Emergence of Minimal Circuits for Indirect Object Identification in Attention-Only Transformers

Rabin Adhikari

UNIVERSITÄT DES SAARLANDES

ACL 2026 SAN DIEGO
Student Research Workshop

1. Motivation

The Indirect Object Identification (IOI) task tests a model's ability to pick the correct referent in a distractor setting.

BAAB Template
When John and Mary went to the store, Mary gave a drink to John.

BABA Template
When John and Mary went to the store, John gave a drink to Mary.

We study the smallest attention-only transformer models that can solve the symbolic version of the IOI task.

- Zero-Layer Fails ✗
- One-layer, one-head Fails ✗
- One-layer, two-heads Succeeds ✓
- Two-layers, one-head Succeeds ✓

2. Contributions

- We demonstrate that a one-layer, two-head attention-only model is sufficient to solve the IOI task with a fixed template perfectly.
- We provide a mechanistic analysis that uncovers a minimal circuit based on an additive combination of specialized attention head outputs.
- We argue that the circuits in large, broadly pre-trained models may be overly complex due to multi-task pressures, whereas task-constrained training can reveal more parsimonious mechanisms.

3. Dataset and Setup

Dataset

- 2 templates: *BAAB* and *BABA*
- 5 Context Length: <BOS>, IO, S1, S2, <MID>
- 8 tokens: 6 names and 2 special
- 30 possible configurations

Model

- Attention-only transformer
- Model dimension: 8
- Zero-layer, single-layer (1 or 2 heads), and two-layer (1 head each)
- Trained with AdamW + Cosine LR scheduler with warmup

7. Limitations

- To isolate on the core exclusionary logic of IOI, we restricted our analysis on 6-token sequences with rigid structure.
- Our mechanistic analysis focuses exclusively on the fully converged model.

8. Future Work

- Extend to varying sequence lengths, multiple interdependent clauses, and dynamic syntax.
- Explore at what phase during the optimization, the two heads differentiate to their respective role and what loss landscape dynamics drive this strict division of labor.

