Investigating Expert Semantic Specialization in Mixture-of-Expert Models

Luyu Wang [1] **Shuhui Wang** [2]

Abstract

Mixture-of-Experts (MoE) models achieve exceptional scalability through selective routing, yet our understanding of what routing actually learns remains limited. In particular, it is unclear to what extent routing induces meaningful expert semantic specialization. Sparse Autoencoders (SAEs) provide a way to extract interpretable semantic feature spaces from model representations. Building upon this capability, we introduce the Semantic Specialization Index (SSI), a quantitative metric designed to measure the degree of expert semantic specialization. Using SSI, we systematically quantify expert semantic specialization in MoE models. We further investigate the relationship between semantic specialization and language modeling performance, and find that specialization exhibits a non-monotonic trajectory during training and does not increase indefinitely with model quality, suggesting the existence of an optimal specialization regime. These findings provide a quantitative foundation for understanding expert organization in sparse models and open new opportunities for specialization-aware optimization of MoE models.

1. Introduction

Mixture-of-Experts (MoE) architectures scale model capacity via conditional computation, routing each token to a sparse subnet of expert feed-forward networks while keeping per-token computation roughly constant (Shazeer et al., 2017). Because each experts is selectively activated, a large body of work investigates whether MoE models exhibit specialization among experts. For example, Gururangan et al. study specialization at the domain level, introducing DEMix layers in which experts are explicitly specialized to distinct domains (2022). Specialization has also been studied at the token and vocabulary level, where particular vocabulary items are disproportionately routed to specific experts (Muennighoff et al., 2025). Recent interpretability studies further identify distinct functional roles among experts, including general, domain-specific, and driver experts that strongly influence model outputs (Hu et al., 2026). Collectively, these studies provide strong evidence that MoE routing is not uniform and that experts can exhibit structured behaviors across tokens, tasks, and domains.

Despite the current progress, a key gap remains between the outstanding performance of MoE and lack of understanding on the behavior of the experts. Existing specialization analyses often operate on non-semantic levels, making it difficult to answer a more direct question: *to what extent are experts specialized with respective to coherent semantic dimensions?* Without a clear quantitative metric, expert specialization cannot be systematically evaluated, making it difficult to establish a link between model performance and semantic specialization. Addressing this requires a semantic basis that is more interpretable than row hidden dimensions, since dense representations in neural networks are frequently polysemantic, where single neurons respond to mixtures of unrelated concepts (Olah et al., 2020). Recent mechanistic interpretability work shows that sparse autoencoders (SAEs) can recover more monosemantic, human-interpretable features from language model activations, offering an explicit representation in which semantic factors are disentangled and measurable (Huben et al., 2024). This motivates measuring expert specialization not only via routing statistics or token identities, but expert-semantic alignment in an interpretable feature space.

In this work, we propose the Semantic Specialization Index (SSI), a quantitative metric that measures expert specialization at the semantic level. Given SAE-based interpretations, we first construct an Expert-Semantic Association Matrix (E-SAM) that presents a conditional distribution of expert-semantic association. From E-SAM, we derive two complementary factors, *i.e.*, Purity, capturing whether each semantic dimension is concentrated within a small subset of experts, and Spread, capturing how evenly semantic responsibilities are distributed across experts. SSI combines these factors and enables principled comparison across models

[1]Georgia Institute of Technology, Atlanta, GA, USA [2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Correspondence to: Shuhui Wang <wangshuhui@ict.ac.cn>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

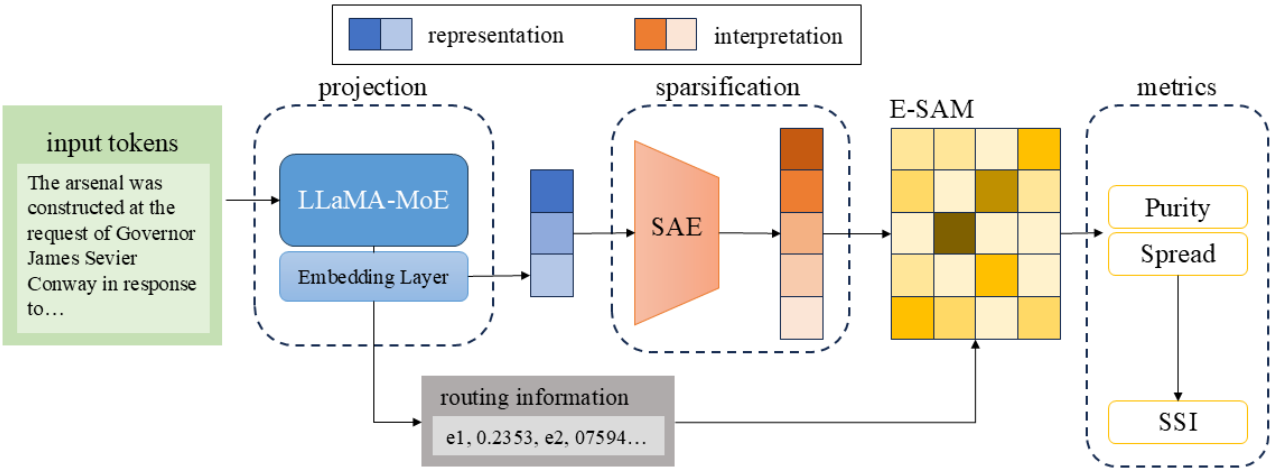

Figure 1. Overview of the proposed Semantic Specialization Index (SSI) pipeline.

with different expert counts and routing K through normalization.

Using SSI, we study the dynamics of expert semantic specialization and its relationship to language modeling performance. Across LLaMA-MoE variants and finetuning checkpoints on WikiText-103, we compare SSI with perplexity and observe a consistent, non-monotonic specialization trajectory: specialization increases during early finetuning, peaks at intermediate checkpoints, and partially declines even as perplexity continues to improve. Overall, the contributions of our work are as follows:

- We propose a semantic-level framework for interpreting expert specialization by explicitly measuring expert-semantic alignment in an SAE derived feature space (Figure 1).
- We provide a decomposable specialization metric (SSI) that supports systematic training-dynamics analysis and cross-model comparison.
- We reveal a non-monotonic relationship between SSI and language modeling performance using the LLaMA-MoE-Base model family and the WikiText-103 dataset, showing that semantic specialization increases during early finetuning and peaks at intermediate stages.

