

Extracting and Utilizing Interpretation in Large Language Models

Xuansheng Wu
University of Georgia
Athens, Georgia, USA
xuansheng.wu@uga.edu

Ninghao Liu
University of Georgia
Athens, Georgia, USA
ninghao.liu@uga.edu

Abstract

Large language models (LLMs) have shown strong promise in many scenarios, yet still fall short in certain situations. Addressing these shortcomings requires explanations that not only reveal failure modes but also guide concrete fixes. We present a unified framework that extracts and utilizes interpretations at four stages of the LLM life cycle: data preparation, training, inference, and post-processing. Our preliminary results have demonstrated that explanations can (i) identify hallucinated responses with post-generation filtering, (ii) steer LLMs to prevent jailbreaks during model inference, and (iii) regularize the reliance on unintended features during model training. Ongoing work extends these ideas to diversity-driven synthetic data generation for model training. Collectively, these contributions chart a practical path toward developing and deploying LLMs that are safer, more transparent, and broadly trustworthy.

Keywords

Large Language Models, Explanation, Interpretation, UsableXAI

ACM Reference Format:

Xuansheng Wu and Ninghao Liu. 2018. Extracting and Utilizing Interpretation in Large Language Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 3 pages. <https://doi.org/XXXXXXXX.XXXXXXX>

1 Introduction

Large Language Models (LLMs) have demonstrated strong capabilities in aligning with general human intentions, underpinning their widespread deployment as helpful, honest, and harmless AI assistants across diverse real-world applications [15, 16]. Meanwhile, we still often observe failed or unexpected responses in certain situations, such as making predictions with shortcuts [19], overlooking critical user constraints [17], hallucinating contents [10], being harmful under attacks [24], and so on. These failures persist even when the model appears confident, making end users less trust on them, limiting their applications to broader scenarios. Consequently, practical deployment now demands a clearer understanding of when and why LLMs fail. These insights will help engineers to develop more robust and trustworthy LLMs.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXXX.XXXXXXX>

However, conventional explanation methods may not be practical useful for LLMs. First, LLMs are auto-regressive **generative** models, while existing explanation techniques are majorly designed for classification models [18]. When applying them to LLMs, they ignore the chain of word predictions and therefore miss long-range dependencies that drive failures. Second, LLM hidden states are **polysemantic**, meaning that each neuron encodes several unrelated concepts, so neuron-level attribution yields ambiguous or even misleading signals [1]. Third, modern LLMs are **large** in scale, containing billions of parameters and operate over thousands of context tokens. Thus, exhaustive probing or enumeration quickly becomes computationally prohibitive. These obstacles request to re-design scalable interpretation methods tailored for modern LLMs and can further integrate into their training and inference pipelines.

In this paper, particularly, we propose developing interpretation techniques to improve LLMs throughout their entire life cycle: **Data Preparation, Model Training, Model Inference, and Post-Processing**. During each stage, we will first propose an explanation method by using the available information from that stage, and then, we will develop a particular method that utilizes the collected explanations to improve their performance and overall reliability. Taken together, these objects form a unified framework that uses interpretation not just to understand LLMs, but to actively improve them at every stage of training and deploying LLMs. The goal of this research aligns with the concept of **Usable XAI**, introduced by our initial work [23], where we explored 10 strategies that make explanations useful in debugging and improving LLMs. In the following, we detail the specific problem, approach, and progress of our four research objectives.

2 Related Works

Mechanistic interpretability is one main stream to understand and explain modern LLMs. They focus on reverse-engineering the internal computations of LLMs into human-understandable circuits. Elhage et al. [7] proposed a foundational framework to decompose transformer models into interpretable components. Building upon this, Olsson et al. [14] revealed induction heads in transformers, crucial for enabling in-context learning across various model scales. Recent studies target on the polysemantic nature of LLMs, where each neuron represents multiple features [6]. To mitigate this, sparse autoencoders (SAEs) [4] have been introduced to discover more interpretable and monosemantic latent features within LLM representations. These advances emphasize the potential of mechanistic interpretation to enhance both the interpretability. This work builds on top of these findings, with a unique perspective on making explanations to be useful for more reliable and trustworthy LLMs.

