

Attention Pattern Based Information Flow Visualization Tool*

Sofia Maria Lo Cicero Vaina Anastasiia Ivanova Liubov Yaronskaya

With
Apart Research and Goodfire

Abstract

Understanding information flow in transformer-based language models is crucial for mechanistic interpretability. We introduce a visualization tool that extracts and represents attention patterns across model components, revealing how tokens influence each other during processing. Our tool automatically identifies and color-codes functional attention head types based on established taxonomies from recent research on indirect object identification (Wang et al., 2022), factual recall (Chughtai et al., 2024), and factual association retrieval (Geva et al., 2023). This interactive approach enables researchers to trace information propagation through transformer architectures, providing deeper insights into how these models implement reasoning and knowledge retrieval capabilities.

Keywords:

Mechanistic interpretability, attention circuits, information flow visualization

1. Introduction

1.1. Problem Statement

Mechanistic interpretability aims to reverse-engineer the algorithms implemented by neural networks. Recent work has discovered specific circuits responsible for capabilities like factual recall and indirect object identification. These discoveries have typically required painstaking manual analysis, including custom diagrams to visualize information flow.

Additionally, attention patterns in the form of heatmaps, a common visualization method, are visually inconvenient to analyze and make it difficult to extract meaningful insights efficiently.

Our tool automates this visualization process, making circuit discovery and analysis more accessible. By building on taxonomies from landmark papers (Wang et al., 2022; Chughtai et al., 2024; Geva et al., 2023), we enable rapid exploration of how different model components interact to implement specific capabilities.

1.2. Background and Motivation

Our approach draws inspiration from three complementary lines of research. The circuit diagrams in "Interpretability in the Wild" (Wang et al., 2022) represent information flow between attention heads with different functional roles. "Summing up the Facts" (Chughtai et al., 2024) identifies specialized attention head types for factual recall. Finally, "Dissecting Recall of Factual Associations" (Geva et al., 2023) reveals a three-step mechanism for factual knowledge retrieval:

*Research conducted at the Women in AI Safety Hackathon, 2025

subject enrichment through early MLP layers, relation propagation, and attribute extraction via attention heads.

By integrating these perspectives, our tool offers a comprehensive visualization system that captures both general language processing mechanisms and specialized factual retrieval processes. This enables researchers to visually identify critical information flow pathways, like the distinct phases uncovered by Geva et al. (2023).

1.3. Threat Model and Safety Implications

Our tool enhances AI safety by exposing potential failure modes in language models. By visualizing attention flow, it helps detect issues such as factual inconsistencies, adversarial vulnerabilities, and bias propagation.

2. Methods

2.1. Approach

We define an information flow graph where each node represents a token at a specific layer. Edges connect tokens across consecutive layers, with weights equal to attention values. Following Geva et al.’s (2023) "attention knockout" methodology, our tool allows users to interactively block specific attention edges and observe the impact on predictions, helping identify critical information pathways.

The interface is divided into several sections:

- Model:** A dropdown menu showing 'gpt2-small'. Below it, a note states: 'Predefined head groups are specific to the selected model.'
- Head Groups:** A section for organizing attention heads. It includes a 'New group name' input field and a 'Create' button. Below this, it says 'Create a new group to organize attention heads'. Two groups are listed:
 - Name Mover:** 'Attend to names and copy them to output. Active at END token position.' It shows three orange bars with values 9.9, 10.0, and 9.6.
 - Negative:** 'Write in opposite direction of Name Movers, decreasing prediction confidence.' It shows two blue bars with values 10.7 and 11.10.
- Individual Heads:** A section for selecting specific heads. It has a 'Selected heads:' input field showing 'No heads selected'. Below it, a note says 'Click + to add a head to a group'. There is a text input field with the pattern 'layer,head or layer,: or :,head or :,,' and an 'Add' button. A validation note at the bottom says 'Valid: Layer (0-12), Head (0-11)'.
- Edge Weight:** A slider control labeled 'Edge Weight:' with a value of 0.47.
- Input Text:** A text area containing the sentence: 'When Mary and John went to the store, John gave a drink to'.

Figure 1: Main interface of the attention pattern visualization tool, showing color-coded attention flows between tokens in GPT-2 small.

