SAE-Based Representations for MoE Experts

Anonymous ACL submission

Abstract

Mixture-of-Experts (MoE) models enable efficient scaling via conditional computation, yet it remains unclear whether experts develop meaningful specializations or simply partition the input space. We propose a framework for representing MoE experts in the space of Sparse Autoencoder (SAE) features and introduce a feature specificity metric that quantifies the association between SAE features and individual experts. The resulting representations achieve approximately 0.8 cosine similarity with routing-based importance estimates and support interpretable expert groupings via density-based clustering. We further propose a pruning strategy that combines load-based filtering with cluster-level redundancy, removing semantically non-essential experts while preserving task-relevant ones. Specifically, at matched expert removal budgets, SAE-guided pruning retains up to 2× more task performance on FinQA and up to 4× more on GSM8K relative to the load-based baseline. On broad-domain benchmarks, gains are more modest, reflecting the dependence of semantic expert selection on the coverage of the SAE training corpus. These results demonstrate that MoE experts exhibit meaningful internal structure, and that SAE feature representations provide an effective basis for capturing and exploiting this structure for both model analysis and compression.

1 Introduction

Mixture-of-Experts (MoE) (Fedus et al., 2021) architectures have become a dominant approach for scaling large language models by enabling conditional computation over a large set of parameters. Instead of activating all model weights for every token, MoE models dynamically route inputs to a small subset of experts, allowing substantial increases in model capacity while maintaining manageable inference costs (Shazeer et al., 2017). This paradigm has been adopted in many recent large-scale language models and has demonstrated strong performance across a wide range of natural language processing tasks (Team, 2024; DeepSeek-AI, 2024a,b).

Despite their practical success, the internal organization of MoE models remains poorly understood. A common hypothesis is that experts learn specialized behaviors and encode distinct semantic patterns (Olson et al., 2025). However, existing analyses of expert specialization are typically based on routing statistics or token frequency (Xue et al., 2024), which provide only limited insight into the actual representational structure of experts. As a result, it remains unclear to what extent experts correspond to coherent semantic functionalities, how redundancy emerges between experts, and whether such structure can be exploited for downstream applications such as pruning (Chen et al., 2022).

In this work, we propose a framework for representing MoE experts in the space of Sparse Autoencoder (SAE) features (Bricken et al., 2023). Sparse autoencoders have recently emerged as an effective tool for extracting interpretable features from transformer representations (Bricken et al., 2023; Huben et al., 2024; Gao et al., 2024). We leverage this property to characterize experts through their activation patterns over SAE features, enabling a semantically grounded description of expert behavior. Based on these representations, we propose Normalized Mean Difference (NMD) and Fold Change (FC), feature specificity metrics that quantifies the association between SAE features and individual experts, allowing us to identify both highly specialized and highly redundant experts.

Using the proposed representation, we analyze the structure of experts in several MoE layers and observe that experts form meaningful groups in SAE feature space. We further show that these representations are consistent with routing behavior and align with expert importance estimates derived from activation statistics. Building on these obser-

vations, we investigate expert redundancy through clustering-based analyses and study pruning strategies informed by the discovered semantic structure.

Our experiments demonstrate that many experts exhibit substantial representational overlap and can be removed with limited impact on downstream performance. In particular, we show that pruning strategies informed by SAE-based expert representations outperform or complement standard load-based pruning approaches across multiple reasoning and question-answering benchmarks.

Our approach yields SAE-based expert representations with approximately 0.8 cosine similarity to routing-based importance estimates, and groups the 60 experts of a representative MoE layer into 14 semantically coherent clusters. At matched expert removal budgets, SAE-guided pruning retains up to $2\times$ more task performance on FinQA and up to $4\times$ more on GSM8K relative to load-based pruning.

Our contributions are as follows:

- We propose a novel framework for representing MoE experts in the space of Sparse Autoencoder (SAE) features, enabling a semantically grounded characterization of expert specialization.
- We introduce a feature specificity metric that quantifies the association between SAE features and individual experts, and show that SAE feature space faithfully reflects the interpretable functional organization of MoE experts, as evidenced by strong alignment with both router behavior and regression-based importance estimates.
- We propose a dual pruning strategy combining low utilization and SAE cluster-level redundancy, enabling removal of different subsets of redundant experts while preserving model performance.