2. Related Works

Mixture-of-Experts for Conditional Computation. Mixture-of-Experts (MoE) is a classic conditional-computation paradigm in which a router selects a sparse subset of expert networks for each token, enabling substantially larger total model capacity while keeping per-token computation relatively bounded (Jacobs et al., 1991; Jordan & Jacobs, 1993). Modern sparsely gated MoE layers revived this paradigm at scale, and early empirical observation suggested that different experts may specialize in distinct semantic patterns (Shazeer et al., 2017; Riquelme et al., 2021). Since then, MoE has been widely adopted and extended across model architectures, routing strategies, and training recipes(Fedus et al., 2022; Lepikhin et al., 2021; Jiang et al., 2024; Dai et al., 2024).

Expert Specialization Analyses. A growing body of work studies expert behavior on vocabulary, domain, or functional levels. OLMoE reports substantial routing specialization, showing that experts differ in both domain preference and vocabulary usage, with later layers exhibiting stronger vocabulary specialization (Muennighoff et al., 2025). MoE Lens further shows that specialization is often concentrated rather than evenly distributed across experts even in models with apparent specialization (Chaudhari et al., 2025). Complementary work distinguishes different functional roles among experts, separating domain-oriented experts from driver experts that have disproportionate causal influence on model outputs (Hu et al., 2026). Other studies examine routing from a structural perspective and show that routing decisions are influenced not only by linguistic content but also by positional information, highlighting the multifaceted nature of expert selection and underscoring the need for explicit semantic-level quantification (Bershatsky & Oseledets, 2025).

More recent work probes routing in a semantic perspective as well, suggesting that routing can vary with semantic context and such semantic sensitivity depends on model scale and layer depths (Olson et al., 2025). However, this line of work mainly focuses on existence proofs of semantic sensitivity rather than what semantic concepts individual expert represents, nor does it provide a general quantitative framework for measuring the degree of semantic specialization across models. As a result, although prior studies

show that routing is non-uniform and meaning-sensitive, we still lack a model-agnostic and interpretable metric that measures specialization strengths and enables systematic study of its relationship to language modeling performance.

Sparse Feature Spaces for Semantic Measurement. Mechanistic interpretability research has emphasized that dense neural representations are often polysemantic: a single neuron may respond to multiple semantically distinct patterns (Olah et al., 2020). One influential explanation is superposition, where models represent more features than available dimensions by encoding them in an overcomplete set of directions rather than in individual neurons (Elhage et al., 2022). SAEs have therefore emerged as a widely used tool for recovering sparse feature dictionaries that reconstruct activations using sparse linear combinations, often yielding features that are more interpretable and more nearly monosemantic than original neuron basis.

Inspired by this perspective, recent work has attempted to build interpretability more directly into MoE architectures. For example, MoE-X reformulates MoE layers as sparse, wide MLPs and introduces sparsity-aware routing to favor experts with sparser activations, thereby aligning routing with interpretability objectives (Yang et al., 2025). In parallel, representation interpretability work has proposed explicit semantic spaces and quantitative scores that measure how much task-relevant semantic information is preserved under projection into an interpretable basis (Shen et al., 2025). Although these approaches arise from different communities, they share a common principle: interpretability can be evaluated by using SAE-derived feature spaces as a semantic basis rather than relying on raw hidden dimensions or token identities.

3. Quantifying Expert Semantic Specialization

In this section, we propose SSI to quantify expert semantic specialization in MoE models. Given a set of input tokens indexed by $n = 1, ..., N$, we obtain token representations $\mathbf{h}_n \in \mathbb{R}^D$ by looking up the output token IDs of a pre-trained MoE language model f in its token embedding matrix W. We map $\mathbf{h}_n$ to a sparse interpretable space $\mathbb{R}^S$ using a SAE (Section 3.1). We then construct the Expert-Semantic Association Matrix (E-SAM) by correlating the interpretations $\mathbf{z}_n \in \mathbb{R}^S$ with per-token routing information, capturing the association length between experts and semantic dimensions (Section 3.2). From the E-SAM, we derive two complementary metrics: Purity, measuring how concentrated each expert is within a specific expert group, and Spread, measuring how evenly semantic responsibilities are distributed across experts. SSI is defined a joint function of Purity and Spread (Section 3.3).

3.1. Interpreting Representations via Sparse Decomposition

To measure semantic specialization, we need to identify which semantic concepts each expert preferentially processes. However, directly reading semantics from dense representation is difficult due to polysemanticity. Therefore, we use an SAE to decompose representations into sparse interpretable features, where each active dimension corresponds to a more coherent semantic concept.

To prevent routing information from contaminating the representation, we obtain the context-free representation by looking up the output token IDs of f in the embedding matrix W. For each input token t_n, we obtain the corresponding representation $\mathbf{h}_n$:

$$\mathbf{h}_n = W[f(t_n)]. \tag{1}$$

The representation are then used to train a Top-K SAE $\phi : \mathbb{R}^D \to \mathbb{R}^S$. We obtain the final representation $\mathbf{z}_n$:

$$\mathbf{z}_n = \phi(\mathbf{h}_n). \tag{2}$$

3.2. Expert-Semantic Association Matrix

Let E denote the set of experts and L denote the set of semantic dimensions in the SAE latent space. We compute expert-semantic associations by aggregating routing-weighted semantic activations over N samples. To avoid domination by samples with large activation magnitude, we first normalize the semantic activation vector:

$$\tilde{\mathbf{z}}_n = \frac{\mathbf{z}_n}{||\mathbf{z}_n||_1}. \tag{3}$$

Let $w_{e,n}$ denote the routing weight assigned to expert $e \in E$ for sample n. We then compute the aggregated association scores:

$$A_{e,l} = \sum_N w_{e,n} \tilde{z}_{n,l}. \tag{4}$$

We define E-SAM $M \in \mathbb{R}^{|E| \times |L|}$ as the latent-wise normalization of the aggregated association scores:

$$M_{e,l} = \frac{A_{e,l}}{\sum_{e' \in E} A_{e',l} + \epsilon}. \tag{5}$$

where ϵ is a small constant for numerical stability. E-SAM is thus the conditional semantic distribution $P(e \mid l)$, capturing how each expert is concentrated over the interpretation space.

