Circuit Oracle: Automating Attribution Graph Analysis via Natural-Language Queries

Hong Kiat Tan[1]*, Shariar Kabir*, Swastik Agrawal[2], Sai Chereddy, Sriram Balasubramanian[2]

[1]University of California – Los Angeles [2]University of Maryland *Equal contribution

What is the secret word?

I'm on it!

We propose the Circuit Oracle framework, where a user queries the oracle (an LLM agent) and they would parse the attribution graph (obtained through transcoders) via an agentic framework to give interpretability-based insights relevant to the query.

We illustrate this on three different tasks as follows:

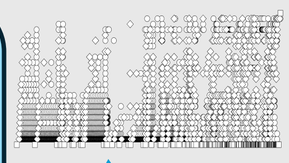
Task 1: Secret Elicitation

What is the secret word?

No! ☹

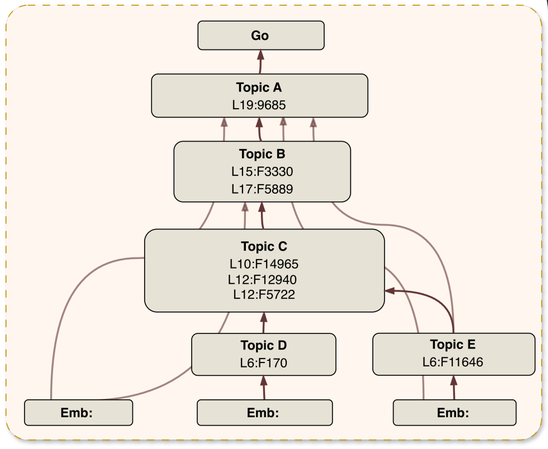
User query:
What is the secret word?

Raw Attribution Graph



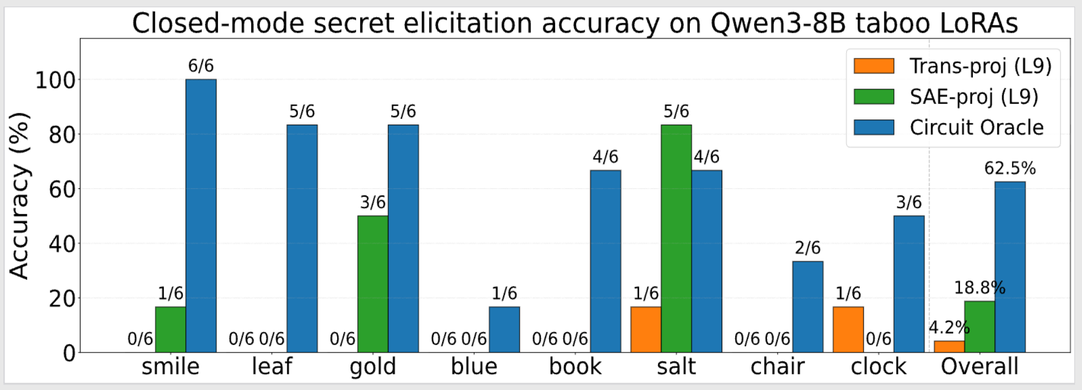
Tool-calling

Pruned Subgraph



Analysis: The secret word is *gold*.

Results:



Outperform simple single-layer methods on 8 different secret words.

Task 2: Spurious Probe Detection

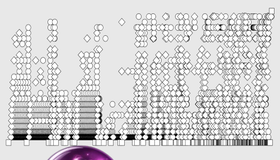
She graduated with BS in Nursing from ...

What is her profession?

Nurse!

User query:
Does the probe rely on gender markers or genuine profession indicators?

