A transparent Mixture Of Experts Architecture For Sequence Level Routing

Björn Pettersson – Exchange from University of Copenhagen,
Computer Science and Economics

What:

What if Mixture of experts, were a mixture of actual domain experts, not just token specialists?

Why: Alignment and modular potential

- **Alignment potential:** Could we gather bad capabilities in some experts and modify routing on inference for better alignment similar to (cf. Fayyaz et al., 2025)
- **Modularity:** Could we more cheaply add, narrow expertise to a larger MoE by narrow training of expert and then slight training of MoE to use that expert.

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

How: could we create domain experts?

- **(SOTA weak experts) Sequence level routing:** We use sequence level routing instead of token level routing
- **(Proposal) + Narrow pretraining of experts:** We pretrain the experts on different narrow subtasks and initialize using those weights

← Shows weak topic specialization Fan et al. (2024)

← This project

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

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?*

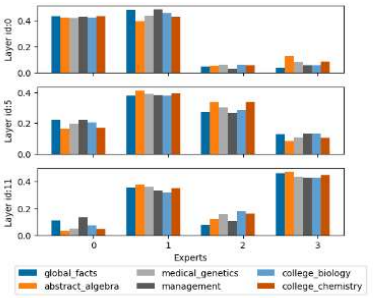
Proposed Project Objectives:

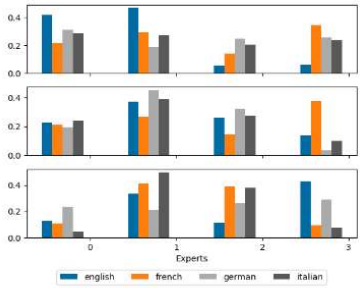
- **C: Control experiment:** Do a smaller version of Fan et Al, try to reproduce weak specialization
- **P: Intervention experiment:** Identical setup as (C) but with the addition of initialization of experts to pretrained narrow weights
- **Analysis:** Show the effect of (P) on a smaller scale. Further discussion about, what expert initializations may collapse or separate more easily during general MoE training and under what circumstances. Experiments with limiting how much initialized weights can change during general training.

Appendix: Reference paper results

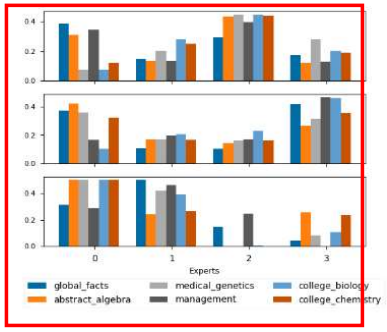
Semantic topics | Languages

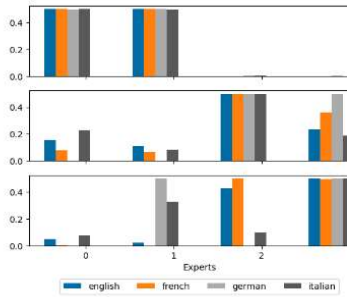
Token level routing





Sequence level routing





Towards an empirical understanding of MoE design choices (Fan et al)

Appendix: Initial experiments

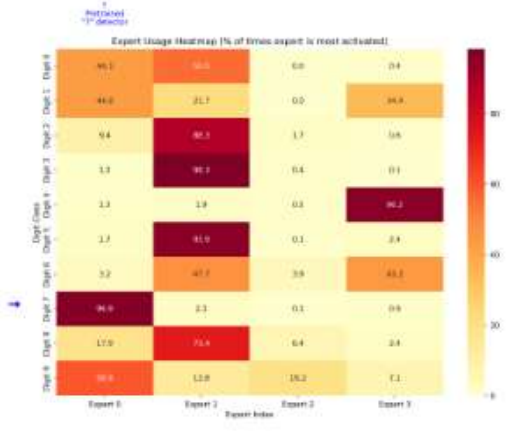


Figure 1: MNIST, pretrained 7 predictor (expert 0) selected 96 percent of the time for this task

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

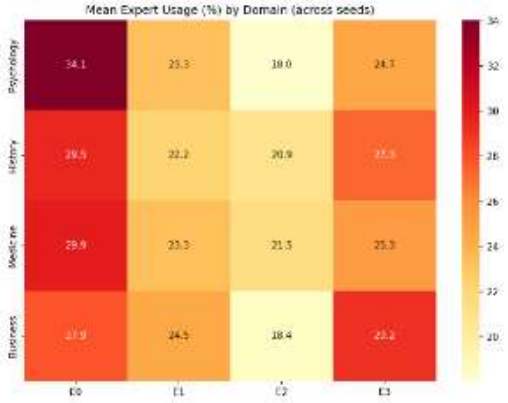


Figure 3: Semantic overlap in domains causes collapse into one majority expert (0) for all fields.