2.2. Attention Head Classification

We incorporate three complementary taxonomies:

1. From Wang et al. (2022): - Name Mover, S-Inhibition, Duplicate Token, Previous Token, Induction, and Negative Name Mover Heads
2. From Chughtai et al. (2024): - Subject, Relation, and Mixed Heads
3. From Geva et al. (2023): - Subject Enrichment Components, Relation Propagation Heads, and Attribute Extraction Heads

These classifications align with Geva et al.’s three-step factual recall process. Each head type receives a distinct color, enabling users to quickly identify potential functional circuits.

2.3. Implementation

Our framework uses PyTorch and TransformerLens, with a Python backend for extracting attention patterns and a React/D3.js frontend for interactive visualization. The implementation is available at: https://github.com/2son1a/information_flow.

3. Results

3.1. Analysis and Findings

Our tool provides interactive visualization of attention pattern-based information flow. A demonstration is available at: <https://youtu.be/XB2eJyJASgo>. Key features include:

- Interactive circuit visualization with automatic detection of functional components
- Support for multiple models (GPT-2 variants and Pythia-2.8)
- Color-coded head type classification based on established research
- Token-level analysis and attention knockout functionality

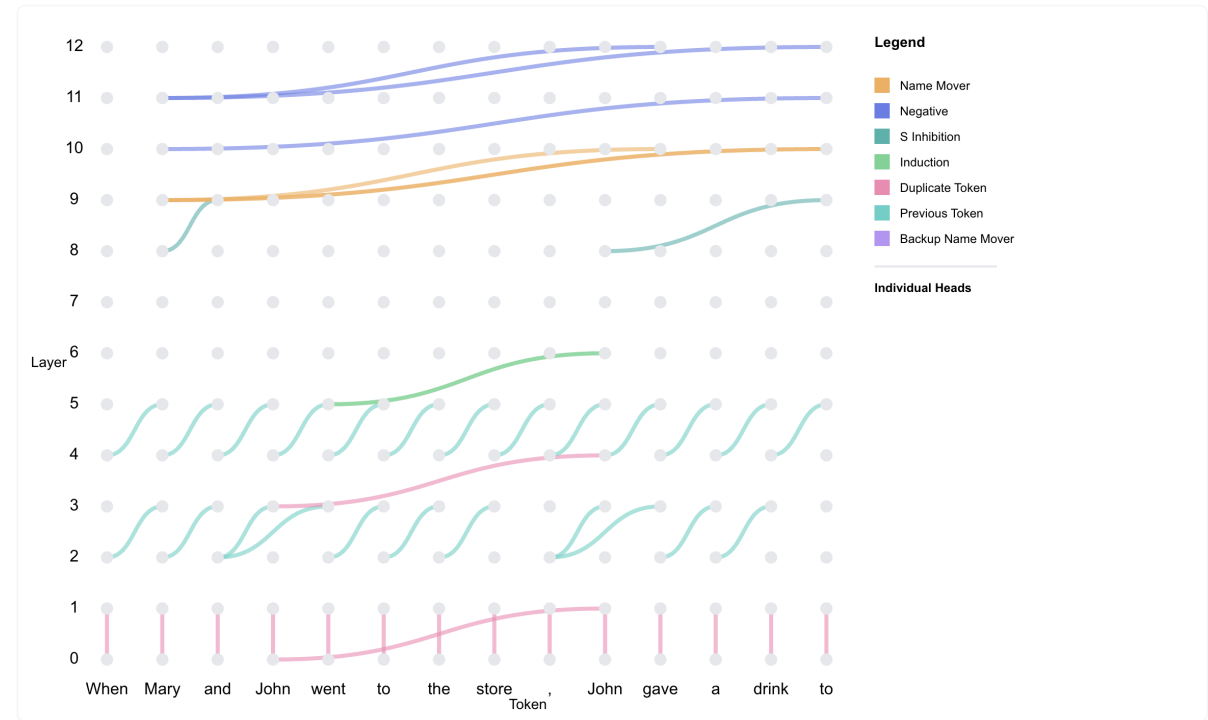


Figure 2: Automatically generated visualization of the indirect object identification circuit in GPT-2 small, showing how information flows from the tokens "Mary" and "John" through various attention heads.

Moreover, no specific expertise is required to integrate the tool into one’s own project, as we integrated it in the CircuitsVis library: <https://github.com/lyaronskaya/CircuitsVis>.

Compared to diagrams in works such as (Wang et al., 2022; Geva et al., 2023; Chughtai et al., 2024), our tool enables simple, interactive visualization with control over edge weights. Figures 3–6 present excerpts from these articles, allowing for a visual comparison with Figure 2.