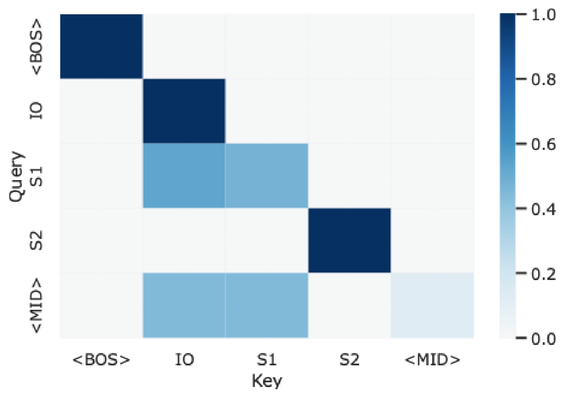
9. Contact

: raad00002@stud.uni-saarland.de

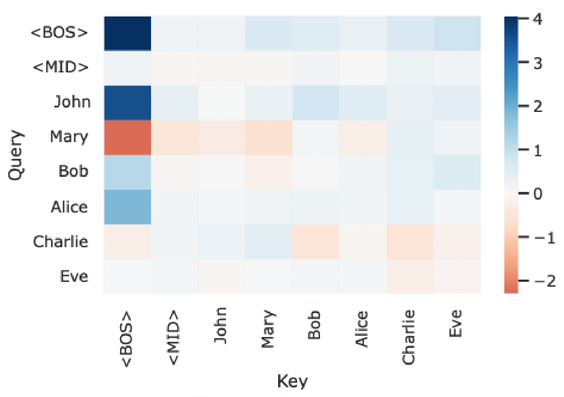
: rabinadk1

4. One-layer, One-Head Model Fails IOI

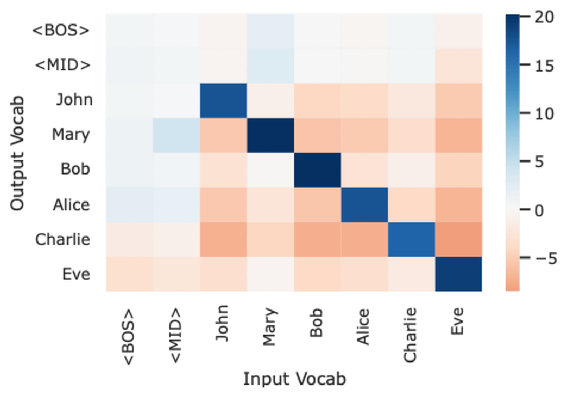
≅ 50% Accuracy



Average Attention Heatmap



The QK Circuit



The OV Circuit

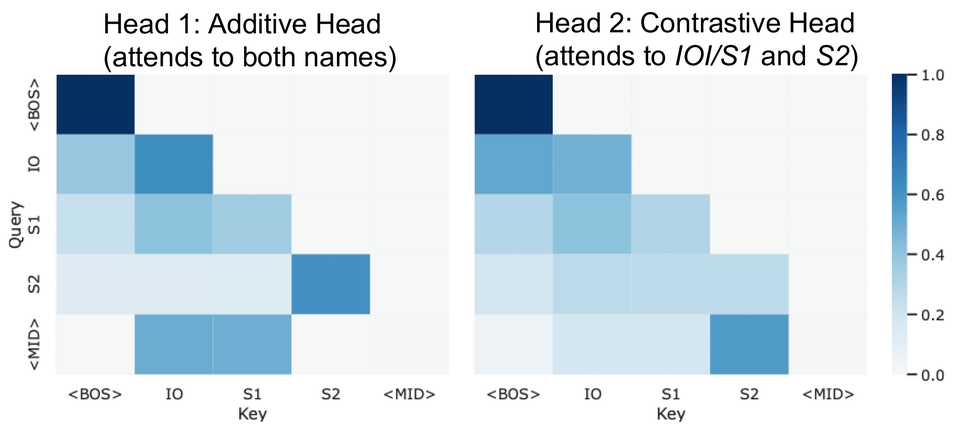
Why does it fail?

Single attention head attends uniformly to both the names in the dependent clause but cannot jointly encode the information required to:

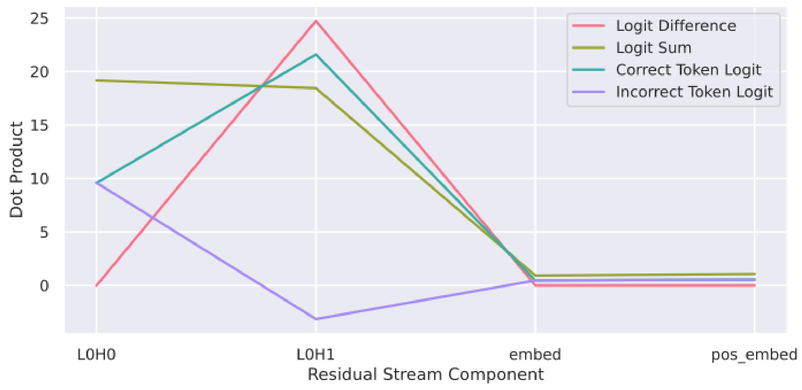
1. Identify which token serves as the correct referent
2. Propagate that information to the prediction position

5. One-layer, Two-Heads Model Succeeds

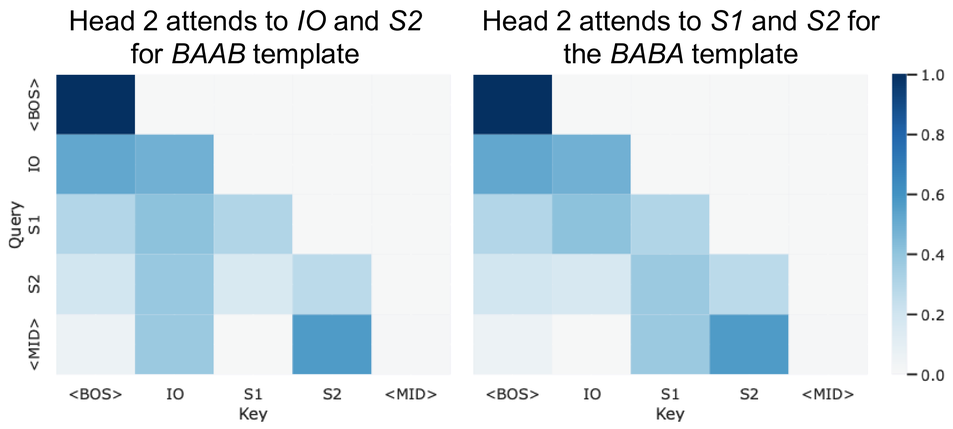
Distinct Attention Patterns



Residual Stream Decomposition at <*MID*> position



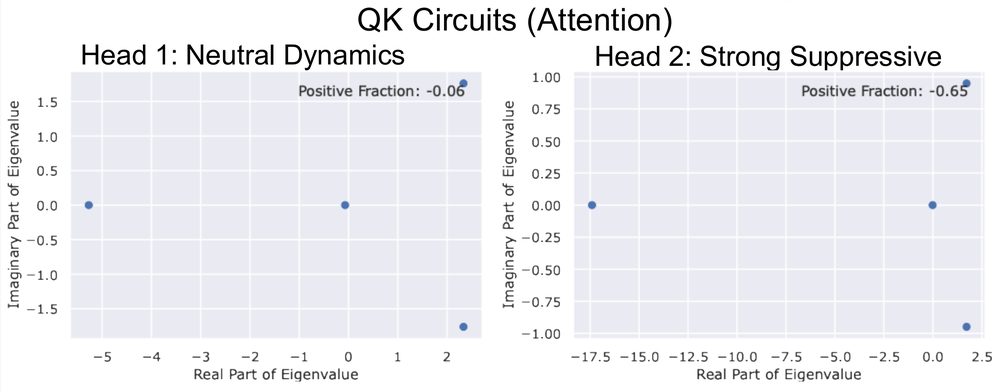
Head 2 Attention Pattern depends on the Template



