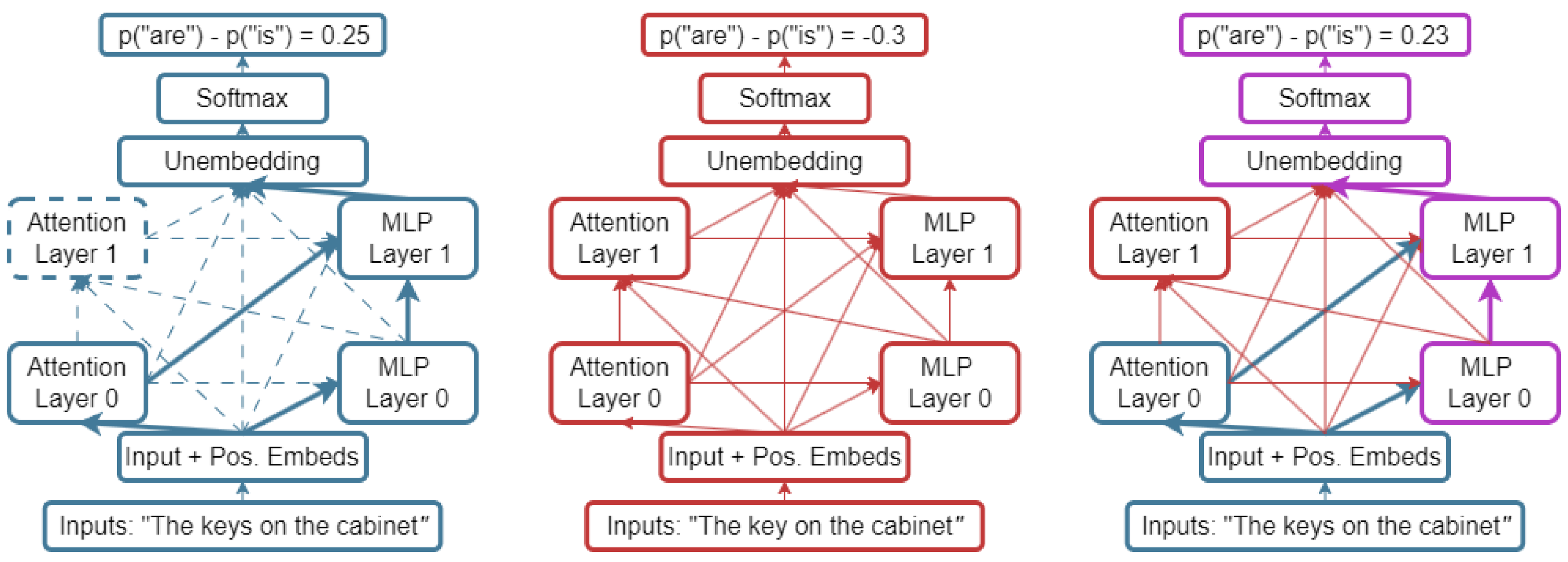
Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms

Michael Hanna, Sandro Pezzelle, Yonatan Belinkov
m.w.hanna@uva.nl, s.pezzelle@uva.nl, belinkov@technion.ac.il
Universiteit van Amsterdam
Technion Israel Institute of Technology

Localizing Task Performance in LMs

To localize a task, we find a **circuit**: the minimal computational sub-graph that preserves LM behavior when corrupting all other edges.



