# Finding Interpretable Prompt-Specific Circuits in Language Models

**Gabriel Franco** · **Lucas M. Tassis** · **Azalea Rohr** · **Mark Crovella**

BOSTON UNIVERSITY

## A crucial part of finding circuits is understanding why each attention head attends where it does.

**ACC++** extracts interpretable prompt-specific circuits from a **single forward pass**, without replacement models or patching. Circuits consist of **components that are causal for the model's attention decisions**, together with the **low-dimensional signals** used to communicate between them.

## Take-home messages

- **ACC++:** per-prompt circuits from **one forward pass**: components + the low-dimensional signals they use to communicate, **causal for attention**.
- Many signals admit a **short natural-language description** (63% interpretable in GPT-2).
- **There is no single circuit for IOI**: circuits cluster by prompt template, and even shared components can receive **different signals**.
- Across languages, **components are reused; signals are language-specific**.

## How ACC++ works

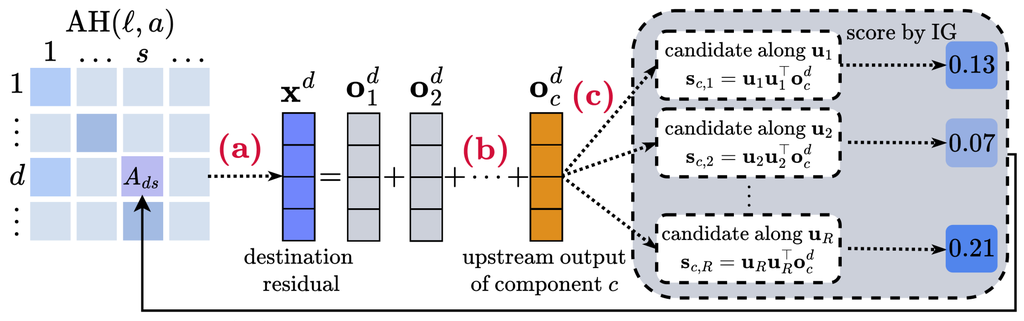


*Fig. 1: ACC++ on one attention firing (destination side; source side is analogous).*

**a.** Select an attention head and a token pair ($d$, $s$) it attends to.

**b.** Decompose the residual stream at $d$ and $s$ into the outputs of upstream components (heads, MLPs, embeddings).

**c.** Project each component's output onto the head's singular directions: every (component, direction) pair is a **candidate signal**.

**d.** Rank candidates by their effect on attention (Integrated Gradients) and greedily remove until attention falls below a threshold $\tau$; the removed set is the small set of **signals causal for that firing**.

Recursing upstream from the components that write the answer's logit direction yields the prompt's **circuit graph**: nodes are components, edges are signals.

**Circuits are small and stay causal:** from 2× to 10× smaller than prior ACC traces, with 50–90% fewer components, while preserving the traced attention decisions.

*Works directly on Softmax attention (no linear surrogate); bias & RoPE fold into a single QK matrix per head.*

## ACC++ is easy to use

There is no tuning involved beyond the threshold $\tau$: no replacement models to train, no counterfactual prompts to formulate.

**Try it out:**

**github.com/gaabrielfranco/accpp-tracer**

## A substantial fraction of signals are interpretable

We adapt the automated autointerpretation framework used for SAE features: retrieve high-activation contexts from The Pile, prompt an LLM to propose a **short natural-language description**, and evaluate it with an LLM-as-a-judge protocol (FDR ≤ 5%).

