Beyond Neuron-Level Sparsity: Achieving Faithful and Interpretable LLMs with Mixture of Decoders

Grigorios Chrysos
University of Wisconsin-Madison
chrysos@wisc.edu

The rapid scaling of large language models (LLMs) (Wu et al. 2023; Geiger et al. 2024) has intensified the need for both interpretability and privacy. A bottom-up, mechanistic understanding of LLMs is critical for addressing safety and transparency concerns, proving helpful for applications ranging from controlling refusal (Arditi et al. 2024) to detecting unsafe code generation (Templeton 2024). However, this goal is challenged by the dense nature of LLM representations, where human-interpretable features are often distributed across many neurons rather than being neatly aligned with individual ones (Olah et al. 2020; Elhage et al. 2022).

One direction for addressing this is specialization through mechanisms like the Mixture-of-Experts (MoE) (Jacobs et al. 1991). While popular sparse MoE approaches enable discrete expert routing, they can suffer from non-differentiable learning and potential instability. If we cast MoE as a series of linear layers, MoE is naturally expressed as a tensor. By leveraging tensor decompositions, the desired increase in the number of experts can be achieved by performing operations in a factorized space (Oldfield et al. 2024). We will show qualitative results demonstrating the expert specialism achieved by this method when pre-training large GPT2 and MLP-Mixer models.

A second, dominant strategy for learning more interpretable representations is to enforce sparsity, as sparser models can aid human explanation (Ramaswamy et al. 2022) and achieve higher scores on auto-interpretability metrics (Juang et al. 2024). Current approaches enforce this constraint at the *neuron-level*. We argue this constraint is too restrictive, leading to a severe trade-off that sacrifices model accuracy for sparsity and results in poor reconstructions of the original model components. Preserving the base model's performance is crucial for both model faithfulness—ensuring the learned subcomputations truly emulate the base model's intricacies (Engels, Riggs, and Tegmark 2024)—and practical adoption, which requires sparse layers that can directly replace their dense counterparts without performance loss. We argue for a paradigm shift from neuron-level to *layer-level* sparsity and propose the Mixture of Decoders (MxD). We will explain how MxDs overcome the sparsity-accuracy trade-off by expanding pre-trained dense layers into tens of thousands of specialized sublayers using a flexible tensor factorization. We will prove that this construction allows each sparsely-activating MxD sublayer to implement a full-rank linear transformation, thereby preserving the original decoder's expressive capacity even under heavy sparsity. Across 108 sparse layers in four LLMs up to 3B parameters, we will demonstrate that MxDs significantly outperform alternatives on the sparsity-accuracy frontier while remaining competitive on 34 sparse probing and feature steering tasks, validating their ability to learn specialized features and opening a new avenue for designing interpretable yet faithful decompositions.

Lastly, moving beyond interpretability, we will address the critical concern of privacy and safety. How can we ensure users of ML models do not leverage predictions based on incorrect personal data to harm others? This question is particularly pertinent given the rise of open-weight models, where simply masking model outputs does not suffice to prevent adversaries from recovering harmful predictions. While the stringent guarantees provided by differential privacy and other rigorous frameworks are desirable, they may not scale to large models. To tackle this issue, we propose to revisit the concept of inducing maximal uncertainty on protected instances while preserving accuracy on all other instances (Ashiq et al. 2025). Towards that end, we introduce a certifiable approximation algorithm and prove a tight bound that characterizes the privacy-utility tradeoff that our algorithm incurs.

References

Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Panickssery, N.; Gurnee, W.; and Nanda, N. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Ashiq, M. H.; Triantafillou, P.; Tseng, H. Y.; and **Chrysos**, G. 2025. Inducing Uncertainty on Open-Weight Models for Test-Time Privacy in Image Recognition. *Technical report*.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; Grosse, R.; McCandlish, S.; Kaplan, J.; Amodei, D.; Wattenberg, M.; and Olah, C. 2022. Toy Models of Superposition. arXiv:2209.10652.

Engels, J.; Riggs, L.; and Tegmark, M. 2024. Decomposing

The Dark Matter of Sparse Autoencoders. *arXiv preprint arXiv:2410.14670*.

Geiger, A.; Wu, Z.; Potts, C.; Icard, T.; and Goodman, N. 2024. Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, 160–187.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Juang, C.; Paulo, G.; Drori, J.; and Belrose, N. 2024. Open Source Automated Interpretability for Sparse Autoencoder Features. https://blog.eleuther.ai/autointerp/. EleutherAI Blog.

Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; and Carter, S. 2020. Zoom In: An Introduction to Circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Oldfield, J.; Georgopoulos, M.; **Chrysos**, G.; Tzelepis, C.; Panagakis, Y.; Nicolaou, M. A.; Deng, J.; and Patras, I. 2024. Multilinear Mixture of Experts: Scalable Expert Specialization through Factorization. In *NeurIPS*.

Ramaswamy, V. V.; Kim, S. S. Y.; Fong, R. C.; and Russakovsky, O. 2022. Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability. *CVPR*, 10932–10941.

Templeton, A. 2024. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic.

Wu, Z.; Geiger, A.; Icard, T.; Potts, C.; and Goodman, N. 2023. Interpretability at Scale: Identifying Causal Mechanisms in Alpaca. In *NeurIPS*.