

MoE Sparsity Begets Monosemanticity: Insights into Understanding and Fine-Tuning

Mingxue Tian^{1,*} Qi Zhang^{2,*} Xiaohan Wang^{3,*} Jiajun Chai³ Guojun Yin^{3,†}
Wei Lin³ Cailian Chen^{1,†} Qimin Xu¹ Siyu Wang¹ Yisen Wang²

¹Shanghai Jiao Tong University ²Peking University ³Meituan

Abstract

Mixture-of-Experts (MoE) architectures enhance the scalability and efficiency of large language models (LLMs) by activating only a subset of parameters. However, the interpretability of individual experts and corresponding strategies for post-training adaptation to domain-specific tasks remain underexplored. In this work, we develop an interpretability framework for expert-level monosemanticity in MoE models, drawing inspiration from sparse autoencoders and offering new insights into the specialization patterns of domain experts. We demonstrate that the inherent sparsity of MoE models encourages stronger monosemantic behavior at the expert level, which allows for the identification of experts responsible for particular downstream tasks and aids in preserving cross-domain performance. Building on this, we propose an expert-frozen fine-tuning method that selectively updates domain-specific experts while freezing the rest, which also alleviates catastrophic forgetting. Experiments on specialized domains, including medicine and finance, show that our approach performs on par with or better than fully fine-tuned models on in-domain tasks, while achieving relative improvements of 19.58% and 47.41% in retaining performance on out-of-domain benchmarks. We also demonstrate consistent advantages over PEFT baselines such as LoRA, achieving relative improvements of 11.49% and 16.74% in other domains.

1 Introduction

Mixture-of-Experts (MoE) architectures have recently emerged as a promising paradigm for scaling foundation models such as Kimi-K2, DeepSeek-V3 and Qwen3 (Kimi Team et al., 2025; DeepSeek-AI et al., 2024; Yang et al., 2025). By activating only a subset of experts in each layer,

MoE substantially reduces computational overhead while preserving the overall capacity of the model. Beyond efficiency, MoE offers a natural mechanism for expert specialization across different domains or tasks, which holds promise for improving the interpretability of these models (Jafar et al., 2025). Nevertheless, systematic exploration into how to better understand and leverage this interpretability remains limited.

A key property in analyzing model interpretability is monosemanticity, where each dimension is activated primarily by a single natural concept (Elhage et al., 2022; Cunningham et al., 2023), such as sentiment polarity, semantic categories, or word senses. While sparse autoencoders (SAEs) are the conventional approach to enhancing monosemanticity, they are post-hoc methods that introduce additional computational overhead. Prior work has shown that sparsity in the SAE feature space is closely linked to improved monosemanticity (Gao et al., 2024; Cunningham et al., 2023). This raises the question of whether MoE’s inherent sparsity promotes monosemantic representations. In this paper, we propose a new analysis framework to evaluate the monosemanticity of MoE models in specific downstream domains. Our findings reveal that increased sparsity significantly strengthens expert-level monosemanticity and drives experts to specialize across different downstream domains.

We further examine how monosemanticity can inspire better post-training strategies for large language models. A well-known challenge in post-training, particularly supervised fine-tuning, is that optimizing for a specific downstream task often comes at the expense of performance in other domains (Dong et al., 2023; Kumar et al., 2022). This raises the question of how to preserve cross-domain generalization while still achieving strong task-specific improvements. Monosemanticity provides a natural lens to address this issue:

*These authors contributed equally.

†Corresponding authors.

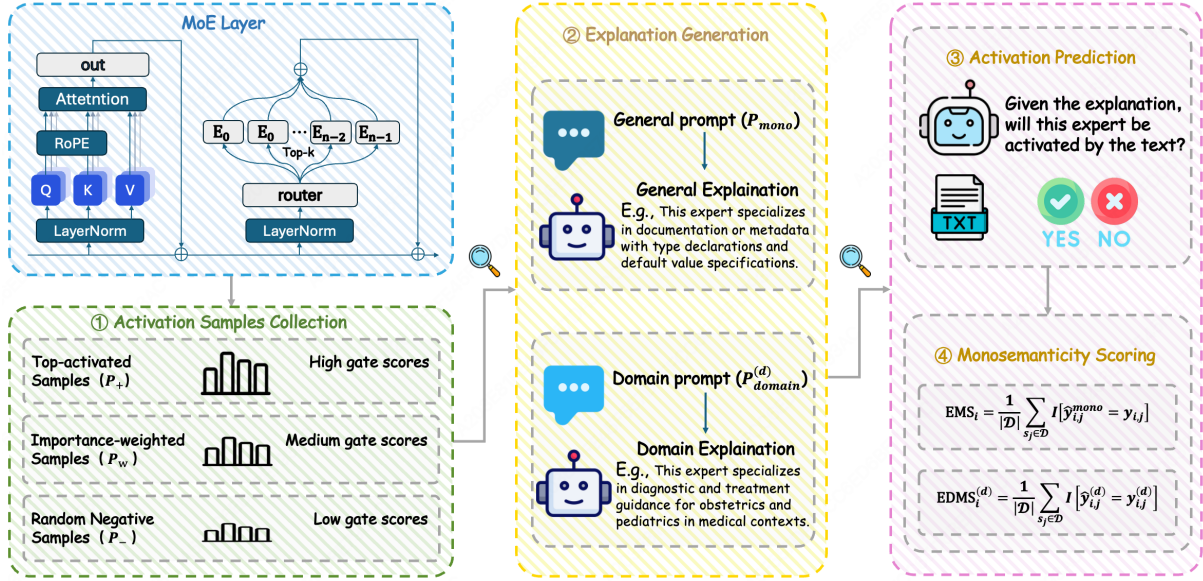


Figure 1: **MoE Monosemanticity Analysis Framework.** Overview of the proposed monosemanticity analysis framework for MoE LLMs, including: activation samples collection, explanation generation, activation prediction, and monosemanticity scoring. This framework advances the understanding of MoE LLMs and lays the groundwork for downstream applications.

with MoE architectures, the increased sparsity encourages experts to specialize in distinct domains, suggesting that only a subset of experts is truly responsible for a given downstream task.

As shown in Figure 1 and Figure 2, we obtain natural language explanations for expert functions and propose a fine-tuning strategy in which only the experts relevant to the target task and decoupled from other domains are fine-tuned, while the others remain frozen. Experimental results demonstrate that EDMRS achieves performance comparable to fully fine-tuned models on target tasks, while better preserving performance in other domains. Our contributions can be summarized as follows:

- We introduce a novel framework to analyze monosemanticity in MoE LLMs, filling the gap in existing research on expert-level monosemanticity in MoE models and providing the infrastructure for further academic research in this area.
- We provide a systematic empirical characterization of expert-level monosemanticity, establishing that sparsity strengthens monosemanticity, middle layers exhibit the highest degree of specialization, and experts differentiate across downstream domains.
- We leverage our analysis framework

and the derived insights to propose a monosemanticity-guided expert selection strategy, achieving competitive target domain performance while alleviating catastrophic forgetting.

2 Related Work & Preliminary

2.1 MoE LLMs

Sparse MoE LLMs have achieved significant scalability by activating only a subset of parameters in each layer through a routing network (Kimi Team et al., 2025; DeepSeek-AI et al., 2024). This modular design enables efficient scaling and introduces additional flexibility for model adaptation. Formally, in the l -th layer of the MoE, the output hidden state of the t -th token, h_t^l , is computed as:

$$h_t^l = \sum_{i=1}^N g_{i,t} \cdot \text{FFN}_i(u_t^l) + u_t^l. \quad (1)$$

Here, N denotes the total number of experts, $\text{FFN}_i(\cdot)$ represents the i -th expert network, and $g_{i,t}$ is the gating value for the i -th expert when processing the t -th token. The gating mechanism is defined as:

$$g_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \text{TopK}(\{s_{j,t}\}_{j=1}^N, K) \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

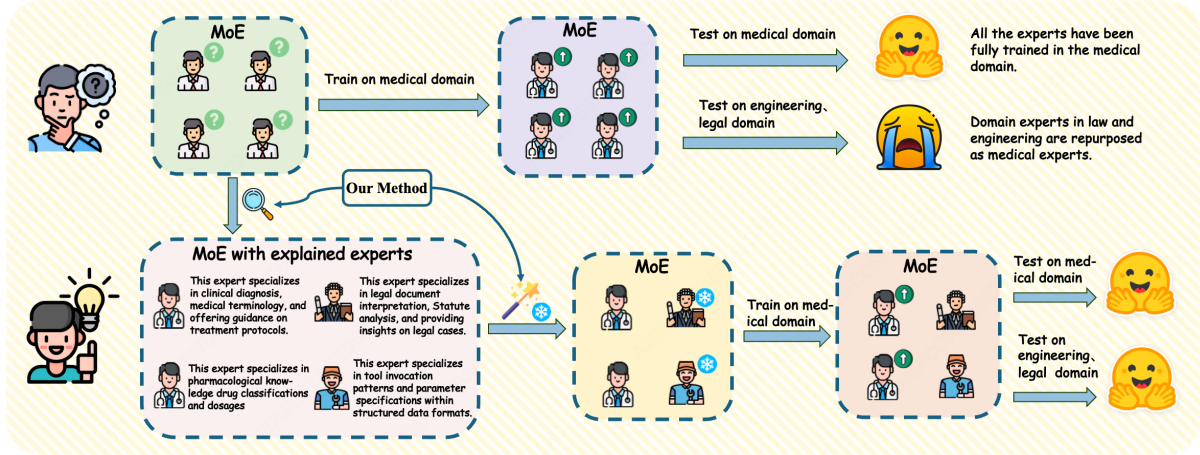


Figure 2: **Monosemanticity-Guided Fine-Tuning Strategy.** (Top) Training without our method: Expert functions remain uncharacterized, causing all experts to be fine-tuned on the target domain. This disrupts their specializations, leading to degraded performance in other domains. (Bottom) Training with our method: We first generate explanations to reveal expert specializations, then leverage fine-grained metrics to selectively fine-tune relevant experts while freezing others. This preserves expert specialization and maintains performance across domains.

where $s_{i,t}$ denotes the token-expert gating score, and $\text{TopK}(\cdot, K)$ selects the K highest scores.

