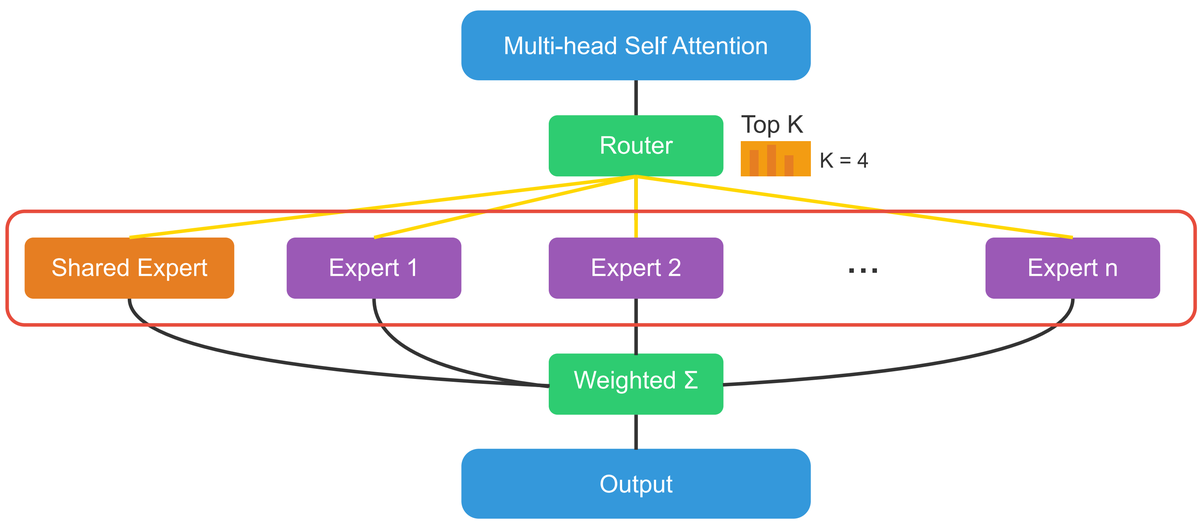
Assessing expertise overlap in Mixture of Experts Architectures

Ádám Divák and Joan Velja

University of Amsterdam

How specialized are experts in Sparse MoEs?

- Mixture of Experts (MoE) promise fast computations at large parameter counts
- **Interpretability analysis of MoEs is more difficult** due to the additional routing mechanism
- How different are MoEs compared to Dense networks on a previously explored task? **Can we prove expert specialization and find circuits?**

