Causal Interventions on Causal Paths: Mapping GPT-2's Reasoning From Syntax to Semantics

Isabelle Lee, Joshua Lum,
Ziyi Liu, Dani Yogatama

Motivation

While interpretability research has shed light on some internal algorithms utilized by transformer-based LLMs, reasoning in natural language, with its deep contextuality and ambiguity, defies easy categorization. As a result, formulating clear and motivating questions for circuit analysis that rely on well-defined in-domain and out-of-domain examples required for causal interventions is challenging.

In this work, we take initial steps to characterize causal reasoning in LLMs by analyzing clear-cut cause-and-effect sentences like "I opened an umbrella because it started raining," where causal interventions may be possible through carefully crafted scenarios using GPT-2 small.

Overview of Methods

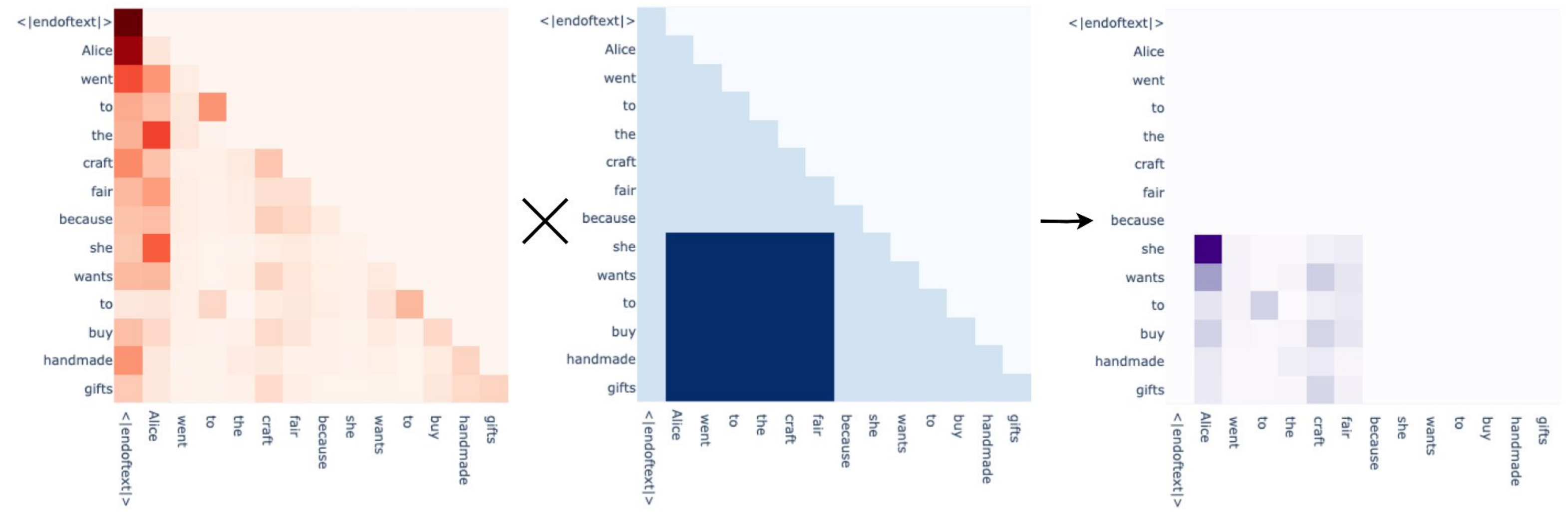
We perform two analyses

- Proportion of attention from cause-to-effect or effect-to-cause
- Semantic relation perturbation with causal interventions