2.2 Finetuning of MoE LLMs

Fine-tuning strategies for MoE LLMs are more complex than those for dense LLMs (Cai et al., 2025). Prior research has explored expert-specific tuning (Bai et al., 2025; Wang et al., 2024b), dynamic data mixing (Zhu et al., 2024), and router adaptation (Liu et al., 2025). A key challenge is identifying relevant experts for the target domain. Existing methods rely on black-box metrics—such as expert activation frequency or router output probabilities (Cai et al., 2025)—but these lack interpretability and fail to distinguish domain-specific from shared experts. As a result, the interpretability of individual experts in MoE LLMs remains underexplored. In contrast, we leverage interpretability principles from sparse autoencoders (Cunningham et al., 2023) to design interpretable expert-level fine-tuning strategies for MoE LLMs.

2.3 Monosemanticity in Large Language Models

LLMs perform well across domains (DeepSeek-AI et al., 2024; Yang et al., 2025), but their decision-making is opaque. Recent studies have introduced SAEs (Gao et al., 2024; Cunningham et al., 2023), inducing monosemantic dimensions that capture specific concepts (Shu et al., 2025). SAEs have been effective in tasks like data selec-

tion (Ma et al., 2025), disentanglement (Pach et al., 2025), and privacy analysis (Frikha et al., 2025). They have also enabled interpretable reward models for transparent feedback (Zhang et al., 2025; Li et al., 2025). While these studies focus on dense LLMs, the expert-level structure of MoE models remains underexplored. This paper explores expert-level monosemanticity in MoE LLMs and proposes interpretable fine-tuning strategies for expert selection.

3 Expert-level Monosemanticity in MoE LLMs

SAEs have recently emerged as powerful tools for improving the interpretability of LLMs (Cunningham et al., 2023). SAEs encourage monosemanticity by enforcing sparsity, such that each hidden dimension tends to represent a distinct natural concept (Gao et al., 2024). Notably, MoE LLMs inherently exhibit sparsity through their expert selection mechanism, activating only a subset of experts for each input. This parallel motivates our central research question: Does the inherent sparsity in MoEs lead to stronger monosemanticity? To address this, Section 3.1 introduces a systematic framework for evaluating monosemanticity in MoE LLMs, while Section 3.2 provides empirical insights into how monosemanticity emerges within MoE architectures.

3.1 Analysis Framework for Monosemanticity in MoE LLMs

To investigate monosemanticity in MoE LLMs, we propose an analysis framework grounded in analogies between SAE and MoE architectures. Specifically, we draw analogies between sparse gating and sparse intermediate representations, top-k and thresholded activation, and MoE experts and SAE neurons, extending monosemanticity from the neuron to the expert level. Inspired by automated SAE monosemanticity evaluation methods (Paulo et al., 2025), we present a systematic, expert-level analysis framework for MoE LLMs.

The proposed framework consists of two stages. First, we employ a three-stage sampling strategy to construct expert activation datasets, including top-activated samples, importance-weighted samples, and randomly selected negatives. Natural language explanations for each expert are then generated using two complementary prompts: P_{mono} , providing a general characterization, and $P_{domain}^{(d)}$, highlighting domain-specific semantics. In the second stage, an external LLM predicts expert activation based on these explanations, formulated as a binary classification task whose accuracy reflects the degree of expert-level monosemanticity. Complete prompt templates and design details are provided in Appendix A.1, with detailed expert explanations in Appendix A.3.

This two-stage pipeline of explanation generation followed by activation prediction is well-established in the SAE interpretability literature, has become a widely accepted consensus in the research community, and has been shown to reliably capture monosemanticity (Paulo et al., 2025; Gur-Arie et al., 2025). To assess the robustness of our metrics to evaluator choice, we conduct controlled experiments replacing GPT-4o with a substantially more economical alternative, finding that the resulting scores exhibit highly consistent relative trends across all layers (Appendix A.9), which demonstrates that the proposed framework is not contingent on frontier model capacity.

To quantitatively assess the monosemanticity of MoE, we propose two key metrics: the expert monosemanticity score (EMS) and the expert domain monosemanticity score (EDMS). EMS measures the degree to which an expert exhibits monosemanticity across a general domain, while EDMS evaluates monosemanticity within a specific downstream domain. Formally, the EMS for

expert E_i is defined as:

$$EMS_i = \frac{1}{|\mathcal{D}|} \sum_{s_j \in \mathcal{D}} \mathbb{I}[\hat{y}_{i,j}^{mono} = y_{i,j}], \quad (3)$$

where $\mathcal{D}$ denotes the test dataset and j indexes its samples. Here, $\hat{y}_{i,j}^{mono}$ represents the binary prediction of whether expert E_i would be activated for sample s_j , based on the explanation derived from prompt P_{mono} , and $y_{i,j}$ is the corresponding ground-truth activation label. Higher EMS values indicate stronger monosemanticity.

For evaluating monosemanticity in a specific domain d , the EDMS is defined as:

$$EDMS_i^{(d)} = \frac{1}{|\mathcal{D}|} \sum_{s_j \in \mathcal{D}} \mathbb{I}[\hat{y}_{i,j}^{(d)} = y_{i,j}^{(d)}]. \quad (4)$$

Here, $\hat{y}_{i,j}^{(d)}$ denotes the prediction of expert E_i for domain d , and $y_{i,j}^{(d)}$ is the corresponding ground-truth label. A high EDMS indicates that the expert is monosemantic with respect to specific concepts in the target domain.

3.2 Empirical Insights of Expert-level Monosemanticity

To advance the understanding of expert monosemanticity and sparsity in MoE models, and to lay the groundwork for further applications, we systematically analyze and uncover three key empirical insights in this section using the proposed analysis framework. We additionally analyze expert-versus-neuron semantic granularity in Appendix A.10.

Insight 1: *The sparse architecture of MoE LLMs not only enhances computational efficiency but also induces a higher degree of expert monosemanticity, distinguishing them from dense LLMs.*

Clarification for Insight 1: Theoretical and empirical findings from SAE research establish that higher sparsity promotes stronger monosemanticity (Gao et al., 2024). In MoE LLMs, this sparsity is intrinsically determined by the number of activated experts, where a greater number of activations corresponds to lower sparsity. Formally, the sparsity of a MoE layer is defined as $\text{Sparsity} = 1 - \frac{k}{N}$, where k is the number of activated experts per token and N is the total number of experts in that layer. We conduct controlled experiments on Qwen3-30B-A3B-Instruct-2507, selecting four representative layers

and varying the sparsity level. Monosemanticity is quantified using the Layer-wise Decoupling Score (LDS), which we define as the average EMS of all experts within a layer:

$$\text{LDS}_l = \frac{1}{N_l} \sum_{i=1}^{N_l} \text{EMS}_i^{(l)}, \quad (5)$$

where LDS_l denotes the LDS for the l -th layer and N_l is the number of experts in that layer. LDS captures the average monosemanticity of experts at a given layer.

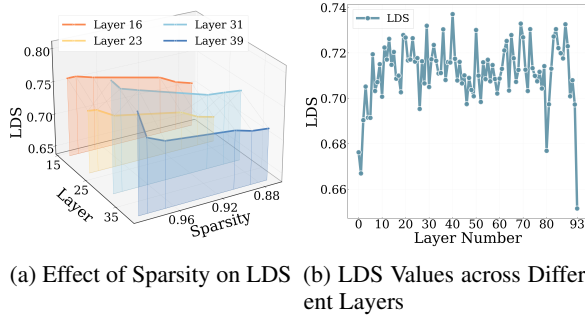


Figure 3: Analysis of sparsity and monosemanticity in MoE. Layer-wise decoupling score (LDS) describes the average monosemanticity of a layer. Panel (a) examines the sparsity-monosemanticity relationship, showing that higher sparsity enhances expert specialization. Panel (b) tracks monosemanticity variation across network depth, revealing that middle layers exhibit optimal monosemanticity.

As shown in Figure 3a, reducing sparsity consistently and substantially decreases LDS across all layers, indicating progressively diminished expert-level monosemanticity. Notably, this trend holds uniformly across all evaluated values of k with no inflection near the pretraining default, ruling out pretraining confounds; as k approaches N , the model degenerates toward a dense LLM, consistent with the lower monosemanticity characteristically observed in dense LLMs. Building directly on this key observation, we proceed to analyze how monosemanticity varies across different layers of the MoE architecture.

