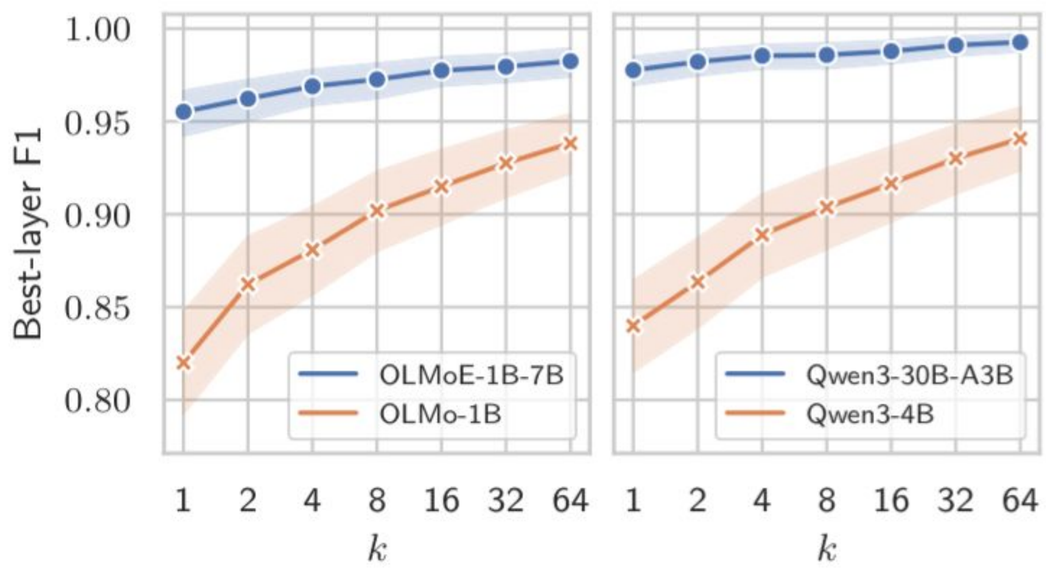
MoE experts are more interpretable than dense FFNs

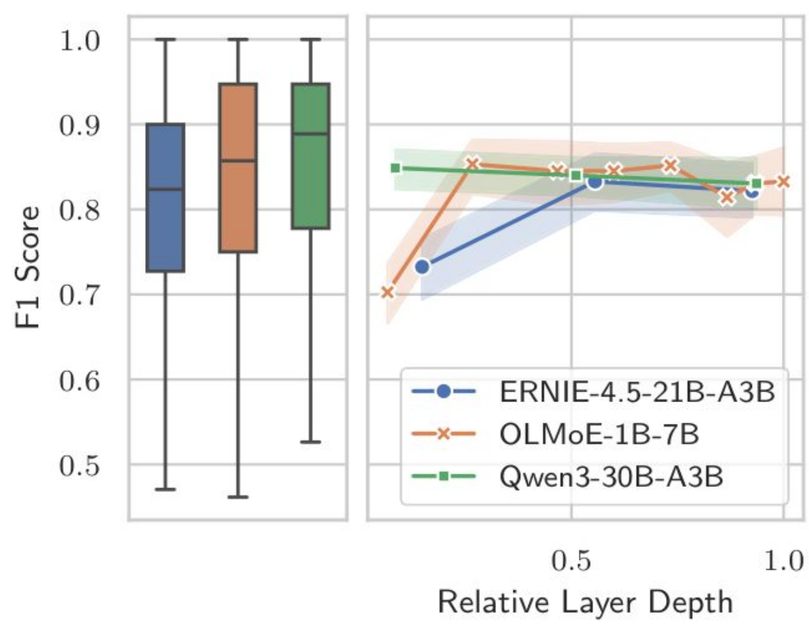
The Expert Strikes Back: Interpreting Mixture-of-Experts Language Models at Expert Level

Background: Almost all SoTA LLMs are MoE models, but mechanistic interpretability has not focused on MoE experts. MoEs might have less polysemanticity because of router + sparsity.

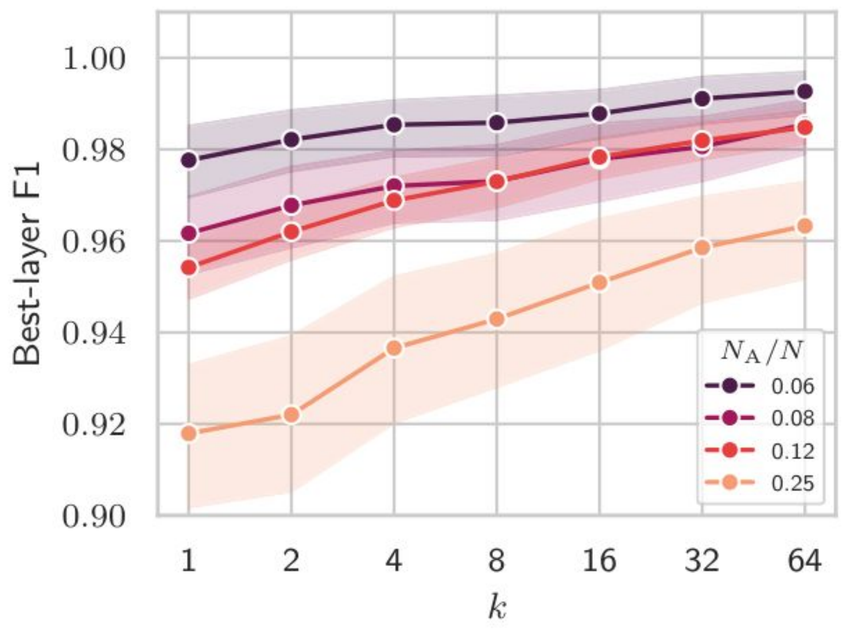
MoE experts are less polysemantic: Probing F1 scores are higher (at k=1) for MoE experts (blue) compared to dense FFNs (orange).



Scaling up interpretability to experts: LLM scorer F1 scores are high across models and layers.



Polysemanticity decreases with routing sparsity: MoE Models with a smaller ratio of active experts to total experts show less polysemanticity.



Methods

Measuring Polysemanticity

Train a linear classifier on **k** neurons (k-sparse probing).
Key idea: Vary **k** to measure how polysemantic neurons are.

Interpreting entire experts

Apply automatic interpretability to experts.
Metric: F1 score of LLM scorer.

Limitation: MoE experts likely have remaining polysemanticity even as routing sparsity increases.

PAPER

Jeremy Herbst, Stefan Wermter, Jae Hee Lee

Mechanistic Interpretability Workshop, ICML 2026

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG