

A UNIFIED XAI TOOLKIT FOR QUESTION-ANSWERING LANGUAGE MODELS IN HIGH-STAKES DOMAINS

Presented by:

Jeff Chong
Anne Ha
Jiyeon Lee
Mathew Leffler
Nairui Lius

Project Advisors:

Dr. Simon Shim

Introduction

As large language models proliferate in high-stakes domains like healthcare, their "black-box" nature creates a critical barrier to trust and adoption. Clinicians and regulators cannot interpret or validate model outputs, and in a domain where a wrong answer can mean a missed diagnosis, this lack of transparency is unacceptable. While XAI research exists, most tools were not designed for medical LLMs and remain inaccessible to non-technical users.

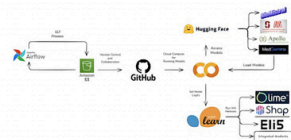


- AI deployed in high-stakes settings requires transparency and accountability
- Without interpretability, practitioners must blindly trust outputs they cannot validate
- Traditional XAI methods were designed/optimized for tabular and vision models, not large language models

Methodology

Toolkit Design and Implementation

A custom MedicalLLMWrapper standardizes inference across four biomedical LLMs of varying scale, integrating four complementary attribution methods: LIME [3], Integrated Gradients [4], TokenSHAP [1], and ELI5 [2], alongside Chain-of-Thought [5] and Best-of-N reasoning methods.



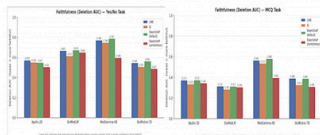
- Custom wrapper enforces constrained decoding and structured outputs across all supported HuggingFace LLMs
- TokenSHAP extended with three custom value functions (correctness-aware, embedding-based, hybrid) and a semantic NER-based splitter
- All methods accessible through a single interactive interface requiring no coding

Analysis and Results

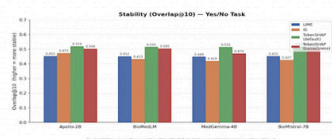
XAI Method Evaluation

Evaluation was conducted across three XAI methods (LIME, Integrated Gradients, and TokenSHAP) over four biomedical LLMs using faithfulness (deletion AUC) and stability (overlap@10) metrics. ELI5 was excluded from quantitative evaluation as its surrogate-based approach produces second-order approximations, making standard metrics unsuitable for direct comparison.

Across both yes/no and multiple-choice tasks, TokenSHAP (correctness) consistently achieved the lowest deletion AUC, indicating the strongest faithfulness overall. Integrated Gradients was competitive across most models, while LIME and TokenSHAP (default) were generally the least faithful. Notably, all methods achieved lower deletion AUC on multiple-choice tasks compared to yes/no, suggesting that attribution methods are better able to identify the tokens driving model decisions in a more constrained answer space. MedGemma-4B showed the highest AUC across both task types, indicating that explanations for larger instruction-tuned models may be harder to faithfully capture.



For stability, TokenSHAP (default) was the most consistent method across all four models, achieving overlap@10 scores between 0.515 and 0.519. The correctness-aware variant was slightly less stable in all cases, pointing to a tradeoff between faithfulness and stability between the two TokenSHAP configurations. LIME showed notably consistent stability across all models at around 0.452, while Integrated Gradients exhibited the most variability, suggesting its attribution behavior is more sensitive to differences in model architecture. Overall, no single method dominated across both faithfulness and stability. This reinforces the value of the toolkit's multi-method design as each method explains model behavior from a fundamentally different perspective, and the appropriate choice depends on the user's priorities rather than a single ranking.



Reasoning Extraction

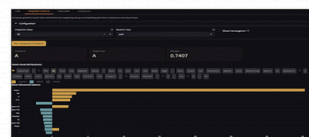
Chain-of-Thought and Best-of-N sampling were implemented as supplementary reasoning methods to complement the core attribution-based XAI methods. Chain-of-Thought generates a reasoning trace based on the prompt before scoring the answer, avoiding post-hoc rationalization by committing to reasoning before an answer is selected. Best-of-N generates multiple independent reasoning branches, scores each by log-probability, and derives the final answer and rationale from the highest scoring branch. Both methods provide a complementary view of model reasoning that token-level attribution alone cannot capture, though verification currently requires a human in the loop.

User Interface

The toolkit is exposed through a unified Streamlit interface designed to make LLM explainability accessible to both technical and non-technical users. Users can load a model, select a task type, and choose between attribution and reasoning modes without writing a single line of code.

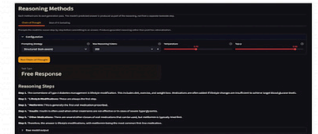


Once a question is submitted, users can run any of the XAI methods to inspect token-level attributions rendered as highlighted text and a ranked bar chart. A cross-method comparison tab aggregates top-k token rankings side by side, giving users a stronger signal of reliability where methods agree.



In reasoning mode, Chain-of-Thought displays step-by-step reasoning traces before committing to an answer.

Best-of-N surfaces multiple reasoning branches scored by log-probability, selecting the highest scoring branch as the final rationale.



Summary/Conclusions

This project delivered a complete, extensible XAI toolkit for auditing and interpreting language model behavior in high-stakes domains. No single method dominated across faithfulness and stability, validating the multi-method design as a comparative framework. Custom QA-aware extensions to TokenSHAP measurably improved faithfulness over the base variant, demonstrating that domain adaptation matters at the explanation level, not just the model level.

Key References

- [1] M. Horowitz and R. Goldschmidt, "TokenSHAP: Interpreting large language models with Monte Carlo Shapley value estimation," arXiv, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2407.10114>
- [2] M. Korobov and K. Lopuhin, "ELI5," Read the Docs, 2021. [Online]. Available: <https://eli5.readthedocs.io/en/latest/overview.html>
- [3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," arXiv, 2016. [Online]. Available: <https://doi.org/10.48550/arXiv.1602.04938>
- [4] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," arXiv, 2017. [Online]. Available: <https://doi.org/10.48550/arXiv.1703.01365>
- [5] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," arXiv, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2201.11903>

Acknowledgements

The authors would like to thank Dr. Simon Shim for his guidance and support throughout the project, and Saurabh Suman, Instructional Student Assistant, for providing valuable feedback throughout the project demos. We also thank San José State University for access to the GPU lab and Google Colab academic account credits, which supported the compute-intensive inference and evaluation runs conducted in this project.

Interested in partnering with us on a capstone project or exploring other ways to collaborate with the College of IDS? Contact CIDS_CAP@sjsu.edu to discuss potential projects and partnership opportunities.