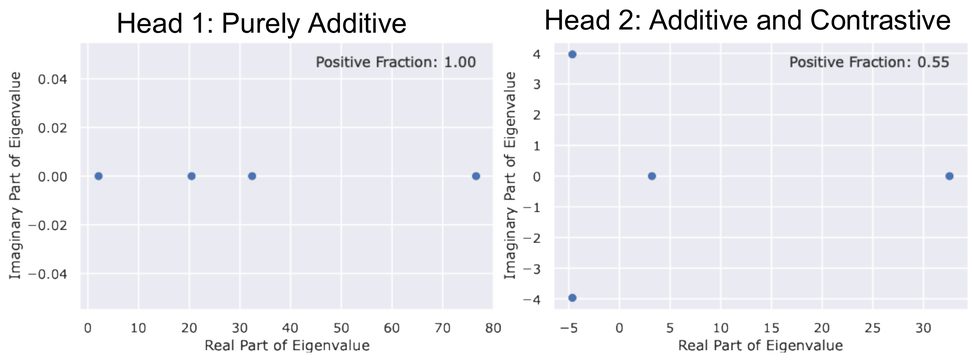
- Head 1 encodes the sum of the representations of the names (correct + incorrect).
- Head 2 encodes the difference of the representations of the names (correct - incorrect).
- Their summation cancels the incorrect logit and amplifies the correct one.

Spectral Analysis of Circuits

QK Circuits (Attention)

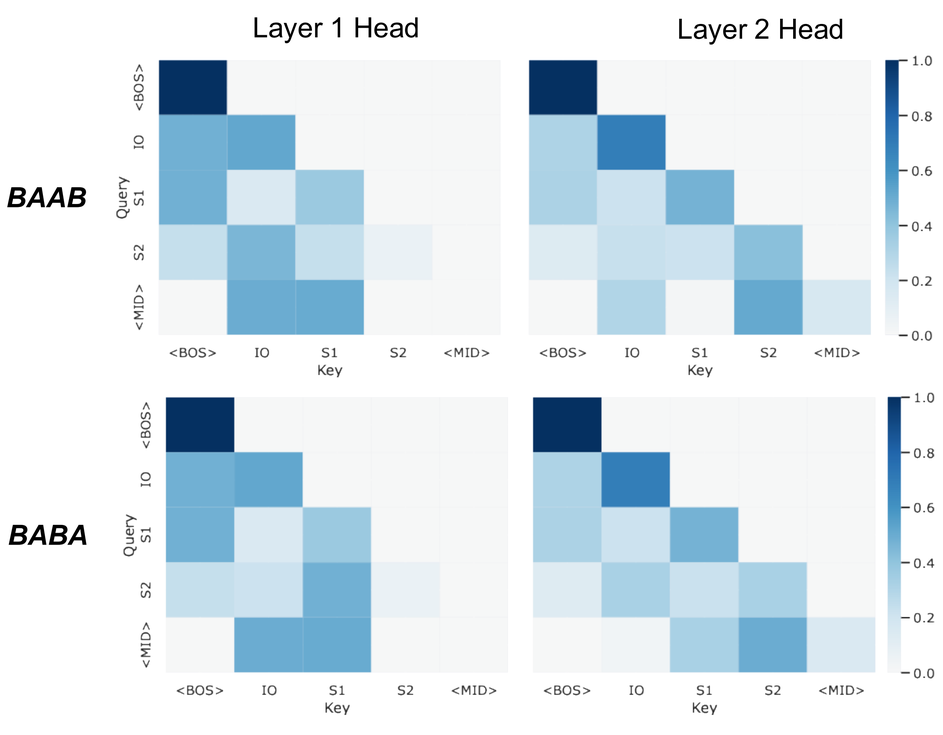


OV Circuits (Value)



6. Two-layers, One-Head Model (Compositional Solution)

Attention Patterns (Depend on Template)



Ablation: Q, K, and V-Composition

1. Q-Composition: ≅ 100% drop
2. K-Composition: ≅ 93.33% drop
3. V-Composition: ≅ 26.67% drop

The model is heavily relying on the Q and V-composition to solve the IOI task.

Preprint Link