Insight 2: *Middle layers in MoE LLMs exhibit the highest degree of monosemanticity, while input and output layers show nearly the lowest. This suggests that experts in the middle layers are most decoupled and specialized, making them especially well-suited for operations that leverage monosemantic representations.*

Clarification for Insight 2: Figure 3b presents the LDS scores across network depth for Qwen3-235B-A22B-Instruct-2507. We observe a rapid increase in LDS in the early layers, a consistently high plateau across the middle layers, and a sharp decline near the output. This trajectory suggests a three-stage organization: early layers focus on disentangling basic features, middle layers achieve the strongest expert decoupling and monosemanticity, and output layers gradually lose this separation. These findings highlight the central role of middle-layer experts in maintaining semantic specialization, i.e., monosemanticity, within the model.

Insight 3: *The degree of monosemanticity among MoE experts varies substantially across domains. Some experts exhibit consistently strong monosemanticity across all settings, while others are highly specialized within particular domains.*

Clarification for Insight 3: To investigate domain-dependent variations in expert monosemanticity, we apply our monosemanticity analysis framework using both the P_{mono} and $P_{domain}^{(d)}$ prompts to the middle layer of Qwen3-30B-A3B-Instruct-2507, computing EMS and EDMS for all 128 experts. As shown in Figure 4a and Figure 4b, EDMS scores are generally lower than EMS scores, suggesting that while some experts maintain high monosemanticity, their activity often pertains to domains outside of the evaluated domain. We also identify polysemantic experts that are activated across multiple domains. Overall, these findings reveal the coexistence of domain-specialized and shared experts, indicating that experts exhibit distinct patterns of activation across different domains.

4 Monosemanticity-Guided Fine-Tuning Strategy for MoE LLMs

Fine-tuning MoE LLMs on a target domain is known to degrade performance in other domains (Dong et al., 2023). The existence of monosemantic experts, as established in Section 3, offers an intuition: selectively fine-tuning only the experts that are highly relevant to the target domain and functionally decoupled from others. The central challenge, then, is how to identify such experts with precision.

In prior work (Bai et al., 2025; Wang et al., 2024b), expert selection for MoE fine-tuning has

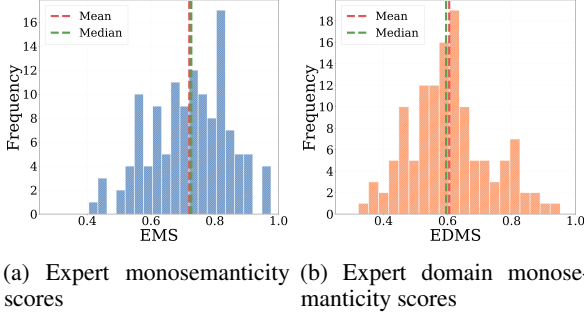


Figure 4: Distribution of expert monosemanticity. Panel (a) evaluates the distribution of general expert monosemanticity scores (EMS) across all domains. Panel (b) evaluates the distribution of domain-specific expert domain monosemanticity scores (EDMS).

relied predominantly on activation frequency (AF) as a proxy for domain relevance:

$$\text{AF}_i^{(d)} = \frac{1}{|T_d|} \sum_{t \in T_d} 1 \cdot \mathbb{I}[E'_i(t)] \quad (6)$$

where E'_i indicates whether expert E_i is activated for token t , T_d denotes the token set of the target domain, and $\mathbb{I}[\cdot]$ is the indicator function.

However, AF measures only how often an expert is activated and does not characterize the domain specificity of the expert’s overall activation pattern. As established in Section 3, polysemantic experts are pervasive in MoE models, being indiscriminately activated across semantically unrelated domains. For such experts, a high AF may reflect broad cross-domain usage rather than target-domain specialization, obscuring the true domain relevance signal. AF is therefore agnostic to expert-level domain specificity and may assign similar importance to frequently activated domain-specialized and polysemantic experts. A more principled selection criterion should therefore build upon activation frequency while incorporating an expert-level measure of domain specialization.

Our monosemanticity analysis framework provides a complementary expert-level quality signal. EDMS quantifies the degree to which an expert’s activation pattern aligns with the semantic profile of the target domain. An expert with high EDMS exhibits an overall activation pattern that is highly relevant to the target domain and functionally decoupled from other domains, thereby capturing information that activation frequency alone does not represent. EDMS thus serves as an expert-level calibration weight in our formulation.

Motivated by this analysis, we propose the Expert Domain Monosemantic Responsibility Score (EDMRS), which unifies activation frequency with monosemanticity-based quality calibration into a single coherent criterion:

$$\text{EDMRS}_i^{(d)} = \frac{1}{|T_d|} \sum_{t \in T_d} \text{EDMS}_i^{(d)} \cdot \mathbb{I}[E'_i(t)] \quad (7)$$

where $\text{EDMS}_i^{(d)}$ serves as an expert-level, monosemanticity-based calibration coefficient that reweights raw activation frequency by domain specificity. Because $\text{EDMS}_i^{(d)}$ is constant with respect to token t , this formulation is equivalently $\text{EDMRS}_i^{(d)} = \text{EDMS}_i^{(d)} \cdot \text{AF}_i^{(d)}$, and therefore performs an expert-level calibration of aggregate activation frequency according to domain specificity. A high EDMRS score thus indicates an expert with both high target-domain activation frequency and high domain specificity, making it a suitable candidate for targeted domain adaptation with reduced cross-domain interference.

A heuristic theoretical analysis of the EDMRS strategy is provided in Appendix A.11. Furthermore, owing to the modular structure of MoE LLMs, our approach is in principle orthogonally composable with parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022).

5 Experiments

This section further validates the reliability of our expert monosemanticity analysis framework, providing evidence that reduces concerns about evaluator-specific effects when the same external LLM is employed for both expert explanation generation and evaluation. It then validates the effectiveness of our proposed EDMRS strategy through two sets of experiments, thereby validating the effectiveness of our expert monosemanticity analysis framework. The first set of experiments compares against the most commonly used full fine-tuning method (Full Strategy) and the expert selection method based on activation frequency (AF Strategy). The second set of experiments compares against commonly used parameter-efficient fine-tuning (PEFT) methods.

We conduct experiments in two domains: **Medicine** and **Finance**. For Medicine, we use MedMCQA, a large-scale medical reasoning dataset. For Finance, we use FinQA, a financial dataset requiring program-based numerical reason-

Method	biology	business	chemistry	computer science	economics	engineering	history	law	math	other	philosophy	physics	psychology	Average
Full	48.91%	40.30%	45.67%	27.07%	53.52%	37.74%	40.68%	29.58%	34.91%	37.12%	39.95%	44.26%	66.67%	42.03%
AF	60.62%	43.09%	43.82%	49.59%	56.16%	37.77%	40.51%	26.85%	48.88%	46.32%	38.68%	47.65%	52.42%	45.57%
EDMRS	62.44%	47.53%	44.99%	54.88%	61.93%	42.31%	51.36%	31.61%	51.84%	50.32%	44.36%	47.93%	61.90%	50.26%
vs Full (pp)	+13.53	+7.23	-0.68	+27.81	+8.41	+4.57	+10.68	+2.03	+16.93	+13.20	+4.41	+3.67	-4.77	+8.23
vs AF (pp)	+1.82	+4.44	+1.17	+5.29	+5.77	+4.54	+10.85	+4.76	+2.96	+4.00	+5.68	+0.28	+9.48	+4.69

(a) Performance comparison of models trained with EDMRS, AF and Full strategies on Medicine in other domains.

Method	biology	law	chemistry	computer science	economics	engineering	health	history	math	other	philosophy	physics	psychology	Average
Full	73.36%	20.59%	41.43%	58.21%	56.44%	31.89%	55.58%	37.71%	59.83%	41.85%	46.83%	51.19%	62.07%	49.00%
AF	86.33%	42.26%	77.27%	78.94%	79.42%	55.56%	71.23%	58.53%	86.48%	65.19%	66.87%	79.24%	76.90%	71.09%
EDMRS	87.12%	43.90%	78.68%	79.76%	79.27%	57.24%	70.99%	62.73%	87.61%	65.55%	67.33%	80.55%	78.32%	72.23%
vs Full (pp)	+13.76	+23.31	+37.25	+21.55	+22.83	+25.35	+15.41	+25.02	+27.78	+23.70	+20.50	+29.36	+16.25	+23.23
vs AF (pp)	+0.79	+1.64	+1.41	+0.82	-0.15	+1.68	-0.24	+4.20	+1.13	+0.36	+0.46	+1.31	+1.42	+1.14

(b) Performance comparison of models trained with EDMRS, AF and Full strategies on Finance in other domains.

Method	biology	business	chemistry	computer science	economics	engineering	history	law	math	other	philosophy	physics	psychology	Average
LoRA	60.62%	40.98%	38.84%	44.72%	59.24%	36.33%	46.02%	25.52%	41.43%	46.61%	43.49%	43.26%	58.98%	45.08%
routerPEFT	70.25%	52.30%	47.44%	51.71%	59.24%	38.77%	41.56%	25.55%	55.39%	43.51%	38.34%	52.32%	58.69%	48.85%
EDMRS	62.44%	47.53%	44.99%	54.88%	61.93%	42.31%	51.36%	31.61%	51.84%	50.32%	44.36%	47.93%	61.90%	50.26%
vs LoRA (pp)	+1.82	+6.55	+6.15	+10.16	+2.69	+5.98	+5.34	+6.09	+10.41	+3.71	+0.87	+4.67	+2.92	+5.18
vs routerPEFT (pp)	-7.81	-4.77	-2.45	+3.17	+2.69	+3.54	+9.80	+6.06	-3.55	+6.81	+6.02	-4.39	+3.21	+1.41

(c) Performance comparison of PEFT-trained models on Medicine across other domains.