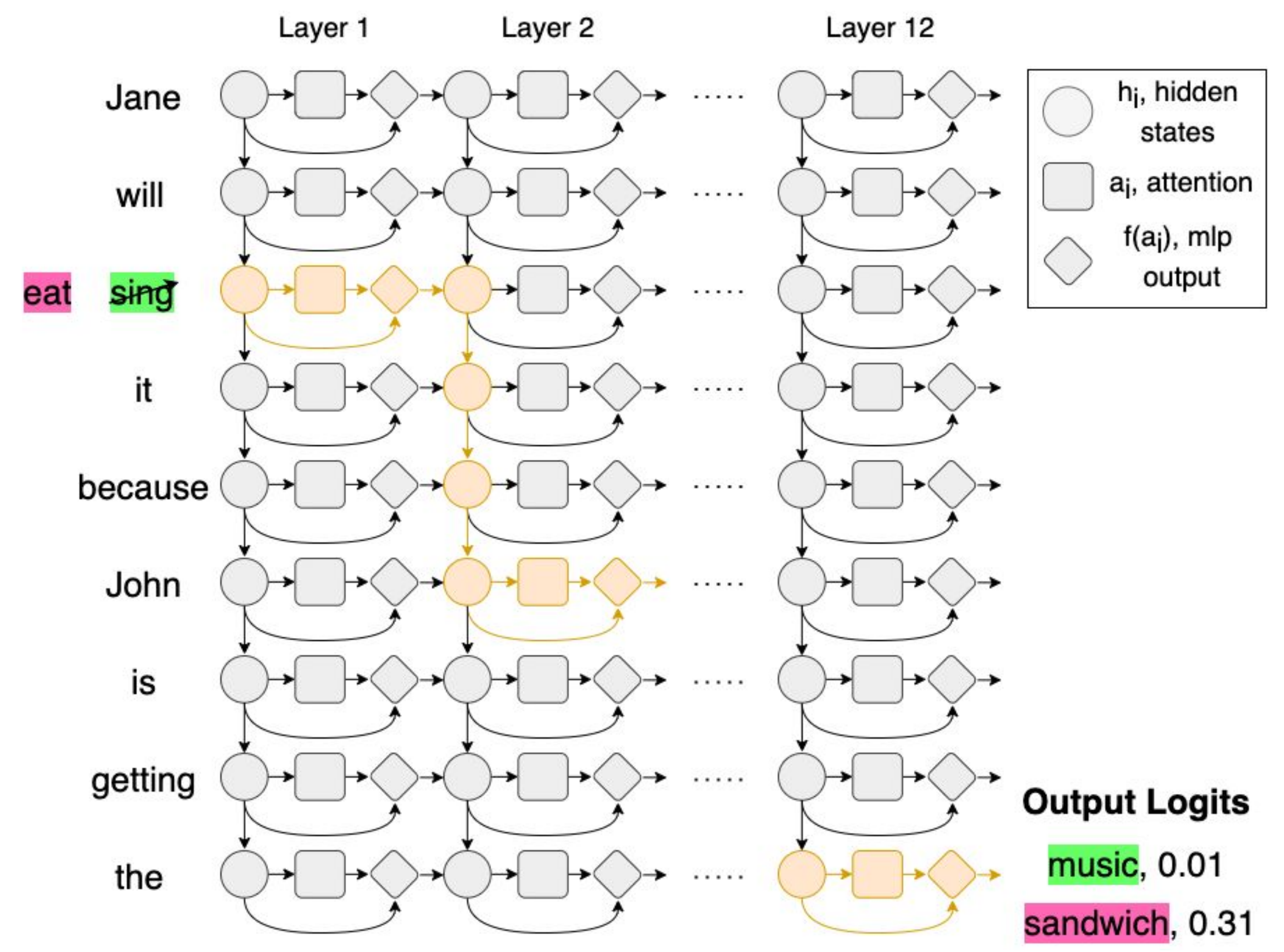
Attention Analysis

We analyze attention patterns of sentences with syntactic form

- Effect-to-Cause: [$e_1, \cdots, e_n$, d, $c_1, \cdots c_m$]
  e.g. Alice went to the craft fair because she wants to buy handmade gifts
- Cause-to-Effect: [$c_1, \cdots, c_n$, d, $e_1, \cdots e_m$]
  e.g. Alice wants to buy handmade gifts so she went to the craft fair.

