Yupei Wang (https://yplarrywang.github.io/)
YpLarryWang wangyupei@mail.bnu.edu.cn

As neural language models (LMs) grow larger to become large language models (LLMs), their internal mechanisms become increasingly complex and challenging to understand. Modern LMs like GPT-4 demonstrate remarkable capabilities through next-token prediction [1]. This raises fundamental questions about how these models acquire and utilize knowledge.

I propose two primary perspectives for investigating LM interpretability: **synchronic interpretability** and **diachronic interpretability**. Synchronic interpretability investigates how trained LMs represent knowledge and construct causal chains in their reasoning process from input tokens to final outputs. Diachronic interpretability, on the other hand, examines how LMs acquire knowledge during training, focusing on the emergence of capabilities when the system as a whole exhibits properties that cannot be directly attributed to its individual components.

Based on these two interpretability perspectives, I will pursue research on causality and emergence in neural LMs. My future research plans focus on two key aspects:

§1 Investigating Causal Mechanisms for LMs' Decision-Making. Neural networks are believed to learn comprehensible algorithms when solving tasks, but they have no inherent tendency to organize these algorithms in a human-readable way. By identifying human-understandable mediating variables from model internals, we can construct causal chains between inputs and outputs that describe the algorithmic mechanisms underlying LMs' decision-making. My goal is to develop effective and scalable methods that reveal these causal mechanisms and leverage these methods to better align LMs with human values. This investigation falls within the scope of *synchronic interpretability*.

§2 Identifying & Predicting Emergence in LMs' Training Process. Even the most powerful LLMs are fundamentally combinations of simple linear and non-linear units. During training, these networks must experience moments where macro-level capabilities emerge that transcend their micro computational units. These qualitative changes in network behavior, known as emergence, are currently difficult to identify and predict. I aim to develop quantitative methods to identify and predict such emergent phenomena, which would enhance our understanding of LMs' training dynamics and potentially reduce the cost of industrial-scale LLM training. This research direction aligns with *diachronic interpretability*.

Next, I will briefly elaborate on some necessary literature on both directions and point out their limitations from my perspective. Then, I will discuss my **Future Plans** *based on these discussions.*

1 Investigating Causal Mechanisms for LMs' Decision-Making

Mechanistic interpretability is often defined as reverse-engineering neural networks to better understand how and why they behave in certain ways (Miller et al., 2024; Nanda et al., 2023a). Within mechanistic interpretability, the branch that adopts a causality-based perspective and emphasizes algorithmic understanding – where an "algorithm" represents a complete causal graph explaining model generalization – is termed *causal interpretability* (Mueller et al., 2024). In this framework, a *mechanism* is defined as a causal chain from a cause to an effect, with intermediate nodes between them termed *mediators* (Lewis, 1973). Research in this field naturally divides into two complementary aspects that build upon each other.

Identifying the Dots. This means finding the mediators that represent specific features or participate in specific phenomena from neural networks. Mediators can be linear or non-linear, but most of current research focuses on linear mediators (Mueller et al., 2024).

[1] During RLHF stage, LLMs do not go through simple next-token prediction, but their knowledge and skills are mostly obtained in pre-training and supervised fine-tuning stages, whose core task is next-token prediction.

Yupei Wang (https://yplarrywang.github.io/)
YpLarryWang wangyupei@mail.bnu.edu.cn

Linear mediators can align to the bases in activation vector, which means they can be neurons, subspaces formed by grouping neurons and attention heads. For example, in LM, Mueller et al. (2022) localized subsets of neurons implementing fundamental linguistic phenomena like syntactic agreement. Todd et al. (2023) found that some heads are directly implicated in encode functions in latent space. It is worth noting that components localized by these methods are not necessarily human-understandable, which makes sparse auto-encoder (SAE, Ng et al., 2011) an inviting option. Although SAE provides correlational rather than causal insights, it learns to decompose a model's internal representations into interpretable components by transforming input activations into a higher-dimensional but sparse representation.

Due to their *polysemanticity* (Elhage et al., 2022), basis-aligned mediators like neurons, attention heads, and their sets do not necessarily correspond to cleanly interpretable features or concepts (Mueller et al., 2024). Since features can be encoded in subspaces that are not aligned to activation bases, mediators can be localized in non-basis-aligned subspaces. Wu et al. (2023) and Geiger et al. (2024) located greater-than and equality relationships encoded in latent space through learned rotation operations.

