# Wired for Overconfidence: A Mechanistic Perspective on Inflated Verbalized Confidence in LLMs

**Tianyi Zhao, Yinhan He, Wendy Zheng, Yujie Zhang, Chen Chen**
University of Virginia
{abs4dj,nee7ne,ncd9cf,qrk9jt,zrh6du}@virginia.edu

## Abstract

Large language models are often not just wrong, but *confidently wrong*: when they produce factually incorrect answers, they tend to verbalize overly high confidence rather than signal uncertainty. Such verbalized overconfidence can mislead users and weaken confidence scores as a reliable uncertainty signal, yet its internal mechanisms remain poorly understood. We present a circuit-level mechanistic analysis of this inflated verbalized confidence in LLMs, organized around three axes: capturing verbalized confidence as a differentiable internal signal, identifying the circuits that causally inflate it, and leveraging these insights for targeted inference-time recalibration. Across two instruction-tuned LLMs on three datasets, we find that a compact set of MLP blocks and attention heads, concentrated in middle-to-late layers, consistently writes the confidence-inflation signal at the final token position. We further show that targeted inference-time interventions on these circuits substantially improve calibration. Together, our results suggest that verbalized overconfidence in LLMs is driven by identifiable internal circuits and can be mitigated through targeted intervention.

## 1 Introduction

Large language models are increasingly deployed in high-stakes settings, where an incorrect answer delivered with unwarranted certainty can cause real harm. When asked to self-assess, these models are not merely wrong; they are often *confidently wrong*, systematically producing inflated verbalized confidence scores that fail to distinguish reliable answers from unreliable ones (see Figure 6). Although prior work indicates that LLMs can express their own uncertainty (Kadavath et al., 2022; Xiong et al., 2024; Lin et al., 2022), their verbalized confidence remains poorly calibrated to factual accuracy (Ji et al., 2025). Meanwhile, the broader calibration literature has mostly treated this as an output-level symptom, emphasizing post-hoc correction rather than understanding the internal mechanisms. A fundamental question therefore remains: *what specific internal mechanisms causally drive this inflated verbalized confidence*, and can they be intervened upon to recalibrate the model?

Mechanistic interpretability (Rai et al., 2024; Conmy et al., 2023) offers a natural framework for answering this question, having successfully localized circuits for tasks such as name prediction (Wang et al., 2023), mathematical reasoning (Hanna et al., 2023), and entity knowledge (Ferrando et al., 2025). We bring this lens to verbalized overconfidence, organizing our study around three complementary challenges. First, **capture**: how can we represent output-level verbalized confidence as a differentiable internal proxy? Second, **discover**: given such a proxy, how can we identify and validate the internal circuits responsible for inflated verbalized confidence? Third, **mitigate**: how can we convert the mechanistic insights into practical inference-time interventions that recalibrate verbal confidence?

We study this problem under a two-step confidence elicitation setup. To address the first challenge, we introduce the *Target Set Logit Difference* (TSLD), a metric that computes the mean logit gap between predefined high-confidence and low-confidence sets. We empirically show that TSLD reliably captures the model's internal confidence propensity, yielding a stable and differentiable metric compatible with subsequent gradient-based attribution. For the second challenge, we devise a *Truth-Injection Counterfactual Design* for

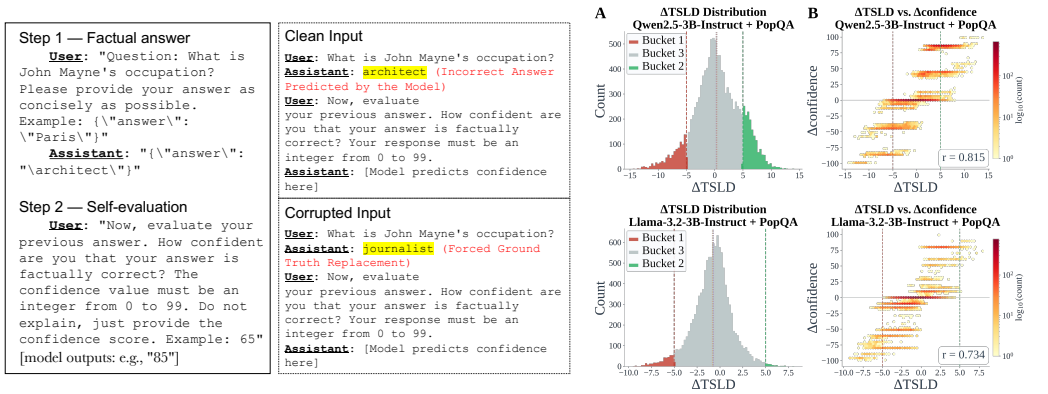



Figure 1: **Left**: Two-step elicitation. The model answers a factual question, then self-reports confidence as an integer (0–99). **Center**: Truth-injection counterfactual design. For each confidently-wrong record, the clean prompt retains the model's incorrect answer, while the corrupted prompt substitutes the ground truth, keeping all other tokens identical. **Right**: ΔTSLD distributions and ΔTSLD vs. Δconfidence scatter plots for Qwen2.5-3B (top) and Llama-3.2-3B (bottom) on PopQA.