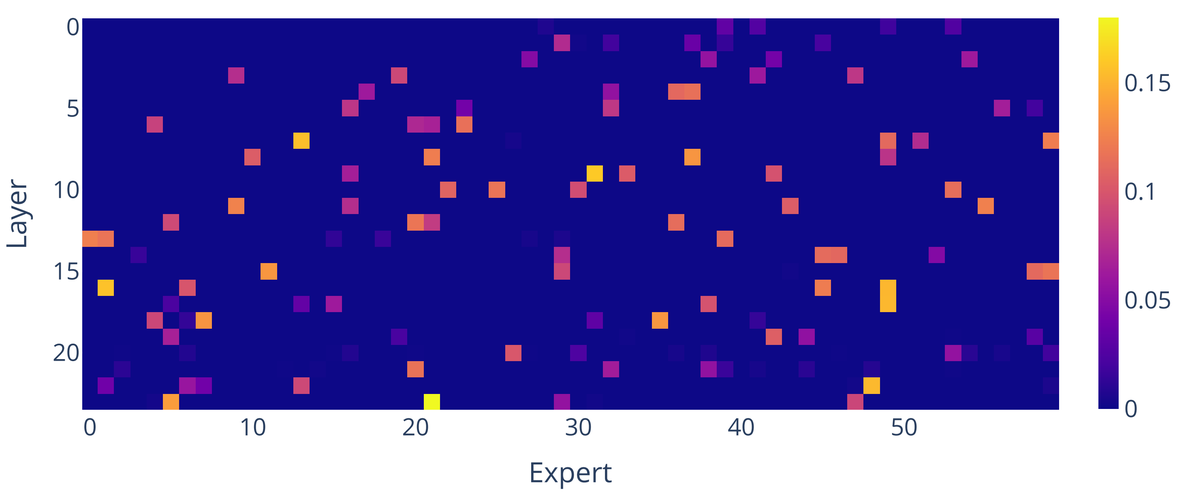


Analysis of the Indirect Object Identification task

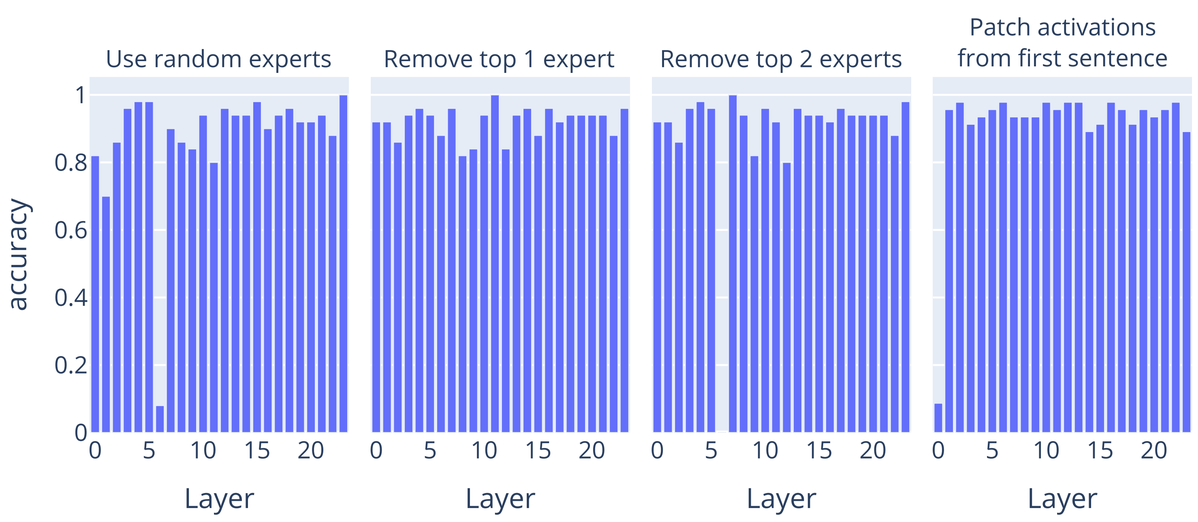
Then, John and Mary went to the a shop. John gave a key to... Mary

- Predict single next token with the correct name, 15 different prompt templates [1]
- **Qwen1.5-MoE, 2.7B parameters** (Chat fine-tuned, Int4 quantized) [2]
- **1 Shared expert + 60 Specialized Experts with Top4 selection**
- Analysis done using `nnsight` [3]

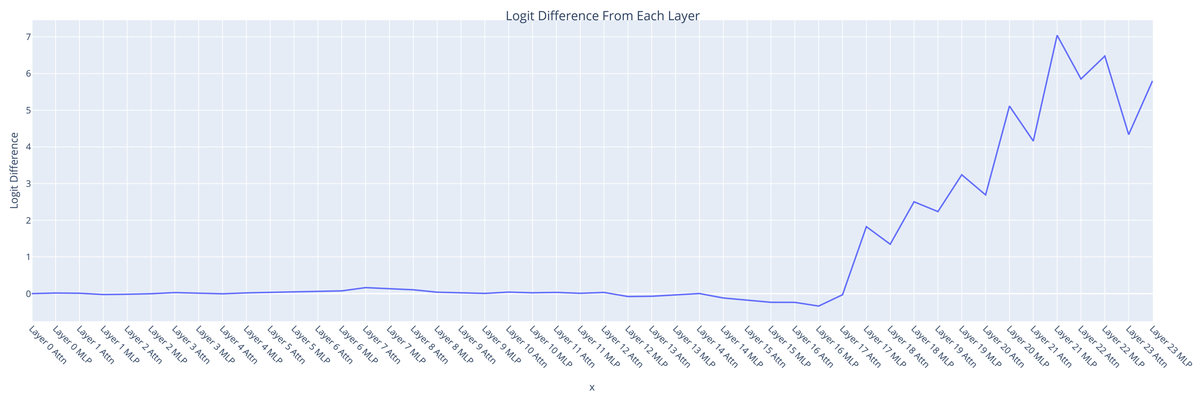
The same few experts are consistently activated in this task



