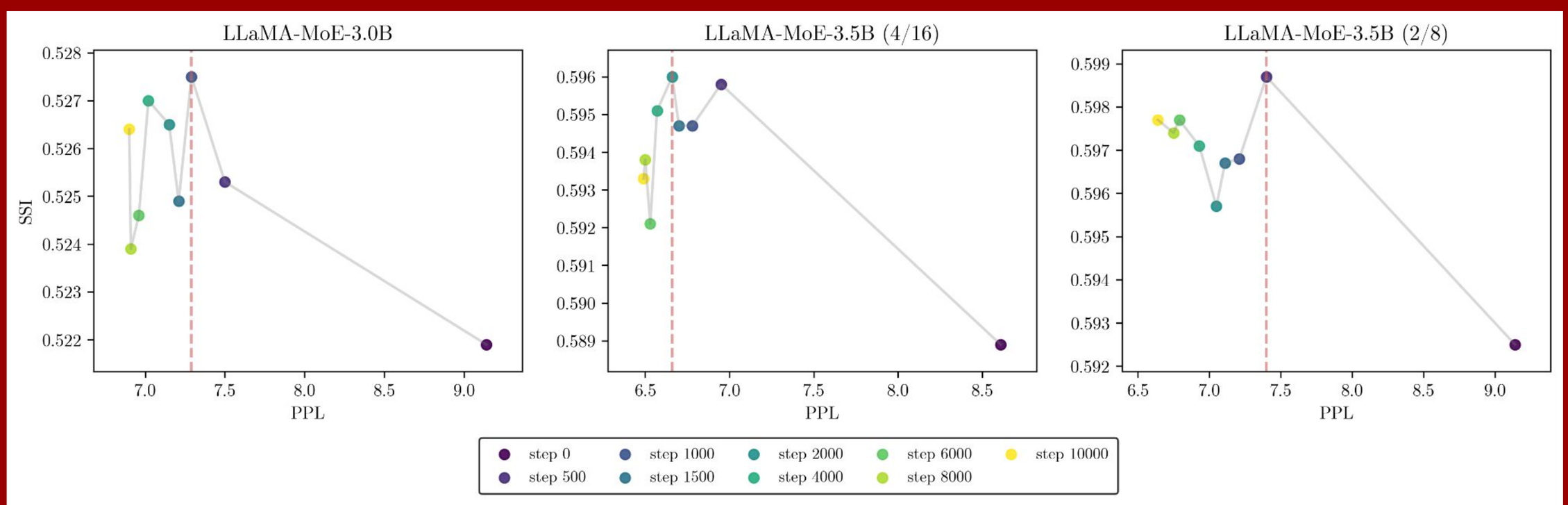
Semantic specialization does not always grow with better language modeling performance.



Investigating Expert Semantic Specialization in Mixture-of-Expert Models

Luyu Wang[1], Shuhui Wang[2]

[1] Georgia Institute of Technology [2] Institute of Computing Technology, Chinese Academy of Sciences

1 TL;DR

- We propose a **semantic-level framework** for interpreting expert specialization.
- 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.

