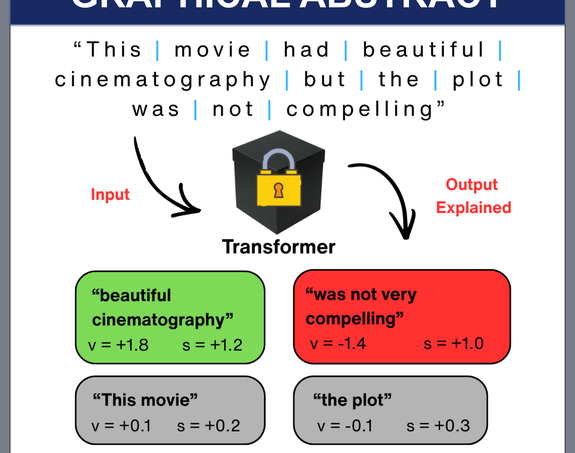
Using Novel Phrase-Level Explanations in a Softmax-Linked Additive Explainability Model for Transformers

Researcher: Neil Gupta
Advisor: Kevin Crowthers Ph.D.

GRAPHICAL ABSTRACT



PROBLEM

Explanations for transformer language models are unclear as they do not look past the token level.

OBJECTIVE

Develop a model that accounts for the transformer architecture and explains its decisions past the token level.

BACKGROUND

Transformers
- Power modern LLMs like ChatGPT and others are deployed in healthcare, law, and finance.
- They process text through a self-attention mechanism
- Function as **"black boxes"**, meaning they have billions of parameters but no transparent reasoning for their outputs.

Explainable AI (XAI) & Feature Attribution:
- XAI methods attempt to open the black box by assigning importance scores to input words
- **LIME** and **SHAP** are the most widely used methods
- LIME fails as it assumes a linear property, and SHAP fails as it assumes an additive property

SLALOM
- Designed specifically for **transformers**: assigns each token two scores:
- **Value** (v): the token's absolute contribution to the output
- **Importance** (s): the token's relative weight among all other tokens.
- These are then combined through **Softmax.**

Key limitation:
- Operates only at the **token-level**
- Cannot capture **phrase-level** meaning

ENGINEERING METHODOLOGY

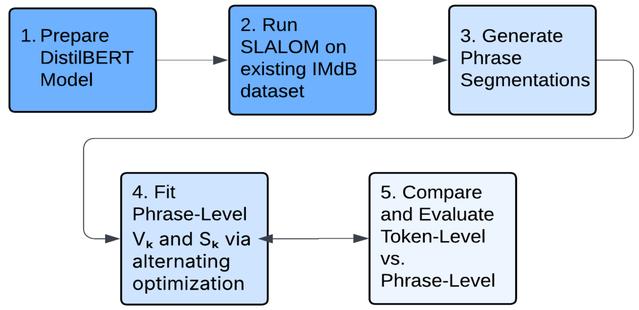
Tools and Libraries:

Python 3.12: Main programming language used for all code implementation, data processing, and model evaluation.

NumPy: Numerical computing library used for matrix operations.

spaCy: Natural language processing library for phrase segmentations

Design Flowchart:



Eq.1: $F(t) = \Sigma_i \alpha_i \cdot v(t_i)$, when $\alpha_i = \exp(s(t_i)) / \Sigma_j \exp(s(t_j))$

Eq. 2: $F(t) = \Sigma_k \beta_k \cdot V_k$, $(\beta_k = exp(S_k) / \Sigma_j exp(S_j))$

Eq. 3: $\Sigma_{i \in P_k} \alpha_i \cdot v(t_i)$

TESTING & RESULTS

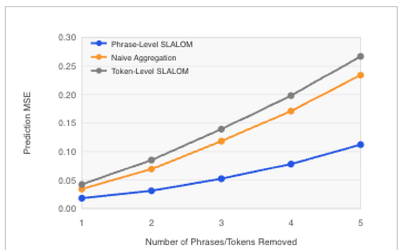


Figure 1: Fidelity MSE vs. number of top-ranked phrases removed. Lower is better. Phrase-level SLALOM predicts transformer output changes with up to 58% less error than token-level methods.

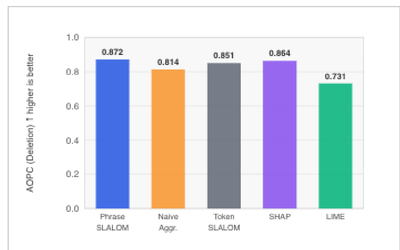


Figure 2: AOPC deletion benchmark across methods. Higher AOPC indicates the method correctly identified the most influential features. Phrase-level SLALOM outperforms all baselines.

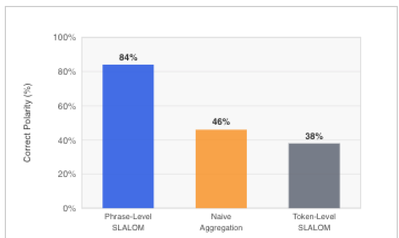


Figure 3: Percentage of compositional phrases (negation, idioms) where each method assigned the correct polarity. Phrase-level re-fitting nearly doubles the accuracy of naive aggregation.

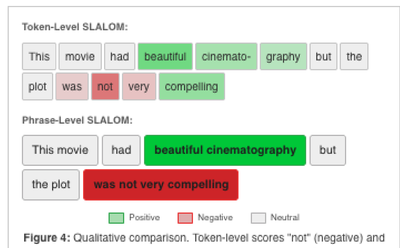


Figure 4: Qualitative comparison. Token-level scores "not" (negative) and "compelling" (positive) separately, missing the negation. Phrase-level correctly identifies "was not very compelling" as a single negative unit.

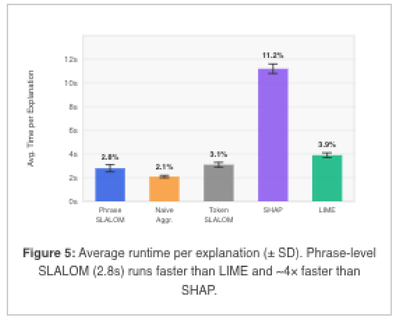


Figure 5: Average runtime per explanation (± SD). Phrase-level SLALOM (2.8s) runs faster than LIME and ~4× faster than SHAP.

CONCLUSION

- Phrase-level model preserves **Softmax-linked** transformer compatibility.
- Re-fitting at the phrase level captures more meaning than the **naive aggregation** of token scores.
- Phrase-level explanations achieve higher **fidelity** with fewer units to interpret.

FUTURE WORK

- Evaluate on diverse datasets beyond **sentiment** analysis, including medical, legal, and multi-topic classification benchmarks.
- Extend the framework beyond **classification** models to **generative** tasks such as text generation and question answering.
- Test on **larger** transformer models to assess whether fidelity gains scale with model size.

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/
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