Method	biology	law	chemistry	computer science	economics	engineering	health	history	math	other	philosophy	physics	psychology	Average
LoRA	83.08%	31.43%	60.13%	69.27%	69.15%	44.38%	66.71%	48.99%	77.50%	54.58%	60.19%	67.39%	71.51%	61.87%
routerPEFT	78.75%	31.67%	32.33%	46.91%	59.91%	24.84%	60.27%	50.13%	54.68%	55.63%	50.50%	47.99%	73.81%	51.34%
EDMRS	87.12%	43.90%	78.68%	79.76%	79.27%	57.24%	70.99%	62.73%	87.61%	65.55%	67.33%	80.55%	78.32%	72.23%
vs LoRA (pp)	+4.04	+12.47	+18.55	+10.49	+10.12	+12.86	+4.28	+13.74	+10.11	+10.97	+7.14	+13.16	+6.81	+10.36
vs routerPEFT (pp)	+8.37	+12.23	+46.35	+32.85	+19.36	+32.40	+10.72	+12.60	+32.93	+9.92	+16.83	+32.56	+4.51	+20.89

(d) Performance comparison of PEFT-trained models on Finance across other domains.

Table 1: Comprehensive performance comparison across different domains and training strategies: (a) Trained on Medicine data, comparing AF and Full strategies; (b) Trained on Finance data, comparing AF and Full strategies; (c) Trained on Medicine data, comparing PEFT strategy; (d) Trained on Finance data, comparing PEFT strategy.

ing. For evaluation, we additionally use 13 out-of-domain reasoning categories from MMLU-Pro, including physics, mathematics, and philosophy (Pal et al., 2022; Chen et al., 2021; Wang et al., 2024a). To further evaluate cross-task generalization, we test the models fine-tuned on Medicine on ten additional out-of-domain benchmarks covering language understanding, knowledge-intensive question answering, lexical and commonsense reasoning, and multi-step reasoning. To further analyze the effects of different expert-selection criteria, we compare EDMRS with EDMS-only, first-order Taylor-, and Fisher-based expert-selection methods. The EDMS-only comparison also evaluates the contribution of incorporating activation frequency into EDMRS. Further experimental details are provided in Appendix A.7.

EDMRS-selected experts	AF-selected experts
Expert 111: Medical terminology and pathology-specific vocabulary including disease mechanisms and clinical references	Expert 116: This expert specializes in processing exact word patterns, numerical sequences, and precise linguistic structures within technical or instructional contexts
Expert 53: This expert specializes in content related to human anatomy and physical movements emphasizing injury mechanisms and clinical examination techniques	Expert 72: The expert specializes in technical and scientific content involving mathematical calculations and physical properties

Table 2: Examples of differences between experts selected by the EDMRS strategy and the AF strategy for the medical domain, where EDMRS favors monosemantic experts while AF selects general experts.

5.1 Reliability Validation of Expert Monosemanticity Analysis Framework

To validate the reliability of our expert monosemanticity analysis framework, we conduct two controlled experiments on Qwen3-235B-A22B.

First, we replace the LLM used in the evaluation stage with GPT-4o-mini and Qwen3.5-35B-A3B, and recompute LDS profiles across all layers. The resulting Spearman correlations with the original scores are 0.815 and 0.774, respectively, indicating that our scores are not an artifact of evaluator self-reinforcement. Second, we sample 32 experts per layer for a total of 3,008 experts, and ask human annotators to judge expert activation under identical criteria, yielding a Spearman correlation of 0.701, which still indicates strong agreement. Taken together, these results demonstrate that our monosemanticity scores are robust to evaluator choice, validating the reliability of our framework.

5.2 Comparison with Full and AF Strategies

We validate the effectiveness of our proposed EDMRS strategy by performing expert-selection fine-tuning on the medical and financial domains using three strategies: Full, AF and EDMRS.

Method	Medicine		Finance	
	Domain Acc	MMLU-Pro Avg	Domain Acc	MMLU-Pro Avg
Full	58.79% ±0.51%	42.03% ±0.02%	42.81% ±0.00%	49.00% ±0.18%
AF	48.39% ±0.61%	45.57% ±0.11%	49.35% ±0.00%	71.09% ±0.16%
EDMRS	60.09% ±0.07%	50.26% ±0.15%	50.04% ±0.00%	72.23% ±0.11%
vs Full (pp)	+1.30	+8.23	+7.23	+23.23
vs AF (pp)	+11.70	+4.69	+0.69	+1.14

(a) Comparison of Full, AF, and EDMRS strategies.

Method	Medicine		Finance	
	Domain Acc	MMLU-Pro Avg	Domain Acc	MMLU-Pro Avg
LoRA	48.80% ±0.43%	45.08% ±0.06%	24.67% ±0.00%	61.87% ±0.17%
routerPEFT	50.16% ±0.26%	48.85% ±0.08%	35.48% ±0.00%	51.34% ±0.12%
EDMRS	60.09% ±0.07%	50.26% ±0.15%	50.04% ±0.00%	72.23% ±0.11%
vs LoRA (pp)	+11.29	+5.18	+25.37	+10.36
vs routerPEFT (pp)	+9.93	+1.41	+14.56	+20.89

(b) Comparison of LoRA, routerPEFT, and EDMRS strategies.

Table 3: Summary of EDMRS performance against baselines on domain accuracy and MMLU-Pro average. (a) compares Full, AF, and EDMRS; (b) compares LoRA, routerPEFT, and EDMRS.

See Table 3(a) for the results. In the medical domain, EDMRS (60.09%) outperforms both AF (48.39%) and Full (58.79%). More importantly, EDMRS demonstrates better performance in other domains (50.26%), surpassing the Full (42.03%) and AF (45.57%) strategies by 8.23 percentage points and 4.69 percentage points, respec-

tively. Table 1(a) shows that EDMRS achieves the best performance among all strategies. Notably, EDMRS outperforms Full by 16.93 percentage points on mathematics, indicating that EDMRS effectively preserves the model’s reasoning capabilities. Table 3(a) and Table 1(b) show that consistent advantages are also observed in the financial domain, surpassing the Full and AF strategies by 23.23 percentage points and 1.14 percentage points, respectively.

Table 2 further reveals that EDMRS selects experts specialized for the target domain, unlike AF, which tends to select general experts. This advantage is consistent with EDMRS favoring domain-specialized experts while reducing updates to more general experts, which can help mitigate capability loss in other domains.

5.3 Comparison of Expert-Selection Criteria

To examine whether the advantage of EDMRS arises from its expert-selection criterion, we compare it with an EDMS-only variant and two loss-based alternatives. Taylor selection ranks experts using the first-order importance $|\theta \odot \partial \mathcal{L} / \partial \theta|$, while Fisher selection uses the expected squared gradient. The scores are aggregated over the parameters of each expert. All methods select the same number of experts per layer and use the same training configuration.

As shown in Table 4, EDMRS improves the Medicine and MMLU-Pro averages over EDMS-only by 8.49 and 5.26 percentage points, respectively, showing the benefit of incorporating activation frequency into semantic specialization. Compared with the loss-based alternatives, EDMRS achieves comparable target-domain performance while improving the MMLU-Pro average by 3.51 and 4.44 percentage points over Taylor and Fisher selection, respectively. These results indicate that loss-based importance identifies experts that influence the target-domain objective, but does not explicitly distinguish domain-specialized experts from general-purpose experts. By jointly considering semantic specialization and activation frequency, EDMRS more effectively preserves out-of-domain capabilities.

5.4 Comparison with PEFT as a Standalone Strategy

To compare against PEFT methods, we evaluate our approach against the most commonly used PEFT method LoRA, as well as router-

Method	Medicine	biology	business	chemistry	computer science	economics	engineering	history	law	math	other	philosophy	physics	psychology	Average
EDMS-only	51.60% ±0.43%	60.62% ±0.56%	34.98% ±0.66%	41.70% ±0.31%	47.89% ±0.56%	53.75% ±0.68%	33.06% ±0.42%	46.46% ±0.00%	28.97% ±0.00%	56.13% ±0.15%	45.31% ±0.23%	38.28% ±0.20%	42.85% ±0.04%	55.05% ±1.05%	45.00% ±0.15%
Taylor	59.46% ±0.38%	61.79% ±0.74%	40.94% ±0.99%	45.73% ±0.05%	32.36% ±0.37%	59.40% ±0.76%	35.84% ±0.39%	42.08% ±1.24%	30.76% ±0.37%	53.66% ±0.13%	48.56% ±0.23%	48.03% ±1.85%	46.29% ±0.12%	62.32% ±0.40%	46.75% ±0.22%
Fisher	59.94% ±0.70%	58.72% ±1.57%	41.95% ±1.34%	44.58% ±0.49%	26.26% ±2.72%	57.07% ±0.81%	34.78% ±1.35%	44.01% ±1.60%	31.76% ±0.14%	56.62% ±0.32%	46.25% ±0.35%	43.95% ±0.64%	46.73% ±0.73%	63.03% ±0.13%	45.82% ±0.40%
EDMRS	60.09% ±0.07%	62.44% ±0.82%	47.53% ±0.55%	44.99% ±0.20%	54.88% ±0.42%	61.93% ±0.45%	42.31% ±1.00%	51.36% ±0.30%	31.61% ±0.16%	51.84% ±1.19%	50.32% ±0.68%	44.36% ±0.23%	47.93% ±0.04%	61.90% ±0.87%	50.26% ±0.15%
vs EDMS-only (pp)	+8.49	+1.82	+12.55	+3.29	+6.99	+8.18	+9.25	+4.90	+2.64	-4.29	+5.01	+6.08	+5.08	+6.85	+5.26
vs Taylor (pp)	+0.63	+0.65	+6.59	-0.74	+22.52	+2.53	+6.47	+9.28	+0.85	-1.82	+1.76	-3.67	+1.64	-0.42	+3.51
vs Fisher (pp)	+0.15	+3.72	+5.58	+0.41	+28.62	+4.86	+7.53	+7.35	-0.15	-4.78	+4.07	+0.41	+1.20	-1.13	+4.44

Table 4: Performance comparison of EDMS-only, Taylor, Fisher, and EDMRS expert-selection criteria on Medicine and MMLU-Pro across domains.