Connecting the Dots. After locating features in models, it is natural to develop methods to determine causal relations among them and build causal graph. There are two ways to look at this aspect, circuit discovery and alignment search.

Circuit discovery means to find sub-networks from the original network that serve for a specific purpose, corresponding to granularity of basis-aligned mediators. For example, Elhage et al. (2021) discovered the existence of *induction head* in two layer attention-only transformers and Wang et al. (2023) found that `GPT-2 small` incorporate induction head as well as attention heads with other functions in a special circuit to conduct indirect object identification (IOI) task.

Corresponding to non-basis-aligned mediators, Wu et al. (2023) and Geiger et al. (2024) developed distributed alignment search (DAS) method that decomposes the search space into learnable subspaces and aligns hypothesized high-level causal variables with these latent representations. This process is called **causal abstraction**, as during which a low-level neural network is aligned to a high-level causal graph. However, DAS-based methods require pre-hypothesizing causal mechanisms, which creates challenges for method automation. Marks et al. (2024) creatively developed an approach that first applies SAE to decompose features for all network components, then performs causal graph search based on counterfactual intervention on the sparse, interpretable feature network. This simultaneously addresses the automation limitations of DAS-based methods and SAE's lack of causality, which marks an important advancement in *synchronic interpretability*.

2 Identifying & Predicting Emergence in LMs' Training Process

According to Anderson (1972), emergence is when quantitative changes in a system result in qualitative changes in behavior. Wei et al. (2022) suggests that an ability is emergent if it is not present in smaller models but appears in larger ones. In fact, during the training process of a neural network model (e.g., LLM pre-training and SFT), the model's capabilities in different domains gradually change, and qualitative transitions can occur at specific moments (Nakkiran et al., 2021; Nanda et al., 2023b).

The quantification of emergence lacked rigorous methods until Hoel et al. (2013) introduced causality into its framework. They proposed that **causal emergence** occurs when a system exhibits stronger causal effects after a specific coarse-graining transformation compared to the original system. They introduced effective information (EI) for quantification. EI measures the strength of causal relationships between successive states in a dynamical system, and causal emergence is formally defined *when the EI at macro level is larger than*

Yupei Wang (https://yplarrywang.github.io/)
YpLarryWang wangyupei@mail.bnu.edu.cn

the EI at micro level:

$$\Delta\mathcal{J} = \mathcal{J}(f_M) - \mathcal{J}(f_m) > 0$$

where $\mathcal{J}(f_M)$ represents the EI at macro level and $\mathcal{J}(f_m)$ is the EI at micro level. The terms f_M and f_m denote the macro-level and micro-level dynamics respectively, while $\Delta\mathcal{J}$ quantifies the causal emergence measured in bits.

However, this theory requires an effective coarse-graining strategy to quantify emergence, which impedes automated emergence quantification algorithms. Rosas et al. (2020) developed a framework based on partial information decomposition theory that eliminates the need for predefined coarse-graining strategies. However, this method's high computational complexity stems from its requirement for exhaustive enumeration of variable combinations. Despite later proposing an approximate method, it still requires predefined macro-variables, limiting automatic causal emergence identification. Later, Zhang and Liu (2022) introduced Neural Information Squeezer (NIS) and Yang et al. (2025) introduced NIS++ models. NIS++ automatically learns coarse-graining strategies and macro-variables by maximizing EI while maintaining accurate micro-state predictions based on time series data of micro-variables.

Since causal emergence applies to any dynamical system evolving over time, it naturally extends to quantifying neural network training processes. Marrow et al. (2020) attempted to quantify the learning process of fully connected neural networks on Iris and MNIST datasets using EI. However, lacking effective coarse-graining strategies, they simply binned node activation levels into discrete states. This simplistic approach failed to capture conceptual representations in neural networks, limiting the interpretability of EI variations. Now, SAE-based causal abstraction strategies (Marks et al., 2024) offer a promising coarse-graining approach, providing an excellent entry point for incorporating EI-based causal emergence quantification into LM training dynamics research, which would significantly contribute to the *diachronic interpretability* line of inquiry.