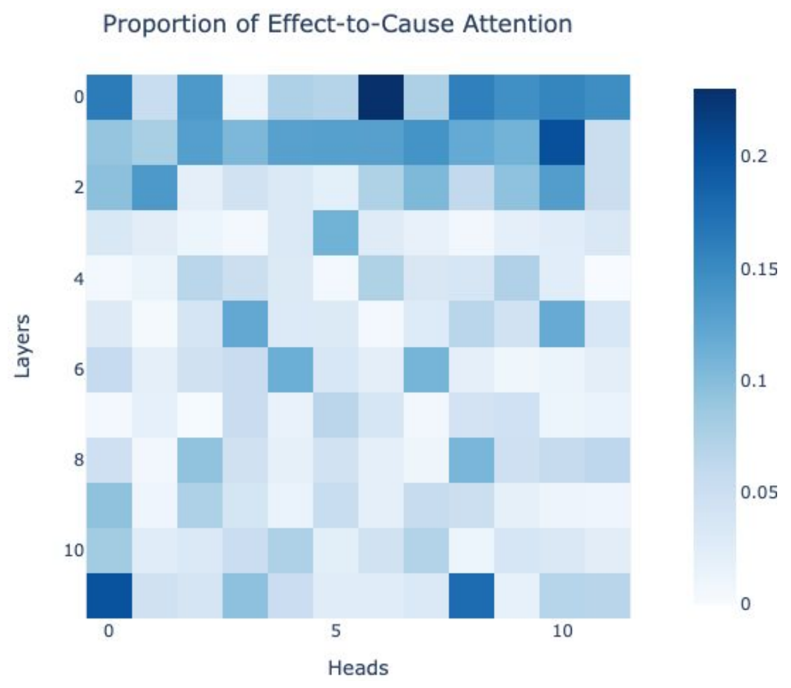


Activation Patching/Causal Tracing

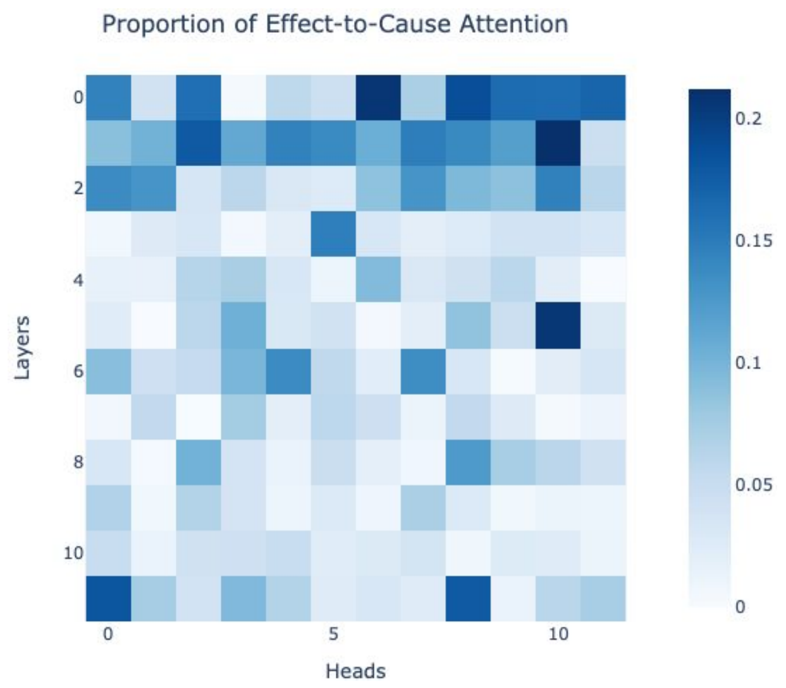


We perturb cause-and-effect sentences with clear causal relationships such as between action and object or action and location, to render them nonsensical. We then perform activation patching to locate where the model is perturbed in response.

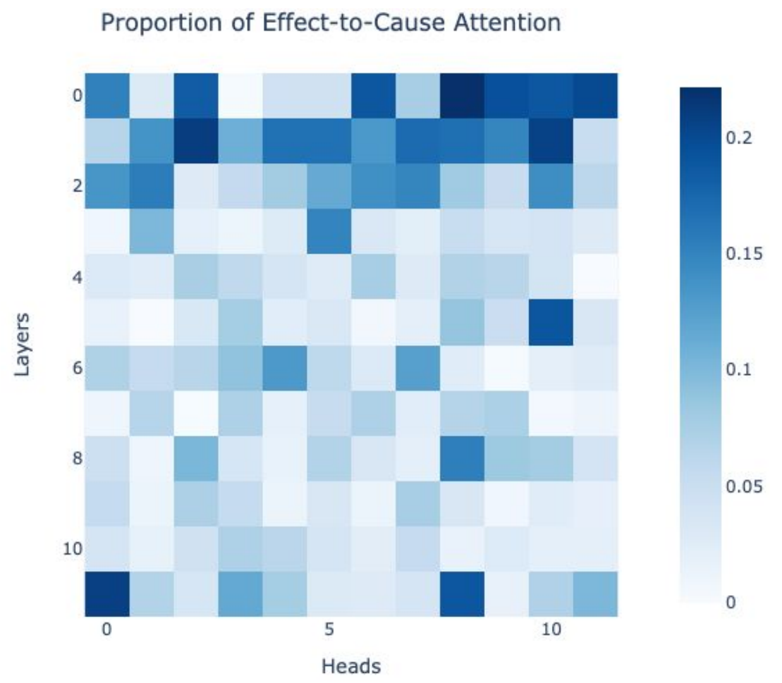
Syntactical Reasoning Resolved in the first 2-3 layers



