P14. Transparent Minds

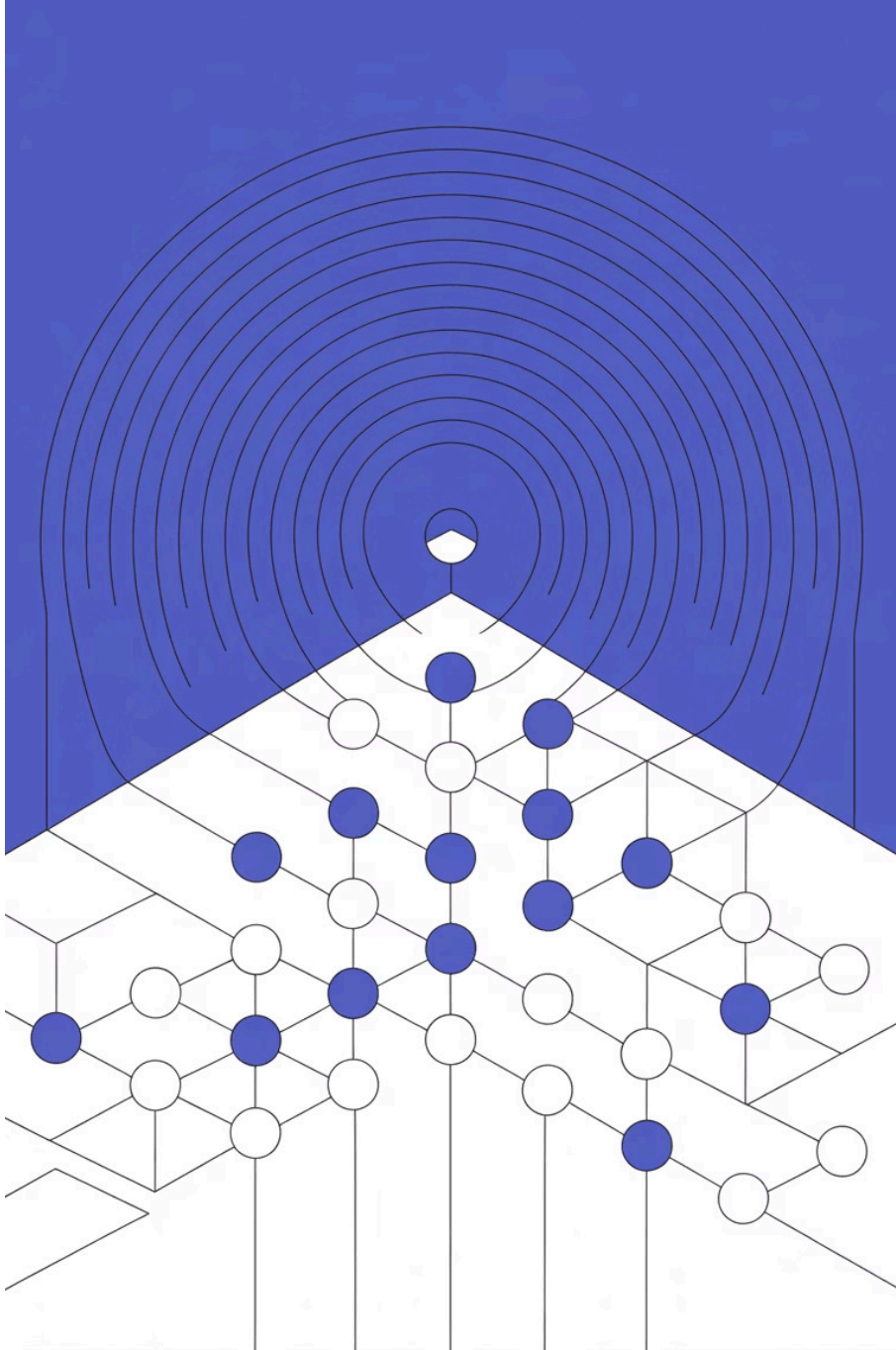
This project aims to design and implement an **interactive toolkit for visualizing and interpreting the reasoning processes** of transformer-based language models. While most explainability studies remain abstract or static, this project focuses on creating a **hands-on, dynamic environment** where users can *see* and *experiment with* the inner mechanics of LLMs.