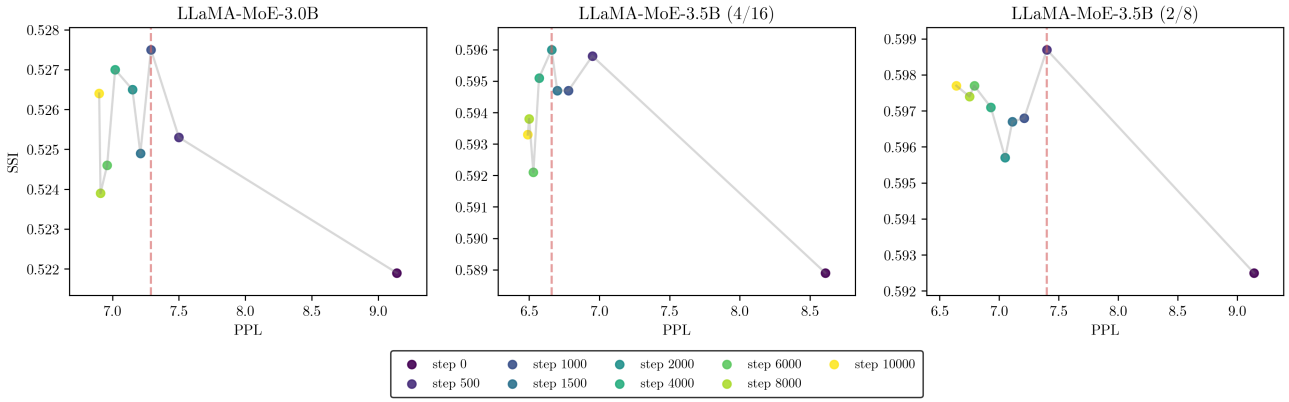

Figure 2. The relationship between SSI and PPL on WikiText-103 in LLaMA-MoE-Base family.

3.3. Semantic Specialization Index

Intuitively, semantic specialization exhibits two distinct structural aspects: whether each dimension is concentrated within a small subset of experts, and whether the overall allocation of semantic responsibilities is even across experts. Thus we derive two complementary factors from the E-SAM matrix: Purity and Spread, to characterize different aspects of expert semantic specialization.

Purity. Purity measures whether each semantic dimension is handled by a stable and concentrated expert subset. Since hard-routing MoE models use Top-K as the routing mechanism, evaluating semantic specialization using the winner expert alone may underestimate the effective participation. Thus, we incorporate the Top-K routing constraint into our purity measurement, using the same K as in the pre-trained MoE language model f.

For each semantic dimension l, we compute the sum of Top-K normalized association score in E-SAM:

$$c_l = \sum_{e \in Top-K(M_l)} M_{l,e}. \tag{6}$$

This cumulative score lies in the range $[\frac{K}{|E|}, 1]$, where the lower bound $\frac{K}{|E|}$ is achieved when association scores are uniformly distributed across experts, indicating no semantic specialization. Since this lower bound depends on both K and $|E|$, Purity across models with different architectures becomes incomparable. We therefore apply a baseline correction that maps c_l to [0, 1], where 0 corresponds to no specialization (uniform association) and 1 corresponds to perfect concentration:

$$b = \frac{K}{|E|}, \ \mathcal{P}_l = \frac{c_l - b}{1 - b}. \tag{7}$$

The overall Purity is computed by averaging across the corrected purity of semantic dimensions:

$$\mathcal{P} = \frac{1}{|L|} \sum_{|L|} \mathcal{P}_l. \tag{8}$$

Spread. Inspired by the auxiliary load-balancing objective introduced in Shazeer et al. (2017), we define Spread to measure how semantic responsibilities are distributed across experts at the global level. We first compute the marginal expert responsibility by averaging over semantic dimensions:

$$r_e = \frac{1}{|L|} \sum_{l \in L} M_{e,l}. \tag{9}$$

We then define Spread as the normalized Shannon entropy of this distribution:

$$\mathcal{S} = \frac{-\sum_{e \in E} r_e \log(r_e + \epsilon)}{\log |E|}. \tag{10}$$

Spread takes values in $[0, 1]$, where higher values indicate more uniform distribution of semantic responsibilities across experts.

SSI. To jointly capture both factors of Purity and Spread, we define the SSI as:

$$SSI = \sqrt{\mathcal{P} \times \mathcal{S}} \in [0, 1]. \tag{11}$$

When both Purity and Spread reach 1, SSI reaches 1, indicating a perfect case for semantic specialization among experts. If either Purity or Spread approaches zero, SSI also approaches zero, reflecting the absence of meaningful semantic specialization.

