et al., 2025b), which may not reflect the challenges of true domain specialization where models initially struggle.

To close this gap, we investigate the representational geometry of LoRA-based domain adaptation, centered on a key question: **Does fine-tuning create new discriminative directions in representation space (emergence), or does it primarily amplify and reshape existing ones (reshaping)?** The answer carries significant practical weight. Reshaping favors light interventions (elicitation, calibration), whereas emergence calls for targeted capacity where it is used (rank/module/layer choices or small DAPT), with large-scale pretraining or architectural changes reserved for cases where targeted steps no longer help. While weight-space analyses show LoRA can introduce new high-ranking singular vectors (Shuttleworth et al., 2024), these findings do not explain where the computation changes or whether those changes are causally necessary for improved performance. Our work bridges this explanatory gap.

We investigate LoRA-based domain adaptation using Word Sense Disambiguation (WSD), a task that provides an ideal lens for connecting behavioral outcomes to geometric mechanisms. First, base LLMs exhibit significant headroom on domain-specific senses, creating a clear and controllable learning problem. Crucially, WSD is token-localized—decisions are grounded in specific lemmas and their context—which permits the precise, position-targeted causal interventions necessary for our analysis. This causal tractability is matched by a geometric one: discrete sense inventories align naturally with the compact, interpretable subspaces required for our diagnostics, enabling us to rigorously distinguish whether LoRA creates new discriminative directions or remaps existing ones.

Our analysis connects behavior to mechanism in three stages: (i) Controlled behavioral evaluation: a $2\times2$ transfer matrix separates in-domain gains from true cross-domain generalization and isolates domain effects from task mechanics (Suresh et al.,



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 2178–2189
11–16 May 2026. 

2023); (ii) causal localization via windowed activation patching to identify the necessary changes in specific layers and token positions; and (iii) geometric characterization using convergent diagnostic tests to determine if LoRA introduces new subspaces or reshapes existing ones. Our experiments are conducted on Llama-3.2-1B, with replications on a 3B model to test for scale dependence.

Our findings provide strong evidence for emergence. We show that LoRA learns new, spatially localized discriminative directions for WSD. These changes are concentrated in the middle layers of the model and are localized to the specific token positions crucial for the task. Post-adaptation, the new discriminative axes show minimal alignment with pre-existing directions, supporting a subspace extension account: LoRA adds new, task-specific dimensions rather than globally reorganizing the representation space. We translate these findings into a mechanistically informed training strategy that restricts LoRA to causally critical layers. This approach improves both efficiency and performance on in-domain and cross-domain tasks, demonstrating the practical value of our analysis.

**Contributions.** Our main contributions are: (1) A principled, three-stage framework connecting behavioral, causal, and geometric analyses to distinguish between emergence and reshaping in domain adaptation. (2) A carefully designed WSD testbed that enables precise, token-level causal analysis. (3) Strong empirical evidence that LoRA adaptation operates via emergence, creating new, localized subspaces that explain its poor cross-domain transfer. (4) A proof-of-concept for mechanistically-informed fine-tuning, showing that our analytical insights can be used to improve both the efficiency and accuracy of LoRA.

## 2. Related Work

**The Limits of Behavioral Analysis in Domain Adaptation.** Domain adaptation is crucial for specializing large language models, with Parameter-Efficient Fine-Tuning (PEFT) methods like LoRA (Hu et al., 2022) now standard practice for enhancing in-domain performance (Huang, 2025; Wang et al., 2025a; Christophe et al., 2024). However, behavioral studies reveal a complex trade-off: fine-tuning often improves in-domain accuracy at the cost of out-of-domain generalization (Chronopoulou et al., 2019). The degree of this specialization depends on factors like task, domain similarity, and model scale (Mosbach et al., 2020). While recent work has benchmarked PEFT methods to measure transferability (Afzal et al., 2025) and catastrophic forgetting (Biderman et al., 2024), these analyses remain focused on what happens to performance metrics, not why the model's internal mechanisms fail to generalize. Moreover, many experimental designs conflate domain shifts with task shifts (Suresh et al., 2023), making it difficult to isolate the effects of domain overfitting from inadequate task learning.

**Mechanistic Interpretability of LoRA Fine-Tuning.** To understand how LoRA fine-tuning operates, researchers have turned to mechanistic interpretability (MI) (Elhage et al., 2021), which aims to explain model behavior by localizing computations to specific components, such as layers or attention heads. Circuit analysis suggests fine-tuning primarily enhances pre-existing capabilities (Prakash et al., 2024). Complementing this, Wang et al. (2025b) find that the most significant changes occur not in the components themselves (nodes) but in the connections between them (edges), characterizing fine-tuning as a circuit "rewiring" process. Other studies successfully identify sparse, task-relevant pathways, pinpointing mid-layer MLPs and select attention heads as critical (Lee, 2025; Nijasure et al., 2025). While insightful, this line of work has largely targeted general-domain tasks rather than domain specialization and has not systematically connected its findings to the behavioral patterns observed during domain adaptation. Crucially, the relationship between geometric changes in the model's representational space and domain specialization remains unexplored.