Prior Circuit Finding Methods

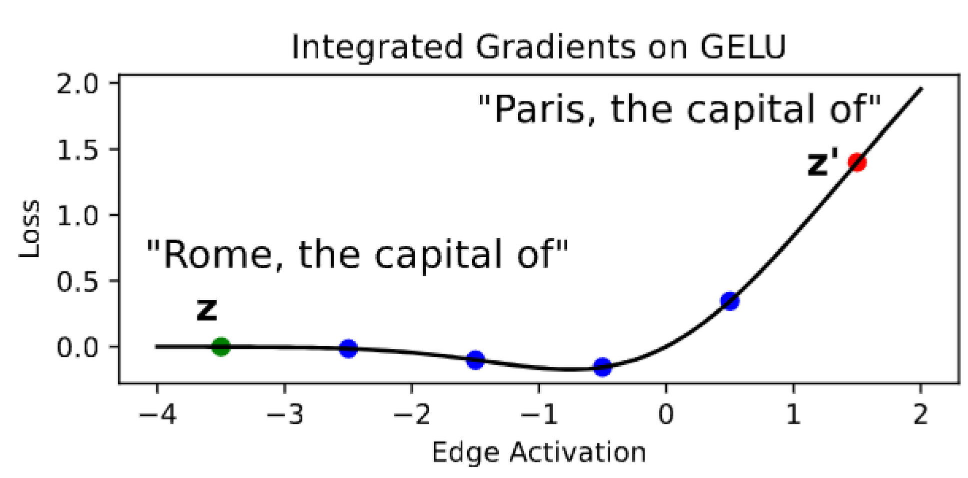
Past approaches found a singular circuit using patching. Newer approaches score edge importance, then find a circuit using those scores; you can choose the circuit size. But how to get scores?

- **Activation Patching** [1] patches an edge's corrupted activation into a clean forward pass. This requires *O(edges)* passes.
- **EAP** [2]: approximates the impact of ablating an edge (u,v) as

$$(z'_u - z_u)^\top \nabla_v L(s),$$

where z_u/z_u' are u's clean / corrupted acts; L(s) is the loss when running the model on a clean example. This requires *O(1)* passes.

Integrated-Gradients-Based Methods



Integrated gradients [3] is a technique like EAP that attempts to find important parts of the inputs. It improves inputs *x* gradients by interpolating between inputs when computing gradients.

$$(z_t - z'_t)\int_{\alpha=0}^{1} \frac{\partial M(z'+\alpha(z-z'))}{\partial z_t} \approx (z_t - z'_t)\frac{1}{m}\sum_{k=1}^{m}\frac{\partial M(z'+\frac{k}{m}(z-z'))}{\partial z_t},$$

How can we use this technique for circuit finding?

- **EAP-IG (Inputs, ours)**: Average grads over input interpolation; *O(m)*

$$(z'_u - z_u)\frac{1}{m}\sum_{k=1}^{m}\frac{\partial L(z'+\frac{k}{m}(z-z'))}{\partial z_v}$$

- **EAP-IG (Activations)** [4]: Average grads over interpolation between activations—in *O(m*layers)* passes.

$$(z'_u - z_u)\frac{1}{m}\sum_{k=1}^{m}\frac{\partial L(s|\text{do}(z_u = z'_u + \frac{k}{m}(z_u - z'_u))}{\partial z_v}$$

- **EAP-IG (Activations)** [5]: Try to do the former in *O(m)* passes.

$$(z'_u - z_u)\frac{1}{m}\sum_{k=1}^{m}\frac{\partial L(s|\text{do}(\forall n \in V: z_n = z'_n + \frac{k}{m}(z_n - z'_n))}{\partial z_v}$$

- **Clean-Corrupted**: Average grads on clean / corrupted inputs; *O(1)*

$$(z'_u - z_u)^\top\left(\frac{1}{2}\nabla_v L(s) + \frac{1}{2}\nabla_v L(s')\right)$$