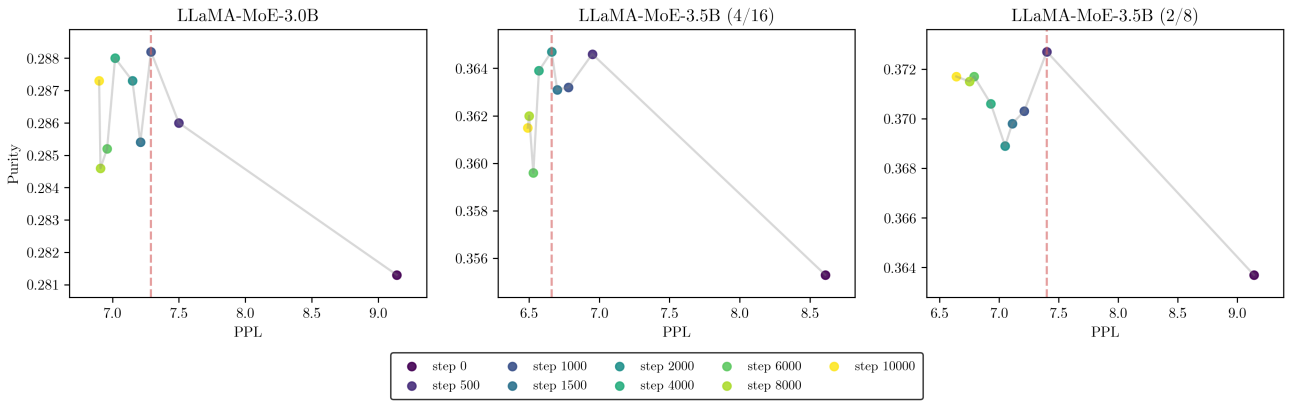

Figure 3. The relationship between Purity and PPL on WikiText-103 in LLaMA-MoE-Base family.

4. Expert Semantic Specialization and Language Modeling Performance

First, we provide the experiment settings in Section 4.1. Then, we investigate the relationship between expert semantic specialization and language modeling performance by comparing SSI and model Perplexity (PPL) across training checkpoints (Section 4.2). We further analyze this relationship by comparing PPL with the components of SSI, namely Purity and Spread (Section 4.3). Finally, we discuss the influence of selecting different representations to compute SSI, and justify why using the embedded model output is the optimal option (Section 4.4).

4.1. Experiment Settings

Model. We quantify the expert semantic specialization quality of three models from the LLaMA-MoE-Base family: LLaMA-MoE-3.0B, LLaMA-MoE-3.5B (4/16), and LLaMA-MoE-3.5B (2/8), where the numbers in parentheses denote the top-k selected experts and the total number of experts respectively (Zhu et al., 2024).

Routing Information. We use the routing information of the last-layer gating network, as it is most directly related to the model output and consistent with prior work showing that later layers exhibit greater interpretability than earlier and middle layers (Zhu et al., 2024; Riquelme et al., 2021).

Datasets. We use WikiText-103 for both finetuning and evaluation (Merity et al., 2017). To study the relationship between PPL and SSI, we obtain checkpoints with varying PPL by finetuning each pretrained model on 80,000 training samples and save checkpoints at 500, 1000, 1500, 2000, 4000, 6000, 8000, and 10,000 training steps. We then evaluate SSI and PPL on the entire validation set for each checkpoint. Datasets are processed by concatenating all samples and chunking into fixed-length blocks (concat-and-chunk), ensuring no padding tokens are introduced during training or inference. The last incomplete block is discarded.

SAE. We use a Top-K SAE following the implementation released by OpenAI (2024). Empirically, we set $K = 50$ and use a latent space of twice the representation size as fixed hyperparameters across all models.

4.2. SSI-Performance Relationship in Language Modeling

We now investigate the relationship between expert semantic specialization and language modeling performance by analyzing the evolution of SSI and PPL across training checkpoints. Figure 2 presents the SSI-PPL trajectories for three models in the LLaMA-MoE-Base family. Each point corresponds to a training checkpoint, and the gray line implicitly traces the temporal evolution from step 0 to step 10,000. The vertical dashed line marks the checkpoint with an SSI peak in all models.

Across all three models, we observe a consistent pattern. SSI generally improves when the model is finetuned on the target dataset. During the early and mid phases of finetuning, as PPL decreases rapidly, SSI increases correspondingly. However, after reaching a peak at intermediate checkpoints, SSI declines to some extent, even though PPL continues to improve in later stages. This non-monotonic behavior indicates that expert semantic specialization does not increase indefinitely with performance. One plausible explanation is that high initial loss gradients drive active routing differentiation, which promotes specialization. Once the model enters a lower-gradient regime, routing stabilizes rather than further differentiates, and SSI declines accordingly. We leave a mechanistic investigation of this phenomenon to future work. Moreover, the absolute SSI values across all models consistently fall within the range of approximately 0.5–0.6. This moderate scale suggests that expert semantic specialization is present but not strongly polarized, which aligns with prior empirical observations that MoE models exhibit

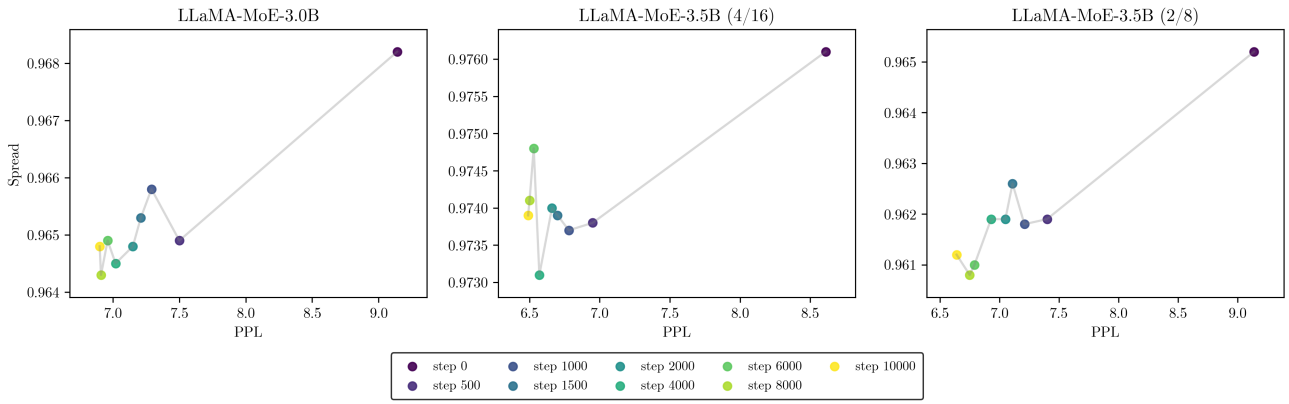

Figure 4. The relationship between Spread and PPL on WikiText-103 in LLaMA-MoE-Base family.

