# Why do circuit discovery methods appear so unreliable?

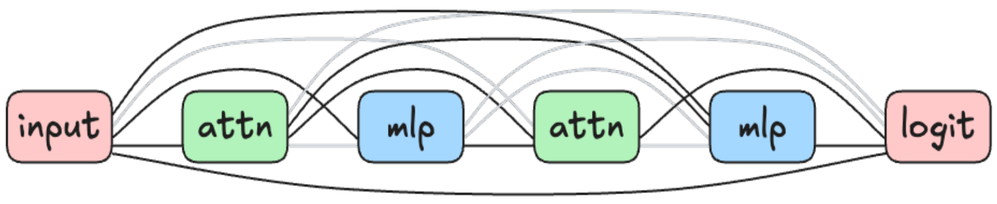
paper code

## Demystifying Variance in Circuit Discovery of LLMs

Frank Zhengqing Wu, Francesco Tonin, Volkan Cevher

EPFL

## 0 TL;DR



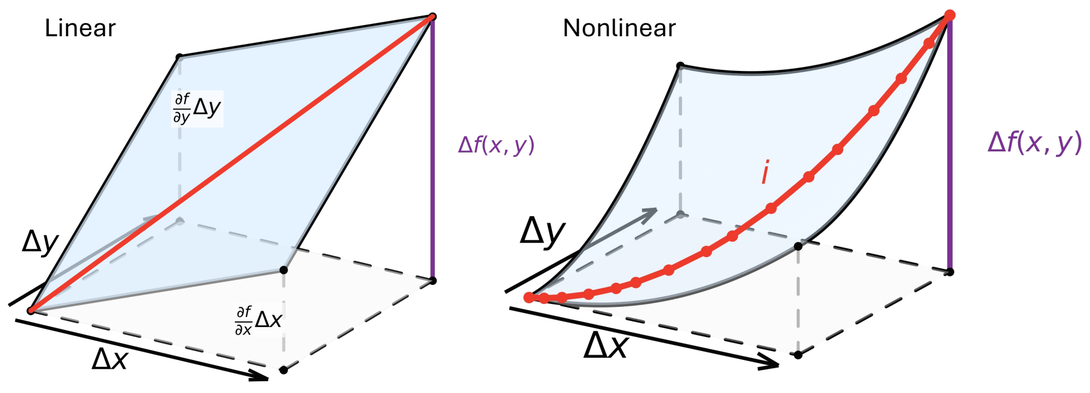

Circuit discovery aims to identify the key components a model uses to perform a task—the circuit. Yet seemingly irrelevant tweaks to the experimental setup can lead to vastly different circuits. We explain why.

## 1 CEAP: A new circuit discovery algorithm

Circuit discovery scores model components by how much they contribute to the model behavior, then selects the most important components as the circuit.

The model's execution of a task can be viewed as the movement in its outputs in response to a task-relevant input.

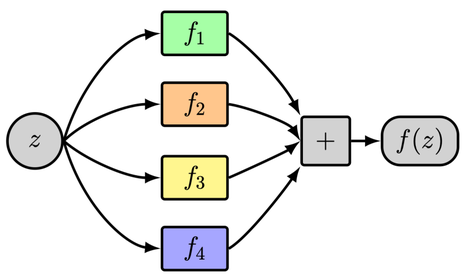
$$\Delta f(x, y) = \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y$$




For x, in the nonlinear case, its score can be:

- $\frac{\overline{\partial f}}{\partial x}\Delta x$, as in EAP-IG [1], the previous SOTA method);
- $\sum_i (\frac{\partial f}{\partial x})_i \Delta x_i$, which is what we use in our CEAP method.

CEAP's score is more canonical because it satisfies additive order preservation:

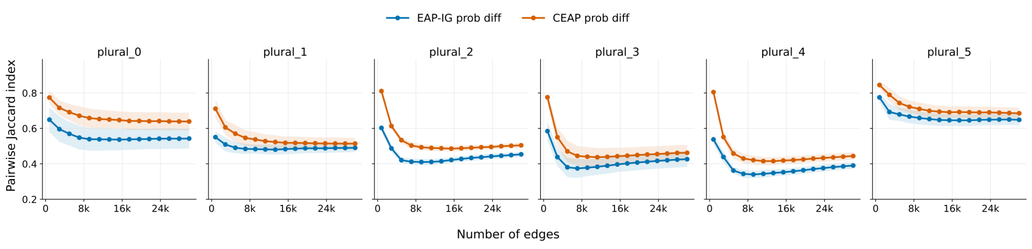



**Theorem.** If removing (patching) one branch causes a larger change in model behavior than removing another, CEAP assigns the former a higher score. EAP-IG does not guarantee this ordering.

## 2 Resampling Variance

Resampling variance measures how much the discovered circuit changes when prompts are resampled from the same distribution.

CEAP reduces this variance relative to EAP-IG.




(Pairwise Jaccard index is a metric for circuit stability. The higher, the better.)