PEFT, which fine-tunes only the MoE router. As shown in Tables 3(b), 1(c), and 1(d), EDMRS exhibits consistent trends across both the medical and financial domains, as observed in Section 5.2, achieving better performance on the target domains while outperforming both LoRA and routerPEFT on almost all other domains and on average, further demonstrating the advantages of EDMRS in these comparisons. One possible explanation is that LoRA still updates all experts, whereas our strategy selectively freezes irrelevant experts and therefore reduces cross-domain interference. Similarly, our method may reduce the interference introduced by routerPEFT, which directly modifies the router shared across all experts.

5.5 Generalization Across Multi-Task Scenarios

To evaluate cross-task generalization after fine-tuning on Medicine, we compare Full, AF, and EDMRS on ten additional out-of-domain benchmarks. These benchmarks cover Chinese language understanding (C3 and OCNLI) (Sun et al., 2020; Hu et al., 2020), scientific and professional knowledge (HeadQA, OpenBookQA, and EXAMS) (Vilares and Gómez-Rodríguez, 2019; Mihaylov et al., 2018; Hardalov et al., 2020), lexical and commonsense reasoning (WiC, HellaSwag, SWAG, and CODAH) (Pilehvar and Camacho-Collados, 2019; Zellers et al., 2019, 2018; Chen et al., 2019), and multi-step reasoning (MuSR) (Sprague et al., 2024). Detailed results are provided in Appendix A.8. Overall, EDMRS demonstrates a clear advantage, indicating that it more effectively preserves cross-task out-of-domain capabilities within the scope of this evaluation and providing further evidence of its effectiveness in multi-task settings.

5.6 Overall Discussion

Considering both target-domain adaptation and out-of-domain capability retention, the results show that EDMRS achieves a more favorable balance between the two under our evaluation settings. By incorporating domain-relevant semantic information into expert selection, EDMRS reduces unnecessary updates to general-purpose or cross-domain experts, thereby mitigating cross-domain interference during fine-tuning. Consequently, EDMRS improves target-domain performance while more effectively preserving the model’s capabilities across other domains and tasks. These results suggest that semantic information can provide a valuable signal for parameter selection in MoE fine-tuning, motivating further investigation into how such information can be more fully exploited to guide expert selection and model adaptation.

6 Conclusion

We present a framework extending monosemanticity to MoE LLMs, introducing quantitative metrics to assess expert-level monosemanticity, and uncovering three key insights: MoE’s inherent sparsity promotes expert-level monosemanticity; middle layers exhibit the strongest specialization; and experts exhibit either domain-specific or general behavior. Motivated by these findings, we propose EDMRS, which combines activation frequency with monosemanticity to fine-tune only domain-relevant experts. Experiments show that our approach achieves competitive in-domain performance while substantially improving out-of-domain retention, effectively alleviating catastrophic forgetting.

7 Limitations

Our method introduces some computational overhead due to the process of generating expert-level explanations and calculating monosemanticity scores. In addition, our heuristic analysis adopts a positive proportional form shared across all experts to obtain the current EDMRS formulation. This simple form may not be optimal across different domains and training settings, and exploring more flexible functional forms may further improve expert selection. Additionally, we have not explored the orthogonal combination with other parameter-efficient fine-tuning methods, which could potentially enhance performance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62432009.

References

- Jun Bai, Minghao Tong, Yang Liu, Zixia Jia, and Zilong Zheng. 2025. Understanding and leveraging the expert specialization of context faithfulness in mixture-of-experts llms. *arXiv preprint arXiv:2508.19594*.
- Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*.
- Michael Chen, Mike D’Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. [CODAH: An adversarially-authored question answering dataset for common sense](#). In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.
- Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.
- Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.
- Ahmed Frikha, Muhammad Reza Ar Razi, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2025. Privacyscalpel: Enhancing llm privacy via interpretable feature intervention with sparse autoencoders. *arXiv preprint arXiv:2503.11232*.
- Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.
- Yoav Gur-Arieh, Roy Mayan, Chen Agassy, Atticus Geiger, and Mor Geva. 2025. [Enhancing automated interpretability with output-centric feature descriptions](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5757–5778, Vienna, Austria. Association for Computational Linguistics.
- Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. [EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. [OCNLI: Original Chinese Natural Language Inference](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online. Association for Computational Linguistics.
- Rajab Jafar, Fawzi Gamal, and Rais Raheem. 2025. Scalable and interpretable mixture of experts models in machine learning: Foundations, applications, and challenges.

- Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, and 181 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.
- Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*.
- Sihang Li, Wei Shi, Ziyuan Xie, Tao Liang, Guojun Ma, and Xiang Wang. 2025. Safer: Probing safety in reward models with sparse autoencoder. *arXiv preprint arXiv:2507.00665*.
- Yilun Liu, Yunpu Ma, Yuetian Lu, Shuo Chen, Zifeng Ding, and Volker Tresp. 2025. Parameter-efficient routed fine-tuning: Mixture-of-experts demands mixture of adaptation modules. *arXiv preprint arXiv:2508.02587*.
- Da Ma, Gonghu Shang, Zhi Chen, Libo Qin, Yijie Luo, Lei Pan, Shuai Fan, Lu Chen, and Kai Yu. 2025. Task-specific data selection for instruction tuning via monosemantic neuronal activations. *arXiv preprint arXiv:2503.15573*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. 2025. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.
- Gonalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. 2025. Automatically interpreting millions of features in large language models. *Preprint*, arXiv:2410.13928.
- 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.
- Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*.
- Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *Preprint*, arXiv:2310.16049.
- Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.
- David Vilares and Carlos G3mez-Rodr3guez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max W.F. Ku, Kai Wang, Alex Zhuang, Rongqi "Richard" Fan, Xiang Yue, and Wenhui Chen. 2024a. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *ArXiv*, abs/2406.01574.
- Zihan Wang, Deli Chen, Damai Dai, Runxin Xu, Zhuoshu Li, and Yu Wu. 2024b. Let the expert stick to his last: Expert-specialized fine-tuning for sparse architectural large language models. *arXiv preprint arXiv:2407.01906*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Shuyi Zhang, Wei Shi, Sihang Li, Jiayi Liao, Tao Liang, Hengxing Cai, and Xiang Wang. 2025. Interpretable reward model via sparse autoencoder. *arXiv preprint arXiv:2508.08746*.

Tong Zhu, Daize Dong, Xiaoye Qu, Jiacheng Ruan, Wenliang Chen, and Yu Cheng. 2024. Dynamic data mixing maximizes instruction tuning for mixture-of-experts. *arXiv preprint arXiv:2406.11256*.

A Appendix

A.1 Details of Prompts Used in Our monosemanticity analysis framework

This section provides a detailed introduction to the prompt designs used in our expert monosemanticity analysis framework, covering both explanation generation and evaluation stages. The prompt contents are crucial for the accuracy of expert explanation generation and evaluation, and facilitate reproducibility and comparison.

A.1.1 Explanation Generation Prompts:

P_{mono} and $P_{domain}^{(d)}$

During the explanation generation stage, we design two types of prompts: P_{mono} and $P_{domain}^{(d)}$. Full details are as follows:

Full Details of P_{mono}

System:

We're studying experts in a large-scale Mixture of Experts (MoE) language model. Each expert specializes in processing certain types of content, knowledge domains, linguistic patterns, or reasoning tasks in documents. The content that activates each expert in documents is indicated with $\langle\langle \dots \rangle\rangle$. We will give you a list of documents on which the expert is activated by the router, in order from most strongly activating to least strongly activating. Look at the parts of the document the expert activates for and summarize in a single sentence what type of content or domain this expert specializes in. Provide a reasonable explanation that captures the key distinguishing features. Note that some experts will specialize in very specific linguistic patterns or substrings, but others will activate on broader content types provided that the text contains particular knowledge domains or reasoning patterns. Your explanation should cover the general specialization pattern across most or all activating content (for example, don't give an explanation specific to a single domain if the expert activates across

multiple related areas). Pay attention to things like the subject matter, linguistic style, technical complexity, or reasoning type of the activating content, if that seems relevant. Keep the explanation as short and simple as possible, limited to 20 words or fewer. Omit punctuation and formatting. You should avoid giving long lists of specific topics. Include specific details about the semantic patterns, syntactic structures, or functional roles that characterize this expert's specialization. Mention key distinguishing features such as specific word types, grammatical constructions, or conceptual themes.