2 Related Work

2.1 Mixture-of-Experts Language Models

Mixture-of-Experts (MoE) architectures are a standard approach for scaling large language models while maintaining efficient per-token computation. The Switch Transformer (Fedus et al., 2021) demonstrated routing each token to a single expert at large scale, while Mixtral (Jiang et al., 2024) extended this idea with top-k routing over larger expert pools, achieving strong performance across benchmarks. Zhou et al. (2022) proposed expert-choice routing, where experts select tokens instead of tokens selecting experts, improving load balancing and providing a direct measure of utilization.

Despite their success, the structure of MoE experts remains not fully understood. Studies such as OpenMoE (Xue et al., 2024) and probing analyses (Olson et al., 2025) suggest partial structure in routing behavior, while Wang et al. (2026) argue that specialization is often overstated due to significant redundancy among experts.

2.2 Sparse Autoencoders for Interpretability

Sparse Autoencoders (SAEs) are widely used for interpreting internal representations of language models (Gao et al., 2024; Huben et al., 2024). By learning sparse decompositions of activations, SAEs extract more monosemantic features than individual neurons (Bricken et al., 2023; Gao et al., 2024). Recent work improves feature quality via decorrelation objectives (Yan et al., 2025) and extends SAE analysis across layers to capture higher-level patterns (Shi et al., 2025).

2.3 Expert Pruning in MoE Models

Pruning MoE models aims to reduce memory footprint while preserving performance by removing experts. Large-scale MoE systems exhibit highly imbalanced expert utilization, indicating substantial redundancy. Load-based pruning, which removes least-used experts based on routing frequency, is a common baseline in sparse MoE architectures (Fedus et al., 2021; Lepikhin et al., 2020).

However, recent pruning strategies move beyond load-based utilization as the sole signal of expert importance (Lu et al., 2024; Lasby et al., 2025). These methods improve selection criteria but still lack interpretability.

3 Expert Representations in SAE Feature Space

We study Mixture-of-Experts (MoE) models through the lens of Sparse Autoencoder (SAE) features. Our central hypothesis is that experts can be represented as structured activation patterns in an interpretable sparse feature space.

Instead of analyzing experts directly in the original hidden activation space, we project token representations into SAE features and aggregate them at the expert level. This produces sparse expert representations that capture characteristic feature

activation patterns associated with different routing behaviors.

3.1 SAE Feature Extraction

We first map dense hidden activations into a sparse interpretable feature space using a Sparse Autoencoder (SAE) (Bricken et al., 2023; Gao et al., 2024). This representation enables analysis of expert behavior in terms of sparse semantic features rather than dense hidden states. The approach is motivated by the superposition hypothesis (Elhage et al., 2022), which suggests that neural networks represent more features than they have dimensions by superimposing them in activation space.

Given a hidden activation vector $h \in \mathbb{R}^d$, the SAE encoder produces

$$z = f(W_{\text{enc}}h + b_{\text{enc}}),$$

where $W_{\text{enc}} \in \mathbb{R}^{k \times d}$ is the encoder matrix, $b_{\text{enc}} \in \mathbb{R}^k$ is the encoder bias, $f(\cdot)$ is a sparsifying nonlinearity (Huben et al., 2024), and $z \in \mathbb{R}^k$ is the resulting sparse feature representation.

We denote individual SAE feature indices by $i \in \{1, \ldots, k\}$, where z_i corresponds to the activation of feature i.

The SAE representation provides a sparse and approximately interpretable basis over model activations (Bricken et al., 2023; Gao et al., 2024). Each token is therefore represented not by a dense hidden state, but by a sparse set of activated semantic features. Multiple SAE architectures have been proposed, differing primarily in the choice of sparsifying nonlinearity $f(\cdot)$: ReLU-based SAEs (Bricken et al., 2023; Huben et al., 2024) use an L1 penalty to encourage sparsity, TopK SAEs (Gao et al., 2024) enforce exactly K active features per input, JumpReLU SAEs (Rajamanoharan et al., 2024) learn per-feature activation thresholds, and Matryoshka SAEs (Bussmann et al., 2025) extend this to hierarchical multi-resolution representations. In this work, we employ TopK SAEs.

3.2 Aggregating Token-Level Features into Expert Profiles

To characterize experts in feature space, we aggregate SAE activations over tokens routed to each expert.