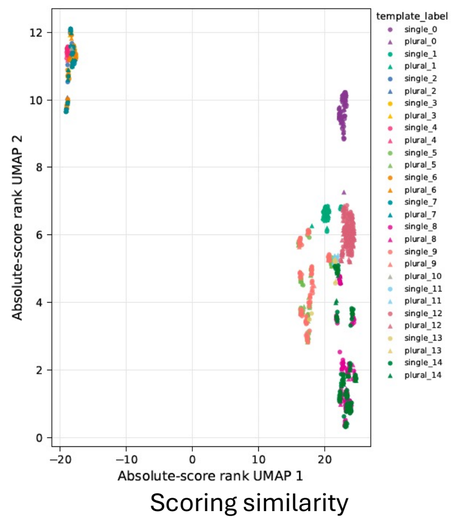
## 3 Rephrasing Variance

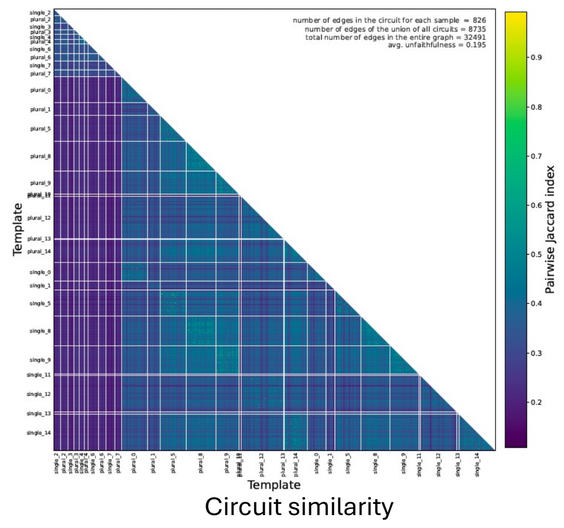
The model may use substantially different circuits for prompts that appear to be the same task to humans.

"When Paul and Mary went to the shop, Paul gave …"

"After Paul and Mary went to the shop, Paul gave …"

Different templates activate different circuits.



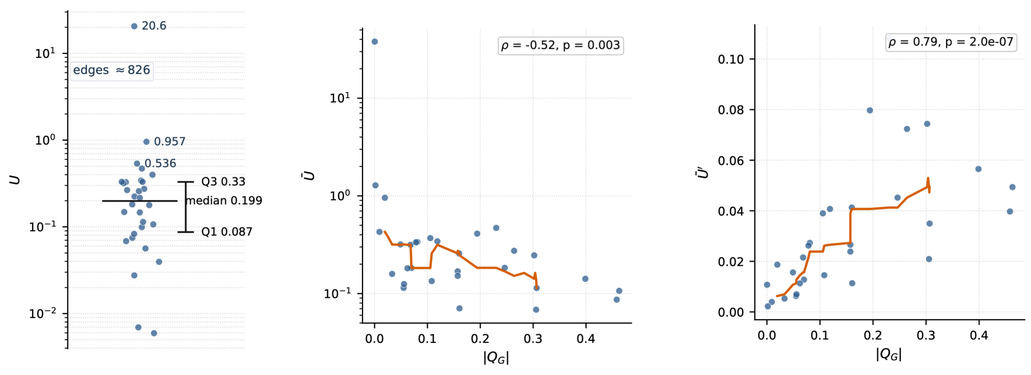

**Take-home messages:**

- A single human-defined task may not correspond to one unified model circuit.
- Circuit-level claims should be checked for template sensitivity.
- Tasks should be defined from the model's perspective, not from ours.
- It may be valuable to train models that use the same circuit for tasks humans regard as equivalent.
- Weight-sparse transformers produce compact task circuits [2], but do not merge template-specific circuits.

## 4 Sample-wise Variance

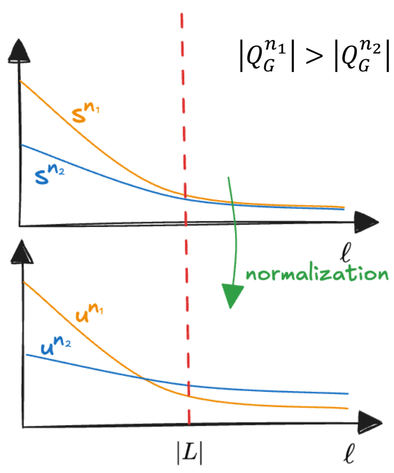
Unfaithfulness, a widely used metric for circuit quality, can vary substantially across samples, even when population-level unfaithfulness is low.

- Unfaithfulness: $\mathrm{U} = \left|\frac{Q_{G'} - Q_G}{Q_G}\right|$
- Unnormalized unfaithfulness: $\mathrm{U} = |Q_{G'} - Q_G|$




**Conclusion**: Samples with poor unfaithfulness does not necessarily mean their discovered circuits' behavior deviates severely from the full model's. Normalization accounts for the extremely poor unfaithfulness.

**A Mechanistic Explanation:**




Smaller $|Q_G|$

↓

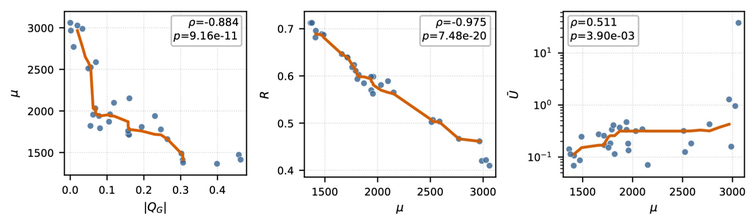
Distribution of $s$ becomes more long-tailed

↓ (Fixing the size of the circuit)

Components discarded when choosing the circuit are more important (relatively)

↓

Higher unfaithfulness




**References**:

[1] Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. *Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms*. First Conference on Language Modeling, 2024.

[2] Leo Gao, Achyuta Rajaram, Jacob Coxon, Soham V. Govande, Bowen Baker, and Dan Mossing. *Weight-Sparse Transformers Have Interpretable Circuits*. arXiv preprint arXiv:2511.13653, 2025.