Your response should be in the form "This expert specializes in..."

User:

Here is the explanation: this expert specializes in

Here are the examples:
{examples}

Here, {examples} refers to our sampled activation examples, with activated tokens marked using $\langle\langle \dots \rangle\rangle$.

Full Details of $P_{domain}^{(d)}$

$P_{domain}^{(d)}$ includes all the content of P_{mono} . Additionally, to constrain $P_{domain}^{(d)}$ to generate explanations within a specific domain, taking medicine and finance as examples, we add the following prompt on the basis of P_{mono} :

System:

For the medicine:

This expert processes medical multiple-choice questions and clinical reasoning, so focus on medical terminology, diagnostic reasoning patterns, clinical decision-making, anatomical concepts, pharmacological knowledge, pathophysiology, treatment protocols, medical examination patterns, or healthcare do-

main knowledge.

IMPORTANT: This expert should **ONLY** specialize in medical and healthcare content. Do not generate explanations about non-medical domains.

For the finance:

This expert processes financial report question-answering, including SEC filings, earnings statements, balance sheets, and income statements. Focus on financial terminology, numerical reasoning over financial data, accounting concepts, revenue/profit analysis, financial ratios, or corporate financial reporting patterns.

IMPORTANT: This expert should **ONLY** specialize in financial report analysis and numerical financial reasoning. Do not generate explanations about non-financial domains.

Your response should be in the form "This expert specializes in..."

For the medicine:

Some examples: "This expert specializes in differential diagnosis reasoning in multiple-choice contexts", "This expert specializes in pharmacological drug mechanism descriptions and interactions", "This expert specializes in anatomical structure identification and spatial relationships", "This expert specializes in clinical symptom pattern recognition and analysis", "This expert specializes in medical procedure descriptions and protocols", "This expert specializes in pathophysiology and disease mechanism explanations", "This expert specializes in medical terminology and nomenclature", "This expert specializes in treatment option evaluation and selection".

For the finance:

Some examples: "This expert specializes in revenue and profit calculation patterns

in financial statements", "This expert specializes in percentage change computation over fiscal periods", "This expert specializes in balance sheet item extraction and arithmetic", "This expert specializes in financial ratio analysis and interpretation", "This expert specializes in multi-step numerical reasoning over tabular financial data", "This expert specializes in SEC filing language and regulatory disclosure patterns", "This expert specializes in interest expense and debt service calculation", "This expert specializes in equity and shareholder value terminology".

A.2 Explanation Evaluation Prompts

In the explanation evaluation stage, the full details of the designed prompts are as follows:

Full Details of Explanation Evaluation Prompts

System:

We're studying experts in a large-scale Mixture of Experts (MoE) language model. Each expert in the MoE layers specializes in processing certain types of content, knowledge domains, linguistic patterns, or reasoning tasks.

You will be given a short explanation of what this expert specializes in, and then be shown N example sequences in random order.

You will have to return a comma-separated list of the examples where you think this expert should be activated by the router, based on the content type, domain knowledge, or linguistic characteristics required.

For example, your response might look like "1, 3, 5".

Try not to be overly specific in your interpretation of the explanation.

If you think there are no examples where this expert will be significantly activated,

you should just respond with "None".

You should include nothing else in your response other than comma-separated numbers or the word "None" - this is important.

User:

The documents on which this expert is activated are given below:

{examples}

Here, {examples} is a set of numbered samples, where each sample is shown in its original form without $\langle \langle \rangle \rangle$ indicating activated tokens.

A.3 Expert Explanations from the Monosemanticity Analysis Framework

To enhance the understanding of expert specialization within MoE, we present the natural language explanations generated by our monosemanticity analysis framework below.

General Expert Explanations by P_{mono}

- This expert specializes in processing geographical coordinates and East Asian country names and language codes.
- This expert specializes in processing scientific and mathematical content such as synaptic input, quantum physics, and neuronal activity rate.
- This expert specializes in content related to gas and electricity prices and environmental impact on air quality.
- This expert specializes in documentation or metadata with type declarations and default value specifications.
- This expert specializes in processing scientific measurements and data conversion parameters in technical domains.
- The expert specializes in technical and scientific content involving mathematical calculations and physical properties.
- This expert specializes in processing exact word patterns, numerical sequences, and precise linguistic structures within technical or instructional contexts.

Medicine Expert Explanations by $P_{domain}^{(d)}$

- This expert specializes in preventive and social medicine textbook content analysis
- This expert specializes in diagnostic and treatment guidance for obstetrics and pediatrics in medical contexts
- This expert specializes in clinical reasoning for medical multiple-choice questions focusing on diagnostic and treatment options
- This expert specializes in medical legal and forensic terminology and concepts in healthcare contexts
- This expert specializes in step-by-step diagnostic reasoning in medical multiple-choice question contexts

Finance Expert Explanations by $P_{domain}^{(d)}$

- This expert specializes in SEC filing language and regulatory disclosure patterns
- This expert specializes in hedging strategies and adjustments in financial valuations using fair value and interest rates
- This expert specializes in fair value measurement and recognition in corporate financial statements
- This expert specializes in financial instrument valuation and accounting, focusing on fair value measurement and reporting standards
- This expert specializes in stock option valuation and compensation expense calculation using financial modeling techniques

A.4 Full Results with Mean and Standard Deviation

Results are reported as mean $\pm$ standard deviation over three runs.

A.5 Comparison with Parameter-Budget-Matched LoRA

To further control for the effect of LoRA's trainable-parameter budget, we increase its number of trainable parameters from the default setting to a level comparable to that of EDMRS, while using the same training hyperparameters.

Method	biology	business	chemistry	computer science	economics	engineering	history	law	math	other	philosophy	physics	psychology	Average
Full	48.91% ±0.82%	40.30% ±0.51%	45.67% ±0.62%	27.07% ±2.35%	53.52% ±1.02%	37.74% ±0.59%	40.68% ±1.84%	29.58% ±0.59%	34.91% ±0.59%	37.12% ±0.47%	39.95% ±0.64%	44.26% ±0.34%	66.67% ±1.32%	42.03% ±0.02%
AF	60.62% ±0.53%	43.09% ±1.41%	43.82% ±1.04%	49.59% ±0.61%	56.16% ±1.25%	37.77% ±0.47%	40.51% ±0.40%	26.85% ±1.78%	48.88% ±0.80%	46.32% ±1.82%	38.68% ±0.87%	47.65% ±1.06%	52.42% ±1.71%	45.57% ±0.11%
EDMRS	62.44% ±0.82%	47.53% ±0.55%	44.99% ±0.20%	54.88% ±0.42%	61.93% ±0.45%	42.31% ±1.00%	51.36% ±0.30%	31.61% ±0.16%	51.84% ±1.19%	50.32% ±0.68%	44.36% ±0.23%	47.93% ±0.04%	61.90% ±0.87%	50.26% ±0.15%
vs Full (pp)	+13.53	+7.23	-0.68	+27.81	+8.41	+4.57	+10.68	+2.03	+16.93	+13.20	+4.41	+3.67	-4.77	+8.23
vs AF (pp)	+1.82	+4.44	+1.17	+5.29	+5.77	+4.54	+10.85	+4.76	+2.96	+4.00	+5.68	+0.28	+9.48	+4.69

(a) Performance comparison of models trained with EDMRS, AF and Full strategies on Medicine in other domains.

Method	biology	law	chemistry	computer science	economics	engineering	health	history	math	other	philosophy	physics	psychology	Average
Full	73.36% ±0.14%	20.59% ±0.59%	41.43% ±0.40%	58.21% ±0.92%	56.44% ±0.30%	31.89% ±0.74%	55.58% ±0.25%	37.71% ±1.06%	59.83% ±0.23%	41.85% ±0.54%	46.83% ±0.23%	51.19% ±0.54%	62.07% ±0.26%	49.00% ±0.18%
AF	86.33% ±0.50%	42.26% ±0.41%	77.27% ±0.44%	78.94% ±0.92%	79.42% ±0.65%	55.56% ±0.70%	71.23% ±0.63%	58.53% ±1.05%	86.48% ±0.38%	65.19% ±0.62%	66.87% ±0.23%	79.24% ±0.18%	76.90% ±0.26%	71.09% ±0.16%
EDMRS	87.12% ±0.16%	43.90% ±2.21%	78.68% ±0.57%	79.76% ±1.06%	79.27% ±0.31%	57.24% ±0.51%	70.99% ±0.78%	62.73% ±0.79%	87.61% ±0.30%	65.55% ±0.54%	67.33% ±1.06%	80.55% ±0.04%	78.32% ±0.38%	72.23% ±0.11%
vs Full (pp)	+13.76	+23.31	+37.25	+21.55	+22.83	+25.35	+15.41	+25.02	+27.78	+23.70	+20.50	+29.36	+16.25	+23.23
vs AF (pp)	+0.79	+1.64	+1.41	+0.82	-0.15	+1.68	-0.24	+4.20	+1.13	+0.36	+0.46	+1.31	+1.42	+1.14

(b) Performance comparison of models trained with EDMRS, AF and Full strategies on Finance in other domains.