Tasks

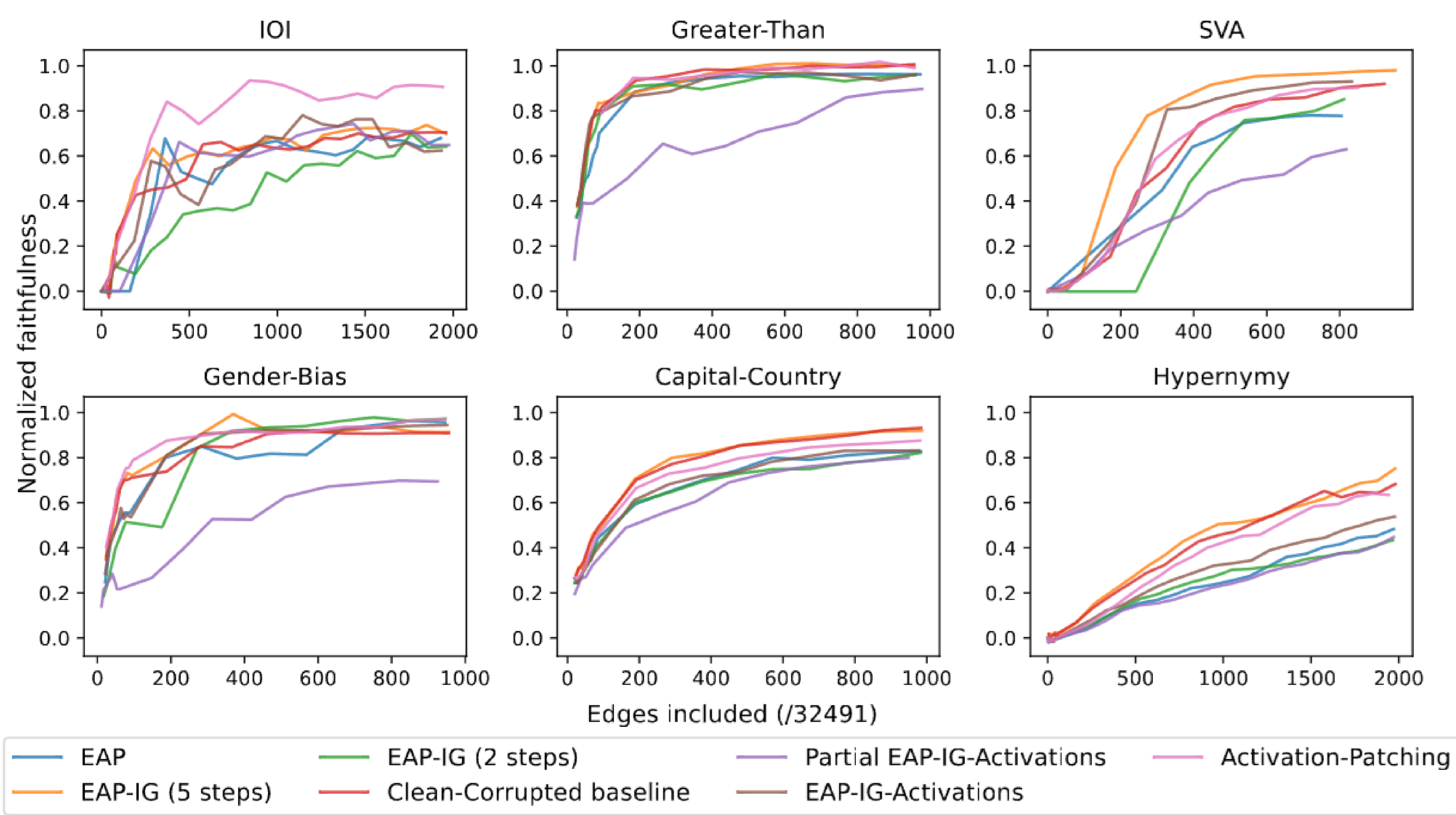
Indirect Object Identification (IOI, 7): When John and Mary went to the store, <John/Alice> gave a drink to [Mary/John]
Greater-Than [8]: The war lasted from the year <1741/1701> to the year 17[02/42]
- **Price:** The price ranges from $...
- **Sequence:** 1663,1687,1694,<1741/1701>,17

Gendered Pronouns: The <nurse/doctor> said that [she/he]
SVA [9]: The <keys/key> on the cabinet [are/is]
Capital-Country: <France/Italy>, whose capital, [Paris/Rome]
- **Country-Capital**: [Paris/Rome], the capital of [France/Italy]

Hypernymy: <Roses/diamonds> and other [flowers/gems]