Let E denote the set of experts in an MoE layer, and let T_e denote the set of token activations routed to expert $e \in E$. We define the expert profile as

$$z_e = \frac{1}{|T_e|} \sum_{h \in T_e} f(W_{\text{enc}}h + b_{\text{enc}}).$$

Here, $|T_e|$ is the number of tokens routed to expert e, and $z_e \in \mathbb{R}^k$ represents the average SAE activation vector associated with that expert.

This aggregation converts token-level sparse activations into an expert-level semantic representation. Intuitively, dimensions with large values in z_e correspond to SAE features that are frequently activated for tokens processed by the expert.

3.3 Feature–Expert Association Metrics

Not all active features are equally informative for describing expert specialization.

We therefore consider two feature-expert association metrics that estimate how selectively a feature activates for a particular expert.

Let $\mu_e(i)$ denote the mean activation of SAE feature i over tokens routed to expert e, and let $\mu_{\text{other}}(i)$ denote the corresponding mean activation over tokens routed to all remaining experts.

Normalized Mean Difference (NMD). To additionally account for feature variability Our primary metric is the Normalized Mean Difference (NMD), defined as

$$\text{NMD}(i, e) = \frac{\mu_e(i) - \mu_{\text{other}}(i)}{\sigma(i) + \varepsilon},$$

where $\sigma(i)$ is the global standard deviation of SAE feature i, and ε is a small numerical constant.

Fold Change (FC). For comparison, we also evaluate Fold Change (FC), which measures relative feature enrichment for a given expert:

$$\text{FC}(i, e) = \frac{\mu_e(i)}{\text{mean}_{e' \neq e} \mu_{e'}(i) + \varepsilon}.$$

Here, $\mu_k(i)$ denotes the mean activation of SAE feature i for expert k, and ε is a small numerical constant added for stability.

High FC values indicate that a feature activates substantially more strongly for a given expert than for the remaining experts. FC is simple and computationally efficient, but does not account for activation variance.

Unlike FC, NMD normalizes activation differences by feature variability and is therefore comparable across features with different activation scales. High NMD values correspond to features that are

both strongly activated and selectively associated with a particular expert.

Empirically, we find that NMD produces more stable and semantically coherent expert representations than FC. Therefore, we use NMD throughout the main experiments, while additional comparisons are provided in the Figure 1.

3.4 Sparse Expert Representations

The full SAE feature space is highly dimensional and contains many weakly informative activations. To obtain compact expert representations, we retain only the most strongly associated features for each expert.

For every expert e, we select the top K features according to an association metric $NMD(i, e)$.

$$\mathrm{TopK}(e) = \arg\operatorname*{topK}_{i} NMD(i, e).$$

Here, $\mathrm{TopK}(e)$ denotes the set of SAE features with the highest association scores for expert e.

Using these feature subsets, we define the sparse expert subspace

$$\mathcal{Z}^{(K)} = \bigcup_{e \in E} \mathrm{TopK}(e).$$

This representation isolates the most informative semantic dimensions of expert behavior while discarding weakly associated or non-specific features. As a result, experts can be represented compactly in terms of a relatively small number of characteristic SAE features.

3.5 Expert Similarity and Geometry

Representing experts in SAE feature space enables direct geometric comparison between experts.

Given two experts $e_a, e_b \in E$ with profiles z_{e_a} and z_{e_b}, we compute cosine similarity as

$$\mathrm{sim}(e_a, e_b) = \frac{z_{e_a}^{\top} z_{e_b}}{\|z_{e_a}\| \, \|z_{e_b}\|}.$$

Here, $\| \cdot \|$ denotes the Euclidean norm.

Experts with similar feature activation patterns occupy nearby regions of SAE space, while experts specializing in different behaviors become geometrically separated. This geometric structure enables downstream analyses such as clustering, redundancy estimation, and structure-aware expert pruning.