Method	biology	business	chemistry	computer science	economics	engineering	history	law	math	other	philosophy	physics	psychology	Average
LoRA	60.62% ±1.26%	40.98% ±0.29%	38.84% ±0.80%	44.72% ±0.56%	59.24% ±1.42%	36.33% ±0.52%	46.02% ±2.00%	25.52% ±0.65%	41.43% ±0.56%	46.61% ±0.06%	43.49% ±0.72%	43.26% ±0.58%	58.98% ±1.67%	45.08% ±0.06%
routerPEFT	70.25% ±1.40%	52.30% ±0.72%	47.44% ±1.24%	51.71% ±0.49%	59.24% ±0.52%	38.77% ±1.37%	41.56% ±0.61%	25.55% ±0.69%	55.39% ±0.64%	43.51% ±0.60%	38.34% ±1.33%	52.32% ±0.55%	58.69% ±1.43%	48.85% ±0.08%
EDMRS	62.44% ±0.82%	47.53% ±0.55%	44.99% ±0.20%	54.88% ±0.42%	61.93% ±0.45%	42.31% ±1.00%	51.36% ±0.30%	31.61% ±0.16%	51.84% ±1.19%	50.32% ±0.68%	44.36% ±0.23%	47.93% ±0.04%	61.90% ±0.87%	50.26% ±0.15%
vs LoRA (pp)	+1.82	+6.55	+6.15	+10.16	+2.69	+5.98	+5.34	+6.09	+10.41	+3.71	+0.87	+4.67	+2.92	+5.18
vs routerPEFT (pp)	-7.81	-4.77	-2.45	+3.17	+2.69	+3.54	+9.80	+6.06	-3.55	+6.81	+6.02	-4.39	+3.21	+1.41

(c) Performance comparison of PEFT-trained models on Medicine across other domains.

Method	biology	law	chemistry	computer science	economics	engineering	health	history	math	other	philosophy	physics	psychology	Average
LoRA	83.08% ±0.69%	31.43% ±0.60%	60.13% ±0.94%	69.27% ±0.98%	69.15% ±0.56%	44.38% ±0.74%	66.71% ±0.19%	48.99% ±0.15%	77.50% ±0.44%	54.58% ±0.60%	60.19% ±0.83%	67.39% ±0.19%	71.51% ±0.51%	61.87% ±0.17%
routerPEFT	78.75% ±0.21%	31.67% ±1.00%	32.33% ±0.85%	46.91% ±1.34%	59.91% ±0.30%	24.84% ±0.67%	60.27% ±1.32%	50.13% ±0.45%	54.68% ±0.09%	55.63% ±0.68%	50.50% ±1.57%	47.99% ±1.33%	73.81% ±0.25%	51.34% ±0.12%
EDMRS	87.12% ±0.16%	43.90% ±2.21%	78.68% ±0.57%	79.76% ±1.06%	79.27% ±0.31%	57.24% ±0.51%	70.99% ±0.78%	62.73% ±0.79%	87.61% ±0.30%	65.55% ±0.54%	67.33% ±1.06%	80.55% ±0.04%	78.32% ±0.38%	72.23% ±0.11%
vs LoRA (pp)	+4.04	+12.47	+18.55	+10.49	+10.12	+12.86	+4.28	+13.74	+10.11	+10.97	+7.14	+13.16	+6.81	+10.36
vs routerPEFT (pp)	+8.37	+12.23	+46.35	+32.85	+19.36	+32.40	+10.72	+12.60	+32.93	+9.92	+16.83	+32.56	+4.51	+20.89

(d) Performance comparison of PEFT-trained models on Finance across other domains.

Table 5: Full results (mean $\pm$ std) across different domains and training strategies: (a) Trained on Medicine, comparing AF and Full strategies; (b) Trained on Finance, comparing AF and Full strategies; (c) Trained on Medicine, comparing PEFT strategy; (d) Trained on Finance, comparing PEFT strategy. Bolded rows report absolute differences between EDMRS and the baselines, measured in percentage points.

As shown in Table 6, EDMRS outperforms the parameter-budget-matched LoRA baseline on both the Medicine target domain and the MMLU-Pro out-of-domain evaluation. This comparison indicates that the performance improvement over LoRA cannot be attributed solely to a larger trainable-parameter budget.

A.6 Empirical Examination of the EDMRS Heuristic

To empirically validate the core intuition of our heuristic analysis, we isolate the effect of a sin-

Method	Medicine	MMLU-Pro Average
LoRA ($r = 64$)	53.39% $\pm$ 0.87%	42.74% $\pm$ 0.24%
EDMRS	60.09% $\pm$ 0.07%	50.26% $\pm$ 0.15%
vs LoRA (pp)	+6.70	+7.52

Table 6: Comparison between EDMRS and parameter-budget-matched LoRA on Medicine.

gle expert by training only one expert at a time, setting the number of training steps inversely proportional to activation frequency to approximately

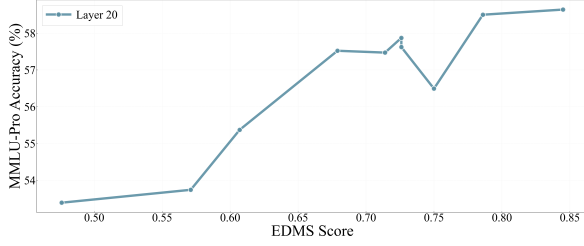


Figure 5: Performance of individual experts with varying EDMS scores after AF-adjusted training. Higher EDMS is generally associated with better performance.

equalize gradient updates across experts. We randomly select ten experts with varying EDMS from a representative middle layer (layer 20). As shown in Figure 5, higher EDMS scores are generally associated with better out-of-domain performance, suggesting that experts with higher EDMS tend to exhibit less forgetting, lending empirical support to our heuristic.

A.7 Experiments Settings

In our experiments, GPT-4o is employed as the external large language model within the analytical framework, while Qwen3-30B-A3B-Instruct-2507 serves as the base model.

To reduce potential effects arising from differences in data formats, we convert both domain datasets into two-turn conversational formats. For MedMCQA, we embed the four answer choices into the A-J option-label space used by MMLU-Pro and remap the correct answer accordingly. For FinQA, we organize the serialized financial tables together with the corresponding report contexts and questions as model inputs, and use the annotated calculation programs and their execution results as target outputs.

For the comparative experiments on Medicine involving three strategies, the hyperparameter configurations are as follows: a batch size of 32 and a learning rate of $5e-5$ are used. For the Finance dataset, all methods utilize a batch size of 32, with an initial learning rate of $1e-5$ for all methods. Results are reported as mean $\pm$ standard deviation over three runs. When applying the expert selection fine-tuning strategy, we select all layers of the model and only fine-tune 16 experts within each layer.

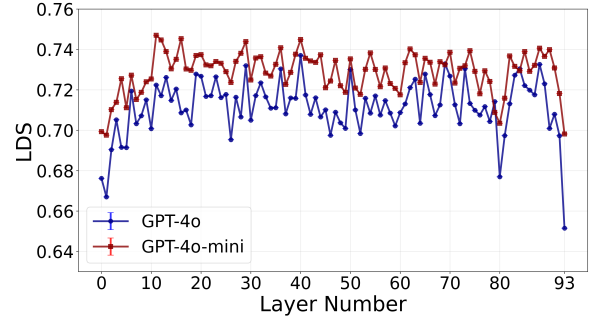


Figure 6: Analysis of metric stability under affordable external LLMs.

A.8 Cross-Task Generalization on Out-of-Domain Benchmarks

The results in Table 7 indicate that, within the scope of this evaluation, EDMRS more effectively preserves cross-task out-of-domain capabilities after domain-specific fine-tuning.

A.9 Stability of Metrics with Affordable External LLMs

To evaluate the stability of monosemanticity metrics when employing more cost-effective external large language models, we conducted experiments on Qwen3-235B-A22B using both GPT-4o and the smaller, lower-cost GPT-4o-mini to assess all layers. Comparative results are presented in Figure 6, where the relative variation trends demonstrate high consistency across different external LLMs, reflecting the robustness of monosemanticity prediction. Thus, even when shifting to a more economical external LLM, our scoring remains highly reliable.

A.10 Case Study: Comparing Semantic Granularity of Experts and Neurons

Intuitively, an MoE expert is a large feed-forward network composed of many neurons, and thus the semantics it encodes should operate at a higher and more abstract granularity. To verify this, we apply our monosemanticity evaluation framework separately to experts and neurons on a randomly selected monosemantic intermediate layer, without specifying any granularity constraints. The results show that expert-level explanations operate at a higher semantic level, while neuron-level explanations remain at a lower level. For example, Expert 114 is described as “specializing in medical textbook content covering preventive medicine, pathology, pharmacology, and physiology,” whereas its internal Neuron 131 is described

Method	C3	OCNLI	HeadQA	OpenBookQA	EXAMS	WiC	HellaSwag	SWAG	CODAH	MuSR
Full	56.95% ±1.03%	40.42% ±0.17%	46.61% ±0.49%	65.67% ±0.64%	39.63% ±0.35%	53.69% ±0.44%	54.24% ±0.13%	25.97% ±0.08%	43.47% ±0.25%	47.49% ±0.35%
AF	50.92% ±0.21%	49.04% ±0.44%	41.41% ±0.52%	80.80% ±0.87%	35.04% ±0.17%	57.48% ±0.29%	73.39% ±0.16%	57.67% ±0.21%	78.15% ±0.32%	48.32% ±0.50%
EDMRS	80.53% ±0.13%	49.93% ±0.70%	49.11% ±0.13%	82.87% ±0.64%	37.48% ±0.28%	63.29% ±0.52%	76.44% ±0.10%	60.23% ±0.10%	80.01% ±0.33%	50.53% ±0.46%
vs Full (pp)	+23.58	+9.51	+2.50	+17.20	-2.15	+9.60	+22.20	+34.26	+36.54	+3.04
vs AF (pp)	+29.61	+0.89	+7.71	+2.07	+2.45	+5.81	+3.05	+2.56	+1.86	+2.20

Table 7: Performance comparison of EDMRS, AF, and Full strategies trained on Medicine and evaluated on multiple benchmarks.

as “words with embedded content between ‘<’ and ‘>’ symbols, often indicating emphasis or technical notation.” Representative cases are shown in Table 8.