Core Pipeline

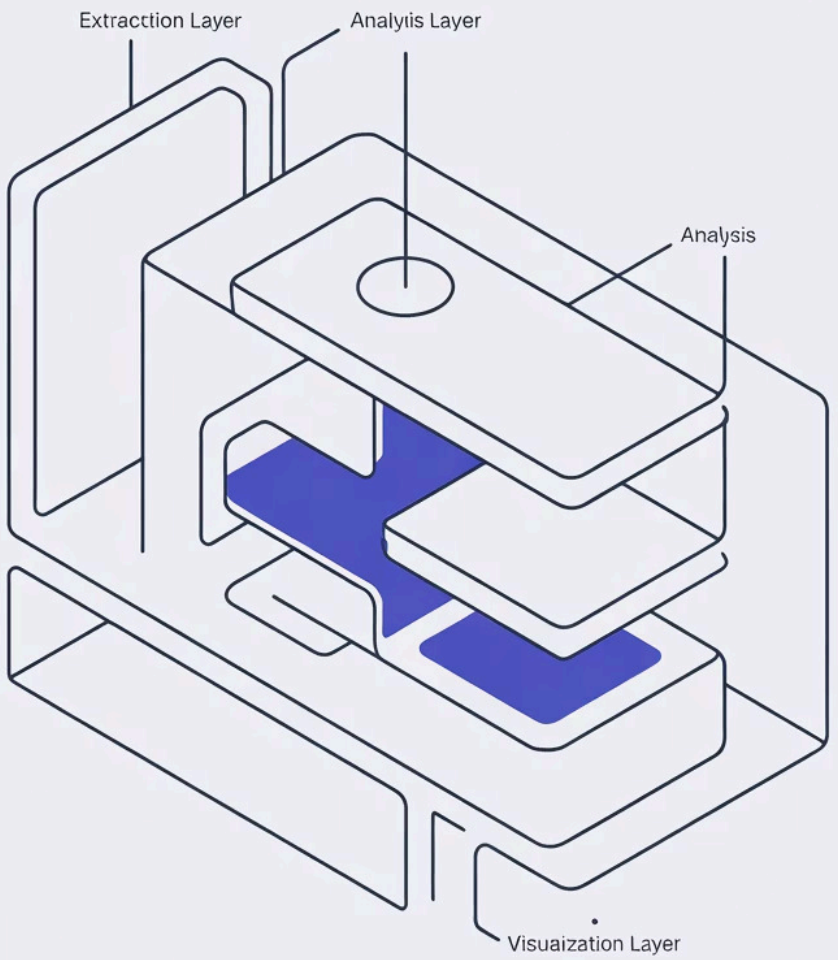
Interactive toolkit for extracting and visualizing internal states (attention weights, gradients, activations) with dynamic interfaces

Expected Outcomes

Working explainability prototype with human-centered evaluation linking technical inspection to cognitive understanding



Explainability Suite Architecture



Methodology

1

Architecture Design

Three-layer system: Extraction (model internals), Analysis (interpretability metrics), Visualization (interactive interface)

2

Model Integration

Select transformer models, implement wrappers for attention matrices, embeddings, activations using existing libraries

3

Interface Development

Build interactive dashboard allowing users to upload text, inspect layers, view attention heatmaps and influence scores

4

Evaluation

Assess visualization methods for human insight, conduct usability experiments, evaluate explanation consistency

Implementation Guidelines: Encourage modular design with Python modules, OOP principles, documented APIs, and at least one interactive notebook demo.

Dataset & References

Dataset: Any text classification or QA dataset suitable for visual experiments and interpretability analysis.

References

- Grimsley, C., et al. (2020). Why attention is not explanation: Surgical intervention and causal reasoning about neural models. *LREC*, 1780-1790.
- Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *arXiv preprint arXiv:1902.10186*.
- Raza, S., et al. (2025). Humanibench: A human-centric framework for large multimodal models evaluation. *arXiv preprint arXiv:2505.11454*.