3.6 Experimental Setup

We conduct all experiments using `Qwen/Qwen1.5-MoE-A2.7B` (Team, 2024), a transformer-based Mixture-of-Experts language model with 60 experts per routed layer. For feature extraction, we employ a Sparse Autoencoder with expansion factor 32, mapping dense hidden states $h \in \mathbb{R}^d$ to sparse feature representations $z \in \mathbb{R}^k$ with $k = 32 \cdot d$. We train multiple SAEs, one per each layer starting from layers 12 through 24 of the same model. Expert representations and routing statistics are computed on the `DKYoon/SlimPajama-6B` dataset (DKYoon, 2023), with a minimum of 50,000 tokens routed to each expert to ensure stable profile estimation. Unless stated otherwise, all analyses are performed over layers 12 through 24.

4 Analysis of SAE Expert Representations

In this section, we evaluate whether the proposed SAE-based expert representations capture meaningful structure in MoE routing behavior. We focus on three complementary aspects: (i) the extent to which SAE features encode routing decisions, (ii) alignment between feature-based importance and learned routing behavior, and (iii) geometric structure of experts in SAE space.

4.1 Routing Probing

We evaluate whether SAE-based feature representations contain sufficient information to approximate routing decisions in a Mixture-of-Experts model. Specifically, we study whether expert assignments can be recovered from compact subsets of SAE features selected using the proposed feature–expert association metrics.

We use activations extracted from layer 15 of the model described in Section 3.6. To construct the evaluation set, we sample 60,000 token-level SAE representations uniformly across all 60 experts of the MoE layer, resulting in 1,000 samples per expert. Each token is labeled using the original routing decision of the model, yielding a supervised dataset of pairs (z, e), where $z \in \mathbb{R}^k$ is the SAE representation and $e \in E$ is the routed expert.

For each expert, we select the top K associated SAE features using either Fold Change (FC) or Normalized Mean Difference (NMD). We then evaluate routing prediction performance using only the selected feature subsets while varying K from 1

to 100 features per expert. Performance is measured using macro-averaged F1 score over expert prediction.

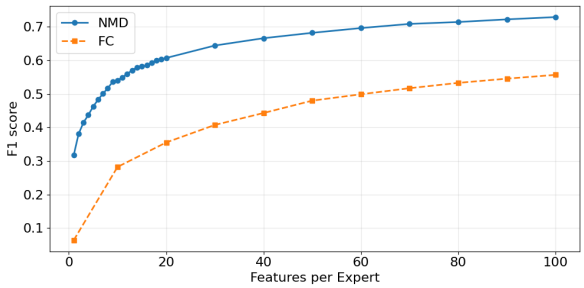


Figure 1: Routing prediction performance as a function of the number of selected SAE features per expert for FC and NMD feature selection.

The results are shown in Figure 1. We observe that routing decisions can be recovered surprisingly well from compact SAE feature subsets, indicating that routing-relevant information is strongly reflected in SAE feature space.

Remarkably, even in the extremely sparse regime of a single selected feature per expert, routing prediction achieves approximately 0.3 F1 score. Increasing the representation size to 100 features per expert improves performance to approximately 0.7 F1 score, suggesting that a relatively small number of SAE features already captures a substantial fraction of routing structure.

Across all values of K, feature subsets selected using NMD consistently outperform those selected using FC. This indicates that variance-normalized feature specificity provides a more reliable estimate of routing-relevant features than raw activation enrichment alone.

4.2 Feature Importance Alignment

We next evaluate whether feature importance inferred from SAE association metrics aligns with feature importance learned by a supervised routing prediction model. The goal of this experiment is to determine whether features identified as expert-specific by the proposed metrics are also important for approximating routing decisions.

Using the same routing prediction setup as in Section 4.1, we train a linear classifier on the Top-100 SAE feature subspace and compare the resulting classifier weights with feature importance scores produced by the NMD metric.

For each expert $e \in E$, we compare the vector of classifier weights $w_e(i)$ with the corresponding NMD feature association scores $\mathrm{NMD}(i, e)$. We evaluate agreement using both Pearson correlation and cosine similarity.

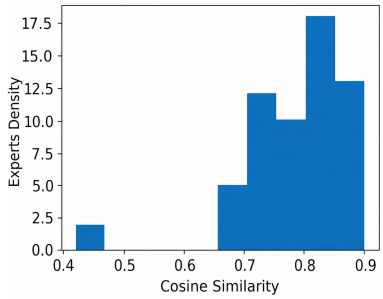


Figure 2: Alignment between NMD feature association scores and classifier feature importance weights in the Top-100 SAE feature subspace.