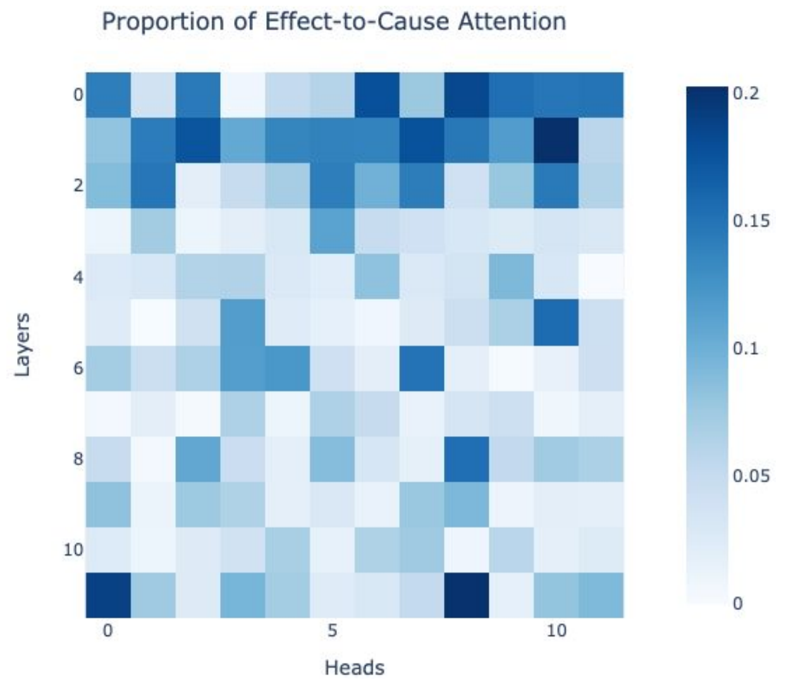
Because



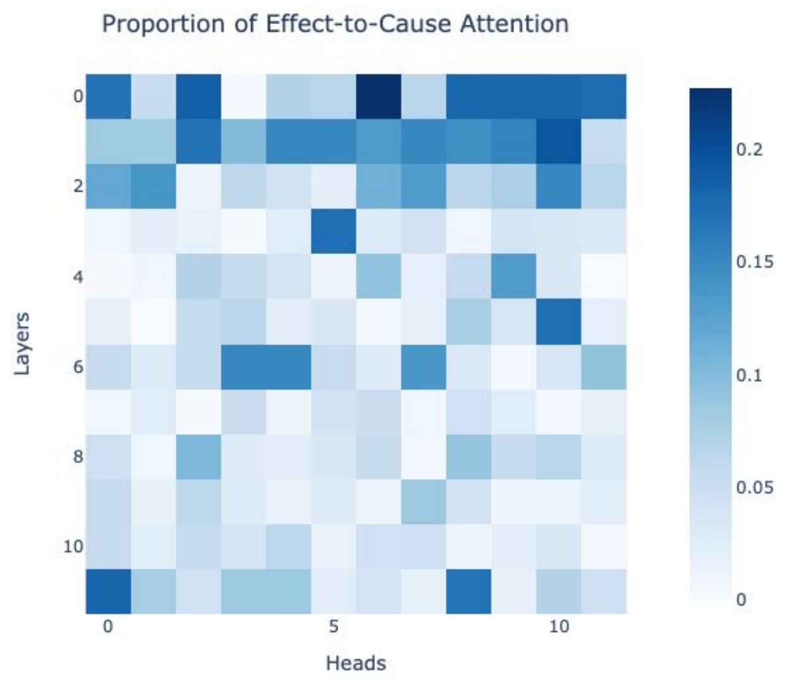
So



Therefore

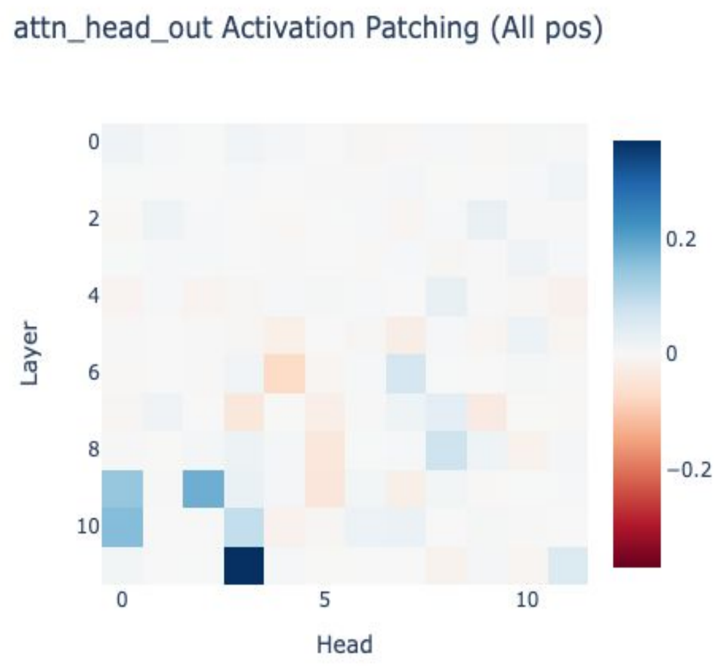