Future Plans

During my PhD, I plan to conduct research on both synchronic and diachronic interpretability, with a particular focus on diachronic interpretability.

Specifically, for *synchronic interpretability*, I propose to extend research in two directions:
(1) More diverse mediators. Most current causal interpretability works employ linear mediators. However, LMs also contain *non-linear concept representations*. For example, days of the week are encoded circularly as a set of 7 directions in a two-dimensional subspace (Engels et al., 2024), yet current methods cannot systematically capture such features. Furthermore, current research treats LM layers in isolation. However, Templeton et al. (2024) discovered a phenomenon called *cross-layer superposition* where features in LLMs are not cleanly isolated to specific layers but are instead "smeared" across multiple layers, similar to how features can be stored in superposition across neurons. Similar observations exist for attention heads, where Jermyn et al. (2023) identified a phenomenon called *attention superposition* where multiple "attention circuits" or "attentional features" can be stored across multiple attention heads in a transformer model. These involve macro-level objects with larger granularity than existing mediators. *I aim to advance our understanding of LMs' synchronic aspects by focusing on non-linear or more macro-level mediators.*

(2) Circuit complexity economics. Current methods like activation patching (Meng et al., 2022) and SAE-based algorithms (Templeton et al., 2024; Huben et al., 2024) have helped locate many features and phenomena, while circuit discovery research has successfully reverse-engineered numerous effective circuits. It is intuitive to hypothesize that more

Yupei Wang (https://yplarrywang.github.io/)
YpLarryWang wangyupei@mail.bnu.edu.cn

complex features or phenomena correspond to circuits of higher complexity, and models share circuits across different tasks (Meng et al. (2022) discovered the presence of induction head (Elhage et al., 2021) in the IOI circuit). However, it is easy to ignore that these features and phenomena occur with varying frequencies, which might also affect circuit complexity. *How do models balance task complexity against usage frequency, and how do they leverage existing circuits to quickly generalize to new tasks?* In multilingual models, for example, do different languages share circuits (potentially corresponding to linguistic universality like universal dependencies)? Understanding how models utilize shared circuits to transfer knowledge from high-resource to low-resource languages remains an important open question.

For *diachronic interpretability*, I plan to focus on exploring the integration of causal emergence theory and causal interpretability from two perspectives:

(1) Identifying causal emergence in LM training. *Can the causal abstraction proposed by Marks et al. (2024) serve as a coarse-graining strategy for causal emergence identification?* If so, what qualitative changes occur in the causal graph when causal emergence is detected? Are these changes related to model performance improvements on specific tasks? If not, do the emergent patterns identified in latent space by optimization-based methods like NIS++ correspond to any human-understandable concepts? For example, during language model training, do model performances on multilingual and numerical computation tasks show significant improvements when causal emergence is detected? Does this indicate the formation of causally structured representations for cross-lingual mapping or numerical computation capabilities in the latent space?

(2) Predicting causal emergence in LM training. Current algorithms can only *identify* causal emergence but cannot *predict* it in advance. Predicting the timing of capability emergence (such as cross-lingual mapping or numerical computation) could help establish relationships between model training dynamics and training data characteristics. When capability emergence becomes a quantifiable metric, we can better measure sample efficiency. This has implications for both theoretical understanding and engineering practices of model training.

I believe these perspectives can support rich doctoral research topics and provide new insights for interpretable natural language processing.

References

J. Miller, B. Chughtai, and W. Saunders, "Transformer circuit faithfulness metrics are not robust," *arXiv preprint arXiv:2407.08734*, 2024.

N. Nanda, A. Lee, and M. Wattenberg, "Emergent linear representations in world models of self-supervised sequence models," in *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, Y. Belinkov, S. Hao, J. Jumelet, N. Kim, A. McCarthy, and H. Mohebbi, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 16–30. [Online]. Available: https://aclanthology.org/2023.blackboxnlp-1.2/

A. Mueller, J. Brinkmann, M. Li, S. Marks, K. Pal, N. Prakash, C. Rager, A. Sankaranarayanan, A. S. Sharma, J. Sun *et al.*, "The quest for the right mediator: A history, survey, and theoretical grounding of causal interpretability," *arXiv preprint arXiv:2408.01416*, 2024.