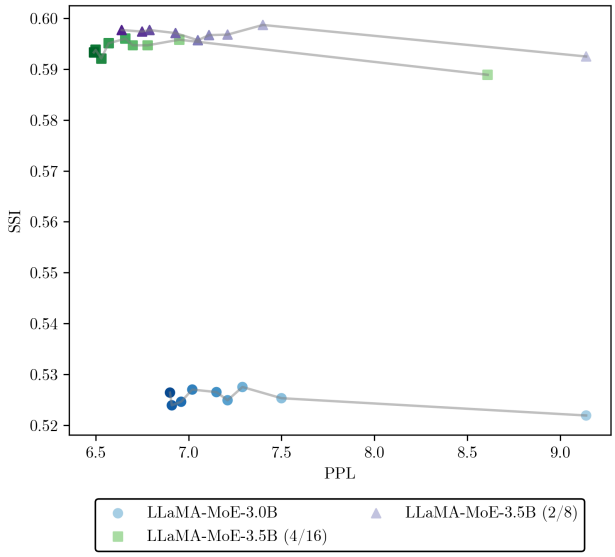

Figure 5. SSI-PPL relationship across the three models.

partial rather than fully disentangled expert specialization (Chaudhari et al., 2025).

Across models, we observe a clear separation in overall specialization levels (Figure 5). The LLaMA-MoE-3.0B model consistently attains substantially lower SSI values than both 3.5B variants throughout training, remaining centered around 0.52-0.53, whereas the 3.5B configurations maintain SSI values near 0.59-0.60. This indicates that expert semantic specialization in the 3.0B model is generally weaker than in the two 3.5B models.

Between the two 3.5B variants, the overall specialization levels are similar, but the 2/8 configuration generally achieves slightly higher SSI values than the 4/16 model. Given that the 2/8 model consistently yields higher PPL compared to 4/16 configuration, this observation suggests that architectural factors beyond the quality of expert semantic specialization also contribute to language modeling performance.

4.3. Factor Analysis

To better understand the dynamics underlying the SSI-performance relationship, we take a closer look into SSI's two constituent components, Purity and Spread, and analyze their individual correlations with perplexity across training checkpoints.

Across all three models, Purity exhibits a trend that is closely consistent with the SSI-PPL relationship (Figure 3): Purity increases as perplexity decreases during early finetuning, and starts to decline even as PPL continues to improve after reaching a peak at intermediate checkpoints. This suggests that the trend observed in the SSI-PPL relationship is largely driven by changes in Purity. Concretely, early finetuning appears to reinforce consistent expert preferences for specific semantic dimensions, tightening the association between semantic content and a stable subset of top-ranked experts.

In contrast, Spread decreases during the early phase of finetuning and then fluctuates within a relatively narrow range at later checkpoints (Figure 4). Unlike Purity, Spread does not show a clear peak aligned with performance improvements. This behavior suggests that global expert usage balance adjusts rapidly at the beginning of finetuning but does not continue to systematically evolve alongside perplexity improvements. In particular, later-stage performance gains do not appear to be strongly associated with further redistribution of expert activation mass.

4.4. Ablation of Representation Selection

We next investigate how representation selection influences SSI computation. We evaluate four candidate representations as SAE inputs: embedded model outputs (Section 3.1), last-layer gating inputs, last-layer inputs, and last-layer expert outputs. For each representation, we train an SAE,

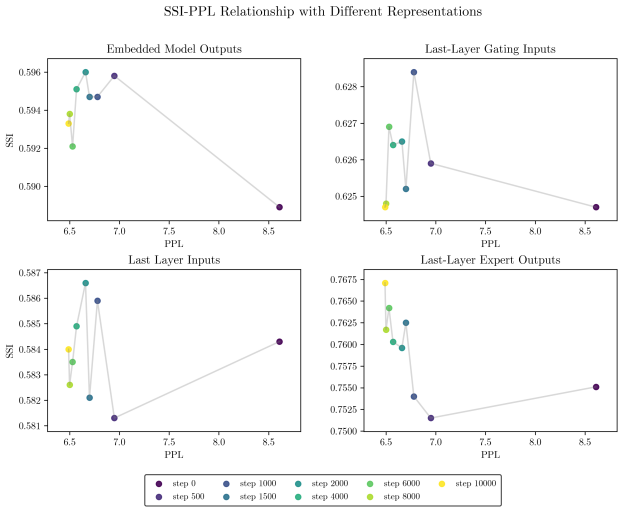

Figure 6. The relationship between SSI and PPL with different representation choices in 4/16 LLaMA-MoE-3.5B model.

Table 1. The comparison of SAE Losses for the last training step among different representation choices. The representations are selected in 4/16 LLaMA-MoE-3.5B checkpoint 10,000.

Representations	**SAE Loss (Final Step)**
Embedded Model Outputs	0.0519
Last-Layer Gating Inputs	0.4445
Last Layer Inputs	0.4347
Last-Layer Expert Outputs	0.0789

compute SSI, and analyze its relationship with PPL.

As shown in Figure 6, the SSI-PPL trajectories vary across different representation choices. This variation reflects the fact that SSI quantifies specialization within the semantic space induced by the underlying representation. Different representations emphasize distinct structural properties of the model, leading to different observed specialization dynamics. This motivates a principled selection of representation to ensure that SSI captures routing-induced semantic structure.

We first examine SAE reconstruction quality across different representation inputs. As shown in Table 1, SAEs trained on last-layer gating inputs and last-layer inputs exhibit substantially higher reconstruction loss than those trained on embedded outputs and expert outputs. We note that SAE hyperparameters are fixed across all representations for controlled comparison; however, the nearly order-of magnitude gap in reconstruction loss suggests a fundamental difference in sparse factorizability between representations, rather than a marginal capacity mismatch. Since SSI is computed over SAE-derived semantic dimensions, poor reconstruction quality undermines the interpretability and validity of the resulting specialization patterns.