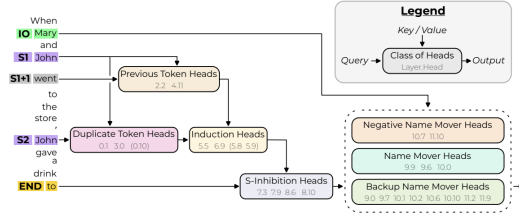


Figure 3: Wang et al., 2022

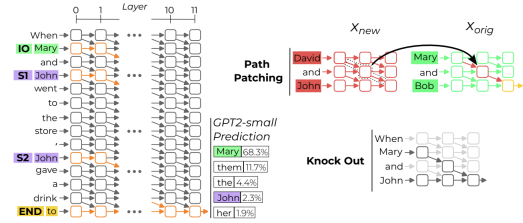


Figure 4: Wang et al., 2022

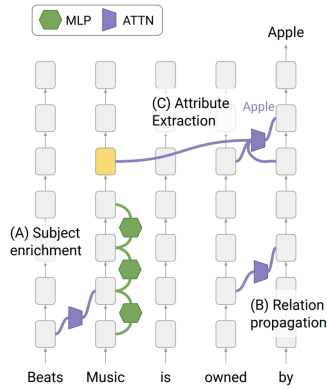


Figure 5: Geva et al., 2023

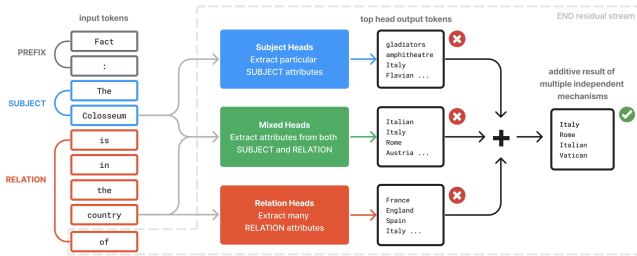


Figure 6: Chughtai et al., 2024

3.2. Application to Known Circuits

We validated our tool using prompts from prior research:

1. **Indirect Object Identification (IOI)**: For prompts like "When Mary and John went to the store, John gave a drink to," our tool successfully visualizes the circuit described by Wang et al. (2022).

2. **Factual Recall**: For prompts like "The Eiffel Tower is in the country of," our visualization reveals the additive mechanisms described by Chughtai et al. (2024).

3. **Subject-Relation Queries**: For factual queries like "Beats Music is owned by," our tool visualizes the three-step mechanism discovered by Geva et al. (2023): subject enrichment, relation propagation, and attribute extraction. Our visualization confirms their findings that blocking critical attention edges at specific stages dramatically reduces prediction accuracy.

Our tool serves both experienced researchers and newcomers. For researchers, it enables comparative analysis across model architectures, integrates with established methodologies, and visualizes key phenomena like the subject enrichment process. For newcomers, it provides an intuitive entry point to understanding transformer mechanics through color-coding, guided examples, and progressive complexity options.

4. Discussion and Conclusion

4.1. Advancing Mechanistic Interpretability

Our tool contributes to mechanistic interpretability by:

1. Enabling rapid hypothesis generation and testing through interactive visualization
2. Facilitating cross-model comparison to identify convergent circuit structures
3. Visually demonstrating the distinct phases of factual recall identified by Geva et al. (2023)

The tool is easy to use due to its integration into the CircuitsVis library and provides a holistic view of attention patterns across multiple heads and layers, serving as both a research instrument and an educational resource. It bridges the gap between technical papers and practical understanding through interactive exploration.

Current limitations include reliance on predefined head types, decreased performance with larger models, and incomplete visualization of MLP contributions (though we do highlight the subject enrichment role of MLPs identified by Geva et al.). Future work should integrate more causal intervention techniques and expand the classification system as new head types are discovered.

5. References

- (Chughtai et al., 2024) - Chughtai, B., Cooney, A., & Nanda, N. (2024). Summing up the facts: Additive mechanisms behind factual recall in llms. arXiv preprint arXiv:2402.07321. <https://arxiv.org/abs/2402.07321>
- (Geva et al., 2023) - Geva, M., Bastings, J., Filippova, K., & Globerson, A. (2023). Dissecting recall of factual associations in auto-regressive language models. arXiv preprint arXiv:2304.14767. <https://arxiv.org/abs/2304.14767>
- (Wang et al., 2022) - Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. arXiv preprint arXiv:2211.00593. <https://arxiv.org/pdf/2211.00593>