constructing clean and corrupted pairs. By using ΔTSLD as the stratification criterion, we partition the data to isolate cases in which truth injection suppresses the inflated confidence. We then apply Edge Attribution Patching with Integrated Gradients (Hanna et al., 2024) to discover internal circuits that causally mediate the overconfidence signal, and validate them at both the edge and component levels. Finally, we intervene on a small set of top-ranked components identified as drivers of overconfidence in LLMs. Specifically, we study two intervention strategies, mean ablation and activation steering, applied during confidence inference, and evaluate calibration performance over the full dataset.

Across two models and three factual QA benchmarks, we find that a compact set of MLP blocks and attention heads, primarily in the middle-to-late layers, consistently write the confidence-inflation signal at the final token position. The strong cross-dataset conservation of the identified circuits indicates that verbalized overconfidence is driven by a stable model-internal mechanism rather than a task-specific artifact. Intervening on these components substantially improves model calibration with respect to verbalized confidence, establishing mechanistic recalibration as a viable complement to post-hoc calibration methods.

## 2 From Verbal Confidence to Internal Signals

**Models, Datasets, and Confidence Elicitation Setup.** We conduct experiments on Qwen2.5-3B-Instruct (Yang et al., 2024) and Llama-3.2-3B-Instruct (Grattafiori et al., 2024) across three factual QA benchmarks: PopQA (Mallen et al., 2023), MMLU (Hendrycks et al., 2021), and NQOpen (Lee et al., 2019)[1]. Verbalized confidence is elicited through a two-step setup (Figure 1, left): the model first answers the question, then reports a confidence score (0–99) for its own answer. This separation isolates confidence verbalization as a distinct stage of computation, allowing us to trace it to internal activations during the Step 2 forward pass.

**Truth-Injection Counterfactual Design.** For each incorrect record, we construct a corrupted prompt by replacing the Step 1 assistant turn with the ground-truth answer, keeping all other tokens identical (Figure 1, center). This yields a counterfactual pair $(x_{\text{clean}}, x_{\text{corrupt}})$ and isolates the causal effect of answer correctness on subsequent confidence generation. We use the ground truth, rather than a random or out-of-vocabulary token, as the corrupted answer to keep both prompts on the natural data manifold; injecting meaningless tokens might contaminate the activation differences with distributional-anomaly signals unrelated to confidence processing.

[1]See model and dataset details in Appendix A

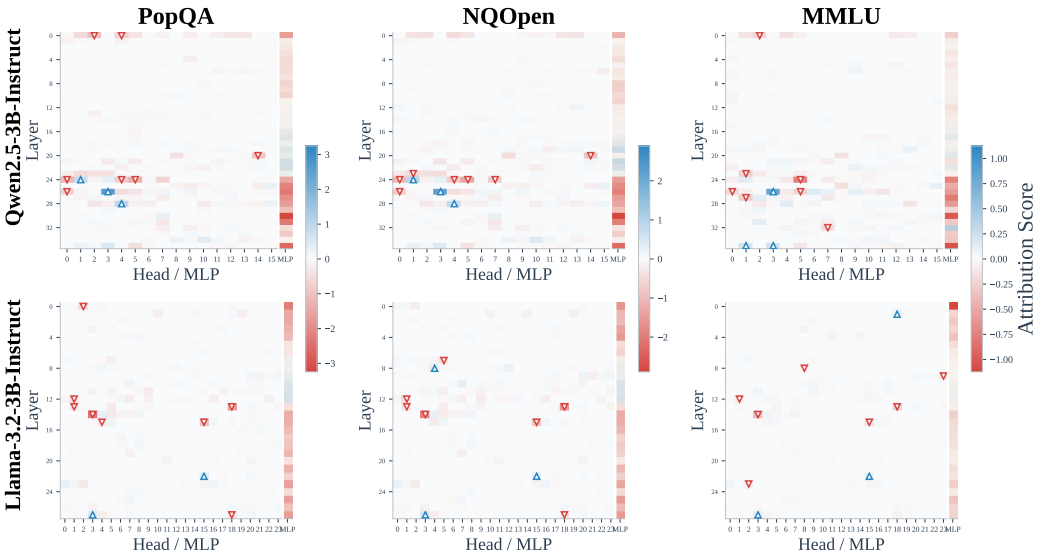



Figure 2: Attribution heatmaps across all six model×dataset configurations.