Resulting

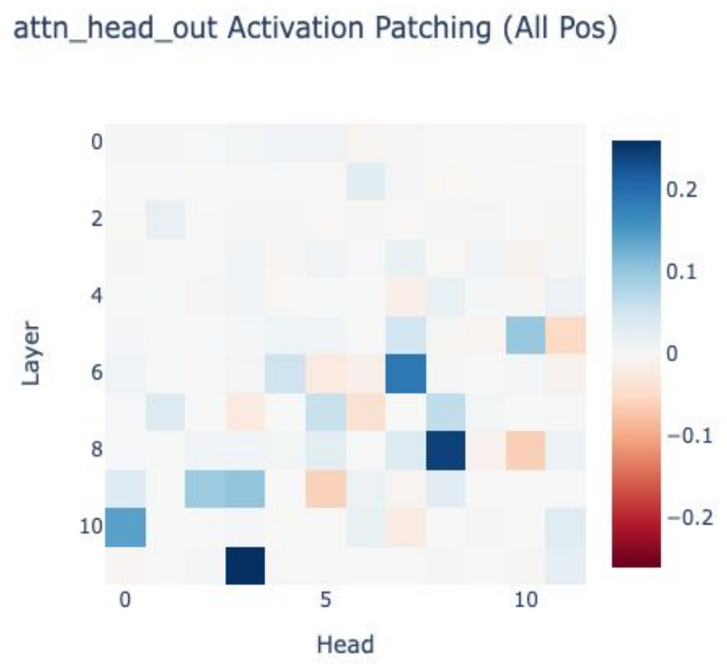


Since

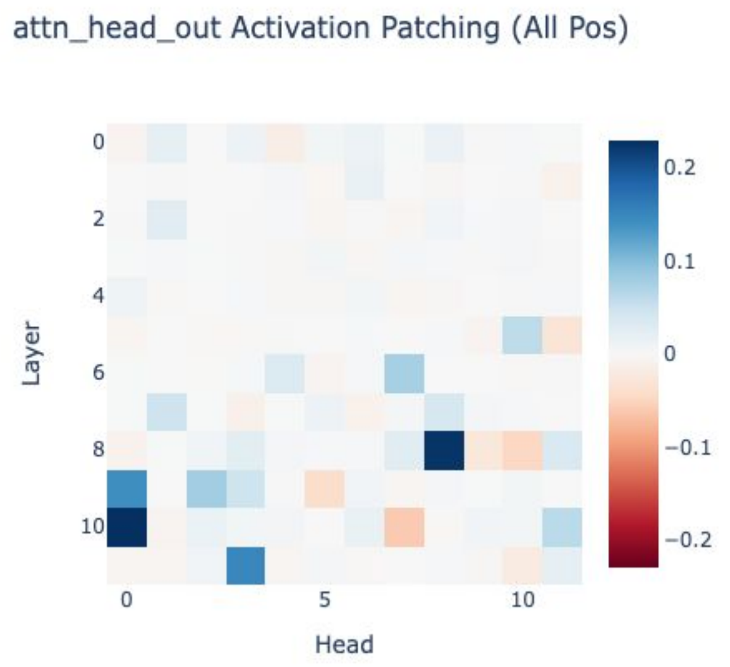
Semantic Reasoning Processed in Later layers