Our work bridges this gap by converging these two lines of inquiry. We introduce a unified methodology that combines: (i) a controlled, same-task cross-domain evaluation to isolate domain-specific effects; (ii) windowed activation patching to identify functional loci with fine-grained positional granularity; and (iii) geometric analysis to distinguish whether LoRA reshapes existing representations or creates new ones. This approach allows us to make causal claims about how LoRA encodes specialized knowledge, paving the way for more robust and efficient domain adaptation strategies.

## 3. EXPERIMENTAL SETUP AND ANALYSIS PIPELINE

We investigate where and how LoRA fine-tuning encodes domain-specific knowledge through word sense disambiguation (WSD). Our analysis follows a three-stage pipeline: (i) establish baseline performance and measure specialization-generalization trade-offs (§3.2), (ii) localize functional changes via drift analysis and causal intervention (§3.3), and (iii) characterize the geometric nature of learned representations (§3.4). This progression moves from behavioral observation to mechanistic understanding, culminating in practical insights for improved fine-tuning strategies.

### 3.1. Task, Datasets, and Base Model

**Model.** We analyze Llama-3.2-1B, a 16-layer decoder-only Transformer. Key findings are replicated on Llama-3.2-3B to ensure scalability.

**Task and Data.** We use two datasets with an identical binary WSD structure: WiC (general-domain English) and BioWiC (biomedical text). This same-task, cross-domain setup allows us to isolate the effects of domain shift. See Table 1 for dataset statistics.

| Dataset | Train | Dev | Test |
|---|---|---|---|
| WiC | 5,428 | 638 | 1,400 |
| BioWiC | 17,156 | 1,000 | 2,000 |

Table 1: Statistics of datasets used for fine-tuning and evaluation.

### 3.2. Stage 1: Specialization versus Cross-Domain Generalization

To understand how fine-tuning works, we first measure what it achieves. We assess whether LoRA promotes general task learning that transfers across domains or creates domain-specific solutions by evaluating three conditions:

1. **Baseline:** Base model performance on BioWiC and WiC in zero-shot and few-shot (4 in-context examples: 2 positive, 2 negative) settings.
2. **BioWiC-tuned:** LoRA fine-tuned on BioWiC, evaluated on both BioWiC (in-domain) and WiC (cross-domain).
3. **WiC-tuned:** LoRA fine-tuned on WiC, evaluated on both WiC (in-domain) and BioWiC (cross-domain).

This bidirectional evaluation controls for dataset-specific artifacts; symmetric transfer failures would indicate that domain specialization is the primary bottleneck.

**Prompting and Evaluation.** We use instruction-style prompts with a constrained "Yes"/"No" output space to minimize decoding confounds:

- **BioWiC:** "Do '{term1}' in Sentence 1 and '{term2}' in Sentence 2 have the same or similar meaning? Answer with 'Yes' or 'No'."
- **WiC:** "Do '{the target word}' in Sentence 1 and '{the target word}' in Sentence 2 have the same or similar meaning? Answer with 'Yes' or 'No'."

At inference, we compare the conditional probabilities of the "Yes" and "No" tokens and report Accuracy and Macro-F1.

**Fine-Tuning Configuration.** We apply LoRA to Llama-3.2-1B with rank $r = 32$, $\alpha = 64$, dropout $0.05$, targeting all attention $(q, k, v, o)$ and MLP projections. We train for 3 epochs using a cosine learning rate schedule $(1 \times 10^{-4})$ and report the final results as the mean of three runs with different random seeds.

### 3.3. Stage 2: Localizing Functional Changes

Having observed the model's behavior, we next ask where in the network specialization occurs. Localization reveals whether adaptation is diffuse or concentrated in specific computational stages and identifies which components to target for regularization or intervention. We adopt an observe-then-intervene workflow (Nanda and Bloom, 2022) to identify the key computational loci.

**Step 2.1: Activation Drift Heatmap.** LoRA perturbs internal computation via low-rank updates. To rapidly triage where changes occur, we compute per-head activation drift across the whole model ($L$ layers, $H$ heads; here $L = 16$, $H = 32$) as a label-free screen. Activations are taken at the pre-`o_proj` concatenation of head outputs to avoid confounds from the `o_proj` matrix and thus isolate changes arising from Q/K/V computation and attention routing (Elhage et al., 2021). Because WSD relies on the target lemma and its local context, we define a lemma window $W(\mathbf{x})$ consisting of all tokens within $\pm R$ (here, $R = 2$) positions of any lemma in either sentence, and average activations within this window (Belrose et al., 2023). Drift is the cosine distance between base and fine-tuned vectors, averaged over the dataset. The resulting heatmap provides a correlation-only overview and prioritizes loci for subsequent causal tests.

### 3.4. Stage 3: Geometric Analysis of Sense Subspaces

Localization tells us where changes happen; geometric analysis tells us how. To distinguish emergence from reshaping, we combine three methods: cross-readout probing, geometric alignment, and causal ablation—to triangulate the answer. These three methods provide complementary perspectives: cross-readout tests functional preservation, alignment measures geometric similarity, and ablation establishes causal necessity.