3 Problem Statement

Let g_θ denote an LLM parameterized by weights θ , mapping an input prompt x into a latent representation space \mathbb{Z} . We formally define an interpretation method f as a mapping function:

$$f : \mathbb{Z} \mapsto \mathbb{H}, \quad (1)$$

where \mathbb{H} is a human-understandable representation space. In this study, the interpretable space \mathbb{H} typically corresponding to the text space $\mathbb{H} := \cup_{n=0}^{\infty} \mathcal{V}^n$ defined on the vocabulary set \mathcal{V} . Our formal definition of interpretation for LLMs aligns with numerous recognized informal definitions of the term “interpretation”, where they also emphasize the central role of humans [5, 8, 12].

4 Explanation for Post-Generation Filtering

Problem. LLMs may generate responses that are inconsistent with factual knowledge, commonly referred to as hallucinations [11]. Existing approaches typically rely on external knowledge bases or retrieval-augmented methods for fact verification [3], limiting their scalability in open-domain scenarios. Therefore, we seek a method to detect hallucinations without external dependencies.

Approach. From our pilot study, we observe that during hallucinations, LLMs disproportionately attribute prediction importance to instruction verbs, while overlooking critical instruction objects. To exploit this finding, we propose a hallucination detection approach that leverages gradient-based attribution scores mapped onto input tokens. By analyzing the distribution of attribution scores according to part-of-speech tags of input tokens, we develop a classifier to identify hallucinated responses based on internal model signals.

Results. Our preliminary results validate this approach empirically. Using a smaller model (Vicuna-7B and Mistral-7B), we successfully detect hallucinated outputs generated by a significantly larger model (ChatGPT-3.5), demonstrating the effectiveness and generality of our internal attribution-based detection method. The foundation of this work [20] has been accepted by NAACL 2024.

5 Explanation for Inference-Time Steering

Problem. LLMs sometimes produce undesirable or harmful content, particularly when subjected to maliciously designed inputs known as jailbreak prompts [24]. Traditional methods to mitigate such risks typically involve reinforcement learning from human feedback, which requires costly human supervision [16]. We aim to explore whether inference-time internal interventions could effectively steer model behaviors in a more efficient way.

Approach. To address this problem, we propose leveraging Sparse Autoencoders (SAEs) combined with mutual information-based explanations to pinpoint latent features responsible for generating safe outputs. First, we train an SAE on LLM hidden representations, producing an interpretable feature set representing semantic concepts. Subsequently, we introduce a mutual information-based objective to extract some words to explain each learned feature vector. During model inference, we zero out these identified harmful features and also enforce those safety-related features be activated, thereby steering LLM outputs away from undesired content.

Results. Our preliminary results show that inference-time steering significantly reduces jailbreak success rates on Mistral-7B from 81.6% to 72.8%. Specifically, by identifying and intervening on only

a small set of safety-related SAE features, harmful completions are notably suppressed, while minimally impacting general task performance. This highlights the practicality of using interpretable internal explanations for efficiently steering model behaviors at inference time. This work [22] is under-reviewing by EMNLP 2025.

6 Explanation for Training Regularization

Problem. Text classifiers based on embedded representations derived from latent representations of LLM may exploit unintended or spurious features, reducing robustness and fairness [2, 9]. Current mitigation approaches cannot explicitly control the usage of certain unintended features. Thus, we seek a regularization method that perform feature engineering on LLM latent spaces to control the usage of unintended features for classifiers.

Approach. We propose a self-regularization framework to improve the controllability and robustness of LLM-based classifiers. Initially, we employ SAEs to decompose LLM latent representations into human-understandable latent features. We then introduce a self-explanation strategy, leveraging advanced LLMs (e.g., GPT-4), to automatically identify unintended latent features based on their associated explanations. During classifier training, we incorporate a self-regularization objective that explicitly penalizes reliance on these automatically identified unintended features.

Results. Preliminary experiments are conducted on some challenging real-world datasets. Compared with baseline methods, classifiers trained using our framework show reduced sensitivity to spurious correlations while maintaining high task accuracy. This work [21] has been accepted by KDD 2025.

7 Explanation for Synthetic Data Generation

Problem. Many empirical works have shown that the diversity of dataset is critical to train a more robust LLM [25]. However, current studies [13] focus on selecting diverse samples from a large source dataset, naturally ignoring the features that absent from the source dataset. Therefore, we seek a method to generate synthetic data covering as diverse concepts as possible to train LLMs.