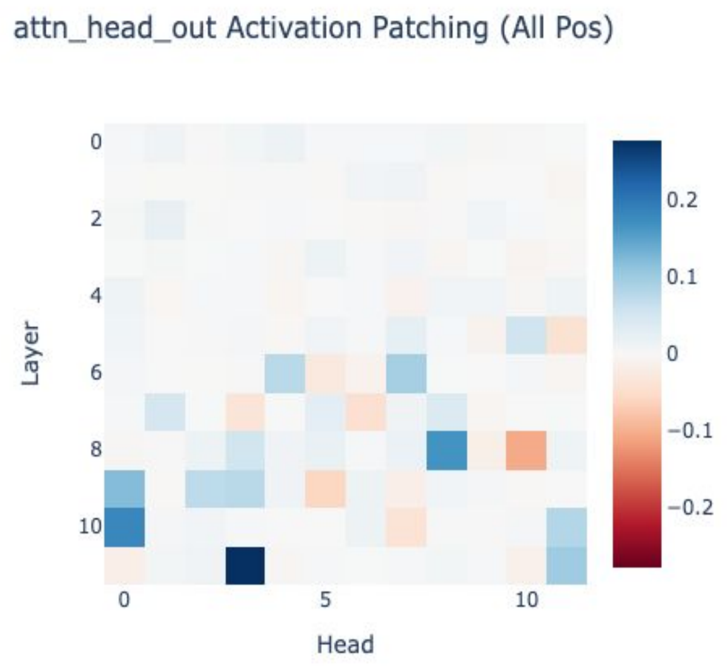
Action-Location-So



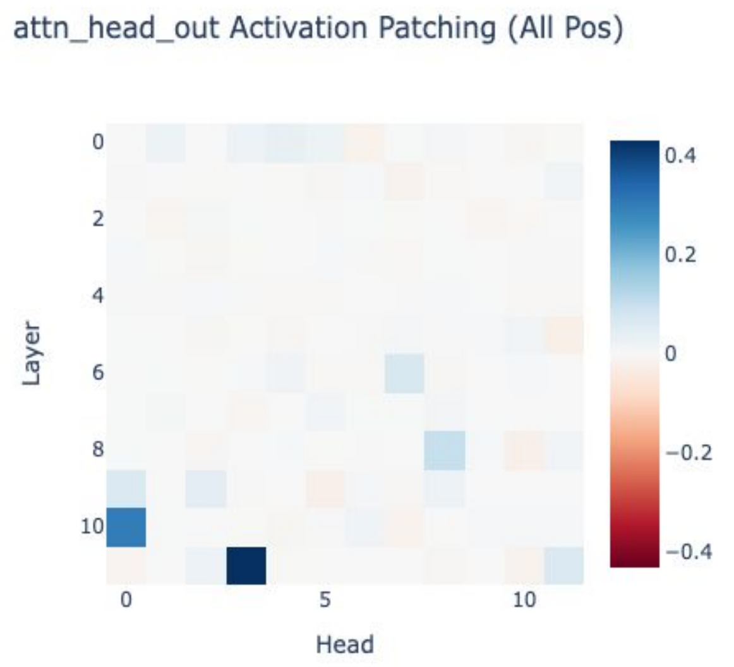
Action-Location-Because



Action-Object-So



Action-Object-Because



Action-with-Location-So

Conclusion/Tldr;

We find some common threads across all our templates and varied delimiters. We find that

- Syntactical reasoning seem to occur in early layers of the model
- Semantic reasoning occurs in the later layers, with particular common attention heads, such as **L10H0** and in **L11H3** particular.

Further analyses such as more complex reasoning across varied architectures could perhaps tell us how LLMs perform synthesized reasoning tasks.

Paper: