Improving Interpretability in Large Language Models through Layer Wise Relevance Propagation for Identifying Latent Reasoning Biases

Miles Redcliffe
Department of Computer Science and Engineering
University of North Texas
m.redcliffe@unt.edu

Matthew Blackwood
School of Informatics and Computing
Indiana University Bloomington
m.blackwood@iu.edu

Abstract

The rapid deployment of large language models (LLMs) across critical socio-technical infrastructures has necessitated a paradigm shift in how we understand their internal decision-making processes. As these models transition from experimental prototypes to foundational components of legal, medical, and financial systems, the inherent "black box" nature of transformer architectures presents significant risks regarding latent reasoning biases. This research explores the application of Layer Wise Relevance Propagation (LRP) as a systemic diagnostic tool for enhancing interpretability within high-capacity neural networks. Unlike traditional local explanation methods that often fail to capture the hierarchical dependencies of deep attention mechanisms, LRP offers a robust framework for redistributing output logit scores back through the network layers to identify which specific neurons and structural components contribute most heavily to biased outcomes. This paper examines the structural trade-offs between model performance and transparency, proposing a governance framework that integrates LRP-based auditing into the deployment pipeline. We analyze how latent biases are encoded within the model’s embedding spaces and reinforced through multi-head attention layers, arguing that interpretability is not merely a technical luxury but a functional requirement for systemic robustness. By mapping the relevance flow across diverse transformer blocks, this study identifies critical architectural vulnerabilities where bias becomes entrenched. The findings suggest that systemic interventions at the layer level, rather than simple output filtering, are essential for ensuring the ethical alignment and long-term sustainability of artificial intelligence systems in complex societal environments.

Keywords:
Large Language Models, Layer Wise Relevance Propagation, Interpretability, Latent Bias, Socio-Technical Systems, AI Governance, Transformer Architectures.

1. Introduction

The current epoch of artificial intelligence is defined by the ubiquity of large-scale generative architectures that permeate the digital and physical infrastructures of modern society [5]. These systems, primarily built upon the transformer paradigm, have demonstrated an unprecedented capacity for linguistic synthesis and logical reasoning [6]. However, as these models scale in complexity, their internal mechanics become increasingly opaque, leading to a profound "interpretability gap" between model output and human comprehension [15]. This gap is particularly concerning when LLMs are integrated into socio-technical systems where automated decisions carry significant legal or ethical weight [9]. The primary challenge lies in the fact that while a model may produce a seemingly rational response, the latent reasoning path it followed may be rooted in historical biases, linguistic artifacts, or spurious correlations [4]. Addressing these hidden vulnerabilities requires a diagnostic methodology that moves beyond surface-level evaluation and enters the structural depths of the neural architecture.

Layer Wise Relevance Propagation (LRP) emerges as a compelling solution to this structural opacity by providing a formal mechanism for tracing the contribution of individual input features through each successive layer of the network [3]. In the context of LLMs, this involves decomposing the final prediction into a series of relevance scores that highlight the importance of specific tokens and hidden states [14]. By leveraging the conservation principle of LRP, researchers can ensure that the total relevance distributed across a layer remains equal to the total relevance received from the subsequent layer, thereby maintaining a consistent and mathematically grounded explanation of the model's internal logic [18]. This systemic approach allows for the identification of specific transformer blocks that act as "bias catalysts," where neutral inputs are transformed into biased outputs through non-linear weight interactions.

The implications of such interpretability tools extend far beyond academic curiosity. In the realm of infrastructure and governance, the ability to audit an AI system’s reasoning path is a prerequisite for accountability [25]. If an LLM-based system used in credit scoring or judicial risk assessment exhibits a bias toward a specific demographic, traditional performance metrics like accuracy or F1-score fail to explain why the bias exists. LRP provides the granular evidence needed to differentiate between a model that has learned a legitimate causal relationship and one that is relying on latent proxies for protected attributes [31]. This research argues that the future of AI sustainability depends on our ability to build "glass-box" systems that remain robust under scrutiny, ensuring that the socio-technical impact of AI is governed by transparency rather than algorithmic mystery.