**Step 3.1: Cross-Readout Probing.** We train L2-regularized logistic-regression probe (scikit-learn, solver=lbfgs, class_weight=balanced) on raw hidden states taken from the residual stream after the MLP sublayer. Probes are fit separately

on pre- and post-fine-tuning activations for each layer/position pair. The regularization strength $C \in \{0.3, 1.0, 3.0, 10.0\}$ is selected on the dev set; we then refit on train+dev with the chosen $C$ (max_iter=2000) and evaluate once on the test set. We assess feature interchangeability by evaluating probes crosswise (pre→post and post→pre) (Bansal et al., 2021). If probes transfer bidirectionally with minimal performance loss, the underlying discriminative geometry is preserved (reshaping). Asymmetric transfer, particularly if post-trained probes fail on pre-trained features, suggests the fine-tuned space contains new dimensions absent from the base model.

**Step 3.2: Geometric Alignment.** We quantify alignment via cosine similarity between probe weight vectors. Since a linear probe's weight defines the discriminative direction for the target concept (Geva et al., 2021a), $\cos \approx 1$ indicates reuse of the same axis, while low cosine implies rotation or novelty.

**Step 3.3: Causal Projection Ablation.** We test causal necessity by removing components of the hidden state along task-relevant linear directions and measuring the resulting performance drop. This projection ablation follows the INLP and amnesic-probing paradigm, which use linear subspace removal to make causal claims about whether a concept/direction is functionally required—offering an intervention-based complement to purely correlational analyses (Ravfogel et al., 2020). To determine whether low alignment signals new, functionally important directions, we ablate along candidate axes and quantify degradation as a function of ablation strength. Selective vulnerability—sharp degradation for POST but not PRE—confirms functionally important *new* post-only directions; comparable drops suggest reshaping of shared structure.

## 4. Experimental Results

We present our findings following the three-stage analysis pipeline. We begin by characterizing the model's behavior through cross-domain transfer experiments (§4.1). Next, we localize the source of these behaviors using causal analysis (§4.2). We then analyze the underlying geometric transformations to distinguish representational reshaping from emergence (§4.3). A replication on a larger model confirms these mechanisms and reveals scale-dependent shifts (§4.4). Finally, we demonstrate how these mechanistic insights can inform more effective fine-tuning strategies (§4.5).

### 4.1. LoRA Specializes to Domains, Not General Tasks

Our cross-domain evaluation reveals a clear pattern of domain specialization. As shown in Table 2, models fine-tuned with LoRA achieve strong in-domain performance but fail to generalize to the same task in a different domain.

For instance, Llama-3.2-1B tuned on BioWiC achieves 76.8% accuracy on its in-domain test set but plummets to 57.1% on WiC—barely outperforming the few-shot baseline. The reverse holds true for a model tuned on WiC, which performs reasonably in-domain (62.1%) but falls to near-chance levels on BioWiC (51.0%). Because this transfer failure is symmetric, we can rule out dataset-specific artifacts. The results confirm that LoRA is learning domain-specific features rather than a general, domain-agnostic strategy for word sense disambiguation.

Notably, fine-tuning on BioWiC is more stable (76.80% ± 0.30 accuracy) than on WiC (62.14% ± 9.85 accuracy). This suggests that the specialized terminology in biomedical text provides a stronger, more concentrated learning signal than the diffuse patterns in general-domain text.

### 4.2. Specialization Localizes to Middle-Layer MLPs

Having established domain specialization, we now identify where in the network this specialization occurs.

**Drift Analysis Identifies Middle Layers as the Locus of Change.** To map the effects of fine-tuning, we first measured activation drift across all layers. The resulting heatmap (Figure 1) shows that fine-tuning predominantly alters representations in the middle layers (7–9), while early and late layers remain relatively stable. This finding provides an initial, correlational signpost, suggesting that these middle layers are the primary site of task-specific processing, a conclusion that aligns with prior work (Nijasure et al., 2025).

**Causal Patching Confirms MLPs Drive Behavior.** To verify that these observed changes are causal, we used activation patching to measure each module's contribution to the model's performance on the task. The results confirm that the MLP modules in layers 8-10 are the primary drivers of the learned behavior, with prediction rescue rates exceeding 80%, as shown in Figure 2. While attention modules have a smaller direct impact, Layer 7 attention is a notable exception, achieving a 57.3% rescue rate.

This suggests a functional division of labor: the attention module in Layer 7 acts as a critical in-

| Method | Train Domain | Test Domain | Shots | Accuracy | Macro-F1 |
|---|---|---|---|---|---|
| Base Llama | – | BioWiC | zero-shot | 50.05% | 34.00% |
| Base Llama | – | BioWiC | few-shot | 50.90% | 38.50 |
| Base Llama | – | WiC | zero-shot | 50.00% | 33.00% |
| Base Llama | – | WiC | few-shot | 50.50% | 39.00% |
| LoRA (BioWiC) | BioWiC | BioWiC | zero-shot | 76.80% | 76.74% |
| LoRA (BioWiC) | BioWiC | WiC | zero-shot | 57.07% | 55.30% |
| LoRA (WiC) | WiC | WiC | zero-shot | 62.14% | 57.01% |
| LoRA (WiC) | WiC | BioWiC | zero-shot | 51.00% | 36.26% |

Table 2: Cross-domain transfer evaluation with Llama-3.2-1B. All results report mean results over three random seeds.