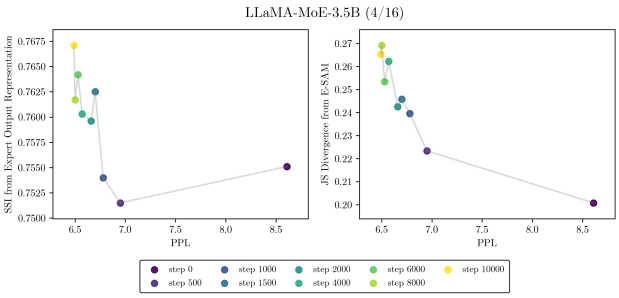

Figure 7. Comparison between SSI-PPL relationship derived from expert output representations and the JS Divergence-PPL relationship based on E-SAM. The model used is 4/16 LLaMA-MoE-3.5B.

Although the expert outputs achieve relatively low SAE loss, their specialization trend behaves fundamentally differently from embedded outputs. Specifically, the expert-output-based SSI exhibits an overall upward trajectory with fluctuations, in contrast to the non-monotonic pattern observed with embedded outputs. Since expert outputs are obtained after routing and aggregation, they inherently encode routing-induced variation alongside semantic content. This suggests that expert-output SSI may reflect inter-expert representational divergence rather than the semantic structure induced by routing decisions.

To verify this hypothesis, we compute the mean pairwise JS divergence across expert-wise semantic distributions, where P_i and P_j denote the semantic distribution of expert i and expert j respectively, obtained by normalizing each expert's column in E-SAM over semantic dimensions:

$$\text{Global Diversity} = \frac{2}{|E|(|E|-1)} \sum_{i<j} JS(P_i \parallel P_j)\,. \quad (12)$$

As shown in Figure 7, the expert-output-based SSI closely tracks this diversity measure, supporting the hypothesis that it captures inter-expert representational divergence rather than the semantic structure induced by routing decisions. This suggests that expert outputs encode not only semantic content but also functional differences in how experts transform input representations, causing the resulting SSI to reflect inter-expert output divergence rather than the semantic structure induced by routing decisions. For completeness, we also conducted the same representation ablation experiments on the other two models in the LLaMA-MoE-Base family, with results reported in Appendix A.

We adopt embedded model outputs for an additional reason. In contextual representations, high-frequency tokens such as 'the' or 'a' produce diverse context-dependent vectors, potentially disproportionately occupying SAE latent capacity with nuisance variance rather than meaningful semantic content. Since embedded outputs are obtained by looking

up token IDs in the embedding matrix, identical tokens are always mapped to identical vectors regardless of context, reducing this variance and allowing SAE latent capacity to be allocated toward semantically meaningful dimensions.

5. Discussion

The experiments presented in this work offer a systematic characterization of expert semantic specialization in MoE language models through the proposed SSI metric. We discuss the limitations of the current study and directions for future work in this section.

Scope of Evaluation. Our analysis is conducted on three models from the LLaMA-MoE-Base family, finetuned on a single homogeneous corpus, WikiText-103. While the consistency of observed patterns across all three models is encouraging, generalization to other MoE architectures such as Mixtral or DeepSeek-MoE, which differ in routing mechanisms and expert granularity, remains to be verified. Similarly, whether the SSI-PPL dynamics observed under finetuning extend to more diverse, domain-shifted datasets is an open question. We view this as a natural direction for future work, and the consistent cross-model patterns observed here provide a reasonable basis for expecting broader generalizability.

Single-Layer Analysis. SSI is currently computed using only the last-layer gating network, motivated by prior work suggesting that later layers exhibit greater semantic interpretability. This choice provides a focused and well-motivated measurement point, but does not capture how semantic specialization evolves across layers. A layer-wise extension of SSI could offer a more complete picture of specialization dynamics within the network.

Cross-Model Comparability. SSI is designed to be architecture-invariant through baseline correction of Purity and entropy normalization of Spread, ensuring values lie in $[0, 1]$ regardless of K and $|E|$. However, while the metric is mathematically normalized, we do not empirically verify that SSI values are semantically comparable across architectures. Establishing a principled empirical validation of cross-model SSI comparability remains an open challenge, and cross-model comparisons in this work should be interpreted with this caveat in mind.

Beyond the limitations outlined above, our findings suggest promising directions for future research.

SSI-Informed Training Objectives. The non-monotonic SSI trajectory observed across all models suggests the existence of an optimal specialization regime: a "sweet spot" at which expert semantic specialization is maximized. This raise the possibility of explicitly optimizing toward this regime, for instance by incorporating SSI-derived signals as auxiliary losses during training or soft constraints on the routing mechanism. Such an approach could potentially improve model interpretability and performance simultaneously and effectively. We leave the design and evaluation of SSI-informed training objectives to future work.

Extension to Other Modalities. While this work focuses on language modeling, the SSI framework is not inherently language-specific, as it requires only a sparse interpretable representation of inputs and per-token routing information, both of which are available in MoE models across modalities. Extending SSI to vision or multimodal MoE architectures could reveal whether this specialization dynamics observed here reflect a universal property of MoE learning or are specific to the linguistic domain.

6. Conclusion

In this work, we study expert specialization in Mixture-of-Experts language models from a semantic perspective. We introduce a semantic-level framework that quantifies how experts align with coherent semantic dimensions, and propose the SSI metric that captures both concentration and distributional balance of semantic responsibility across experts. Using SSI, we identify a consistent non-monotonic relationship between semantic specialization and language modeling perplexity, which suggests that expert semantic differentiation does not grow indefinitely with performance. Overall, our results provide a quantitative and interpretable lens for studying expert semantic organization, and suggest that specialization-oriented regularization may be most effective during the early phases of MoE training, while later optimization stages may benefit from shifting toward alternative objectives beyond continual semantic differentiation. Building on these findings, we plan to further investigate the underlying mechanisms behind the non-monotonic trend of semantic specialization, aiming to better understand the dynamics of expert organization in MoE models. We hope this work can serve as a foundation for future research on specialization-aware optimization and semantic analysis in MoE systems.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bershatsky, D. and Oseledets, I. On the spatial structure of mixture-of-experts in transformers. In *Sparsity in LLMs (SLLM): Deep Dive into Mixture of Experts, Quan-*

tization, Hardware, and Inference, 2025. URL https://openreview.net/forum?id=Ntxdj8OmOv.