Across the full Top-100 feature subspace, we observe an average Pearson correlation of approximately 0.47 and an average cosine similarity of approximately 0.47 between NMD scores and classifier weights. This indicates moderate agreement between unsupervised feature specificity estimates and supervised routing importance.

We additionally perform the same analysis restricted to expert-specific features, i.e., features assigned to expert e according to the NMD criterion. In this setting, the average cosine similarity increases substantially to approximately 0.786, while the average Pearson correlation decreases to approximately 0.185.

This behavior suggests that both approaches identify highly similar discriminative directions in SAE feature space, despite differences in the relative scaling of feature importance values. In particular, the strong directional agreement indicates that SAE association metrics recover feature subsets that are closely aligned with those used by the supervised routing approximation model.

4.3 Expert Clustering

Finally, we analyze the geometric structure of experts in SAE feature space, focusing on layer 20 of `Qwen/Qwen1.5-MoE-A2.7B` as a representative mid-layer where both syntactic and semantic information are expected to coexist.

Each expert is represented by its aggregated SAE feature profile $z_e \in \mathbb{R}^k$, computed in the Top-K feature subspace with $K = 100$, where features are selected according to the NMD association metric introduced in Section 3.

The resulting representation space is highly sparse and high-dimensional (approximately 65k

features across all SAE dimensions for 60 experts), which motivates dimensionality reduction prior to clustering. We apply PCA to reduce the Top-100 feature representations to 10 dimensions. On the resulting embeddings, we evaluate multiple clustering methods, including k-means and HDBSCAN, selecting HDBSCAN as the best-performing approach.

Under this configuration, HDBSCAN identifies 14 clusters with a noise ratio of 0.233. The clustering quality is supported by a silhouette score of 0.482 and a Davies–Bouldin index of 0.814 in the reduced space. Additionally, the density-based clustering validation (DBCV) score computed in the original feature space is 0.283.

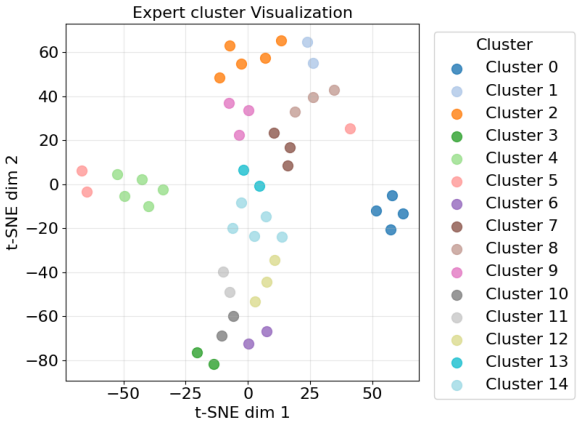


Figure 3: Visualization of expert clusters obtained from SAE feature profiles in layer 20 using Top-100 NMD-selected features, PCA dimensionality reduction, and HDBSCAN clustering.

These results indicate that expert representations in SAE feature space exhibit a non-trivial internal organization: experts are neither uniformly distributed nor perfectly separable, but instead form a moderate number of structured clusters with a non-negligible fraction of noise points.

To further validate the clustering structure, we evaluate the interpretability consistency of experts within each cluster. For every expert, the top-20 NMD-ranked SAE features are interpreted, then aggregated into expert-level descriptions via LLM summarization. Cluster semantics are obtained by merging the descriptions of experts belonging to the same cluster.

The resulting interpretations show strong semantic coherence inside clusters. In particular, clusters 1, 2, 8, and 9 share a common technical and factual orientation, covering analytical, informational, and report-style discourse. In contrast, cluster 0 is dominated by personal and conversational content, while clusters 11 and 12 primarily capture narrative, cultural, and media-oriented semantics rather than technical knowledge.

Overall, these results suggest that clustering in SAE feature space reflects meaningful functional specialization patterns rather than purely geometric similarity.

5 Applications to Expert Pruning

Mixture-of-Experts (MoE) architectures often exhibit significant redundancy in expert utilization, where multiple experts learn overlapping or weakly differentiated functions. This motivates the study of structured pruning strategies that operate in a semantically grounded representation space rather than relying solely on routing statistics. In this work, we propose a Sparse Autoencoder (SAE)-guided framework for expert pruning and post-pruning refinement, where experts are analyzed and filtered in the space of SAE-derived features. Our goal is to improve pruning robustness across heterogeneous downstream tasks while preserving task-relevant expert diversity.