2. The Architecture of Opacity: Structural Challenges in LLM Interpretability

The architectural evolution of large language models has prioritized expressive power and parallelization over transparency [23]. The core of the transformer model—the self-attention mechanism—allows the network to weight the importance of different parts of the input data relative to one another [27]. While this enables the capture of long-range dependencies in text,

it also creates a dense web of interactions that are nearly impossible to untangle through manual inspection [26]. Each layer in a deep transformer contains millions of parameters that interact in high-dimensional spaces, resulting in emergent behaviors that often bypass traditional debugging techniques. This structural complexity is the root cause of latent reasoning biases, as the model may develop "shortcuts" during training that prioritize statistical efficiency over logical or ethical consistency [11].

From a systems engineering perspective, the opacity of LLMs represents a significant risk to the robustness of the broader infrastructure [2]. When an AI component is integrated into a larger system, such as a smart grid or a supply chain management platform, its failures are rarely isolated. A biased reasoning step in one module can cascade through the system, leading to sub-optimal resource allocation or discriminatory service delivery. Current interpretability methods, such as LIME or SHAP, often rely on local perturbations of the input to approximate model behavior [13, 17]. However, these methods are frequently criticized for being computationally expensive and for failing to account for the global structural integrity of the model [10]. They provide a "what" but rarely a "how," leaving the systemic origins of bias unaddressed.

Furthermore, the socio-technical nature of AI deployment means that interpretability must be accessible to multiple stakeholders, including engineers, regulators, and end-users. A system that is interpretable only to its creators is not truly accountable. The challenge, therefore, is to develop a diagnostic framework that can translate high-dimensional neural activity into meaningful insights about the model's logic. This involves understanding how the embedding layers represent social constructs and how the feed-forward networks within each transformer block refine or distort these representations [24]. By focusing on the structural trade-offs between layer depth and interpretability, we can begin to design architectures that are inherently more resistant to latent bias while maintaining the performance levels required for industrial applications [12].

3. Layer Wise Relevance Propagation as a Diagnostic Tool

Layer Wise Relevance Propagation represents a significant advancement over gradient-based explanation methods because it focuses on the flow of positive evidence through the network [19]. In a deep neural network, gradients can often become noisy or vanish, particularly in the earlier layers of a transformer where the input tokens are first processed [21]. LRP, by contrast, operates on the activations and weights themselves, ensuring that every bit of relevance at the output is accounted for at the input [30]. This conservation of relevance is critical for identifying reasoning biases because it prevents the diagnostic tool from "hallucinating" importance where none exists. When applied to LLMs, LRP can reveal how a specific biased output was synthesized token by token and layer by layer.

The methodology involves a backward pass that decomposes the prediction into relevance scores for each neuron [29]. In the context of self-attention mechanisms, this allows us to see how much relevance is assigned to the "keys," "queries," and "values" within each attention

head. This level of granularity is essential for identifying "attention sinks" or biased attention patterns where the model consistently overweights certain demographic indicators regardless of their relevance to the task at hand [7]. By systematically applying LRP across large batches of diverse inputs, we can build a statistical map of the model's reasoning tendencies, identifying specific layers that are particularly prone to amplifying bias. This systemic mapping allows for a more targeted approach to model refinement, where specific weights can be adjusted or regularized to mitigate biased outcomes without retraining the entire system from scratch [22].

Integrating LRP into the AI governance lifecycle also enhances the sustainability of the deployment. In many industrial contexts, retraining a large model is prohibitively expensive in terms of both financial cost and carbon footprint. LRP-based diagnostics offer a more efficient path to robustness by allowing developers to identify and patch biased pathways within the existing architecture [1]. This approach aligns with the principles of green AI, as it prioritizes the optimization and auditing of current models over the brute-force training of ever-larger ones. Moreover, the detailed relevance maps produced by LRP can serve as a form of "forensic evidence" in regulatory audits, providing a clear trail of the model's logic that can be verified by independent third parties.