Chaudhari, M., Gulati, I., Hundia, N., Karra, P., and Raval, S. Moe lens - an expert is all you need. In *Sparsity in LLMs (SLLM): Deep Dive into Mixture of Experts, Quantization, Hardware, and Inference*, 2025. URL https://openreview.net/forum?id=GS4WXncwSF.

Dai, D., Deng, C., Zhao, C., Xu, R., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., Xie, Z., Li, Y., Huang, P., Luo, F., Ruan, C., Sui, Z., and Liang, W. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1280–1297, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.70. URL https://aclanthology.org/2024.acl-long.70/.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Carter, S., Mineault, P., Tamkin, A., Christiano, P., and Hubinger, E. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23(1), 2022.

Gururangan, S., Lewis, M., Holtzman, A., Smith, N. A., and Zettlemoyer, L. DEMix layers: Disentangling domains for modular language modeling. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5557–5576, Seattle, United States, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.407. URL https://aclanthology.org/2022.naacl-main.407/.

Hu, G., Li, M., Peng, Q., Hu, L., Xu, B., and Cai, R. What gets activated: Uncovering domain and driver experts in moe language models, 2026. URL https://arxiv.org/abs/2601.10159.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Jordan, M. and Jacobs, R. Hierarchical mixtures of experts and the em algorithm. In *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*, volume 2, pp. 1339–1344 vol.2, 1993. doi: 10.1109/IJCNN.1993.716791.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Muennighoff, N., Soldaini, L., Groeneveld, D., Lo, K., Morrison, J., Min, S., Shi, W., Walsh, E. P., Tafjord, O., Lambert, N., Gu, Y., Arora, S., Bhagia, A., Schwenk, D., Wadden, D., Wettig, A., Hui, B., Dettmers, T., Kiela, D., Farhadi, A., Smith, N. A., Koh, P. W., Singh, A., and Hajishirzi, H. OLMoe: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xXTkbTBmqq.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 5(3), 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in/.

Olson, M. L., Ratzlaff, N., Hinck, M., Luo, M., Yu, S., Xue, C., and Lal, V. Probing semantic routing in large mixture-of-expert models. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18263–18278, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.991. URL https://aclanthology.org/2025.findings-emnlp.991/.

OpenAI. sparse_autoencoder, 2024. URL https://github.com/openai/sparse_autoencoder. GitHub repository, accessed 2026-03-08.

Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Pinto, A. S., Keysers, D., and Houlsby, N. Scaling vision with sparse mixture of experts. *CoRR*, abs/2106.05974, 2021.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G. E., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *The Fifth International Conference on Learning Representations*, 2017.

Shen, S., Qi, Z., Sun, J., Huang, Q., Tian, Q., and Wang, S. Enhancing pre-trained representation classifiability can boost its interpretability. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GjfIZan5jN.

Yang, X., Venhoff, C., Khakzar, A., de Witt, C. S., Dokania, P. K., Bibi, A., and Torr, P. Mixture of experts made intrinsically interpretable. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=6QERrXMLP2.

Zhu, T., Qu, X., Dong, D., Ruan, J., Tong, J., He, C., and Cheng, Y. Llama-moe: Building mixture-of-experts from llama with continual pre-training. *arXiv preprint arXiv:2406.16554*, 2024. URL https://arxiv.org/abs/2406.16554.

A. Ablation of Representation Selection for Additional LLaMA-MoE-Base Models

We repeat the representation ablation study on LLaMA-MoE-3.0B and LLaMA-MoE-3.5B (2/8). All experimental settings are identical to those described in Section 4.4.

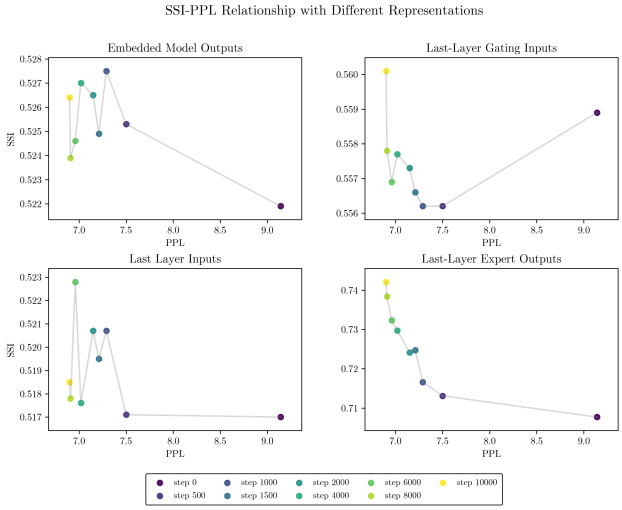

Figure 8. The relationship between SSI and PPL with different representation choices in LLaMA-MoE-3.0B model.

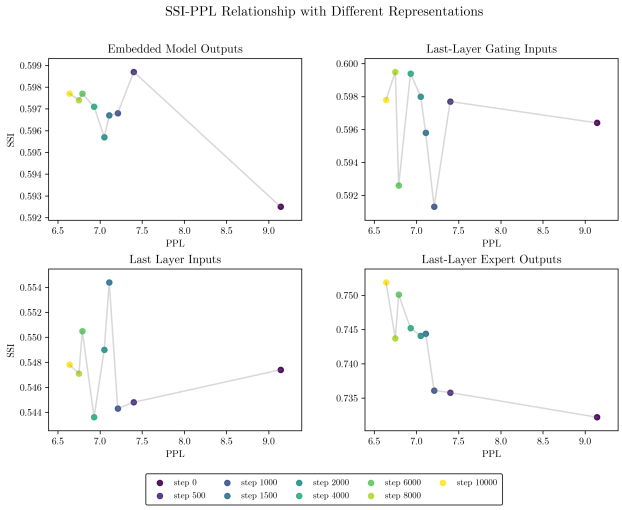

Figure 9. The relationship between SSI and PPL with different representation choices in 2/8 LLaMA-MoE-3.5B model.