5.1 Method

We formalize our two-stage pruning strategy in Algorithm 1. The method first constructs a compact semantic representation of experts using SAE features, then applies load-based filtering followed by cluster-aware refinement.

The resulting pruning strategy combines routing-based importance estimation with semantic structure derived from sparse SAE feature representations.

5.2 Experimental Setup

We follow the model, SAE, and data configuration described in Section 3.6. The overall expert removal ratio reported in our results is aggregated across layers 12 through 24. For Stage 2 semantic refinement, we construct dataset-specific anchor representations using four evaluation benchmarks: FinQA (Chen et al., 2021), PubMedQA (Jin et al., 2019), GSM8K (Cobbe et al., 2021), and MMLU (Hendrycks et al., 2021). These datasets collectively span mathematical reasoning, scientific question answering, and general knowledge domains.

Hyperparameters. We explore two key pruning hyperparameters via grid search:

Algorithm 1 SAE-Guided Two-Stage Expert Pruning

Require: Experts E, SAE encoder, load prune ratio R, anchor set $\mathcal{D}_{\text{anchor}}$, cluster prune ratio P_{cluster}

Ensure: Set of expert indices to remove E_{remove}

Stage 0: Preprocessing
for each expert $e \in E$ **do**
  Compute routing frequency $|T_e|$ and sparse SAE profile z_e (Top-100 NMD features)
end for
Project $\{z_e\}$ to 10D via PCA; cluster with HDBSCAN
Stage 1: Load-Based Filtering
$E_{\text{remove}} \leftarrow$ bottom-R experts by $|T_e|$
Stage 2: Semantic Refinement
Compute anchor z_{anchor} as mean SAE activation over $\mathcal{D}_{\text{anchor}}$
Rank clusters by ascending similarity to z_{anchor}
for each cluster in bottom-P_{cluster} fraction **do**
  Retain only the expert closest to z_{anchor}; add the rest to E_{remove}
end for
return E_{remove}

- **Load prune ratio R:** the fraction of experts removed based on lowest routing frequency $|T_e|$ in Stage 1. We evaluate $R \in \{0.1, 0.2, \ldots, 0.9\}$.

- **Cluster prune ratio P_{cluster}:** the fraction of lowest-similarity clusters subjected to representative-only retention in Stage 2. We evaluate $P_{\text{cluster}} \in \{0.1, 0.2, \ldots, 0.9\}$.

For anchor computation, we sample $N_{\text{anchor}} = 5$ instances per dataset, except for MMLU where we use a larger sample due to its multi-domain structure.

Baseline. We compare against a load-based pruning baseline that iteratively removes experts with the lowest routing frequency $|T_e|$, evaluating performance across increasing pruning ratios at matched expert removal budgets.

Evaluation Metrics. Model performance after pruning is evaluated using task-specific metrics: exact match for GSM8K and FinQA, accuracy for PubMedQA and MMLU. We report average performance across the four benchmarks to assess overall pruning effectiveness.

5.3 Main Results

We evaluate our method against the baseline across four benchmarks: FinQA (4a), GSM8K (4b), MMLU (5a), and PubMedQA (5b). Performance is measured using number_match for FinQA, exact_match for GSM8K, and accuracy for MMLU and PubMedQA. In all figures, the X-axis represents the pruning ratio across MoE layers, while the Y-axis shows the corresponding task performance metric.