**Target-Set Logit Difference as Metric.** Unlike standard circuit analysis tasks, verbalized confidence has no single golden token to track: it is expressed as integers spanning a continuous range. We therefore introduce Target-Set Logit Difference (TSLD), defined as the gap between predefined high-confidence and low-confidence sets at the confidence prediction position:

$$L_{\text{TSLD}}(x) = \frac{1}{|\mathcal{H}|} \sum_{c \in \mathcal{H}} \text{logits}_{[-1, c_1]} - \frac{1}{|\mathcal{L}|} \sum_{c \in \mathcal{L}} \text{logits}_{[-1, c_1]}, \tag{1}$$

where $\mathcal{H} = \{70, 75, 80, 85, 90, 99\}$ and $\mathcal{L} = \{0, 10, 15, 20, 25, 30\}$ are fixed integer sets, $c_1$ denotes the first token of candidate $c$, and logits are read at the final prompt position (POS_END) in a single forward pass. TSLD captures the model's internal preference for high versus low confidence before decoding, while remaining fully differentiable and thus compatible with subsequent gradient-based attribution. Here, we track the *difference* $\text{Mean}(\mathcal{H}) - \text{Mean}(\mathcal{L})$, rather than $\text{Mean}(\mathcal{H})$ alone, for two reasons. First, because raw logits are unnormalized, subtracting the low-confidence mean neutralizes global shifts in activation magnitude and captures relative confidence preference. Second, it avoids an inhibition trap: an intervention might suppress $\text{Mean}(\mathcal{H})$ simply by disrupting the model's generation capability, not by resolving overconfidence. A genuine confidence shift should simultaneously decrease high-confidence logits and increase low-confidence logits, which the difference captures directly. For a clean-corrupted pair, we further define

$$\Delta_{\text{TSLD}} = L_{\text{TSLD}}(x_{\text{corrupt}}) - L_{\text{TSLD}}(x_{\text{clean}}) \tag{2}$$

which measures how truth injection shifts the model's confidence propensity. As shown in Figure 1 (right, panels B), $\Delta_{\text{TSLD}}$ correlates strongly with the change in raw verbalized confidence ($r = 0.815$ for Qwen2.5-3B, $r = 0.734$ for Llama-3.2-3B), indicating that TSLD serves as a reliable internal proxy for the confidence signal of interest. We emphasize that TSLD is a proxy for the model's *verbalized* confidence, not for epistemic uncertainty; the misalignment between the two is precisely the phenomenon under study.

**Data Stratification.** Incorrect records are heterogeneous in how they respond to truth injection: some exhibit overconfident suppression ($\Delta_{\text{TSLD}} \ll 0$), others show truth-sensitive increases in confidence ($\Delta_{\text{TSLD}} \gg 0$), still others remain near-neutral and are uninformative ($\Delta_{\text{TSLD}} \approx 0$). Running attribution on this mixed set can cause gradient cancellation, as oppositely directed response types average toward zero and obscure the circuits of interest. To improve attribution fidelity, we therefore stratify records into three buckets, as shown in Figure 1 (right, panels A) and Table 10. Bucket 1 forms the key discovery subset: records in

| Qwen2.5-3B-Instruct | | | | Llama-3.2-3B-Instruct | | | |
|---|---|---|---|---|---|---|---|
| Rank | PopQA | MMLU | NQOpen | Rank | PopQA | MMLU | NQOpen |
| 1 | m30 | m35 | m30 | 1 | a14.h3 | m0 | a14.h3 |
| 2 | m35 | m30 | m35 | 2 | m0 | a14.h3 | m0 |
| 3 | m26 | m27 | m31 | 3 | m27 | m4 | a13.h18 |
| 4 | m31 | a24.h5 | m26 | 4 | a13.h18 | m25 | m4 |
| 5 | m25 | m24 | m25 | 5 | m23 | m2 | m27 |
| 6 | m27 | m28 | m27 | 6 | m25 | a13.h18 | m25 |
| 7 | m0 | m26 | m24 | 7 | m1 | m14 | m23 |
| 8 | m28 | m25 | a24.h5 | 8 | m2 | m23 | m3 |
| 9 | m24 | a27.h1 | m28 | 9 | m14 | m27 | m14 |
| 10 | a24.h5 | m34 | a24.h0 | 10 | m15 | m26 | m15 |

Table 1: **Top-10 components by attribution score.** Components in blue are shared across all three datasets. Prefix m: MLP layer; prefix a: attention head (e.g. a14.h3 = layer 14, head 3).

which the model is initially confidently wrong, but truth injection sharply suppresses that confidence. We use this subset for subsequent attribution, since it most cleanly correlates with the mechanism underlying inflated verbalized confidence.