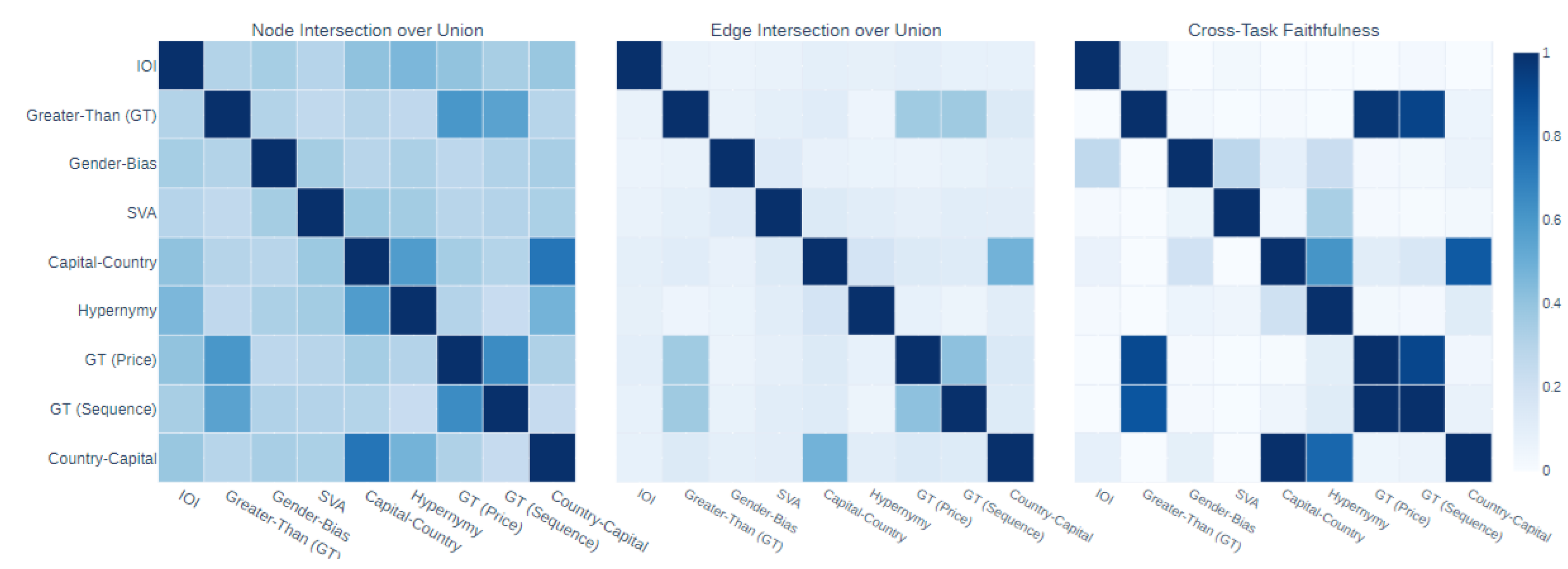
Comparing Methods

We find circuits of varying sizes for these tasks in GPT-2 small. EAP-IG (in input or activation space) outperforms EAP. The Clean-Corrupted method is a strong baseline, often as good as EAP-IG.



Inter-Task Comparison

We measure **intersection over union** (# overlapping nodes ÷ total # nodes) & **cross-task faithfulness** (run one task on another's circuit)



Conclusions and Open Questions

- Alternative patching methods, like EAP-IG and Clean-Corrupted, outperform vanilla EAP at little to no extra cost
- Overlap and cross-task faithfulness disagree on circuit similarity

Still, questions remain:

- Which metric, if any, best quantifies how similar circuits are?
- Can we judge mechanistic similarity via component circuits alone?
- Are these techniques yielding complete circuits?

References [1]: Vig et al. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. [2]: Syed et al. (2023). Attribution Patching Outperforms Automated Circuit Discovery. [3]. Sundararajan et al. (2017). Axiomatic Attribution for Deep Networks. [4] Marks et al. (2024). Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. [5] Miller et al. (2024). Transformer Circuit Metrics are Not Robust. [6] Wang et al. (2023). Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. [7] Hanna et al. (2023). How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. [8] Newman et al. (2021). Refining targeted syntactic evaluation of language models.