5.4 Analysis

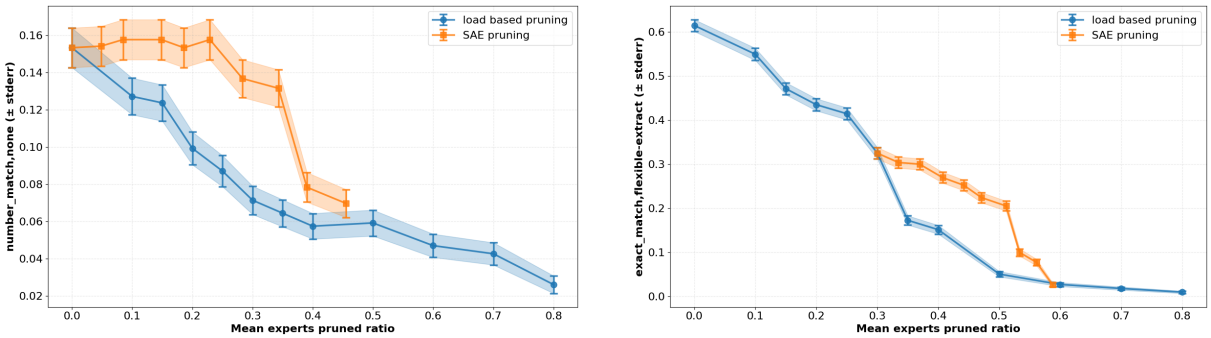
The results show a clear difference between domain-specific and broad-domain benchmarks. On FinQA (4a) and GSM8K (4b), our SAE-guided pruning method significantly outperforms the load-based baseline, preserving more than twice the performance at moderate pruning ratios (~ 0.3 for FinQA and ~ 0.5 for GSM8K). This suggests that SAE representations successfully capture experts associated with mathematical and financial reasoning, allowing the method to retain semantically important experts even when their routing frequency is low.

To further validate this behavior, we analyze expert specialization through SAE feature representations on layer 20 using the FinQA dataset. The interpretation of the corresponding SAE clusters for selected experts is summarized in Table 1. Among the first 20% of experts selected for removal in the FinQA-oriented pruning setting, experts 0 and 11 are primarily associated with conversational and narrative semantics, whereas experts 5 and 8 exhibit numerical and factual semantics and are preserved until more than 80% of clusters are pruned.

This qualitative analysis supports the hypothesis that the proposed method removes experts associated with non-relevant semantics while preserving those critical for financial and mathematical reasoning tasks.

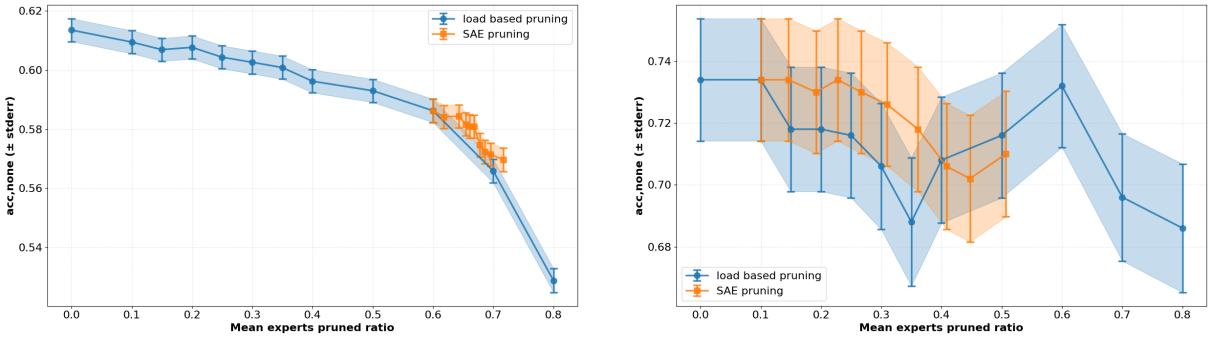
In contrast, PubMedQA (5b) and MMLU (5a) show only marginal improvements over the baseline. For PubMedQA, this limitation is likely caused by the SAE training corpus, `DKYoon/SlimPajama-6B`, which contains relatively little biomedical content. As a result, the SAE does not learn sufficiently specialized medical semantic features, reducing the effectiveness of semantic expert selection.

For MMLU, the limitation stems from the benchmark design itself. MMLU covers a broad range



(a) FinQA (b) GSM8K

Figure 4: Comparison of pruning results on domain-specific reasoning benchmarks.



(a) MMLU (b) PubMedQA

Figure 5: Comparison of pruning results on broad-domain benchmarks.

Cluster	Semantic Interpretation
0	Conversational and personal discourse: narratives, dialogue, advice, and socially oriented text with strong personal voice.
11	Narrative and reflective discourse: personal experiences, emotions, dialogue, and retrospection.
5	Numerical and quantitative content: dates, measurements, identifiers, and technical numerical information.
8	Factual and report-style discourse: events, institutions, chronology, and technical factual content.