## 3 Locate and Validate Confidence Mover Circuits

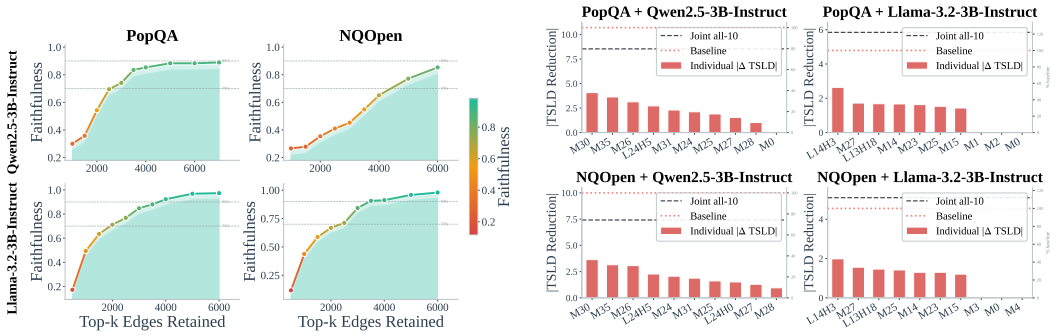
### 3.1 Circuit Attribution via EAP-IG

With target records selected and the computation graph (Appendix B) constructed, we instantiate the EAP-IG framework (Appendix C) for overconfidence circuit discovery. $L_{\text{TSLD}}$ (Eq. 1) defines both the discovery subset (via $\Delta_{\text{TSLD}}$) and the attribution objective, aligning sample selection with the signal being attributed and ensuring the attribution specifically locates components responsible for the overconfidence suppression. Concretely, for each Bucket 1 record and each edge $e_{a\to b}$ in the computation graph, the attribution score is:

$$\text{EAP-IG}(e_{a\to b}) \;=\; \frac{1}{m}\sum_{k=0}^{m-1} \frac{\partial L_{\text{TSLD}}}{\partial\, \text{input}_b}\bigg|_{x_k} \cdot \big(\text{out}_a(x_{\text{clean}}) - \text{out}_a(x_{\text{corrupt}})\big), \tag{3}$$

where $x_k$ interpolates between the corrupted and clean activations as defined in Eq. 12, and we use $m = 5$ integration steps. The gradient term captures how sensitive $L_{\text{TSLD}}$ is to perturbations at component $b$'s input; the activation difference term captures how much component $a$'s output shifts between the two counterfactual conditions. We then aggregate edge-level attributions to the component level by summing the scores of edges incident on each node, and average the resulting component scores across all Bucket 1 examples. We refer to the resulting subcircuit as the *Confidence Mover Circuit* (CMC).

**Inflated verbalized confidence is localized to mid-to-late-layer components.** Across all six model×dataset settings, high attribution components are concentrated primarily in middle-to-late layers (Figure 2; Table 1), although the attribution pattern differs across models. **Qwen2.5-3B-Instruct** shows a strongly MLP-dominant circuit: the top-ranked components are almost exclusively MLP layers in layers 24–35, with only a few attention heads (e.g., L24H5 and L27H1) entering the top-10 on individual datasets. **Llama-3.2-3B-Instruct** initially appears more distributed, with attention heads L14H3 and L13H18 and several early- and mid-layer MLPs ranking highly. However, subsequent single-component ablation (§3.2) shows that many of the early-layer MLPs (M0–M4) yield approximately zero individual TSLD reduction, indicating that they function as relay nodes rather than causal bottlenecks. After this causal disambiguation, the Llama circuit also concentrates in middle-to-late layers, with M27, M25, M23, M14, L14H3, and L13H18 emerging as the individually necessary components. Thus, despite model-specific differences in attribution footprint, the causally validated picture is consistent across models: inflated verbalized confidence is primarily supported by a compact set of mid-to-late-layer components.




(a) Faithfulness Curves. (b) Single-component Ablation.

Figure 3: (a) Faithfulness as a function of retained top-$k$ edges. (b) TSLD reduction when each of the top-10 components is individually ablated.