D. Lewis, "Counterfactuals, blackwells," 1973.

Yupei Wang (https://yplarrywang.github.io/)
YpLarryWang wangyupei@mail.bnu.edu.cn

A. Mueller, Y. Xia, and T. Linzen, "Causal analysis of syntactic agreement neurons in multilingual language models," in *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, A. Fokkens and V. Srikumar, Eds. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 95–109. [Online]. Available: https://aclanthology.org/2022.conll-1.8/

E. Todd, M. L. Li, A. S. Sharma, A. Mueller, B. C. Wallace, and D. Bau, "Function vectors in large language models," *arXiv preprint arXiv:2310.15213*, 2023.

A. Ng *et al.*, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, no. 2011, pp. 1–19, 2011.

N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah, "Toy models of superposition," *Transformer Circuits Thread*, 2022. [Online]. Available: https://transformer-circuits.pub/2022/toy_model/index.html

Z. Wu, A. Geiger, T. Icard, C. Potts, and N. Goodman, "Interpretability at scale: Identifying causal mechanisms in alpaca," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=nRfClnMhVX

A. Geiger, Z. Wu, C. Potts, T. Icard, and N. Goodman, "Finding alignments between interpretable causal variables and distributed neural representations," in *Causal Learning and Reasoning*. PMLR, 2024, pp. 160–187.

N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah, "A mathematical framework for transformer circuits," *Transformer Circuits Thread*, 2021. [Online]. Available: https://transformer-circuits.pub/2021/framework/index.html

K. R. Wang, A. Variengien, A. Conmy, B. Shlegeris, and J. Steinhardt, "Interpretability in the wild: a circuit for indirect object identification in GPT-2 small," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=NpsVSN6o4ul

S. Marks, C. Rager, E. J. Michaud, Y. Belinkov, D. Bau, and A. Mueller, "Sparse feature circuits: Discovering and editing interpretable causal graphs in language models," *arXiv preprint arXiv:2403.19647*, 2024.

P. W. Anderson, "More is different: Broken symmetry and the nature of the hierarchical structure of science." *Science*, vol. 177, no. 4047, pp. 393–396, 1972.

J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, 2021.

N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt, "Progress measures for grokking via mechanistic interpretability," *arXiv preprint arXiv:2301.05217*, 2023.

Yupei Wang (https://yplarrywang.github.io/)
YpLarryWang wangyupei@mail.bnu.edu.cn

E. P. Hoel, L. Albantakis, and G. Tononi, "Quantifying causal emergence shows that macro can beat micro," *Proceedings of the National Academy of Sciences*, vol. 110, no. 49, pp. 19 790–19 795, 2013.

F. E. Rosas, P. A. Mediano, H. J. Jensen, A. K. Seth, A. B. Barrett, R. L. Carhart-Harris, and D. Bor, "Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data," *PLoS computational biology*, vol. 16, no. 12, p. e1008289, 2020.

J. Zhang and K. Liu, "Neural information squeezer for causal emergence," *Entropy*, vol. 25, no. 1, p. 26, 2022.

M. Yang, Z. Wang, K. Liu, Y. Rong, B. Yuan, and J. Zhang, "Finding emergence in data by maximizing effective information," *National Science Review*, vol. 12, no. 1, p. nwae279, 2025.

S. Marrow, E. J. Michaud, and E. Hoel, "Examining the causal structures of deep neural networks using information theory," *Entropy*, vol. 22, no. 12, p. 1429, 2020.

J. Engels, E. J. Michaud, I. Liao, W. Gurnee, and M. Tegmark, "Not all language model features are linear," *arXiv preprint arXiv:2405.14860*, 2024.

A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan, "Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet," *Transformer Circuits Thread*, 2024. [Online]. Available: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

A. Jermyn, C. Olah, and T. Henighan, "Attention head superposition," *Transformer Circuits Thread: Circuits Updates — May 2023*, May 2023. [Online]. Available: https://transformer-circuits.pub/2023/may-update/index.html#attention-superposition

K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in gpt," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 359–17 372, 2022.

R. Huben, H. Cunningham, L. R. Smith, A. Ewart, and L. Sharkey, "Sparse autoencoders find highly interpretable features in language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=F76bwRSLeK