Using Novel Phrase-Level Explanations in a Softmax-Linked Additive Explainability Model for Transformers

Neil Gupta

Massachusetts Academy of Math and Science @ Worcester Polytechnic Institute

Worcester, MA

Author Note

I thank Dr. Kevin Crowthers for assisting me in the brainstorming, development, and implementation of my project.

Executive Summary

The growth of transformer-based language models in important domains such as healthcare, law, and finance has created an urgent need for interpretability and understanding, in order to feel confident about results, thus resulting in the emergence of the field of explainable artificial intelligence(XAI). Existing XAI methods attempt to explain the predictions of transformers, they suffer from a fundamental limitation: they assume the additive property of most neural networks also holds true for transformers which isn't the case, thus rendering them incompatible with the transformer softmax architecture. State-of-the-art methods like the Softmax-Linked Additive Log Odds Model (SLALOM) are designed specifically for the transformer architecture, yet they also contain an inherent flaw that they operate exclusively at the token level. This token-level granularity creates a critical gap between how models explain decisions and how humans understand language. This project proposes to extend the existing operations of SLALOM at the token-level to the phrase-level to help take into account the higher-order contexts and meanings of human language. By preserving SLALOM's guarantees of transformer compatibility while addressing its acknowledged limitations, this work enables trustworthy AI deployment in high-stakes applications.

Keywords: Transformers, Softmax, Feature Attribution, Explainable AI, Tokens, Interpretability

Intro/Need, Methods, Results, and

Discussion

Using Novel Phrase-Level Explanations in a Softmax-Linked Additive Explainability Model for Transformers

Natural Language Processing

The industry of artificial intelligence has been fundamentally transformed by the creation of transformer architectures and large language models. Since the introduction of ChatGPT in November of 2022, generative AI applications have achieved groundbreaking adoption, reaching 100 million users faster than any technology in history. These systems, built upon the transformer architecture introduced by Vaswani et al. (2017) have demonstrated amazing capabilities across a plethora of diverse natural language processing tasks. The transformer's so-called “self-attention” mechanism allows the model to weigh the importance of different words in a sentence simultaneously, rather than processing them sequentially as earlier recurrent neural networks did, enabling both a better performance and a more efficient parallelization during training (Vaswani et al., 2017). The impact of transformers extends far beyond chatbots and consumer applications. In healthcare, transformer-based models have a variety of different impacts, ranging from analyzing medical records to predicting patient outcomes (Kalyan et al., 2024). Legal professionals employ these systems to review contracts, identify precedents, and draft legal documents. Financial institutions also utilize transformers for tasks such as fraud detection and risk assessment (Financial Times Visual Journalism Team, 2025). The widespread deployment of these models in high stakes domains, however, introduces a critical challenge: these systems function as "black boxes," making decisions through complex mathematical operations involving billions of parameters without providing any transparent rationale for their outputs (Adadi & Berrada, 2018). This opacity conflicts with fundamental requirements for human trust.

The Critical Need for Explainable Artificial Intelligence

The field of Explainable Artificial Intelligence (XAI) has emerged to address the needs for greater transparency in AI systems, developing methods to make the decision-making processes of complex models transparent and interpretable to humans (Van Lent et al., 2004). The goal is to create explanations that describe which input features, in the case of language models, which words or tokens, most strongly influenced the model's output. These explanations allow users to verify that models are making reasonable decisions and build their trust in systems that are deployed in important applications.