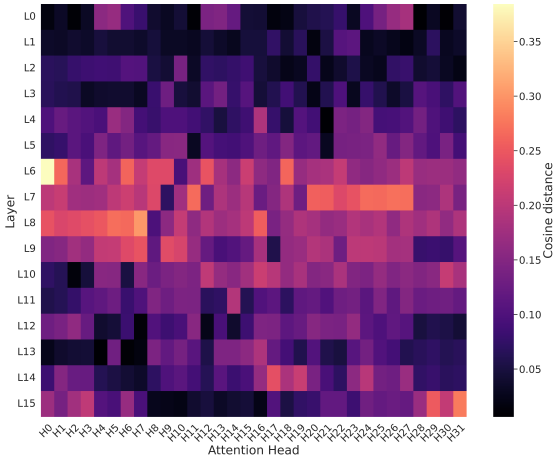


Figure 1: Per-head activation drift measured at pre-o_proj position within lemma window (±R tokens from target terms). Warm colors indicate larger representational changes. Drift concentrates in middle layers (L7-L9), while early and late layers remain stable.

formation router, identifying relevant context and channeling it to the subsequent MLP layers (8-10) for domain-specific processing. The dominance of MLPs is consistent with their established role in storing factual and task-specific knowledge (Geva et al., 2021b).

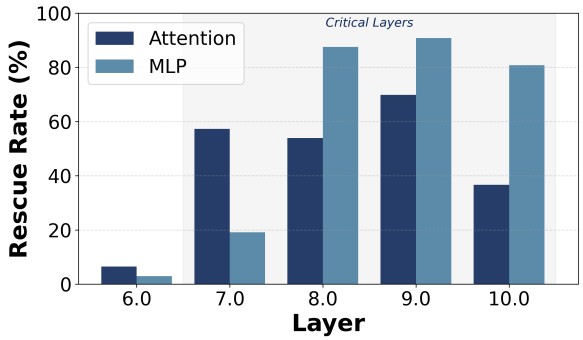


Figure 2: Layer-wise rescue rates for attention and MLP modules when patching all tokens.

**Effects are Functionally and Positionally Localized.** Further analysis using window-specific patching reveals that these changes are highly localized to specific token positions. For instance, Layer 7 attention is most impactful when patching the ANSWER window, supporting its role as a router that directs information toward the final classification position (Figure 3). This positional asymmetry supports a targeted-rewriting hypothesis: under the residual-stream view, attention aggregates context and the readout classifies at the decision point (Elhage et al., 2021); our causal tests indicate that LoRA primarily modifies these decision-point computations. In contrast, the causally critical MLPs (Layers 8-10) show more distributed effects, with the highest rescue rates achieved when patching all tokens, indicating that while their function is specialized, their computation still leverages the full sequence.

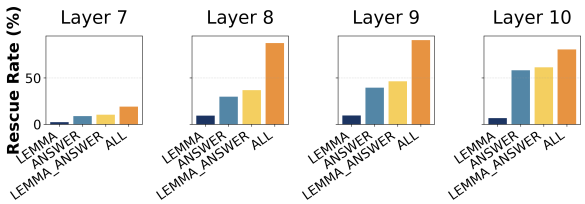


Figure 3: Rescue rates across different token windows for key layers.

Finally, we find that the effectiveness of these mechanisms depends on local context. As shown in Figure 4, expanding the context window around the target lemma from 0 to 4 tokens substantially improves rescue rates for the MLPs. This radius-dependence confirms that the model has learned a core principle of WSD: leveraging surrounding words to resolve ambiguity (Zhong and Ng, 2010).

### 4.3. Geometric Analysis Reveals Emergent Discriminative Axes

Our localization analysis identified where adaptation occurs; we now investigate how this adaptation

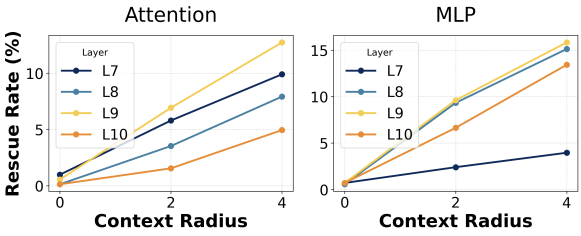


Figure 4: Effect of context radius (R) on rescue rates for LEMMA window.

transforms the model's internal geometry. To distinguish between reshaping and emergence, we employ a series of three convergent diagnostic tests.

**Diagnostic 1: Low Alignment Suggests New Discriminative Directions.** We first assessed the geometric alignment between the pre- and post-fine-tuning spaces. After confirming that linear probes could effectively capture sense distinctions in both, we treated their learned weight vectors as descriptors of the primary "discriminative directions" for the task.