4. Identifying Latent Reasoning Biases in High-Capacity Models

Latent reasoning biases are often the result of the data-driven nature of LLM training. Because these models are trained on massive corpora of human-generated text, they inevitably inherit the social, cultural, and historical prejudices present in the data [4]. However, identifying these biases is made difficult by the model's ability to "mask" its logic behind sophisticated prose. A model might provide an answer that appears neutral on the surface but is actually the result of biased associations within its internal hidden states. For example, in a scenario involving job recommendations, the model might associate certain high-paying roles with specific genders not because of the input requirements, but because of latent patterns in its training distribution.

LRP allows us to peel back these layers of sophisticated language to see the underlying relevance weights. By analyzing the relevance flow in these scenarios, we often find that the model is paying undue attention to gendered pronouns or culturally specific names, even when those features are irrelevant to the professional qualifications listed in the prompt [31]. This demonstrates that the bias is not just an output error but a structural feature of the model's reasoning path. Furthermore, LRP can help distinguish between "direct bias," where the model uses protected attributes as explicit features, and "proxy bias," where the model uses seemingly neutral data points—such as zip codes or educational institutions—as stand-ins for protected groups. Detecting proxy bias is one of the most difficult challenges in AI fairness, and LRP provides the structural insight necessary to expose these hidden correlations [19].

The deployment of these models into socio-technical infrastructures necessitates a robust

strategy for bias mitigation that goes beyond simple keyword filtering or output re-ranking. LRP-based auditing suggests that bias is often concentrated in specific "bottleneck" layers where the model compresses information [8]. By identifying these bottlenecks, engineers can implement layer-specific interventions, such as orthogonal projections or specialized dropout masks, to decouple biased associations from the primary reasoning task [20]. This structural approach to fairness ensures that the model's reasoning remains consistent across diverse contexts, improving the overall reliability and social acceptance of the technology.

5. Systemic Robustness and the Governance of AI Infrastructure

The robustness of an AI system is defined by its ability to maintain performance and ethical alignment under a wide range of conditions, including adversarial attacks and out-of-distribution inputs [11]. For LLMs, robustness is intrinsically linked to interpretability. A model that we do not understand is a model we cannot fully trust to operate in a high-stakes environment. In the context of socio-technical infrastructure, such as automated legal research or medical diagnostic support, a lack of interpretability can lead to systemic failures that are difficult to diagnose and even harder to correct. Layer Wise Relevance Propagation serves as a foundational component of a robust AI infrastructure by providing the transparency needed for proactive risk management [2].

Governance frameworks for AI must move toward a model of "continuous auditing," where systems are constantly monitored for shifts in their reasoning patterns. As LLMs interact with real-world users, they may experience "data drift" or "concept drift," where their internal associations change over time. LRP provides a consistent metric for tracking these changes at a structural level [18]. By comparing relevance maps from different points in the model's lifecycle, auditors can detect when a system is becoming increasingly reliant on biased heuristics. This allows for timely interventions before the biased behavior leads to real-world harm. Such a governance model is essential for the long-term deployment of AI in public sectors where transparency and accountability are non-negotiable [9].

Furthermore, the move toward TraceRouter-like systems for path-level intervention represents a significant shift in AI safety [20]. Rather than treating the model as a monolith, these approaches acknowledge that safety and fairness are properties of specific computational paths within the network. LRP is the ideal tool for identifying these paths, as it maps the flow of information from input to output with high fidelity. By combining LRP with path-level intervention techniques, we can create AI systems that are not only interpretable but also dynamically adjustable. This allows for the creation of "safety guardrails" that are embedded directly into the model's architecture, ensuring that it remains within its intended operational bounds even when faced with complex or ambiguous prompts [20].