**Cross-dataset evaluation reveals a shared, compact circuit underlying overconfidence, indicating a conserved model-internal mechanism.** Despite differences in dataset format and content, the overconfidence circuit is strongly conserved within each model. Figure 2 shows that high-attribution regions exhibit similar spatial patterns across all three datasets, while Table 1 reveals a substantial shared core among the top-ranked components. For **Qwen2.5-3B-Instruct**, 8 of the 10 PopQA components recur in the top-10 on both NQOpen and MMLU. For **Llama-3.2-3B-Instruct**, a consistent set of attention heads and MLP blocks is shared across datasets. This $\geq$80% cross-dataset overlap across both model families suggests that verbalized overconfidence is supported by stable model-internal circuit rather than task-specific artifact. The same locus recurs at larger scale: on Qwen2.5-7B-Instruct, 9 of the top-10 components, again mid-to-late MLP blocks, are shared across PopQA and NQOpen, and steering on this circuit reduces ECE from 0.647 to 0.068 on PopQA and from 0.626 to 0.109 on NQOpen at $\alpha = 0.6$ (Appendix J).

### 3.2 Causal Validation

**Edge-level validation: The discovered circuits are both sufficient and necessary.** We evaluate along two standard dimensions, corresponding to the two central questions:

- **Faithfulness** (*Sufficiency*): Can the model reproduce the specific behavior (e.g., overconfidence) relying solely on the isolated sub-circuit with top-$k$ edges retained?

$$\text{Faithfulness} = \frac{\Delta_{\text{circuit_only}} - \Delta_{\text{corrupt}}}{\Delta_{\text{clean}} - \Delta_{\text{corrupt}}} \times 100\% \tag{4}$$

- **Completeness** (*Necessity*): To what extent does severing the sub-circuit remove the signal, rather than allowing it to persist through alternative pathways?

$$\text{Completeness} = \frac{\Delta_{\text{clean}} - \Delta_{\text{ablate_circuit}}}{\Delta_{\text{clean}}} \times 100\% \tag{5}$$

*Sufficiency.* The full computation graph contains 527,743 possible edges for Qwen2.5-3B-Instruct and 693,267 for Llama-3.2-3B-Instruct. Retaining only the top-3,000 edges isolates an extremely sparse subcircuit - less than 0.6% of all possible edges for Qwen and less than 0.5% for Llama. Despite this sparsity, as shown in Figure 3a, all faithfulness curves cross 70% by $k \approx$ 2,000–3,000 and plateau above 85%.This characteristic S-shaped saturation confirms that the overconfidence signal is concentrated in a compact edge-level circuit rather than diffusely distributed across the network.

*Necessity.* Completeness provides the converse test: we remove the top-3,000 circuit edges from the full computation graph and measure how much of the original TSLD remains. If substantial backup pathways existed, the model could still produce the overconfidence

| Model | Dataset | Baseline TSLD | Compl. Drop | Completeness% |
|---|---|---|---|---|
| Qwen2.5-3B-Instruct | PopQA | 10.05 | 7.92 | 78.8% |
| | NQOpen | 9.68 | 7.25 | 74.9% |
| Llama-3.2-3B-Instruct | PopQA | 4.79 | 3.56 | 74.3% |
| | NQOpen | 4.55 | 3.37 | 74.1% |

Table 2: **Completeness validation.** Completeness drop = full-model TSLD − ablated-sub-network TSLD, where the ablated sub-network retains all edges *except* the top-3,000.

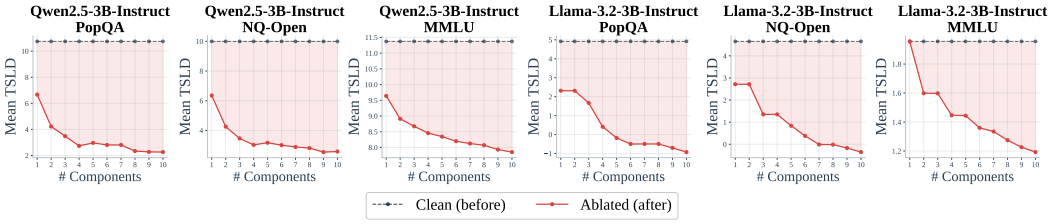



Figure 4: **Incremental component ablation**: mean TSLD as the top-$k$ components are jointly ablated, for $k = 1$ to 10. Dashed lines mark the clean (pre-ablation) mean TSLD.

behavior through these alternative routes, and the completeness drop would be small. Table 2 shows that this is not the case. Ablating the circuit eliminates 74–79% of the baseline TSLD across all four settings. This provides strong evidence that the discovered circuits are not merely correlated with overconfidence but carry the bulk of the causal signal in the studied settings.

