Tracing Neural Pathways in Political Tweet Classification using Mechanistic Interpretability

Garance Colomer
Ai Scientist

Luc Chan
Data Engineer

Sacha Lahlou
Ai Engineer

Leina Corporan
Data Engineer

With
In collaboration with Apart Research.

Abstract

This paper explores mechanistic interpretability in large language models (LLMs) by examining the classification of political tweets. Using Llama 3.3 70B, we infer the political alignment (Pro-Biden, Pro-Trump, or Neutral) of users based on their tweets, descriptions, and locations. We then employ the Goodfire framework to extract the most highly activated features corresponding to each class. By re-running the model on the same dataset while activating these extracted features, we analyze the influence of these neuron activations on classification. Our approach provides insights into the pathways taken by the LLM during classification, shedding light on the internal mechanisms behind political bias in model predictions. We validate our findings through accuracy comparisons and neuron pathway coverage graphs.

Keywords:
mechanistic interpretability, sparse autoencoders, feature visualization, feature attribution, neural pathways, model steering, interpretability tools, neural features, activation patterns, model internals, model activation analysis, decision boundary analysis, latent space analysis, political alignment classification, LLM interpretability, activation clustering, causal feature analysis, ai safety

1. Introduction

a. Problem Statement

Understanding how large language models classify political content is crucial for AI interpretability. Our study aims to map the neural pathways used by Llama 3.3 70B in classifying tweets and user information into political categories. By analyzing activation patterns, we seek to determine whether specific features drive classification decisions.

After an initial analysis using extracted features (see figure 1 and figure 2), we observe that the language model employs different vocabularies when performing classification. Notably, we identify a clear differentiation in the activated features that guide the classification process.

Figure 1: Biden's political features

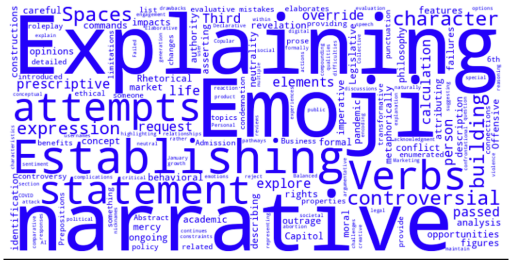
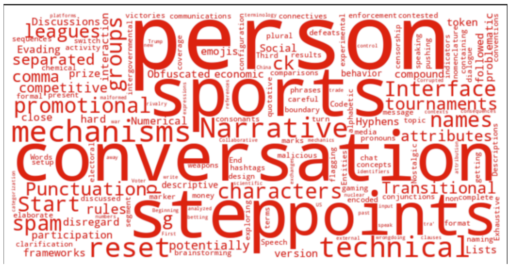


Figure 2: Trump's political features



b. Background and Motivation

Existing work in interpretability focuses on feature attribution and activation analysis, yet little research has been done on tracing neuron pathways in LLM classification tasks. By applying mechanistic interpretability techniques, we aim to reveal decision-making patterns that can improve trust and transparency in AI models.

c. Threat Model and Safety Implications

Bias in AI models can lead to misinformation and polarization. Understanding how models categorize political affiliations enables researchers to detect and mitigate unintended biases in AI-driven decision systems.

2. Methods

a. Approach

1. **Initial Classification:** We first use Llama 3.3 70B to classify tweets, user descriptions, and locations into three categories: Pro-Biden, Pro-Trump, or Neutral.

2. **Feature Extraction:** Using Goodfire, we extract the most highly activated features for the Biden and Trump-labeled users. We then extract the top 10 features from each conversation.

