Unraveling the Influence of Training Data and Internal Structures in Large Language Models for Enhanced Explainability (Student Abstract)

Lingfang Li[1], **Procheta Sen**[1]

[1]Department of Computer Science, University of Liverpool, Ashton Street, Liverpool, L69 3BX, United Kingdom
{sglli24,procheta.sen}@liverpool.ac.uk

Abstract

Recent advances in deep learning have expanded the application of large language models (LLMs) across fields such as medicine, finance, and education. Understanding the mechanisms underlying these models is essential to mitigate issues like hallucinations and bias. This study provides deep learning practitioners with insights into how specific training data points and internal structures influence model behaviour. Using influence functions and mechanistic interpretability, we will analyze the impact of data on model predictions across various tasks. Preliminary findings indicate that semantic search techniques, such as FAISS, enable efficient identification of influential training points in GPT-2 small. Future work will extend these methods to additional tasks and more complex models, with a focus on further elucidating LLM structures to improve interpretability.

Introduction

LLMs have gained widespread adoption across various domains, largely due to the transformer architecture introduced in Attention is All You Need (Vaswani et al. 2017). However, as LLMs are increasingly deployed in fields such as finance and medicine, where the benchmarks for evaluating results are often ambiguous, there is an urgent need to understand their prediction behaviour and internal mechanisms better. This understanding is essential for interpreting how these models make decisions and fostering trust in their use, especially in critical applications.

Several studies have explored explanations for LLMs, such as analyzing token importance (Goldshmidt and Horovicz 2024; Tanneru, Agarwal, and Lakkaraju 2023; Mohanty 2023), generating contrastive explanations (Yin and Neubig 2022; Yao 2024; Liu et al. 2024), and using attention as an explanation (Bibal et al. 2022; Ethayarajh and Jurafsky 2021; Rigotti et al. 2022; Sun and Marasović 2021). However, limited research investigates the roles of internal components or training data in mitigating harmful effects, like hallucinations (Ji et al. 2023; Manakul, Liusie, and Gales 2023; Huang et al. 2023; Perković, Drobnjak, and Botički 2024; Xu, Jain, and Kankanhalli 2024; McKenna et al. 2023; Farquhar et al. 2024)—where LLMs produce factually incorrect information—and biases related to demographic factors like gender and nationality (Gallegos et al. 2024). Few studies have explored the relationships between training data, model architecture, and performance. Additionally, while the importance of attention heads and layers in LLMs has been studied (McDougall et al. 2024), their generalizability, robustness, and effectiveness remain unclear. Depending on their effectiveness, there could be potential ways to develop LLMs that address these shortcomings. In this research, we aim to focus specifically on these underexplored areas.

The end users of the explanation discussed above are the deep learning practitioners. The primary motivation for explaining LLMs with the above-mentioned components is to develop more effective LLMs for different kinds of tasks. For example, hallucination and bias are of the major issues in LLMs. If for a particular type of task (e.g. classification) we can identify the internal components responsible for hallucination, then it can enable the design of LLMs where the issue of hallucination can be reduced.

Problem Formulization

This abstract is centred around three main research questions:

Training LLMs involves vast datasets, making it computationally challenging to assess the influence of each data point. This leads to our first research question:

- **RQ1:** How effectively can we estimate the influence of individual training data points in LLMs?

LLMs are generally used in three different modes a) pretrained, b) fine-tuned and c) in-context learning. Performance varies across these modes for different tasks, yet there is no clear understanding of why a model excels in a particular mode. This prompts our second question:

- **RQ2:** How can we explain the role of training data in LLM performance across different usage modes (pretraining, fine-tuning, and in-context learning)?

LLMs generally perform well across tasks like classification and summarization, showing strong generalization capabilities. However, no methodology currently explains this generalizability. To address this, we need to identify critical components for various tasks, leading to our third question:

			Evaluation	
Model	Dataset	NN Type	Before- Influence	After-Influence
GPT-2 Small	Movie Review	Random	82%	84%
GPT-2 Small	Movie Review	Tf-IDf	82%	82%
GPT-2 Small	Movie Review	FAISS	82%	80%

Table 1: Faithfulness of Influence Functions With the variation of Nearest Neighbor Technique

- **RQ3:** What key internal components of LLMs enable their generalization across different tasks?

Methodology

For RQ1 and RQ2, we utilize influence functions to evaluate the importance of individual training data points on model predictions and guide our experimental design.

To address the computational challenges in RQ1, we reduce the training dataset size while maintaining the relevance of identified points. Traditional methods (Koh and Liang 2017) are computationally expensive due to Hessian matrix calculations, and alternatives often depend on the computing progress and model architecture. We propose using the Facebook AI Similarity Search (FAISS) to identify training samples most similar to test samples. This narrows influence computation to a relevant subset rather than the entire dataset, improving efficiency without sacrificing accuracy.

We further simplify computations by approximating large LLMs with smaller models (fewer layers and parameters) trained on subsets identified by FAISS. These simpler models can then be tested for their capacity to approximate larger LLMs, as measured by influence functions. Though simpler models may not fully replicate complex LLMs, this approach allows us to explore the potential of computationally efficient alternatives while evaluating data point influence across training stages for deeper insights. We will evaluate the influence of data points across different training stages, offering deeper insights into their impact.

For RQ2, we aim to explain how training data impacts LLM performance across usage modes by measuring LLM predictions before and after removing the most influential samples. Each influence value of the sample, as calculated by influence functions, will reveal its importance. By observing the drop in accuracy following the removal of these samples, we can determine their significance. The explanation will highlight influential samples, the overlap of top-k influential samples, and variations in influence values across usage modes, clarifying why LLMs perform better in certain modes for specific tasks.