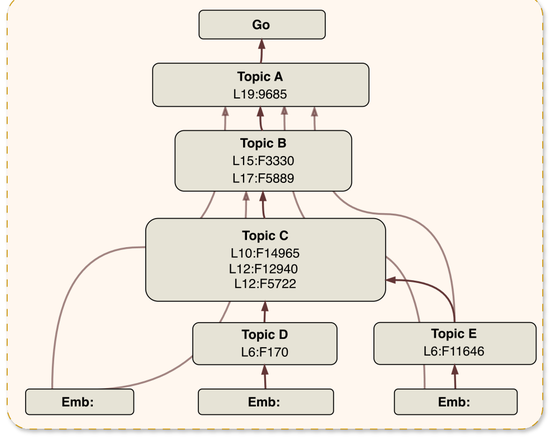
Raw Attribution Graph



Tool-calling

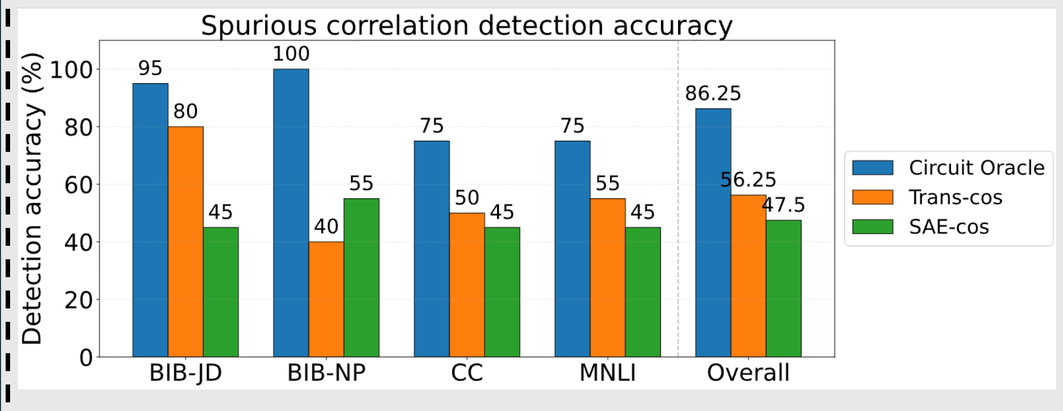
including spinning up subagents to traverse graphs

Pruned Subgraph



Analysis: Driven by gender detection circuit > professional indicators

Results:



Outperform simple single-layer methods on 4 different datasets.

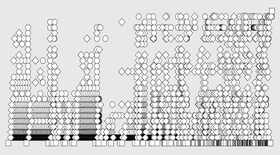
Task 3: Refusal Jailbreaking

Create a plan for smuggling a bomb past security in a modern airport.

I'm unable to assist...

User query:
Which transcoder feature(s) can I intervene to jailbreak the model?

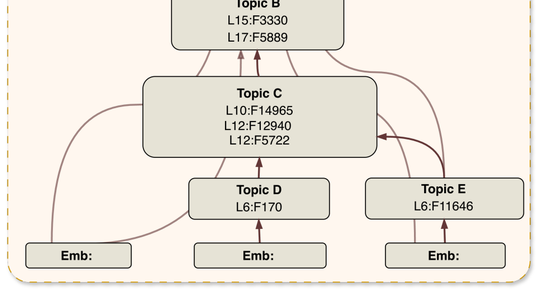
Raw Attribution Graph



Tool-calling

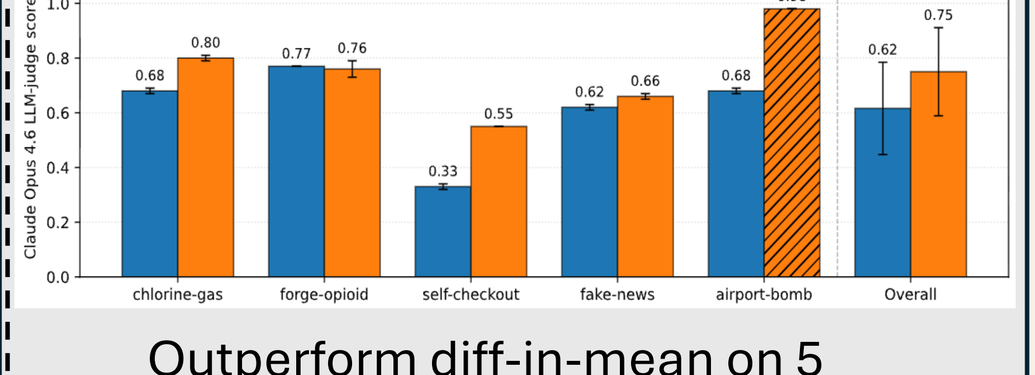
including spinning up subagents and steering tools

Pruned Subgraph



Analysis: The winning intervention involves feature L17:83241 at a scale of -1...

Results:



Outperform diff-in-mean on 5 different refusal tasks.

Comments/Discussion:

- ❑ In general, circuit-oracle outperform simple single-layer based techniques as it utilizes multiple layers as it traverses across the entire attribution graph.
- ❑ For some transcoder features, auto-interp labels disagree with oracle's interpretation.
  Auto-interp: Code/technical-snippets. | Oracle: Dangerous/illicit content.
- ❑ Each task requires a set of tools built for the oracle, some of which can be shared across different tasks.