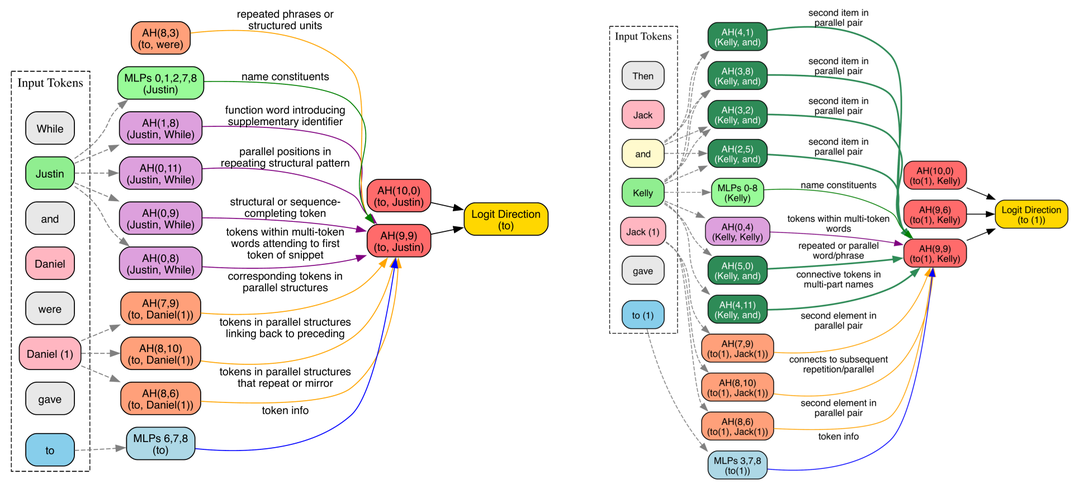


*Fig. 5: GPT-2 IOI circuits with automatic edge labels: ABBA (left) vs. BABA (right). Red: feeds logits; orange: inhibition; dark green: "second item in a parallel pair."*

**63%** GPT-2 **50%** Pythia-160M **31%** Gemma-2 2B

of signals are interpretable; autointerpretation performance **approaches SAE-feature baselines**, with no SAE training.

## There is no single circuit for IOI

Tracing every IOI prompt separately, prompt-specific circuits form **well-defined clusters, dominated by prompt templates**: GPT-2 is strongly sensitive to role order (ABBA vs. BABA), whereas Pythia is more sensitive to surface wording.

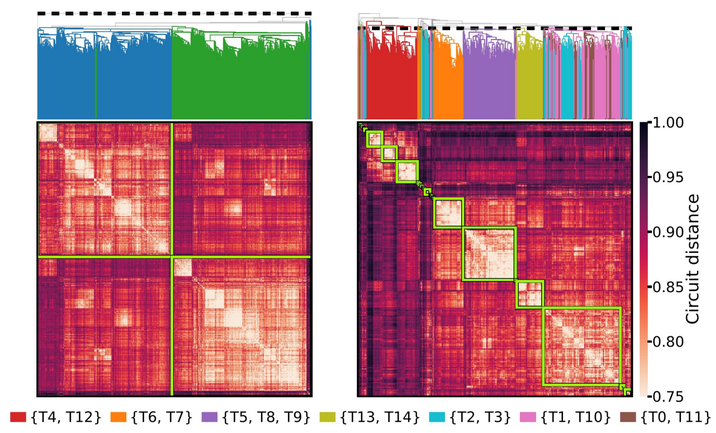


*Fig. 3: pairwise circuit distances. GPT-2 (left) splits **ABBA** vs. **BABA**; Pythia (right) forms wording-level clusters.*

**Across clusters, heads receive systematically different signals:** the name-mover head (9,9) fires in both formats, but in BABA the indirect object is tagged as the "second item in a parallel pair", while ABBA relies on more generic structural cues.

## Shared mechanisms, language-specific signals

Multilingual IOI (en, es, fr, pt; mGPT): **components are reused across languages**, but the **signals they carry are often language-specific**. Circuits cluster by high-level template first, then by language.

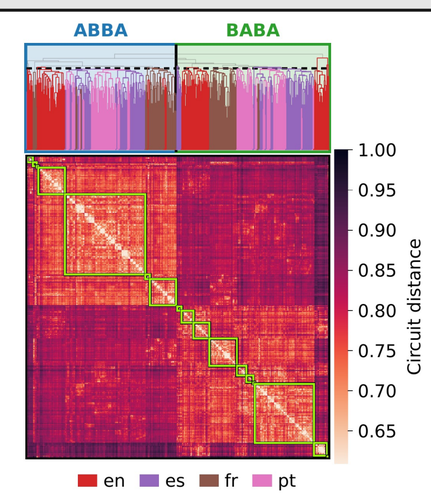


*Fig. 6: clusters by template, then language*

**Code & experiments:** github.com/gaabrielfranco/finding-highly-interpretable-circuits • Contact: gvfranco@bu.edu • Paper: scan the QR code above