PROJECT REPORT

Project Title: Mechanistic Interpretability of GPT-2 Small: A Layer-by-Layer Analysis of Semantic Abstraction

Author: Venkateswarlu Thatha

Tools: PyTorch, TransformerLens, Plotly

1. Executive Summary

This project aimed to investigate the phenomenon of "Polysemanticity" (superposition) in Large Language Models by analyzing internal activations of GPT-2 Small. By identifying the neuron with the maximum activation ("The Champion") at each of the 12 layers, we discovered a distinct "Hierarchy of Abstraction" and a "U-Shaped" activation energy pattern. The findings suggest that GPT-2 processes information linearly from syntactic recognition to structural reasoning, and finally to semantic decision-making.

2. Key Research Insight: The "U-Shaped" Thinking Process

A quantitative analysis of maximum activation strengths revealed a distinct energy pattern across the model's depth:

- Input Phase (Layers 0-2): High Energy (5.7 to 3.8)
  The model exhibits high activation ("shouting") as it recognizes raw tokens and fundamental syntax rules.
- Deep Thinking Phase (Layers 3-7): Low Energy (2.9 to 3.4)
  Activation levels drop significantly. This suggests a distribution of workload where information is routed, dependencies are tracked, and logical scaffolding is built ("whispering"). No single neuron dominates, indicating distributed processing.
- Output Phase (Layers 8-11): High Energy (5.3 to 9.4)
  Energy levels explode as the model collapses complex logic into a final prediction. Layer 11 shows the highest activation (9.36), likely indicating the "Logit Lens" effect where the final token is selected.

3. The Hierarchy of Abstraction

Our "Champion Neuron" analysis verified that neuron specialization evolves with depth:

- **The Librarians (Layers 0-2):**
  - *Representative:* Neuron 1846 (Layer 0).
  - *Behavior:* Monosemantic focus on specific syntactic tokens (e.g., the preposition "by").
  - *Function:* Sorting words and enforcing grammatical rules.
- **The Engineers (Layers 3-7):**
  - *Representative:* Neuron 1395 (Layer 3).

    - *Behavior:* Spikes on structural markers like periods, commas, and contrastive conjunctions ("but").
    - *Function:* Managing sentence flow and logical dependencies.
- **The Philosophers (Layers 8-11):**
    - *Representative:* Neuron 1253 (Layer 8).
    - *Behavior:* High activation for "Action-Oriented Concepts" (e.g., "gradient", "fruits", "action") across disparate domains (Tech & Spirituality).
    - *Function:* Semantic abstraction. What initially appeared as polysemantic confusion was identified as high-level abstraction (grouping different topics under a single conceptual header).

4. Conclusion

The project concludes that GPT-2 Small is highly organized. It does not "think" at a constant volume but follows a distinct recognition-reasoning-decision loop. Furthermore, "superposition" in later layers is often a result of efficient semantic abstraction rather than random polysemanticity.