3.**Feature Scoring Exp-1:** We store the extracted features in a map where the keys are the features and the values are lists of activation values. For each feature, we compute the mean activation, adjust it with a variance term weighted by 0.3 to prioritize stable activations, and sort them to extract the most relevant ones.

$$S(f) = \mu_f + \lambda \cdot \sigma_f^2,$$

where:

$$\mu_f = \frac{1}{N}\sum_{i=1}^{N} a_i, \qquad \sigma_f^2 = \frac{1}{N}\sum_{i=1}^{N}(a_i - \mu_f)^2, \qquad \lambda = \frac{\mu_{\text{global}}}{\sigma_{\text{global}}^2}$$

4. **Feature Scoring Exp-2:** In a second experiment, we compute a mean-adjusted feature activation score, which balances the average activation strength of a feature with its relative frequency of activation.

$$S(f) = \left(\frac{1}{N_f}\sum_{i=1}^{N_f} a_i\right) \times \frac{N_f}{\bar{L}}, \quad \text{with} \quad \bar{L} = \frac{1}{M}\sum_{j=1}^{M} L_j$$

5. **Re-inference with Feature Activation:** We prompt Llama 3.3 70B to reclassify the dataset while instructing it to "rely on these sentences," where "sentences" refer to the extracted features. This process is applied separately to both the Biden-aligned and Trump-aligned datasets.

6. **Merge Datasets:** To consolidate the reclassified datasets, we compare the outputs from both Biden-aligned and Trump-aligned feature-based inferences. If both classifications agree, the label is

retained. If one classification is neutral while the other is not, we assign the non-neutral label. In cases where classifications conflict, we assign a neutral label to capture uncertainty. This step ensures a balanced integration of both perspectives while preserving classification consistency.

7. **Comparative Analysis:** We compare the newly generated classifications with the initial ones to analyze neural pathway overlap and assess classification consistency.

b. Implementation

- Dataset: Tweets labeled with user descriptions and locations
- Model: Llama 3.3 70B
- Framework: Goodfire for feature extraction
- Metrics: Classification accuracy, neuron activation pathways
- Visualization: Graphs of neural pathway coverage and confusion matrix

3. Results

a. Analysis and Findings

We present activation-based classification results and analyze how extracted features influence model decisions. Key findings include:

- Overlapping features contributing to both Biden and Trump classifications, indicating shared linguistic or contextual patterns.
- Feature activation impact on classification accuracy, showing how selected features steer model predictions.
- Distinct neural pathways associated with each political category, revealing differences in activation patterns between Biden-aligned and Trump-aligned classifications.

b. Impact Assessment

Evaluating our results indicate that specific neuron pathways are responsible for political classification. This insight could inform bias mitigation strategies in AI and enhance model transparency.

Figure 4 - LLM Classification Consistency: Feature-Based vs. Tweet-Based Predictions