A.11 Heuristic Analysis and Motivation for the EDMRS Strategy

As discussed earlier, we propose a new strategy to identify monosemantic experts related to a certain downstream domain. In this section, we provide a heuristic theoretical analysis of our strategy.

Specifically, we denote d as the fine-tuned target domain, and consider a layer of the MoE model containing the expert set $E_1, E_2, \dots, E_N$. For each expert E_i , we define its activation frequency in domain d as $\text{AF}_i^{(d)}$ and its Expert Domain Monosemanticity Score as $\text{EDMS}_i^{(d)}$. To formalize the objective for expert selection, we define the performance change resulting from fine-tuning expert E_i in domain d for a target-domain training token t as:

$$C_i(t) = \Delta_{\mathcal{P}}^{(d)}(E_i, t) - \Delta_{\mathcal{N}}^{(\text{other})}(E_i, t) \quad (8)$$

where $\Delta_{\mathcal{P}}^{(d)}(E_i, t)$ represents the target-domain performance improvement attributable to the update induced by token t , while $\Delta_{\mathcal{N}}^{(\text{other})}(E_i, t)$ denotes the other-domain performance degradation attributable to the same update.

Within our MoE monosemanticity analysis framework, a high $\text{EDMS}_i^{(d)}$ score indicates that the token activation pattern of expert E_i aligns well with the domain-specific explanation. In other words, the token features processed by the expert are well-matched to target d , contributing substantially to domain-specific performance while remaining largely decoupled from other domains. Conversely, a low $\text{EDMS}_i^{(d)}$ implies that the expert’s activations do not correspond to the domain explanation, with token features poorly aligned to d , resulting in minimal domain-specific

contribution and increased coupling with other domains.

Building upon this theoretical intuition, we posit the following heuristic: the expected performance change of expert E_i on a single token, conditioned on its activation, satisfies $\mathbb{E}[\Delta_{\mathcal{P}}^{(d)}(E_i, t) \mid E_i'] = h_1(\text{EDMS}_i^{(d)})$ and $\mathbb{E}[\Delta_{\mathcal{N}}^{(\text{other})}(E_i, t) \mid E_i'] = h_2(1 - \text{EDMS}_i^{(d)})$, where $h_1(\cdot)$ and $h_2(\cdot)$ are monotonically non-decreasing functions. Applying the law of total expectation, the expected total performance change for expert E_i across all domains is given by:

$$\begin{aligned} \mathbb{E}[C_i] &= \mathbb{E}[C_i(t)] \\ &= \mathbb{E}[\Delta_{\mathcal{P}}^{(d)}(E_i, t)] - \mathbb{E}[\Delta_{\mathcal{N}}^{(\text{other})}(E_i, t)]. \end{aligned}$$

Within the sparse activation mechanism of the MoE architecture, an expert only contributes to the model output when it is activated. If the expert is not activated, it does not induce any performance change, i.e.,

$$\mathbb{E}[\Delta_{\mathcal{P}}^{(d)}(E_i, t) \mid E_i''] = 0, \quad (9)$$

$$\mathbb{E}[\Delta_{\mathcal{N}}^{(\text{other})}(E_i, t) \mid E_i''] = 0, \quad (10)$$

where E_i'' indicates that expert E_i is not activated during the output of token t .

Therefore, by the law of total expectation, the expectation can be decomposed into the cases of activation and non-activation:

$$\begin{aligned} \mathbb{E}[\Delta_{\mathcal{P}}^{(d)}(E_i, t)] &= P(E_i') \cdot \mathbb{E}[\Delta_{\mathcal{P}}^{(d)}(E_i, t) \mid E_i'] \\ &\quad + P(E_i'') \cdot \mathbb{E}[\Delta_{\mathcal{P}}^{(d)}(E_i, t) \mid E_i''] \\ &= \text{AF}_i^{(d)} \cdot \mathbb{E}[\Delta_{\mathcal{P}}^{(d)}(E_i, t) \mid E_i'] \\ &\quad + (1 - \text{AF}_i^{(d)}) \cdot 0 \\ &= \text{AF}_i^{(d)} \cdot \mathbb{E}[\Delta_{\mathcal{P}}^{(d)}(E_i, t) \mid E_i']. \end{aligned} \quad (11)$$

Similarly,

$$\mathbb{E}[\Delta_{\mathcal{N}}^{(\text{other})}(E_i, t)] = \text{AF}_i^{(d)} \cdot \mathbb{E}[\Delta_{\mathcal{N}}^{(\text{other})}(E_i, t) \mid E_i'']. \quad (12)$$

Expert ID	Neuron ID	Explanation
114	—	This expert specializes in medical textbook content covering preventive medicine pathology pharmacology and physiology
114	131	Words with embedded content between ‘<’ and ‘>’ symbols, often indicating emphasis or technical notation
114	710	The neuron is activating on various short function words and suffixes within word fragments, indicating a focus on small linguistic units rather than specific concepts or content
114	419	The neuron is activating on numerical values such as numbers and percentages, as well as mathematical or quantitative indicators
114	264	The neuron is activating on numeric percentages and proportions, such as 10%, and concepts related to measurement or quantification, like body mass index
114	690	The neuron is activating on common words and connecting words which are not specific to any particular concept, such as “of”, “and”, “in”, “between”, and “is”
114	557	Common short English words such as “and,” “of,” “in,” and “to.”

Table 8: Expert and neuron explanations.

Thus, the expected total performance change is given by:

$$\mathbb{E}[C_i] = \text{AF}_i^{(d)} \cdot \left[\mathbb{E}[\Delta_P^{(d)}(E_i, t) | E'_i] - \mathbb{E}[\Delta_N^{(\text{other})}(E_i, t) | E'_i] \right]. \quad (13)$$

Therefore, under the preceding heuristic, the expected total performance change can be approximated as

$$\mathbb{E}[C_i] \approx \text{AF}_i^{(d)} h(\text{EDMS}_i^{(d)}), \quad (14)$$

where $h(x) = h_1(x) - h_2(1 - x)$. This approximation motivates the following surrogate objective for selecting m experts:

$$\tilde{\mathcal{J}}(S) = \sum_{i \in S} \text{AF}_i^{(d)} h(\text{EDMS}_i^{(d)}), \quad |S| = m. \quad (15)$$

We adopt the positive proportional instantiation $h(x) = kx$, where $k > 0$ is a constant shared across experts. Under this instantiation,

$$\begin{aligned} \tilde{\mathcal{J}}(S) &= k \sum_{i \in S} \text{AF}_i^{(d)} \text{EDMS}_i^{(d)} \\ &= k \sum_{i \in S} \text{EDMRS}_i^{(d)}. \end{aligned} \quad (16)$$

Because the positive constant k does not affect expert ranking, selecting the m experts with the highest EDMRS scores maximizes this surrogate objective, providing a heuristic justification for our EDMRS strategy.

A.12 Analysis of Word Limit in Expert Explanation Generation

We conduct an analysis on a randomly selected monosemantic middle layer, varying the word limit for expert explanations from 20 to 100 and

observing the corresponding changes in LDS. Results are shown in Table 9. A word limit of 20 achieves the highest LDS, reflecting a trade-off between underfitting and overfitting: when the description is too long, it tends to overfit to the sampled activation examples. For instance, an expert that should be described as “epidemiology” may overfit to specific instances such as “influenza and tuberculosis” when given a longer description, thereby narrowing the scope of the explanation and causing LDS to decrease. This finding suggests that sparsity-induced expert monosemanticity can be effectively captured by concise natural language descriptions, further supporting the structural analogy between MoE experts and SAE neurons in our framework: MoE experts are monosemantic and exhibit high-level semantic specialization.

Word limit	20	30	40	50	60	80	100
LDS	0.7081	0.6911	0.6961	0.6939	0.7008	0.6869	0.6974

Table 9: LDS under different explanation word limits.

A.13 Use of AI Assistants

The present manuscript’s Introduction and Conclusion underwent limited language refinement through a general-purpose large language model, strictly for improving grammar, wording, and readability. The model played no role in conceptualization, data or code generation, image production, mathematical derivation, data analysis, literature search, or citation management. Its use was confined to light textual polishing of pre-written content, with no confidential or external data disclosed. All AI-proposed changes were carefully examined and manually approved by the authors, who retain sole accountability for the scholarly content of the work.