Feature Attribution Methods for Transformers

One central approach to XAI is feature attribution, which assigns importance scores to input features based on their contribution to model predictions. For language models, feature attribution identifies which words or tokens in the input most strongly influenced the output. For instance, in sentiment analysis (determining whether a text expresses positive or negative emotion), feature attribution might reveal that the word "excellent" contributed positively to a positive prediction, while "terrible" contributed negatively. Multiple approaches to feature attribution have been developed. Gradient-based methods compute how changes to input features affect the model's output by calculating mathematical derivatives called gradients (Azarkhalili & Libbrecht, 2025). Attention-based methods leverage information about which words the model "attends to" during processing (Abnar & Zuidema, 2020). Surrogate model approaches create simpler, more interpretable models that approximate the complex model's behavior, and then extract explanations from these simpler models. Two widely-used surrogate model approaches are LIME (Local Interpretable Model-agnostic Explanations), introduced by Ribeiro et al. (2016), and SHAP (Shapley Additive exPlanations), proposed by Lundberg and Lee (2017). Both methods work by creating simplified models that can be easily interpreted. However, both LIME and SHAP operate under a fundamental assumption: they assume that model predictions can be approximated as additive combinations of input features, meaning that each feature contributes independently to the final prediction without interactions with other features. For

many types of models, this assumption is reasonable. However, transformers have a specific architectural feature that makes this assumption problematic.

Transformers and Softmax Normalization

To understand why traditional explanation methods fail for transformers, it is necessary to understand how transformers process information. The main mechanism of transformers is the self-attention mechanism, which computes relationships between all words in a sequence. At a high level, self-attention works as follows: for each word in a sentence, the model computes a relevance score with every other word, and then it applies a mathematical function called softmax to convert these scores into weights that sum to 1.

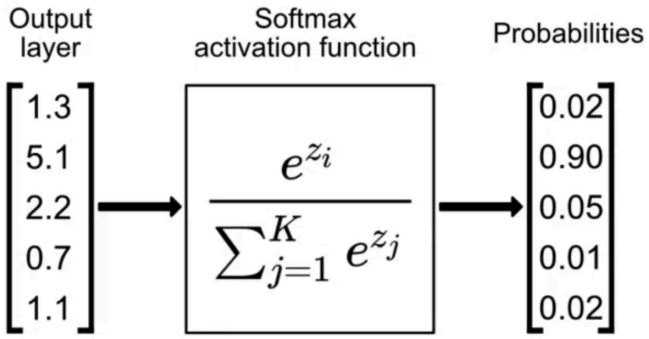

Figure 1: Representation of Softmax Function (Leemann et al., 2024)

These weights indicate how much each word should contribute to understanding the current word. The softmax function is crucial for understanding the limitations of traditional explanation methods. Softmax creates dependencies across all words in a sequence. This means that when a word is added to or removed from a sentence, the softmax function redistributes the weights across all remaining words to ensure that they still sum to 1. This redistribution does not follow the assumption of independence that underlies LIME and SHAP: these methods assume each feature's contribution remains constant regardless of which other features are present, but with softmax normalization, each word's contribution depends on the entire sequence. As a result, applying traditional explanation

methods to transformers produces incorrect explanations that fail to predict how the model would behave if words were added or removed.

SLALOM: A Breakthrough in Transformer Explainability

This mismatch in fundamental architecture was addressed by Leemann et al. (2024) in their recent work introducing SLALOM (Softmax-Linked Additive Log-Odds Model). The authors proved the theoretical result of how transformers are fundamentally incapable of representing additive models due to softmax normalization. More importantly, they designed SLALOM to specifically work with this architecture rather than against it. SLALOM computes two scores for each word: a token value representing its absolute contribution, and token importance representing its relative weight within the sequence.

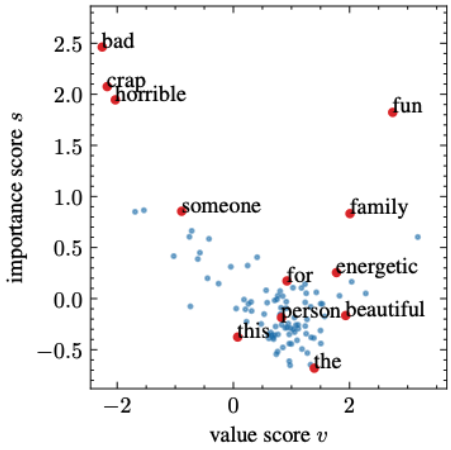

Figure 2: Plot of two scores that are assigned to each token(Leeman et. al, 2024)