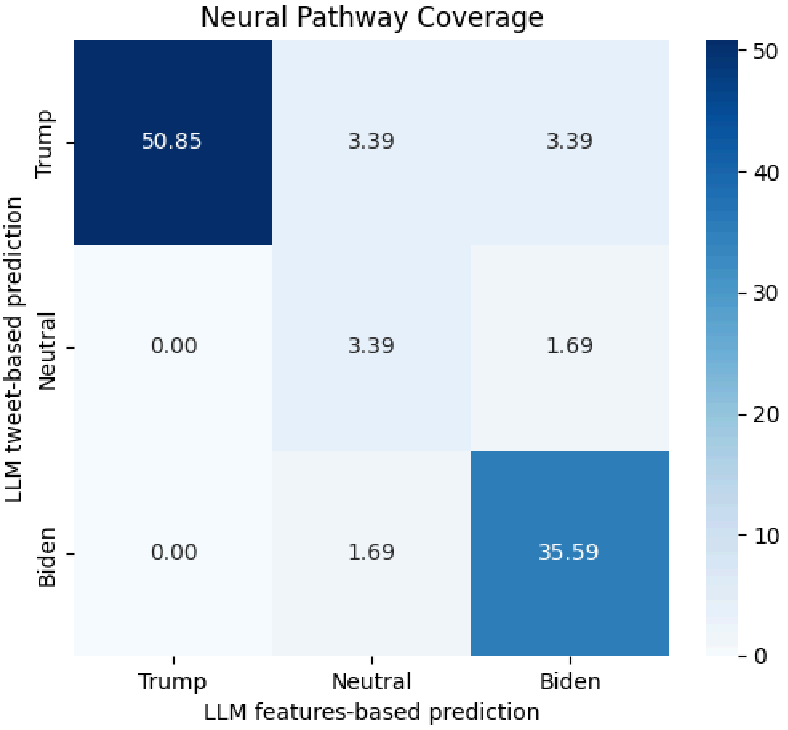
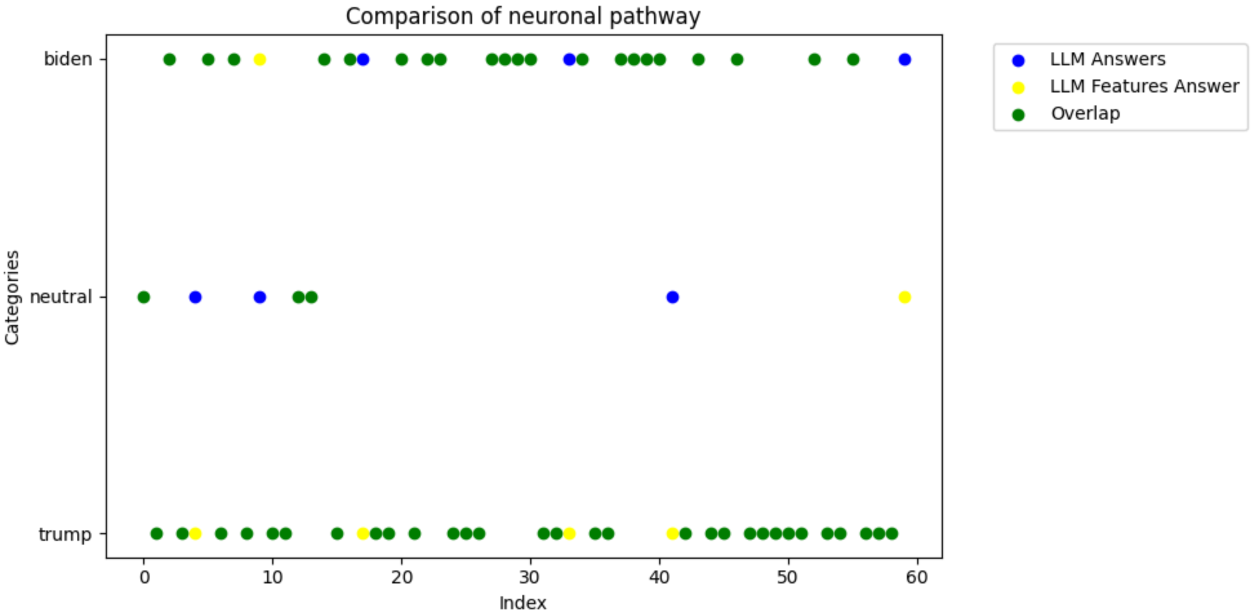


Figure 3 – Representation of answer's overlap according to different sets of features.



4. Discussion and Conclusion

By leveraging mechanistic interpretability, we have identified and visualized neural pathways involved in political tweet classification. Our findings highlight how specific feature activations influence classification outcomes and demonstrate that LLMs rely on distinct linguistic patterns when making decisions.

This methodology provides a deeper understanding of how LLMs process information, offering a structured way to evaluate the quality of extracted features. By identifying the neural mechanisms that drive classification, this approach can help detect and mitigate biases unintentionally embedded in pre-training data.

Future work will focus on refining feature selection techniques to improve interpretability and applying this framework to other classification tasks. Advancing the understanding of model decision pathways is a step toward developing more transparent and fair AI systems.