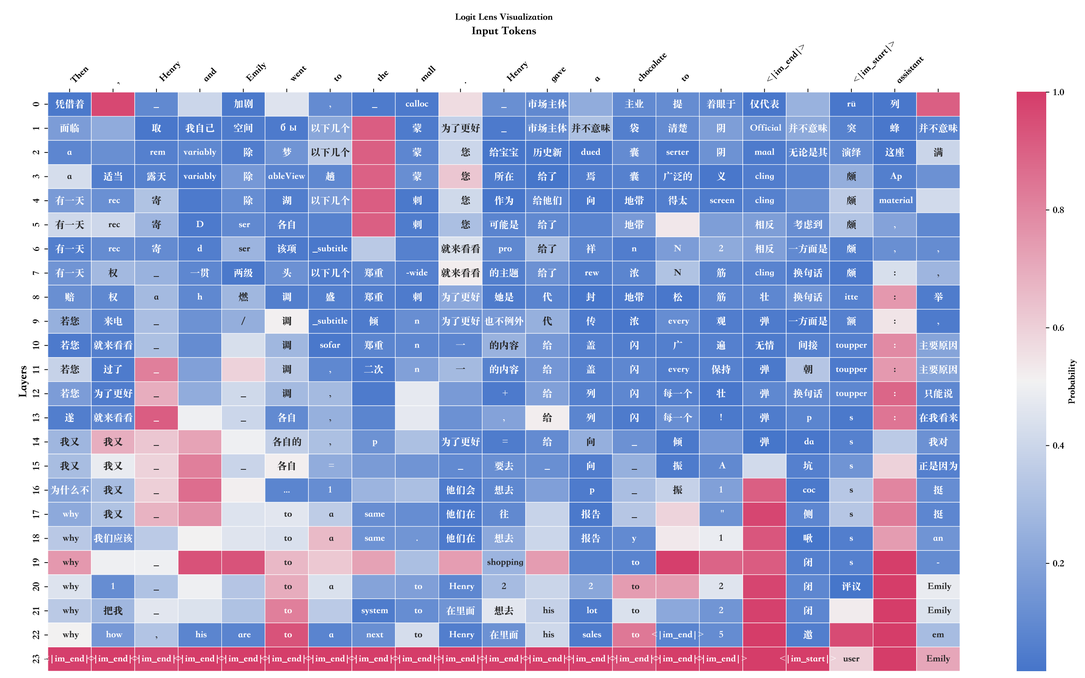
Intervening in expert routing is effective at early layers



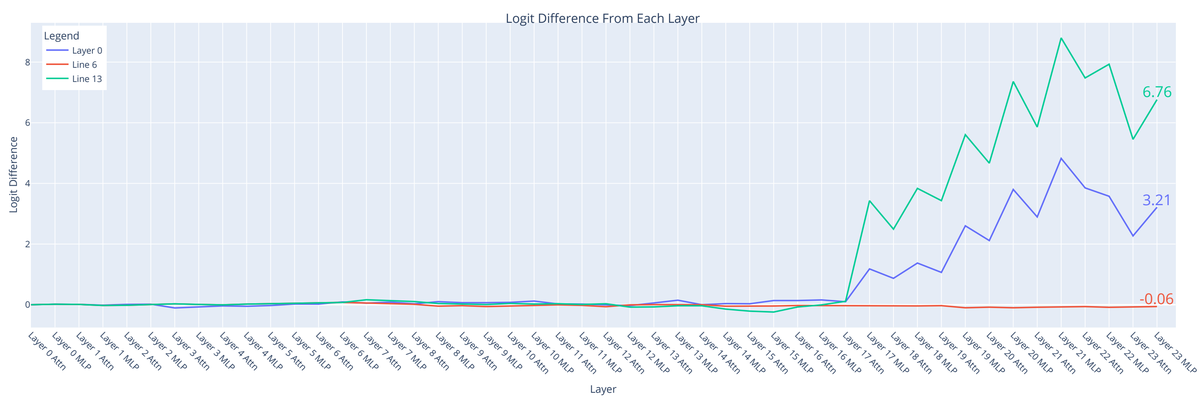
Logit difference however shows the importance of later layers



..due to name tokens appearing late in the residual stream



Logit difference after routing intervention confirms the strong role of the previously identified early layer



Conclusion

- **Expert sparsity is achieved by this model** (on this task)
- **MoEs can be analyzed with the same interpretability techniques** as Dense models
- **Although we find ”specialized” experts for the task**, the results of activation patching are still **inconclusive**, highlighting the requirement for further work

References

[1] K. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt, *Interpretability in the wild: A circuit for indirect object identification in gpt-2 small*, 2022. arXiv: 2211.00593 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2211.00593.

[2] Q. Team, *Introducing qwen1.5*, Feb. 2024. [Online]. Available: https://qwenlm.github.io/blog/qwen1.5/.

[3] J. Fiotto-Kaufman, *nnsight: The package for interpreting and manipulating the internals of deep learned models*. [Online]. Available: https://github.com/JadenFiotto-Kaufman/nnsight.

adam.divak@student.uva.nl

joan.velja@student.uva.nl