These scores are combined using softmax, mirroring the transformer's own architecture. The authors proved mathematically that transformers can represent this functional form. However, despite these significant advances, the authors of SLALOM explicitly acknowledge a limitation: the method operates exclusively at the token level. Tokens, as mentioned earlier, are the basic units that transformers process: subword fragments produced by a tokenizer rather than complete words. For

example, the word "cardiovascular" might be split into "cardio" and "vascular." SLALOM can assign importance scores to these fragments, but doesn't provide any mechanism for grouping them into meaningful units like complete medical terms ("cardiovascular disease"), phrases, or concepts. This limitation creates a discrepancy between how the model explains its reasoning, through fragments and individual words, and how humans naturally think about language, which mainly occurs at the level of phrases and meaningful concepts.

The Gap Addressed by This Project

This project proposes to extend SLALOM from solely token-level explanations to explanations that encompass phrases as well. The result aims to be an explanation system that provides interpretability at the level of granularity that humans naturally use when understanding language, which shows improved trust.

Section II: Specific Aims

This proposal's objective is to extend SLALOM into a higher-order explanation framework that produces meaningful attribution scores for transformer models at the phrase-level, while preserving SLALOM's theoretical guarantees about fidelity. The long-term goal is to develop explanation tools that align heavily with the internal workings of transformer models while also ensuring that a high fidelity is maintained, thus rendering the explanation both interpretable and accurate.

Specific Aim 1: Design and implement higher-order SLALOM algorithms.

The first aim is to develop algorithms that aggregate SLALOM token value and importance scores into phrase-level attributions while maintaining the softmax-linked log-odds structure. This will involve defining which linguistic units I will use: primarily off-the-shelf NLP tools (e.g., sentence segmentation, noun/verb phrase chunking, named entities) and testing multiple aggregation strategies.

Specific Aim 2: Evaluate fidelity and efficiency on real transformer models.

The second aim is to test whether my model will maintain high fidelity and practical runtime when applied to real transformer models on standard NLP tasks. This evaluation will compare my model and SLALOM on tasks such as sentiment analysis and simple medical text classification, measuring how well each method predicts model behavior under token or phrase perturbations and how fast each method runs. The model could also be compared conceptually to attention-based approaches such as attention flow to clarify its position in other methods that are being developed for transformer explainability.

Section III: Project Goals and Methodology

Relevance/Significance

Transformers have become the dominant architecture for modern natural language processing(NLP), enabling large language models that power applications from chatbots to code assistants. These models are increasingly found in high-stakes work environments, including clinical decision support, medical report generation, and analysis of electronic health records, where errors or biases can be detrimental and directly affect patient care. Beyond healthcare, transformer-based generative AI systems are rapidly being adopted in finance and law, where they assist with tasks such as contract review, risk assessment, and content generation (Financial Times Visual Journalism Team, 2025;). At the same time, these systems function as black boxes: they learn complex internal representations over billions of parameters, making it difficult for users to understand why a particular output was produced. The field of explainable artificial intelligence emerged to address this lack of understanding.

Innovation

This project introduces a new approach by implementing an aggregation layer that utilizes the spaCy natural language processing library to segment input sequences into meaningful phrases. By mathematically aggregating validated token-level value and importance scores into these higher-level

units, the model preserves the high-fidelity guarantees of the original softmax-linked architecture while displaying results that align better with human interpretability. Unlike traditional additive models (LIME, SHAP) that fail to capture transformer interactions, or current token-only methods, this approach provides a dual-benefit: it maintains structural alignment with the transformer's internal logic while also increasing interpretability for experts in high-stakes fields like healthcare and law.

Methodology

1. Run the existing SLALOM on a small transformer text classifier to generate token-level value and importance scores.
2. Use the spaCy library to segment inputs into phrases, and implement simple aggregation strategies such as LogSumExp that combine token-level scores into phrase-level scores.
3. Test these aggregation strategies on small synthetic datasets with known phrase/sentence importance and choose the best one using correlation and rank-based metrics.
4. Apply the chosen strategy to one or two real text classification datasets (e.g. IMDb), generating both token-level and phrase-level explanations for test examples.
5. Evaluate fidelity by removing or masking top-ranked tokens or phrases and measuring changes in predicted probabilities and label flips
6. Compare fidelity and runtime between token-level and phrase-level explanations.