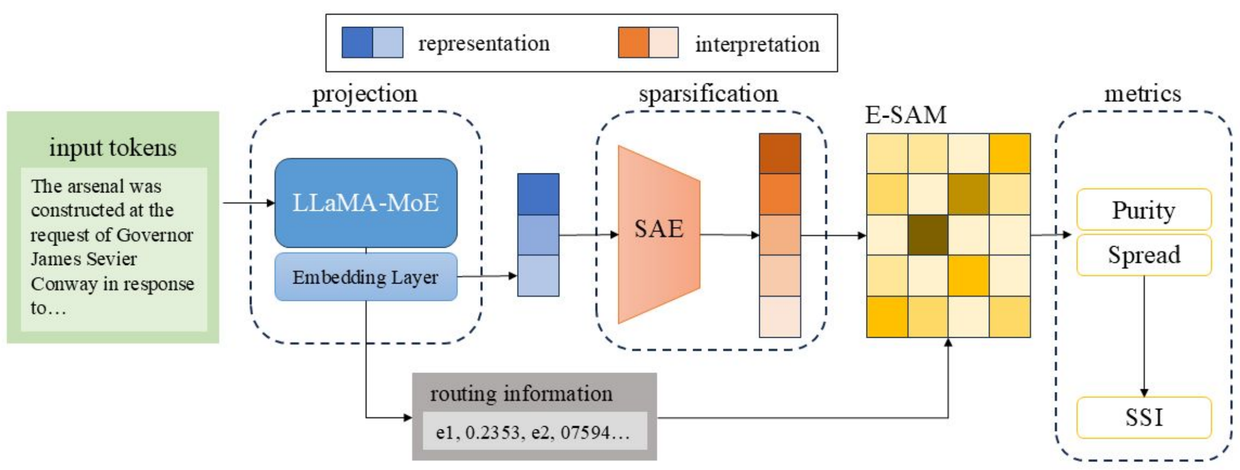
2 Background / Motivating Question

- MoE models route each token to a small subset of experts, enabling scalable conditional computation.
- Prior work shows experts can specialize by token, function, or domain.
- However, these analyses do not directly measure whether experts align with coherent semantic dimensions.

Question: *How can we measure semantic expert specialization, and how does it evolve during finetuning?*

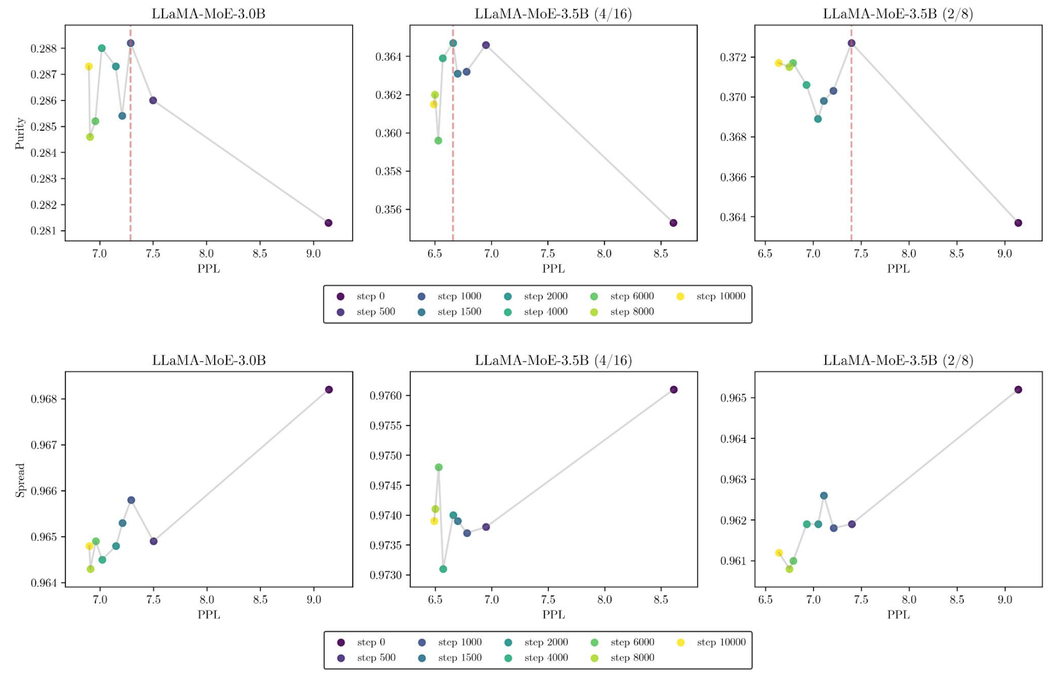
3 Measuring Semantic Specialization with SSI

- We introduce the Semantic Specialization Index (SSI), which measures expert-semantic alignment by combining SAE semantic features with MoE routing weights.
- It captures both semantic concentration among experts (Purity) and balanced use of experts across semantic dimensions (Spread).



4 Specialization Peaks Before Best PPL

- We evaluate three LLaMA-MoE-Base models on WikiText-103, finetuning them across multiple checkpoints and computing SSI using last-layer routing information and SAE-derived semantic features.
- As shown in the figure above, semantic specialization increases early in finetuning, peaks midway, and then partially declines even as language modeling continues to improve.



- This non-monotonic SSI trend is mainly driven by Purity: semantic dimensions become more concentrated in stable expert subsets during early finetuning, but this concentration weakens after the specialization peak.
- Spread shifts mainly at the beginning of finetuning, but does not strongly track later improvements in language modeling.

5 Limitations and Discussion

- Scope: We evaluate three LLaMA-MoE-Base models on WikiText-103; broader MoE architectures and datasets remain to be tested.
- Layer coverage: SSI is computed using last-layer routing, so layer-wise specialization dynamics are left for future work.
- Future Work: use SSI to guide specialization-aware routing regularization, improving model interpretability and performance simultaneously.

PAPER

wangshuhui@ict.ac.cn