A transparent MoE Architecture For Sequence Level Routing

Björn Pettersson

03-10-2025

Abstract

While it is well known that MoE routing is done mostly at a token level, Fan et al. (2024) propose that this may limit expert specialization, suggesting sequence-level routing could foster weak topic specialization when models are trained from scratch. Initial experiments with token-level routing show that pretraining a domain expert (e.g., digit recognition or coding tasks), transferring its weights into a larger MoE, and retraining on broader tasks leads to the pretrained expert being selected for its domain. However, this effect vanishes in overlapping domains such as business, psychology, and history, where a general expert dominates.

Based on Fan et al., the suspected cause is that token-level routing centralizes similar semantic structures into one expert. This separation is easier for code/math vs. language tasks but collapses for overlapping fields. The hypothesis is that training larger models with sentence-level routing may mitigate this by enabling clearer domain boundaries. Token-level results indicate centralizing tendencies in pretrained experts and suggest predictability in routing outcomes under certain conditions.

Experiments are therefore proposed with sentence-level routing to reinforce specialization predictably using pretraining. Goals include: (i) building architectural transparency for studying routing behavior, (ii) exploring alignment applications by concentrating misaligned behaviors in identifiable experts, enabling their suppression during inference (cf. Fayyaz et al., 2025).

Challenges include catastrophic forgetting during MoE training, where initialized expert capabilities may dissipate or relocate. Conversely, pretraining may create a "path of least resistance," reinforcing specialization. Regularization via weight distance constraints (Ludziejewski et al., 2025) may mitigate drift. Further, success may depend on whether tasks are narrow or broad, as complex tasks naturally involve multiple experts (Huang et al., 2024).

Key research questions:

- Will sequence-level routing combined with pretraining strengthen domain expertise?
- How well does pretraining predict expert routing in large MoEs?
- How do outcomes differ between narrow vs. broad tasks?

Literature

Fayyaz, M., Modarressi, A., Deilamsalehy, H., Dernoncourt, F., Rossi, R., Bui, T., ... & Peng, N. (2025). *Steering MoE LLMs via Expert (De) Activation.* arXiv preprint `https://arxiv.org/abs/2509.09660`

Fan, D., Messmer, B., & Jaggi, M. (2024). *Towards an empirical understanding of MoE design choices.* arXiv. `https://doi.org/10.48550/arXiv.2402.13089`

Ludziejewski, J., Pióro, M., Krajewski, J., Stefaniak, M., Krutul, M., Małaśnicki, J., Cygan, M., Sankowski, P., Adamczewski, K., Miłoś, P., & Jaszczur, S. (2025). *Joint MoE scaling laws: Mixture of experts can be memory efficient.* arXiv. `https://arxiv.org/abs/2502.05172v1`

Huang, Q., An, Z., Zhuang, N., Tao, M., Zhang, C., Jin, Y., Xu, K., Chen, L., & Feng, Y. (2024). *Harder tasks need more experts: Dynamic routing in MoE models.* arXiv. `https://arxiv.org/abs/2403.07652v1`

Appendix Initial Experiments

Some initial simple experiments were run on standard token level routing to see if some predictions of routing behaviour based on pretraining, the idea is to extend these experiments to larger sentence level routing models.

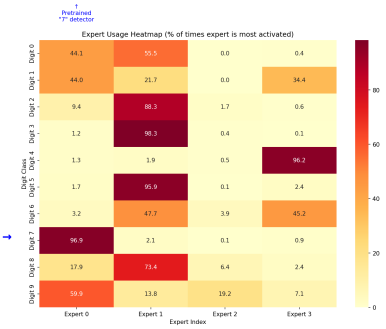

Figure 1: MNIST, pretrained 7 predictor (expert 0) selected 96 percent of the time for this task

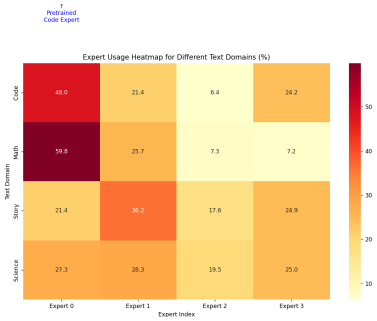

Figure 2: Code and math, centralises in pretrained code expert 0 for these tasks

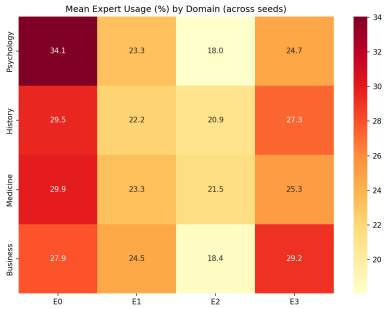

Figure 3: Semantic overlap in domains causes collapse into one majority expert (0) for all fields.