**Component-level validation: Individual and incremental ablation confirm a small causal core.** Edge-level validation establishes that the overconfidence signal concentrates in a sparse subgraph, but edges are abstractions of the computational graph formalism rather than actual units within the model. Moreover, edge attribution can highlight both causal bottlenecks and relay pathways, so high edge-level importance does not necessarily imply that a component actually drives overconfidence. We therefore complement edge-level validation with component-level analyses, ablating the top-ranked CMC components one at a time and then incrementally to test whether these top components are genuinely causal drivers of the behavior. Specifically, we adopt a paired resample ablation procedure: for each record, the target component's activation at POS_END in the clean forward pass is replaced with the corresponding activation from the corrupted forward pass. The resulting TSLD change, averaged across Bucket 1 examples, quantifies the component's causal contribution: a strongly negative value indicates active contribution, whereas a near-zero value indicates a relay node that does not independently drive the behavior.

*Single-component ablation.* Ablating the top-10 components individually reveals a concentrated causal core (Figure 3b). For Qwen, M30 produces the largest TSLD reduction on both PopQA and NQOpen, followed by M35 and M26; for Llama, L14H3 is the dominant component, followed by M27 and L13H18, with the same ordering on both datasets. In contrast, several top-ranked early-layer MLPs, M1, M2, and M0 on PopQA, and M3, M0, and M4 on NQOpen, yield zero individual reduction, indicating that they function as relay nodes rather than causal bottlenecks for inflated verbalized confidence. Joint ablation of all top-10 components yields a much larger effect than any single ablation, but remains markedly sub-additive relative to the sum of individual effects (e.g., ∼20.3 vs. 8.5 for Qwen/PopQA). This non-additivity suggests that the discovered components operate as an interconnected ensemble with shared pathways, rather than as independent parallel channels.

*Incremental joint ablation.* We progressively ablate the top-ranked components, from the top-1 up to the top-10, and track the resulting mean TSLD (Figure 4). Across all six model×dataset settings, the curves show a clear diminishing-return pattern: ablating the first few components produces the largest drop in TSLD, while additional components contribute progres-

| Model | Method | $\alpha$ | ECE | Brier | ECE Impr. | Brier Impr. |
|---|---|---|---|---|---|---|
| Qwen2.5-3B-Instruct | baseline | – | 0.492 | 0.430 | – | – |
| | mean ablation | – | 0.295 | 0.244 | 40.0% | 43.1% |
| | steering | 0.3 | 0.245 | 0.193 | 50.3% | 55.1% |
| | steering | 0.4 | 0.166 | 0.148 | 66.4% | 65.7% |
| | steering | 0.5 | **0.111** | **0.130** | **77.5%** | **69.7%** |
| | steering | 0.6 | 0.145 | 0.146 | 70.7% | 66.1% |
| | steering | 0.7 | 0.167 | 0.166 | 66.0% | 61.4% |
| | steering | 0.8 | 0.170 | 0.170 | 65.5% | 60.5% |
| Llama-3.2-3B-Instruct | baseline | – | 0.570 | 0.495 | – | – |
| | mean ablation | – | 0.018 | 0.161 | 96.9% | 67.5% |
| | steering | 0.3 | 0.338 | 0.287 | 40.7% | 42.0% |
| | steering | 0.4 | 0.196 | 0.189 | 65.6% | 61.9% |
| | steering | 0.5 | 0.063 | **0.146** | 88.9% | **70.5%** |
| | steering | 0.6 | **0.020** | 0.161 | **96.5%** | 67.4% |
| | steering | 0.7 | 0.027 | 0.163 | 95.3% | 67.0% |
| | steering | 0.8 | 0.031 | 0.163 | 94.6% | 67.1% |

Table 3: **Intervention results on PopQA.** Best ECE and Brier per model in **bold**.

sively smaller gains. This confirms that the overconfidence signal is concentrated in a small causal core rather than distributed uniformly across many components.