Table 1: Interpretation of SAE clusters for selected experts in layer 20 (FinQA analysis).

of domains, including STEM, humanities, and social sciences. Since many different expert specializations are required simultaneously, our method cannot reliably identify experts with irrelevant semantics, making semantic pruning less effective.

6 Conclusion

We introduced a framework for representing MoE experts in the space of SAE features, proposing the Normalized Mean Difference metric to quantify feature-expert association. Compact SAE feature subsets faithfully recover routing decisions, align with supervised importance estimates, and reveal semantically coherent geometric structure among experts. Building on this structure, we proposed a two-stage pruning strategy combining load-based filtering with cluster-level semantic refinement, which substantially outperforms load-based pruning on domain-specific benchmarks while offering modest gains on broad-domain tasks due to corpus coverage limitations. These results establish SAE feature representations as an informative and practical basis for analyzing and compressing MoE models.

7 Limitations

The effectiveness of the proposed method depends on the coverage of the SAE training corpus: domains underrepresented in the training data, such as biomedical text, yield less discriminative SAE features and consequently reduce the benefit of semantic pruning. All experiments are conducted on a single model architecture, `Qwen1.5-MoE-A2.7B`, and it remains unclear how well the findings generalize to models with different expert counts or routing mechanisms. Additionally, the method is less effective on broad-domain benchmarks such as MMLU, where the simultaneous coverage of diverse subject areas prevents reliable identification of semantically irrelevant experts, limiting the practical benefit of semantic pruning in such settings.

References

Trenton Bricken, Adly Templeton, Joshua Batson, and 1 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. 2025. Learning multi-level features with Matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*.

Tianyu Chen, Shaohan Huang, Yuan Xie, and 1 others. 2022. Task-specific expert pruning for sparse mixture-of-experts. *arXiv preprint arXiv:2206.00277*.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Shengyun Ma, Iqbal Hussain, Felix Stoppel, Alexander Kaufmann, Sameena Shah, 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*.

Karl Cobbe and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

DeepSeek-AI. 2024b. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

DKYoon. 2023. Slimpajama-6b dataset. https://huggingface.co/datasets/DKYoon/SlimPajama-6B.

Nelson Elhage and 1 others. 2022. Toy models of superposition. *Transformer Circuits Thread*.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.

Leo Gao, Tom Dupré la Tour, Henk Tillman, and 1 others. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Dan Hendrycks and 1 others. 2021. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Robert Huben, Hoagy Cunningham, Logan Riggs, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Mike Lasby, Ivan Lazarevich, Nish Sinnadurai, Sean Lie, Yani Ioannou, and Vithursan Thangarasa. 2025. Reap the experts: Why pruning prevails for one-shot moe compression. *arXiv preprint arXiv:2510.13999*.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, and 1 others. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Lu and 1 others. 2024. Not all experts are equal: Revisiting mixture-of-experts in large language models. *arXiv preprint arXiv:2402.14800*.

Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Man Luo, Sungduk Yu, Chendi Xue, and Vasudev Lal. 2025. Probing semantic routing in large mixture-of-expert models. *arXiv preprint arXiv:2502.10928*.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Wei Shi, Sihang Li, Tao Liang, Mingyang Wan, Guojun Ma, Xiang Wang, and Xiangnan He. 2025. Route sparse autoencoder to interpret large language models. *arXiv preprint arXiv:2503.08200*.

Qwen Team. 2024. Qwen1.5-moe: Matching 7b model performance with 1/3 activated parameters. *arXiv preprint arXiv:2402.15179*.

Yan Wang, Yitao Xu, Nanhan Shen, Jinyan Su, Jimin Huang, and Zining Zhu. 2026. The illusion of specialization: Unveiling the domain-invariant "standing committee" in mixture-of-experts models. *arXiv preprint*.

Fuzhao Xue, Xiaoqing Zheng, Yang Cheng, and 1 others. 2024. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*.

Xinyuan Yan, Shusen Liu, Kowshik Thopalli, and Bei Wang. 2025. Visual exploration of feature relationships in sparse autoencoders with curated concepts. *Preprint*, arXiv:2511.06048.

Yanqi Zhou and 1 others. 2022. Mixture-of-experts with expert choice routing. *arXiv preprint arXiv:2202.09368*.