Figure 5 shows the cosine similarity between these directions. The result is striking: at the crucial answer positions, the pre- and post-fine-tuning directions are nearly orthogonal (cosine similarity of 0.22–0.26 across layers 8–10). In contrast, lemma positions retain moderate-to-high alignment. This provides the first signature of emergence: the model isn't changing how it thinks about a word like "bank" in general; it's creating a new, specialized mechanism at the end of the sentence to decide which meaning of "bank" is correct for this specific task.

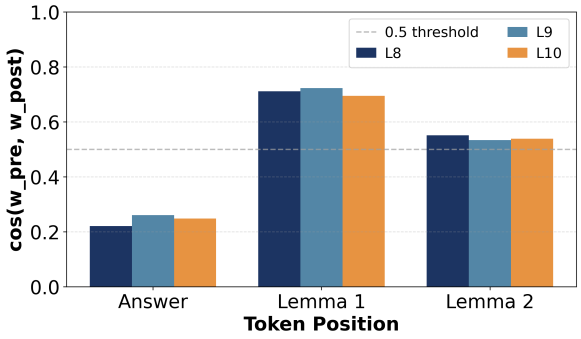


Figure 5: Geometric alignment between pre and post probe weights. Cosine similarity between probe weight vectors across layers and positions.

**Diagnostic 2: Asymmetric Transfer Confirms a Dimensional Extension.** Cross-readout analysis provides the second, independent line of evidence for emergence. We tested whether probes trained on one representation space could generalize to the other. The results reveal a critical asymmetry (Figure 6):

- **Pre→Post transfer**: Pre-trained probes maintain 72-74% accuracy on post-fine-tuned features.
- **Post→Pre transfer**: Post-trained probes collapse to 55-56% on pre-fine-tuned features.

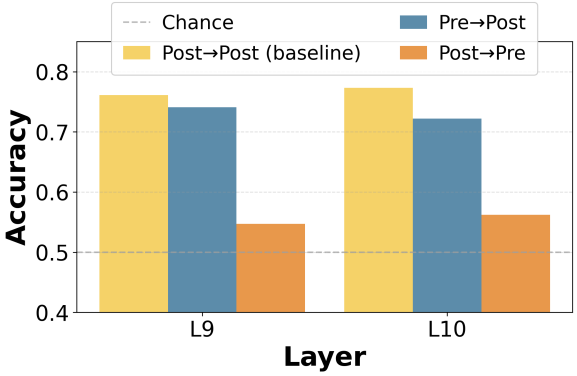


Figure 6: Cross-readout probe transfer at answer position Transfer accuracy of probes trained on pre-fine-tuning (PRE) versus post-fine-tuning (POST) representations at L9-L10 answer positions.

This asymmetry is diagnostic. If fine-tuning were simple reshaping, transfer would succeed in both directions. The failure of the post-trained probe on pre-trained features demonstrates that it relies on new, task-critical information that is simply not linearly accessible in the base model's representations—the definitive signature of emergence.

**Selective Ablation Establishes Causal Primacy of New Directions.** Finally, a causal ablation experiment confirms that the model's behavior depends on these new directions. We selectively removed vector components along the pre-trained and post-trained discriminative directions and measured the impact on model accuracy. As shown in Figure 7, the results are unambiguous:

- **POST direction ablation**: 76% to 50% at $\alpha = 1$, reaching chance by $\alpha = 2$ for L9 and L10.
- **PRE direction ablation**: 76% to 74% at $\alpha = 4$ for L9 and to 72% for L10.

This dramatic selectivity demonstrates that the fine-tuned model's predictions are causally dependent on the newly learned geometric axes, not on a refinement of the original ones.

The convergence of three independent analyses provides strong evidence for emergence: LoRA

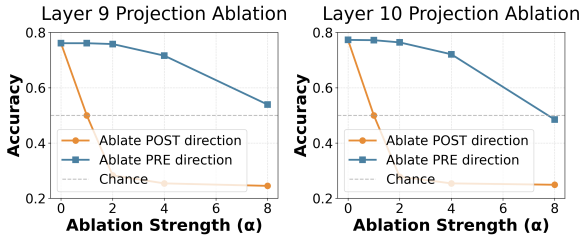


Figure 7: Causal projection ablation results. Accuracy degradation when ablating components along POST-trained versus PRE-trained directions at L9-L10.

creates emergent discriminative subspaces that extend rather than reshape the base model's representation space. This "subspace extension" mechanism differs from both pure feature amplification (which would preserve alignment) and complete rotation (which would destroy transfer in both directions). The finding that pre-trained probes partially succeed on post-features while post-trained probes fail on pre-features indicates that fine-tuning adds information dimensions absent from the original representation space.

Importantly, this provides a mechanistic contrast to Ilharco et al. (2023), who characterize macroscopic weight-space "task vectors" arising from full fine-tuning and study their compositional properties across diverse tasks. In our setting, parameter-efficient domain adaptation via LoRA manifests as highly localized geometric change: the emergent discriminative subspace is concentrated in critical middle layers and decision token positions, and is largely orthogonal to pre-adaptation probe directions. This geometry also complements accounts of in-context learning (ICL): whereas ICL can yield transient task alignment during the forward pass (Hendel et al., 2023), LoRA induces persistent specialization by instantiating new discriminative directions in the model's internal representations.