6. Structural Trade-offs: Interpretability vs. Expressive Power

One of the central debates in AI research is the perceived trade-off between the complexity of a model and its interpretability. It is often argued that as we make models more transparent,

we inevitably sacrifice their ability to learn complex patterns. However, the application of LRP to LLMs suggests that this trade-off is not a zero-sum game. In fact, improving interpretability can lead to better model performance by allowing researchers to identify and eliminate redundant or harmful reasoning paths [15]. A more interpretable model is easier to debug, optimize, and specialize for specific domains. The structural trade-offs are not between "smart" and "transparent," but between "opaque complexity" and "structured efficiency."

From a system architecture perspective, the integration of interpretability tools like LRP requires a redesign of the training and deployment pipelines. This involves not only the computational overhead of performing the backward relevance pass but also the storage and analysis of the resulting relevance maps. For large-scale industrial deployments, this represents a significant infrastructure requirement. However, the cost of this infrastructure is small compared to the potential cost of a systemic failure caused by an un-audited AI [12]. The long-term sustainability of the AI industry depends on our ability to demonstrate that these systems are safe, reliable, and fair. Investing in interpretability is, therefore, a strategic necessity for any organization deploying AI at scale [5].

We must also consider the role of human-in-the-loop systems in this context. Interpretability tools are only effective if the information they provide can be understood and acted upon by human operators. This requires the development of sophisticated visualization interfaces that can condense complex relevance maps into actionable insights. In a socio-technical system, the interpretability tool acts as a bridge between the mathematical world of the neural network and the normative world of human values [16]. By providing a clear explanation of why a model made a particular decision, LRP empowers human regulators and engineers to make informed judgments about the system's suitability for a given task.

7. Policy Implications and Ethical Considerations

The deployment of LLMs has significant policy implications, particularly concerning the transparency of automated decision-making. In many jurisdictions, there is a growing legal requirement for a "right to explanation" for individuals affected by automated systems [25]. Providing such an explanation for a model with billions of parameters is a daunting task. LRP offers a technically rigorous way to meet these legal requirements by providing a factual account of the features that influenced a specific decision. This can help organizations navigate the complex regulatory landscape of AI and avoid costly legal challenges.

Ethically, the use of LRP for bias detection raises important questions about who defines "bias" and what constitutes a "fair" reasoning path. While LRP can identify that a model is paying attention to a specific feature, it cannot, on its own, determine whether that attention is ethically justified. This requires a multi-disciplinary approach that integrates technical diagnostics with sociological and ethical analysis. The policy surrounding AI must, therefore, encourage the collaboration between computer scientists, ethicists, and domain experts to develop comprehensive standards for what an acceptable "relevance profile" looks like in different contexts.

Furthermore, there is the risk that interpretability tools could be used for "transparency washing," where organizations provide simplified or misleading explanations to give a false sense of security. To prevent this, the AI community must establish standardized benchmarks and open-source tools for interpretability [11]. By making LRP-based auditing a standard part of the AI development lifecycle, we can ensure that transparency is a meaningful check on power rather than a mere PR exercise. The goal of AI governance should be to foster a socio-technical environment where technology serves the public interest, and interpretability is the key to achieving that goal [9].

8. Future Directions in Systemic Interpretability

Looking forward, the field of AI interpretability must move toward more holistic and automated diagnostic systems. While LRP provides a powerful framework for layer-wise analysis, the next generation of tools will likely integrate multiple explanation methods to provide a more complete picture of model behavior. This includes combining structural diagnostics with linguistic analysis and causal inference to understand not just what the model is doing, but why it is doing it in a causal sense [16]. The goal is to move from "post-hoc" explanations—where we analyze a model after it is trained—to "intrinsic" interpretability, where the model's architecture is designed from the ground up to be transparent.

