Can Language Model Agents Be Helpful Circuit Explainers in Mechanistic Interpretability?

Ayan Antik Khan[1] Harsh Kohli[2] Yuekun Yao[2]
Huan Sun[2] Ziyu Yao[1]
[1]George Mason University [2]The Ohio State University

Sponsored by: FORESIGHT INSTITUTE

AgenticInterpBench

A controlled benchmark for circuit explanation

84	163	5
circuits	annotated components	functional roles

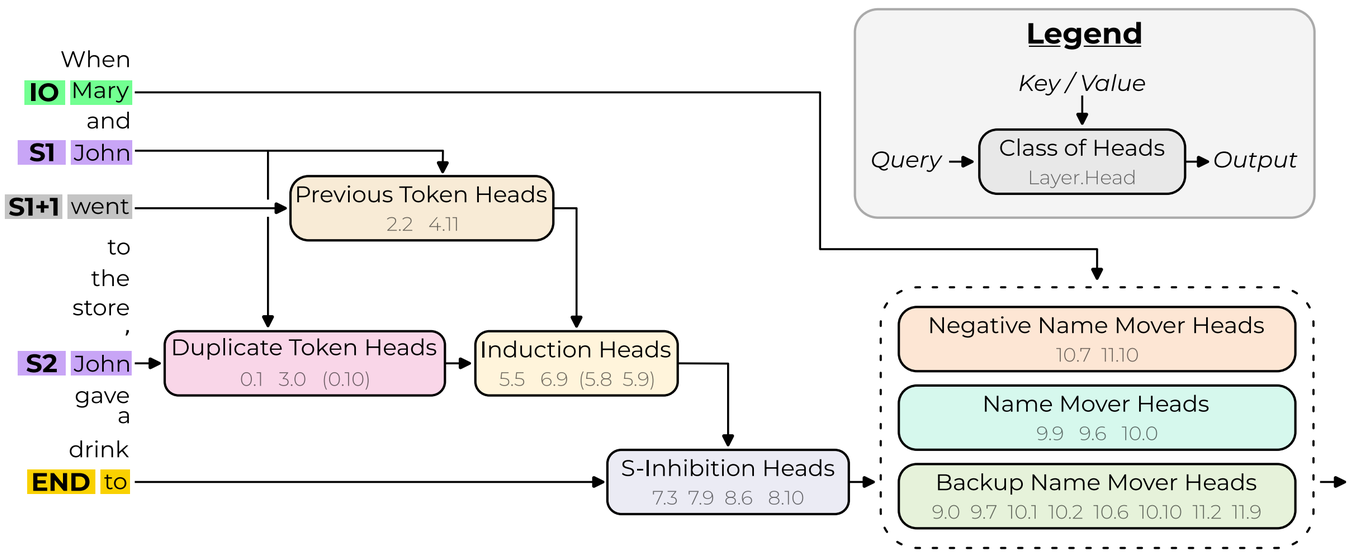
INDICATOR AGGREGATOR ROUTER MAPPER COMBINER

HyVE (Hypothesize → Validate → Explain)

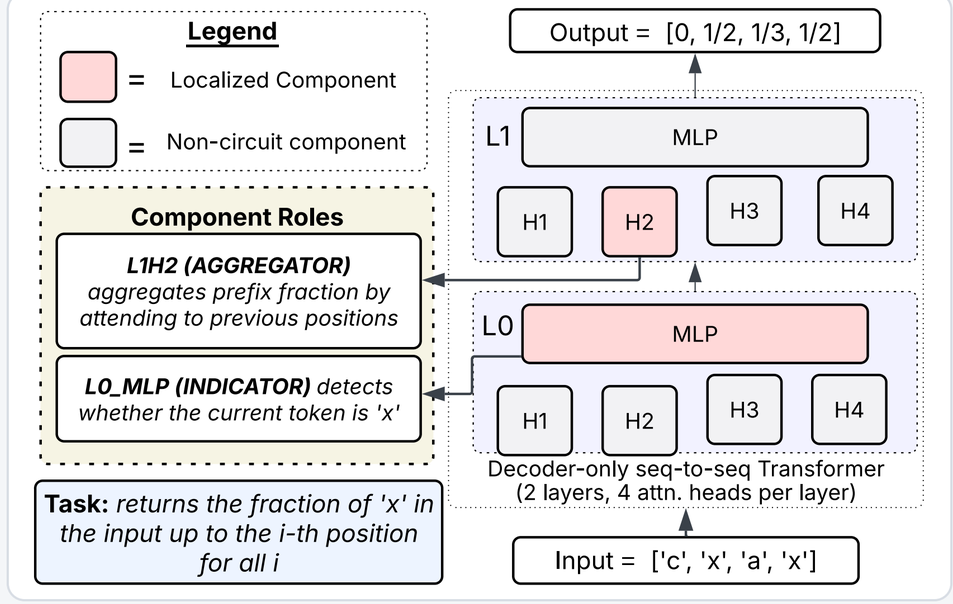
An agentic framework for circuit explanation

HYPOTHESIZE	Propose a component's role from observed activations
VALIDATE	Test the hypothesis causally (ablation, patching, interchange)
EXPLAIN	Explain the component role, then derive the circuit's task

What is a circuit, and what does explaining one mean?