Specific Aim 1: Design and implement higher-order SLALOM algorithms.

The first aim is to develop algorithms that aggregate SLALOM token value and importance scores into phrase-level attributions while maintaining the softmax-linked log-odds structure. This will involve defining which linguistic units I will use: primarily off-the-shelf NLP tools (e.g., sentence segmentation, noun/verb phrase chunking, named entities) and testing multiple aggregation strategies.

Justification and Feasibility. SLALOM already provides high-fidelity token-level explanations for transformers, and its open-source implementation makes it practical and simple to extend. Using NLP python libraries for phrase segmentation, alongside small synthetic datasets, keeps this aim computationally manageable.

Expected Outcomes. We expect to identify one or more aggregation strategies that produce phrase scores closely aligned with ground-truth importance on synthetic data. This will result in a concrete procedure for computing phrase-level SLALOM explanations that preserve transformer compatibility.

Potential Pitfalls and Alternative Strategies. A potential pitfall is that simple aggregation strategies may not recover phrase importance accurately. If this happens, we will experiment with alternative phrase definitions and small learned aggregation functions that take token values and importances as inputs.

Specific Aim 2: Evaluate fidelity and efficiency on real transformer models.

The second aim is to test whether my model will maintain high fidelity and practical runtime when applied to real transformer models on standard NLP tasks. This evaluation will compare my model and SLALOM on tasks such as sentiment analysis and simple medical text classification, measuring how well each method predicts model behavior under token or phrase perturbations and how fast each method runs. The model could also be compared conceptually to attention-based approaches such as attention flow to clarify its position in other methods that are being developed for transformer explainability.

Justification and Feasibility. Prior work has shown that SLALOM can be applied to real transformer-based classifiers and evaluated via perturbation tests, which I will adapt to phrase-level

explanations. Focusing on small transformer models and small classification datasets makes the experiments feasible within the time and hardware available.

Expected Outcomes. We expect phrase-level explanations to achieve fidelity comparable to token-level SLALOM, as measured by probability drops and label flips when key tokens or phrases are removed. This would demonstrate that higher-level explanations can remain faithful while being more aligned with human language understanding.

Potential Pitfalls and Alternative Strategies. A possible pitfall is that phrase-level explanations may show lower fidelity or higher runtime than token-level explanations. If this occurs, we will refine the aggregation scheme by potentially restricting to the most important phrases and report the trade-offs clearly. We would also highlight scenarios where phrase-level explanations are most beneficial despite small fidelity losses.

Section III: Resources/Equipment

Software:

- Python 3.12
- PyTorch
- spaCy (for phrase segmentation)
- scikit-learn, NumPy, Pandas (for data processing)
- Model codebase repository: https://github.com/tleemann/slalom_explanations (Leeman et. al, 2024)
- Google Colab

Hardware:

- Laptop
- 16GB RAM, 500GB storage

Datasets:

The datasets intended for use were chosen with the following criteria:

1. inputs are natural language sentences or short documents,
2. labels correspond to a single, well-defined prediction (e.g., sentiment, topic, or diagnosis)

3. The dataset size is manageable for fine-tuning the model on the hardware that is available on hand.

Current examples:

- IMDb movie reviews
- GLUE benchmark (sentiment: SST-2; QA: SQuAD)
- MIMIC-III clinical notes (public medical text)

Section V: Ethical Considerations

This project includes a specific aim of the model having a high fidelity, which measures how well the explanation of a given "black-box" model, in this case being a transformer, reflects the model's true decision-making process. An explainability model with a lack of high fidelity is dangerous if it is applied within any high-stakes application as a human interpreting the explainability model might accidentally trust the model, while in fact, the model is unable to represent the “black-box” phenomenon well. Thus, the ethical concerns of this project arise from the potential application of the explainability model without the assurance of high fidelity.