## 4.4. Replication on Llama-3.2-3B

To assess the robustness of our findings, we replicated the entire analysis on a larger Llama-3.2-3B model. The results confirm that the core mechanism of emergence is conserved across model scales, while also exposing intriguing shifts in how the computation is implemented.

**Core Behavioral and Geometric Patterns are Conserved.** The 3B model recapitulates the fundamental patterns observed in the 1B model. Behaviorally, the trade-off between specialization and generalization persists: fine-tuning on BioWiC yields strong in-domain accuracy (78.2%) but limited transfer to the general WiC domain (60.5%). Mechanistically, the locus of change remains concentrated in the middle layers and is most pronounced at the decision-critical answer position.

The three primary diagnostics for emergence hold true in the 3B model, providing strong evidence that this mechanism is not an artifact of scale:

- **Geometric Misalignment at the Decision Token:** Pre- and post-fine-tuning probe directions show low cosine similarity at the answer position ($\approx 0.22$) but remain moderately aligned at lemma positions ($\approx 0.68$), consistent with new, localized discriminative axes emerging post-adaptation.
- **Asymmetric Cross-Readout Transfer:** Transfer is successful in the pre-to-post direction (71-75% accuracy) but fails in the post-to-pre direction (53-54% accuracy), confirming that fine-tuning adds new, indispensable information.
- **Causal Ablation:** The model's performance is causally dependent on the new, post-fine-tuning directions. Ablating along these axes collapses accuracy to chance levels, while ablating along pre-existing directions has a negligible effect.

**Scale-Dependent Shift: From MLP Dominance to Attention Co-dominance.** While the overall mechanism is conserved, its implementation shifts with scale. In the 1B model, MLPs were the dominant causal drivers of the new behavior. In the 3B model, however, attention modules in the critical layers achieve comparable or even superior rescue rates (e.g., L12 attention at 94.8%, surpassing MLPs at 80-86%).

This finding suggests that as model capacity increases, the computational workload for sense disambiguation becomes more distributed between attention and MLP sublayers. Larger models may leverage their increased capacity to implement more sophisticated attention-based routing mechanisms, whereas smaller models rely more heavily on MLPs for feature transformation.

## 4.5. Can Mechanistic Insights Improve LoRA? A Proof-of-Concept

Having localized the causal mechanisms of domain specialization, we now test whether these insights can yield practical benefits. We propose a simple, mechanism-informed LoRA configuration, MI-LoRA, as a proof-of-concept to demonstrate that placing adaptive capacity precisely where it is causally required can improve performance. This is intended as an illustration of our findings, not a novel algorithmic contribution; the training budget

**Algorithm 1** MI-LoRA: Mechanism-Informed LoRA Adaptation

**Require:** pretrained model $M$; scale $s \in \{1\text{B}, 3\text{B}\}$; causal scores $\mathcal{I}$; geometry $\mathcal{G}$; critical rank $r_{\text{crit}}$

**Phase 1: Critical-Layer Selection**
**if** $s = 1\text{B}$ **then**
  $\mathcal{L}_{\text{crit}} \leftarrow \{9, 10\}$ ▷ Causal hotspots in 1B model
**else**
  $\mathcal{L}_{\text{crit}} \leftarrow \{11, 12, 13, 14\}$ ▷ Broader causal region in 3B model
**end if**
**Phase 2: Scale-Aware Module Configuration**
**for all** $\ell \in \{1, \ldots, L\}$ **do**
  **if** $\ell \in \mathcal{L}_{\text{crit}}$ **then**
    **if** $s = 1\text{B}$ **then**
      Modules$_\ell \leftarrow \{\texttt{q_proj}, \texttt{k_proj}\}$ ▷ Target attention for routing
      **if** CollapseRisk($\mathcal{G}$) **then**
        EnableRankDropout($B$-matrix, $0.10$)
      **end if**
    **else**
      Modules$_\ell \leftarrow \{\texttt{gate}, \texttt{up}, \texttt{down}\}$ ▷ Target MLPs for 3B
    **end if**
    $r_\ell \leftarrow r_{\text{crit}}$
  **end if**
**end for**
**Phase 3: Position-Aware Calibration**
$\Delta \leftarrow$ EstimateVerbalizerOffsets($\mathcal{D}_{\text{dev}}$, answer token)
ApplyCalibration($\Delta$) at the answer token during evaluation
**return** updated model $M^\star$

and parameter count are kept comparable to, or smaller than, a standard LoRA baseline.

**Mechanism-guided recipe.** Our MI-LoRA configuration is built on three principles derived directly from our analysis: (i) concentrate capacity in the *middle third* where causal rescue and drift peak; (ii) promote *rank diversity* when geometry indicates narrow post-only subspaces; (iii) apply *answer-token* calibration at inference to remove small domain priors on the "Yes/No" verbalizers. The procedure is summarized in Algorithm 1.