## 4 Targeted Circuit Intervention for Confidence Recalibration

Having identified and validated the compact set of components that drive overconfidence, we now ask: can intervening on these components at inference time recalibrate the model's verbalized confidence? We study two simple intervention strategies applied to the top-10 CMC components and evaluate their effect on the *full dataset* using Expected Calibration Error (ECE) (Pavlovic, 2025), Brier score (Glenn et al., 1950), and reliability curves[2]. Importantly, although the intervention references are estimated from a small set of Bucket 1 examples, evaluation is performed on the full dataset, allowing us to test whether circuit-level intervention generalizes beyond the overconfident discovery subset. Both strategies share the same deployment principle: a global reference is computed once from a small set of Bucket 1 examples and then applied uniformly at inference time. As a result, neither method requires constructing per-sample counterfactuals during deployment. Calibration results before and after intervention are reported in Tables 3, 11, and 12. Reliability curves under different intervention strategies are presented in Figures 5 and 7.

Mean ablation suppresses the contribution of a target component by replacing its activation at POS_END with a global reference mean computed from corrupted-prompt activations:

$$\hat{h}_{\text{out}}(x, \text{POS_END}) \leftarrow \mu_{\text{ref}} = \frac{1}{N}\sum_{i=1}^{N} h_{\text{out}}(x^{(i)}_{\text{corrupt}}, \text{POS_END}). \tag{6}$$

Activation steering provides a softer and tunable alternative. Instead of overwriting the activation, it shifts the target component along a precomputed overconfidence direction:

$$\mathbf{v}_{\text{conf}} = \frac{1}{N}\sum_{i=1}^{N}\left[h_{\text{out}}(x^{(i)}_{\text{clean}}, \text{POS_END}) - h_{\text{out}}(x^{(i)}_{\text{corrupt}}, \text{POS_END})\right]. \tag{7}$$

During confidence inference, we subtract a scaled version of it from each target activation:

$$\hat{h}_{\text{out}}(x, \text{POS_END}) \leftarrow h_{\text{out}}(x, \text{POS_END}) - \alpha \cdot \mathbf{v}_{\text{conf}}, \tag{8}$$

where $\alpha \in [0, 1]$ controls the intervention strength. Unlike mean ablation, which fully replaces the activation, steering preserves the instance-specific signal while nudging it away from the overconfident direction, providing a continuous recalibration knob.

[2]See illustrations in Appendix F and Appendix G

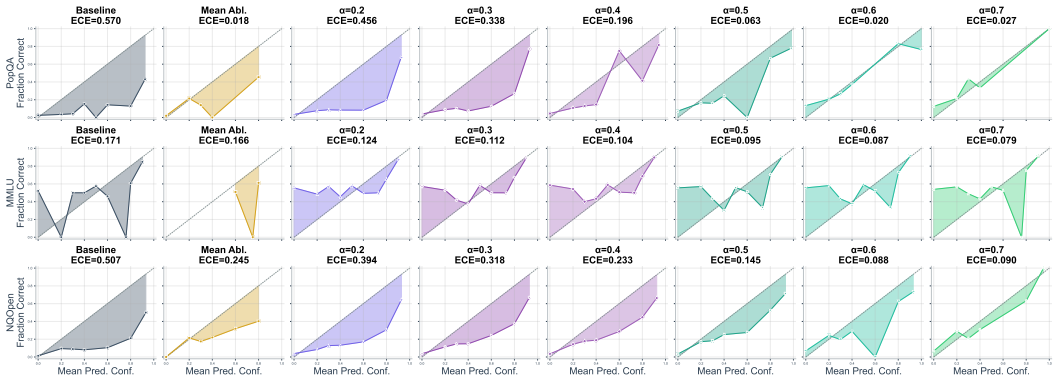



Figure 5: **Reliability curves for Llama-3.2-3B-Instruct** across all three datasets under baseline (before intervention), mean ablation, and steering at $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$.

**Targeted intervention substantially improves calibration.** Direct intervention on the top-ranked CMC components yields substantial calibration gains across models and datasets. This improvement is evident both in aggregate metrics and in bin-level calibration behavior. In terms of ECE and Brier score (Tables 3, 11, 12), both mean ablation and activation steering are most effective on PopQA and NQOpen, where the best settings reduce ECE by roughly 78–97% on PopQA and 81–83% on NQOpen; improvements on MMLU are smaller but still consistent, reaching about 33% for Qwen and 57% for Llama. Reliability curves provide a complementary and more intuitive view. For Llama-3.2-3B-Instruct (Figure 5), both interventions move the confidence–accuracy curves substantially closer to the perfect calibration diagonal on PopQA and NQOpen. On MMLU, the curves improve more gradually, consistent with the weaker metric gains. Together, these results show that the discovered CMCs are not merely descriptive artifacts, but practical intervention targets for recalibration.