Section VI: Timeline

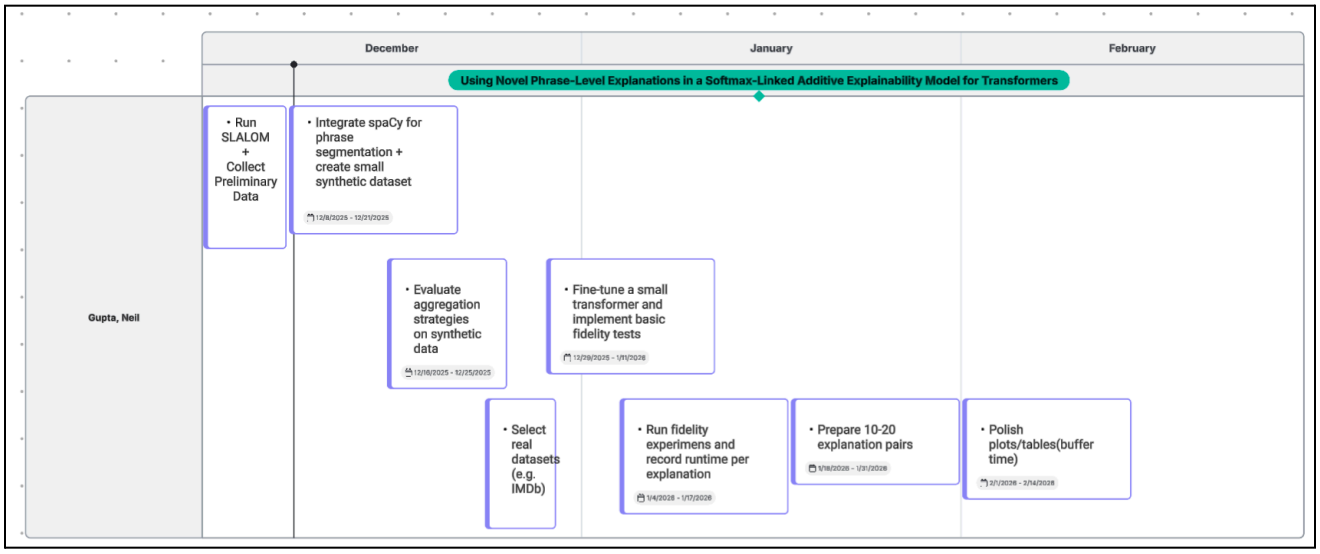

Figure 3: A screenshot of a Gantt Chart which details the steps I have taken and am taking to complete this project in a timely manner by February.

Section VIII: References

Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4190–4197). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.385

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Azarkhalili, B., & Libbrecht, M. (2025). Generalized attention flow: Feature attribution for transformer models via maximum flow. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (pp. 1–18). Association for Computational Linguistics. https://aclanthology.org/2025.acl-long.980/

Financial Times Visual Journalism Team. (2025, January 21). Generative AI exists because of the transformer. *Financial Times.* https://ig.ft.com/generative-ai/

Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2024). Transformers and large language models in healthcare: A review. *Artificial Intelligence in Medicine, 150,* 102578. https://www-sciencedirect-com.ezpv7-web-p-u01.wpi.edu/science/article/pii/S0933365724001428?via%3Dihub

Leemann, T., Fastowski, A., Pfeiffer, F., & Kasneci, G. (2024). Attention mechanisms don't learn additive models: Rethinking feature importance for transformers. *Transactions on Machine Learning Research, In press*. https://doi.org/10.48550/arXiv.2405.13536

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774. https://dl-acm-org.ezpv7-web-p-u01.wpi.edu/doi.org/10.5555/3295222.3295230

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*. https://doi.org/10.1145/2939672.2939778

Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. *In Proceedings of the 16th Conference on Innovative Applications of Artificial Intelligence* (pp. 900–907). https://www.researchgate.net/publication/221606722_An_Explainable_Artificial_Intelligence_System_for_Small-unit_Tactical_Behavior

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://arxiv.org/abs/1706.03762