Even this simple, targeted adaptation strategy yields consistent performance gains across both model scales. On Llama-3.2-1B, it improves in-domain (BioWiC) accuracy from 76.80% to 78.35% and cross-domain (WiC) accuracy from 57.07% to 58.93%. On the 3B model, similar gains are observed for both in-domain (78.2% $\rightarrow$ 79.0%) and cross-domain (60.5% $\rightarrow$ 61.0%) tasks. While modest, these improvements validate the central thesis of our work: a precise mechanistic understanding can guide a more principled and effective approach to model adaptation.

## 5. Conclusion

This paper presented a unified, causally-validated framework for analyzing LoRA-based domain adaptation, successfully linking model behavior, computational localization, and geometric transformation. Our findings consistently demonstrate that for Word Sense Disambiguation, LoRA operates via emergence: it achieves strong in-domain performance by creating new, specialized mechanisms, which explains its limited cross-domain transfer. We causally localized these functional changes to the model's middle layers, with a pronounced specificity to computations occurring at the final answer token. Convergent geometric analyses confirmed that this adaptation creates new, nearly orthogonal discriminative directions rather than simply reshaping pre-existing ones.

To demonstrate the practical value of these insights, we designed a lightweight, mechanism-informed LoRA variant. By concentrating adaptive capacity in causally-critical layers, regularizing the emergent subspace, and calibrating the decision token, this approach delivered modest but consistent improvements to both in-domain and out-of-domain performance. This confirms that a precise mechanistic understanding can directly inform more principled and effective fine-tuning strategies. Future work should extend this analytical pipeline to a broader range of tasks and model scales, compare the mechanisms of alternative PEFT methods, and develop more advanced interventions to capture richer, multi-token circuit dynamics.

## 6. Limitations

Our analysis, while detailed, is subject to several limitations that define important avenues for future research.

**Scope of Generalization.** Our findings are currently grounded in the task of Word Sense Disambiguation (WSD), using the domain pairing of biomedical and general text, and evaluated on two models from the Llama-3.2 family (1B and 3B). While WSD provides an ideal, token-localized testbed for isolating causal mechanisms, further investigation is required to determine whether the "emergence" mechanism we identified is a universal principle of domain adaptation or specific to this task and architecture.

**Methodological Granularity.** Our study employs two key abstractions. First, we use linear probes to analyze representation geometry. While their high accuracy (72-77%) suggests that sense distinctions are largely linearly separable in this case, they cannot capture more complex, non-linear geometric transformations. Second, our causal analysis operates at the layer level, providing a coarse-grained localization. Finer-grained interventions, such as those at the head or neuron level, would be necessary to reverse-engineer the precise micro-circuits responsible for the emergent behavior.

**Adaptation Methods and Prompt Sensitivity.** Our analysis focuses exclusively on Low-Rank Adaptation (LoRA). A question remains whether the observed subspace extension is a specific artifact of LoRA's low-rank updates or a general property shared across other Parameter-Efficient Fine-Tuning (PEFT) methods, such as prompt tuning, prefix tuning, or standard adapters. Furthermore, our behavioral and geometric evaluations rely on a fixed set of instruction-style prompts. Because large language models often exhibit sensitivity to prompt formatting, future work should rigorously test the robustness of these emergent discriminative subspaces against prompt variations and adversarial phrasing to ensure the stability of the identified geometric structures.

## 7. Acknowledgments

This work was supported by grants from the National Science and Technology Major Project (No. 2023ZD0121103). We also thank the reviewers for their constructive feedback, which significantly improved the quality of this manuscript.

## 8. Bibliographical References

Anum Afzal, Mehul Kumawat, and Florian Matthes. 2025. Can smaller llms do better? unlocking cross-domain potential through parameter-efficient fine-tuning for text summarization. *arXiv preprint arXiv:2509.01314*.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

DM Anisuzzaman, Jeffrey G Malins, Paul A Friedman, and Zachi I Attia. 2025. Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health*, 3(1):100184.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. 2021. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. 2024. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Nitay Calderon, Naveh Porat, Eyal Ben-David, Alexander Chapanin, Zorik Gekhman, Nadav Oved, Vitaly Shalumov, and Roi Reichart. 2024. Measuring the robustness of NLP models to domain shifts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 126–154, Miami, Florida, USA. Association for Computational Linguistics.

Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip Yu, and Lichao Sun. 2025. A survey of ai-generated content (aigc). *ACM Computing Surveys*, 57(5):1–38.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. 2024. Med42–evaluating fine-tuning strategies for medical llms: full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*.

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. In *Proceedings*

*of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2089–2095, Minneapolis, Minnesota. Association for Computational Linguistics.

Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Lisa C Adams, and Keno K Bressem. 2025. Evaluating the effectiveness of biomedical fine-tuning for large language models on clinical tasks. *Journal of the American Medical Informatics Association*, 32(6):1015–1024.

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.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021a. Transformer feed-forward layers are key-value memories. In *Empirical Methods in Natural Language Processing (EMNLP)*, page 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021b. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021c. Transformer feed-forward layers are key-value memories. In *Empirical Methods in Natural Language Processing (EMNLP)*, page 5484–5495, Online and Punta Cana, Dominican Republic.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36:76033–76060.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333, Singapore. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, Online. Poster.