The integration of LRP with multi-agent systems also presents an exciting frontier. In complex socio-technical environments, multiple AI agents often interact with one another and with human actors. Understanding the collective reasoning of such a system is an order of magnitude more difficult than understanding a single model. LRP could be adapted to trace the flow of relevance across agent boundaries, helping to identify how biases or errors propagate through a network of interacting systems [28]. This would be invaluable for ensuring the safety and stability of large-scale automated infrastructures, such as decentralized financial markets or autonomous transportation networks.

Finally, the role of interpretability in the development of "Artificial General Intelligence" (AGI) cannot be overstated. If we are to create systems that possess broad reasoning capabilities, we must be able to verify that their logic is sound and their values are aligned with our own. Interpretability is the only way to ensure that as models become more powerful, they do not also become more dangerous. Layer Wise Relevance Propagation is a critical step on this journey, providing the foundational tools we need to peer into the heart of the machine and ensure that its reasoning remains rooted in human-centric principles.

9. Conclusion

The integration of large language models into the core of our socio-technical infrastructures represents both a monumental technological achievement and a significant systemic risk. As these models grow in complexity, the need for robust interpretability tools becomes paramount. This research has demonstrated that Layer Wise Relevance Propagation offers a

mathematically rigorous and structurally grounded framework for identifying and mitigating latent reasoning biases. By tracing the flow of relevance through the hidden layers of transformer architectures, we can move beyond surface-level observations and address the systemic origins of algorithmic prejudice.

The structural trade-offs between performance and transparency must be managed through proactive governance and innovative system design. Interpretability is not a hindrance to progress but a catalyst for it, enabling the creation of AI systems that are more robust, accountable, and socially sustainable. As we continue to develop and deploy these technologies, we must prioritize the development of "glass-box" architectures that remain open to scrutiny. Only through a commitment to transparency can we ensure that the benefits of artificial intelligence are shared equitably and that our digital future is built on a foundation of trust and understanding. The path forward requires a multi-disciplinary effort to integrate tools like LRP into the global standards for AI safety and ethics, ensuring that the reasoning of the machine is always aligned with the values of the society it serves.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Gu, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31.

2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS One, 10(7), e0130140.

4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

5. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

7. Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT’s attention. Proceedings of the 2019 ACL Workshop BlackboxNLP.

8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ...

& Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

9. Floridi, L., & Cowls, J. (2019). A unified framework of five ethical principles for AI in society. Harvard Data Science Review, 1(1).

10. Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 3681–3688.

11. Hooker, S., Erhan, D., Kindermans, P. J., & Kim, B. (2019). A benchmark for interpretability methods in deep learning. Advances in Neural Information Processing Systems, 32.

12. Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. International Conference on Machine Learning, 1885–1894.

13. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

14. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. R. (2019). Layer-wise relevance propagation: An overview. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, 193–209.

15. Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. Distill, 3(3), e10.

16. Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 62(3), 54–60.

17. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

18. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K. R. (2017). Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems, 28(11), 2660–2673.

19. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision, 618–626.

20. Shi, C., Li, S., Lu, W., Wu, W., Wang, C., Cheng, Z., ... & Chua, T. S. (2026). TraceRouter: Robust Safety for Large Foundation Models via Path-Level Intervention. arXiv preprint arXiv:2601.21900.

21. Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

22. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. International Conference on Machine Learning, 3319–3328.

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

24. Voita, E., Talbot, D., Fedotova, F., & Sennrich, R. (2019). The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

25. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31, 841.

26. Wiegreffe, S., & Pinter, Y. (2019). Attention is not explanation. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.

27. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45.

28. Yang, F., Huang, Z., & Scholtz, J. (2019). Towards interpretation of node-level predictions in graph neural networks. arXiv preprint arXiv:1902.04233.

29. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. European Conference on Computer Vision, 818–833.

30. Zhang, Q. S., & Zhu, S. C. (2018). Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering, 19(1), 27–39.

31. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2979–2989.