For RQ3, we will employ mechanistic interpretability with a focus on edge attribution(Conmy et al. 2023; Hanna, Pezzelle, and Belinkov 2024) to reverse-engineer LLM processes. This approach aims to identify internal structures—such as neurons, layers, or attention heads—responsible for specific functions and behaviors (Bereska and Gavves 2024). By analyzing these components, we will investigate LLM generalization and adaptability. The explanation will highlight critical elements and their combinations, offering insights into how these components contribute to model performance and behavior.

Potential Results

Table 1 illustrates the performance of influence functions with different nearest neighbor techniques. Due to the high computational cost of calculating influence functions for all training points, we explored random selection, TF-IDF, and Approximate Nearest Neighbor (FAISS) (Johnson, Douze, and Jégou 2019). Each method identified the top 100 nearest neighbors of each test sample within the training set, and influence functions were computed on these subsets.

The results demonstrate notable differences in the impact of these methods on model performance. Randomly removing 1% of the training data increased accuracy by 2%, likely due to the elimination of noisy or non-informative samples. The removal of samples identified by Term Frequency–Inverse Document Frequency (TF-IDF), based on word-level similarity, showed no measurable effect, indicating these samples were not critical. However, removing samples selected by FAISS, which leverages semantic similarity, led to a 2% decrease in accuracy. This suggests semantically relevant training samples are crucial for maintaining performance and highlights FAISS's effectiveness in isolating influential data points. While a 2% drop may seem small, its disproportionate impact underscores the model's reliance on specific data points rather than the entire training set, demonstrating GPT-2 Small's ability to learn semantic associations during training.

Conclusion and Future Work

This study demonstrates the potential of influence functions to analyze the impact of training data on LLM predictions, particularly in fine-tuned transformers like GPT-2. Our results show that for smaller models, FAISS significantly improves the efficiency of influence function computation compared to existing methods. Furthermore, our findings reveal that LLM predictions are predominantly influenced by semantically related training points, providing a foundation for investigating issues such as bias and hallucination.

Future work will extend this approach to more complex models and larger datasets. Additionally, we aim to explore mechanistic interpretability to identify key internal components responsible for tasks such as machine translation and emotion recognition.

Acknowledgments

This research was supported by the University of Liverpool and the Engineering and Physical Sciences Research Council (EPSRC). We thank the Department of Computer Science at the University of Liverpool for providing access to computational resources. The authors also acknowledge the anonymous reviewers for their constructive suggestions.

References

Bereska, L.; and Gavves, E. 2024. Mechanistic Interpretability for AI Safety–A Review. *arXiv preprint arXiv:2404.14082*.

Bibal, A.; Cardon, R.; Alfter, D.; Wilkens, R.; Wang, X.; François, T.; and Watrin, P. 2022. Is Attention Explanation? An Introduction to the Debate. 3889–3900.

Conmy, A.; Mavor-Parker, A. N.; Lynch, A.; Heimersheim, S.; and Garriga-Alonso, A. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. arXiv:2304.14997.

Ethayarajh, K.; and Jurafsky, D. 2021. Attention Flows are Shapley Value Explanations. arXiv:2105.14652.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and Fairness in Large Language Models: A Survey.

Goldshmidt, R.; and Horovicz, M. 2024. TokenSHAP: Interpreting Large Language Models with Monte Carlo Shapley Value Estimation.

Hanna, M.; Pezzelle, S.; and Belinkov, Y. 2024. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms. arXiv:2403.17806.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 1–38.

Johnson, J.; Douze, M.; and Jégou, H. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3): 535–547.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, 1885–1894. PMLR.

Liu, F.; Xu, P.; Li, Z.; Feng, Y.; and Song, H. 2024. Towards Understanding In-Context Learning with Contrastive Demonstrations and Saliency Maps. arXiv:2307.05052.

Manakul, P.; Liusie, A.; and Gales, M. J. F. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. .

McDougall, C. S.; Conmy, A.; Rushing, C.; McGrath, T.; and Nanda, N. 2024. Copy Suppression: Comprehensively Understanding an Attention Head.

McKenna, N.; Li, T.; Cheng, L.; Hosseini, M. J.; Johnson, M.; and Steedman, M. 2023. Sources of Hallucination by Large Language Models on Inference Tasks. arXiv:2305.14552.

Mohanty, S. 2023. A Surprisingly Effective Way to Estimate Token Importance in LLM Prompts. https://www.watchful.io/blog/a-surprisingly-effective-way-to-estimate-token-importance-in-llm-prompts. Accessed: 2024-11-16.

Perković, G.; Drobnjak, A.; and Botički, I. 2024. Hallucinations in LLMs: Understanding and Addressing Challenges. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, 2084–2088.

Rigotti, M.; Miksovic, C.; Giurgiu, I.; Gschwind, T.; and Scotton, P. 2022. Attention-based Interpretability with Concept Transformers. In *International Conference on Learning Representations*.

Sun, K.; and Marasović, A. 2021. Effective Attention Sheds Light On Interpretability. arXiv:2105.08855.

Tanneru, S. H.; Agarwal, C.; and Lakkaraju, H. 2023. Quantifying Uncertainty in Natural Language Explanations of Large Language Models. arXiv:2311.03533.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xu, Z.; Jain, S.; and Kankanhalli, M. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv:2401.11817.

Yao, L. 2024. Large Language Models are Contrastive Reasoners. arXiv:2403.08211.

Yin, K.; and Neubig, G. 2022. Interpreting Language Models with Contrastive Explanations. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 184–198. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.