Yuan Huang. 2025. Fine-tuning of large language models for domain-specific cybersecurity knowledge. *arXiv preprint arXiv:2509.25241*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations (ICLR)*, Kigali Rwanda. Poster.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Yoon Pyo Lee. 2025. Mechanistic interpretability of lora-adapted language models for nuclear reactor safety applications. *arXiv preprint arXiv:2507.09931*.

Huiyi Leong, Yifan Gao, Shuai Ji, Yang Zhang, and Uktu Pamuksuz. 2024. Efficient fine-tuning of large language models for automated medical documentation. In *2024 4th International Conference on Digital Society and Intelligent Systems (DSInS)*, pages 204–209, Sydney, Australia. IEEE.

Chiyu Ma, Lin Shi, Ollie Liu, Wenhua Liang, Jiaqi Gan, Ming Cheng, Willie Neiswanger, and Soroush Vosoughi. 2024. Mechanistic insights: Circuit transformations across input and fine-tuning landscapes.

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.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*.

Neel Nanda and Joseph Bloom. 2022. Transformerlens: A framework for mechanistic interpretability of transformers.

Atharva Nijasure, Tanya Chowdhury, and James Allan. 2025. How relevance emerges: Interpreting lora fine-tuning in reranking llms. *arXiv preprint arXiv:2504.08780*.

Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun Chen. 2025. How do LLMs acquire new knowledge? a knowledge circuits perspective on continual pre-training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19889–19913, Vienna, Austria. Association for Computational Linguistics.

Karmvir Singh Phogat, Sai Akhil Puranam, Sridhar Dasaratha, Chetan Harsha, and Shashishekar Ramakrishna. 2024. Fine-tuning smaller language models for question answering over financial documents. *arXiv preprint arXiv:2408.12337*.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *Proceedings of the 2024 International Conference on Learning Representations*, pages 8057–8082, Vienna, Austria. Poster.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

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.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Hossein Rouhizadeh, Irina Nikishina, Anthony Yazdani, Alban Bornet, Boya Zhang, Julien Ehrsam, Christophe Gaudet-Blavignac, Nona Naderi, and Douglas Teodoro. 2024. A dataset for evaluating contextualized representation of biomedical concepts in language models. *Scientific Data*, 11(1):455.

Pasi Shailendra, Rudra Chandra Ghosh, Rajdeep Kumar, and Nitin Sharma. 2024. Survey of large language models for answering questions across various fields. In *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 520–527, Tamil Nadu, India. IEEE.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. 2024. Lora vs full fine-tuning: An illusion of equivalence. *arXiv preprint arXiv:2410.21228*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 7035–7052, Singapore.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

Shilpa Suresh, Nazgol Tavabi, Shahriar Golchin, Leah Gilreath, Rafael Garcia-Andujar, Alexander Kim, Joseph Murray, Blake Bacevich, and Ata Kiapour. 2023. Intermediate domain finetuning for weakly supervised domain-adaptive clinical NER. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 320–325, Toronto, Canada. Association for Computational Linguistics.

Yimin Tian, Bolin Zhang, Zhiying Tu, and Dianhui Chu. 2025. Adapters selector: Cross-domains

and multi-tasks LoRA modules integration usage method. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 593–605, Abu Dhabi, UAE. Association for Computational Linguistics.

Dannong Wang, Jaisal Patel, Daochen Zha, Steve Y Yang, and Xiao-Yang Liu. 2025a. Finlora: Benchmarking lora methods for fine-tuning llms on financial datasets. *arXiv preprint arXiv:2505.19819*.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. 2025b. Towards understanding fine-tuning mechanisms of llms via circuit analysis. In *ICML 2025*, Vancouver Convention Center. Association for Computational Linguistics. Poster.

Yucheng Wang, Ziyang Chen, and Md Faisal Kabir. 2025c. Explaining fine tuned llms via counterfactuals a knowledge graph driven framework. *arXiv preprint arXiv:2509.21241*.

Hongyuan Xu, Yuhang Niu, Ciyi Liu, Yanlong Wen, and Xiaojie Yuan. 2025. Taxopro: A plug-in lora-based cross-domain method for low-resource taxonomy completion. *Transactions of the Association for Computational Linguistics*, 13:557–576.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Zhuoyi Yang, Ming Ding, Yanhui Guo, Qingsong Lv, and Jie Tang. 2022. Parameter-efficient tuning makes a good classification head. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7576–7586, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qi Zhang, Yifei Wang, Jingyi Cui, Xiang Pan, Qi Lei, Stefanie Jegelka, and Yisen Wang. 2025. Beyond interpretability: The gains of feature monosemanticity on model robustness. In *ICLR 2025*, Singapore. Poster.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden. Association for Computational Linguistics.

Zhan Zhuang, Yulong Zhang, Xuehao Wang, Jiangang Lu, Ying Wei, and Yu Zhang. 2024. Time-varying lora: Towards effective cross-domain fine-tuning of diffusion models. *Advances in Neural Information Processing Systems*, 37:73920–73951.