Figures 10–12 show the last-step training loss trends across all checkpoints for the LLaMA-MoE-Base family. Across the training process, the loss for last-layer gating inputs and last-layer inputs remains substantially higher than that of the other two representations, while embedded model outputs consistently achieve the lowest loss, consistent with the findings reported in Section 4.4.

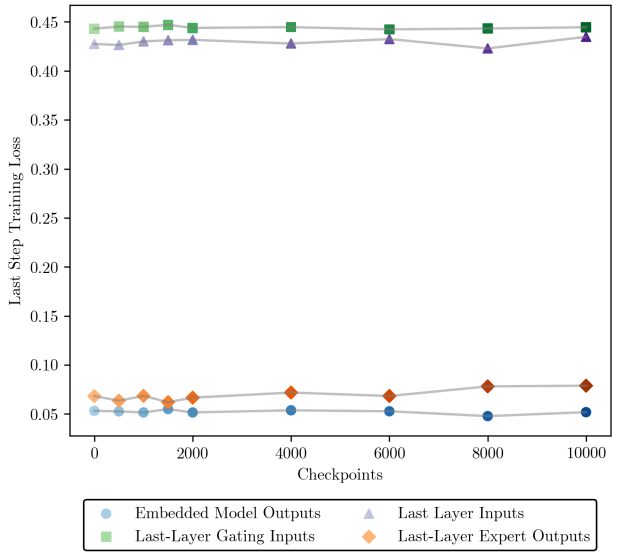

Figure 10. Last-step training loss for all four representation choices on the 4/16 LLaMA-MoE-3.5B model.

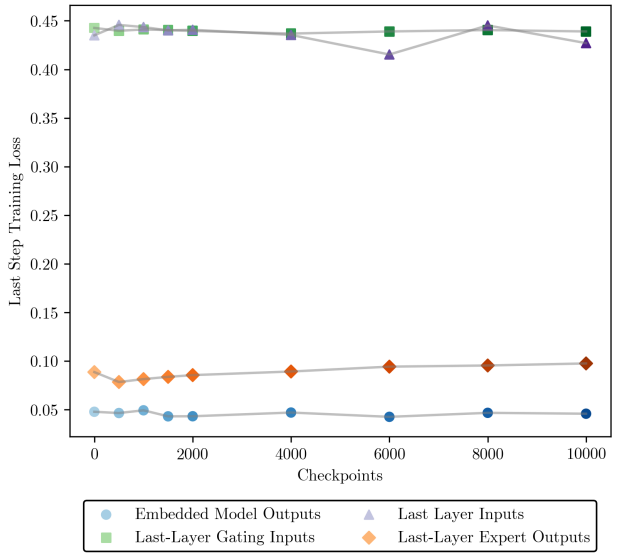

Figure 11. Last-step training loss for all four representation choices on the LLaMA-MoE-3.0B model.

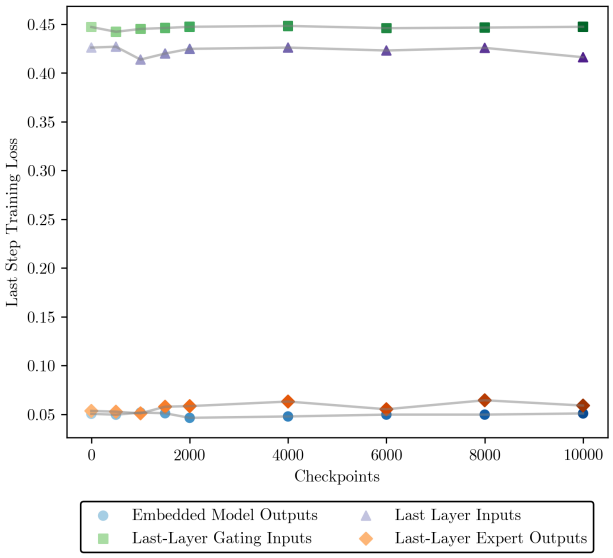

Figure 12. Last-step training loss for all four representation choices on the 2/8 LLaMA-MoE-3.5B model.

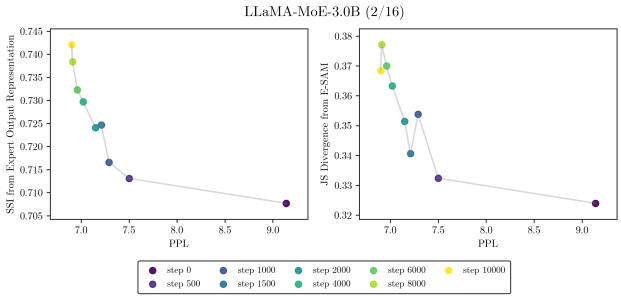

Figure 13. Comparison between SSI-PPL relationship derived from expert output representations and the JS Divergence-PPL relationship based on E-SAM. The model used is LLaMA-MoE-3.0B.

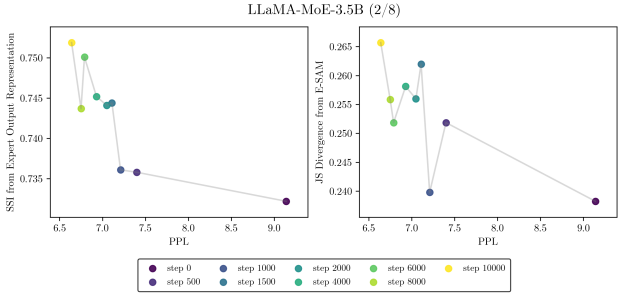

Figure 14. Comparison between SSI-PPL relationship derived from expert output representations and the JS Divergence-PPL relationship based on E-SAM. The model used is 2/8 LLaMA-MoE-3.5B.