Approach. To start with, we first train a SAE for a pre-trained LLM to analyze all encoded features within that model. We then identify the features that are missing from the source dataset by projecting the latent representations of the source data on our trained SAE. Given a missed feature from SAE, we could leverage an LLM to modify a source data to satisfy the missed feature. The modified samples and the source data will jointly train the pre-trained LLM.

Progress. We are developing the theoretical foundation to solidify the proposed approach. This work is target at ICLR 2026.

8 Conclusion

In this paper, we propose a unified interpretation framework that leverages explanations to actively improve LLMs throughout their entire life cycle. By systematically integrating interpretable methods into data preparation, model training, model inference, and post-processing stages, we demonstrate preliminary success in detecting hallucinations, defending jailbreak attacks, mitigating shortcut biases, and enhancing dataset diversity. The proposed interpretation-driven strategies lay a robust foundation for practical deployment of trustworthy LLMs in broader scenarios.

References

- [1] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics* 6 (2018), 483–495.
- [2] Trenton Bricken, Jonathan Marcus, Siddharth Mishra-Sharma, Meg Tong, Ethan Perez, Mrinank Sharma, Kelley Rivoire, Thomas Henighan, and Adam Jermyn. 2024. Using Dictionary Learning Features as Classifiers. <https://transformer-circuits.pub/2024/features-as-classifiers/index.html>.
- [3] I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. FacTool: Factuality Detection in Generative AI—A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528* (2023).
- [4] Katharine Cunningham et al. 2023. Sparse Autoencoders Find Highly Interpretable Directions in Transformers. <https://arxiv.org/abs/2301.04709>
- [5] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608* (2017).
- [6] Nelson Elhage, Neel Nanda, Catherine Olsson, et al. 2022. Toy Models of Superposition. *Transformer Circuits Thread* (2022). https://transformer-circuits.pub/2022/toy_model/index.html
- [7] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Das-Sarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread* (2021). <https://transformer-circuits.pub/2021/framework/index.html>.
- [8] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *Proceedings of the 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 80–89.
- [9] Max Hort, Zhenpeng Chen, Jie M Zhang, Mark Harman, and Federica Sarro. 2024. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing* 1, 2 (2024), 1–52.
- [10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.
- [11] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys* 55, 12 (2023), 1–38.
- [12] Zachary C Lipton. 2018. The Mythos of Model Interpretability. *Commun. ACM* 61, 10 (2018), 36–43.
- [13] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. [n. d.]. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *The Twelfth International Conference on Learning Representations*.
- [14] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, et al. 2022. In-Context Learning and Induction Heads. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>
- [15] R OpenAI. 2023. GPT-4 technical report. *arXiv* (2023), 2303–08774.
- [16] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
- [17] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating Instruction Following Ability in Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*. 13025–13048.
- [18] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
- [19] Xuansheng Wu, Padmaja Pravin Saraf, Gyeonggeon Lee, Ehsan Latif, Ninghao Liu, and Xiaoming Zhai. 2025. Unveiling scoring processes: Dissecting the differences between llms and human graders in automatic scoring. *Technology, Knowledge and Learning* (2025), 1–16.
- [20] Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. 2024. From Language Modeling to Instruction Following: Understanding the Behavior Shift in LLMs after Instruction Tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2341–2369.
- [21] Xuansheng Wu, Wenhao Yu, Xiaoming Zhai, and Ninghao Liu. 2025. Self-regularization with latent space explanations for controllable llm-based classification. *arXiv preprint arXiv:2502.14133* (2025).
- [22] Xuansheng Wu, Jiayi Yuan, Wenlin Yao, Xiaoming Zhai, and Ninghao Liu. 2025. Interpreting and steering llms with mutual information-based explanations on sparse autoencoders. *arXiv preprint arXiv:2502.15576* (2025).
- [23] Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, et al. 2024. Usable XAI: 10 strategies towards exploiting explainability in the LLM era. *arXiv preprint arXiv:2403.08946* (2024).
- [24] Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. 2024. LLM Jailbreak Attack versus Defense Techniques—A Comprehensive Study. *arXiv e-prints* (2024), arXiv–2402.
- [25] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems* 36 (2023), 55006–55021.