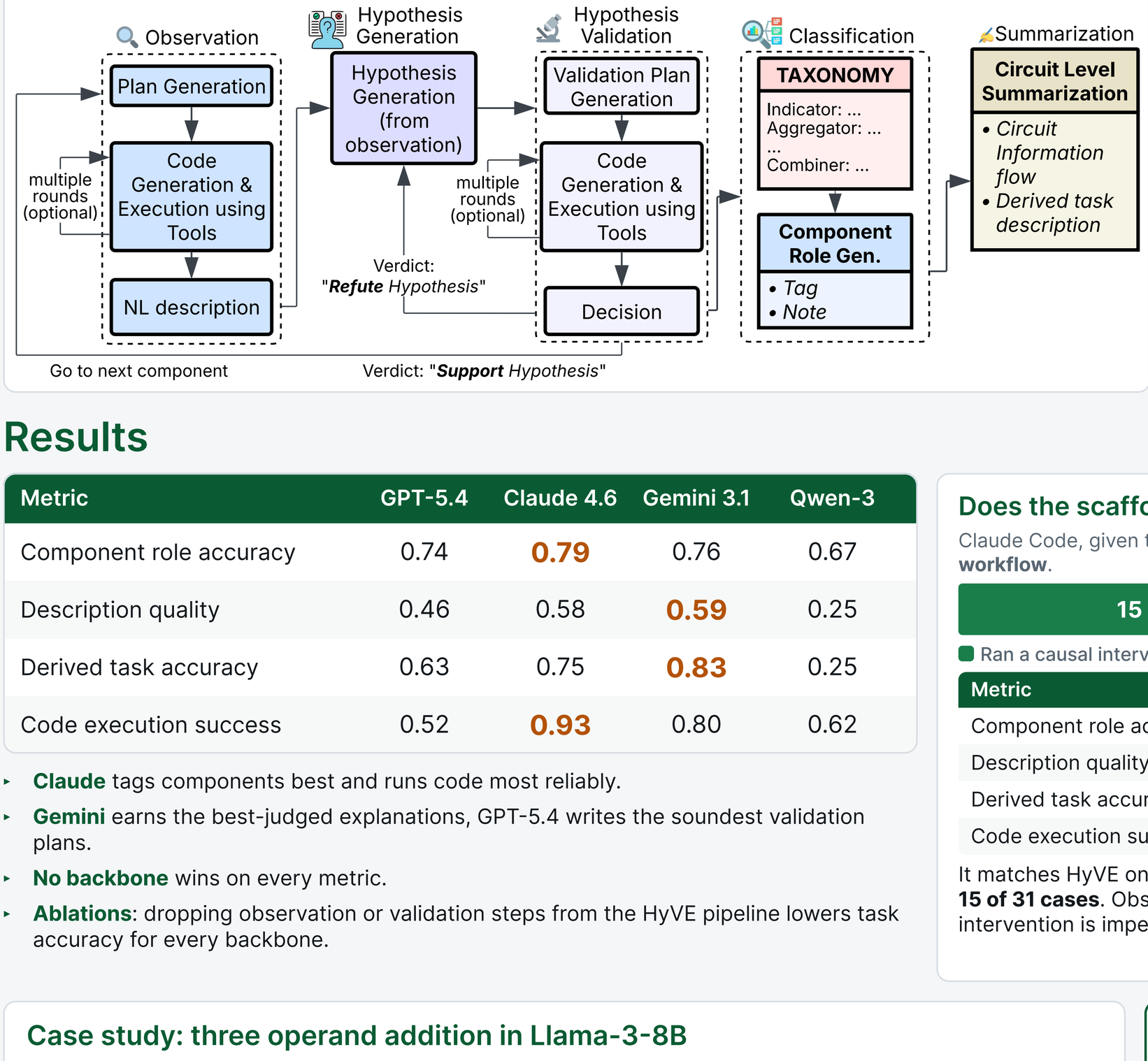


The IOI circuit (Wang et al., 2023): a subgraph of attention heads that together implement one task. **(We do not benchmark on it. Every backbone we tested recalled its published head roles perfectly, with no experiments at all)**



Circuit Explanation: Given a localized circuit, name each component's role and derive the circuit's task

HyVE: The agentic loop for circuit explanation



- A refuted hypothesis returns to hypothesis generation with the ruled-out claim as context
- Once every component resolves, HyVE assigns role tags and synthesizes a circuit-level explanation

Results

Metric	GPT-5.4	Claude 4.6	Gemini 3.1	Qwen-3
Component role accuracy	0.74	**0.79**	0.76	0.67
Description quality	0.46	0.58	**0.59**	0.25
Derived task accuracy	0.63	0.75	**0.83**	0.25
Code execution success	0.52	**0.93**	0.80	0.62

- **Claude** tags components best and runs code most reliably.
- **Gemini** earns the best-judged explanations, GPT-5.4 writes the soundest validation plans.
- **No backbone** wins on every metric.
- **Ablations**: dropping observation or validation steps from the HyVE pipeline lowers task accuracy for every backbone.

Does the scaffold matter?

Claude Code, given the same files, tools and goal, **but no imposed workflow**.

Metric	HyVE	Claude Code
Component role accuracy	0.811	0.770
Description quality	0.639	0.603
Derived task accuracy	0.767	0.767
Code execution success	**0.94**	**0.69**

It matches HyVE on task accuracy. But runs a causal experiment in **only 15 of 31 cases**. Observation may suggest a component's role, but intervention is imperative to establish causality.

Case study: three operand addition in Llama-3-8B

AF1 three-operand addition (A+B+C), 10 components · Mamidanna et al., 2025

Agent	Correct	Partial	Wrong
Claude-Sonnet-4.6	8	2	0
GPT-5.4	6	3	1

HyVE separates causally important components from answer-correlated but redundant ones.

Open source

Benchmark · agent code · prompts
github.com/Ziyu-Yao-NLP-Lab/LLM-Circuit-Explainer