**Steering provides more precise and controllable recalibration than mean ablation.** Both interventions improve calibration, but steering is more controllable. Mean ablation fully overwrites the target activation, which can be highly effective when the overconfidence signal is tightly concentrated, but brittle otherwise: for example, it yields a 96.9% ECE reduction for Llama/PopQA but only 3.1% for Llama/MMLU. Steering instead subtracts only the overconfidence direction while preserving the instance-specific residual. As a result, it exhibits a consistent dose–response pattern: ECE improves from $\alpha = 0.3$ to an optimum around 0.5–0.6, then degrades as stronger intervention suppresses useful confidence-related signals. This optimal range is stable across four of six configurations, making $\alpha$ a practical hyperparameter requiring minimal tuning. Steering therefore provides a more stable recalibration mechanism overall. The full calibration curves over steering strengths are provided in Appendix H.

**Comparison with post-hoc calibration baselines.** Appendix I compares our interventions with temperature scaling (Guo et al., 2017), Platt scaling (Platt et al., 1999), isotonic regression (Zadrozny & Elkan, 2002), and histogram binning (Zadrozny & Elkan, 2001). More specifically, our aim is not to outperform existing calibration methods but to establish circuit-level recalibration as a mechanistic complement that edits the internal computation rather than remapping outputs, and whose full-dataset ECE reduction from editing only the discovered components is itself further causal evidence for the discovered circuit.

## 5 Discussion

**Verbalized overconfidence likely reflects a mismatch between confidence verbalization and factual uncertainty.** Our results can be interpreted in light of Teplica et al. (2025), which argues that a model's actual uncertainty is generated largely in the same parts in LLMs that support factuality. If verbal confidence faithfully reported this uncertainty, one would expect the circuit underlying verbal confidence to align closely with factual retrieval

circuit and to vary with the task-dependent computation that supports answering. Instead, the circuit driving inflated verbal confidence is highly conserved across datasets within each model, and the answer-substitution controls (Application E) show that it responds to the committed answer largely independently of that answer's correctness, consistent with behavioral evidence that verbal confidence is largely answer-independent (Seo et al., 2026). The direct circuit comparison (Figure 9) reinforces this mismatch: the verbal confidence circuit is more concentrated in middle-to-late layers and denser near the logits, whereas the factual retrieval circuit spans a longer pathway through earlier and middle layers. Taken together, these findings suggest that verbalized overconfidence arises from a compact, correctness-insensitive confidence-writing core that is structurally misaligned with the model's factuality-sensitive computation and may impose a default high-confidence bias.

**CMCs form a compact causal core of verbalized overconfidence, but not the full confidence-estimation system.** The intervention results show that the discovered CMCs account for a substantial share of verbalized overconfidence: manipulating only the top-10 components already yields large calibration gains. At the same time, the best results occur at moderate steering strengths rather than under maximal intervention. One interpretation is that these components form only part of a broader confidence-estimation system, so stronger intervention begins to disrupt other useful confidence-related computation. Another is that the steering direction $\mathbf{v}_{\text{conf}}$, estimated as an average over Bucket 1 samples, only approximates the true per-example overconfidence axis, so excessive scaling amplifies this approximation error. Under either interpretation, the CMCs are best understood as a compact causal core that disproportionately biases the model toward inflated confidence, while the full system of confidence estimation remains an open question.

**Post-training may strengthen a prior toward overconfident behavior.** The models studied here are instruction-tuned variants, raising the possibility that post-training contributes to the overconfidence mechanism we identify. Prior work shows that RLHF can exacerbate verbalized overconfidence and that aligned language models are often more overconfident than their pretrained counterparts (Leng et al., 2025; He et al., 2023). A natural next step is to compare pretrained, supervised fine-tuned, and RLHF-tuned checkpoints, or to track how verbal confidence-related circuits emerge over the course of post-training.

## 6 Related Work

**LLM calibration and verbal confidence.** LLMs can be prompted to verbalize confidence, sometimes with reasonable calibration (Lin et al., 2022; Tian et al., 2023), but subsequent work found that verbalized confidence is often miscalibrated (Xiong et al., 2024; Yona et al., 2024), a problem that post-training and RLHF can further exacerbate (He et al., 2023; Leng et al., 2025). We complement this literature by asking *what internal mechanism causally drives* inflated verbal confidence, and whether intervening on it can improve calibration.

**Mechanistic interpretability** seeks to reverse-engineer the internal circuits that support model behavior. Prior work has localized circuits for behaviors such as indirect object identification, mathematical reasoning, and factual/entity knowledge (Wang et al., 2023; Hanna et al., 2023; Ferrando et al., 2025; Conmy et al., 2023). More recent work has emphasized scalable or practical workflows for applications to larger models and more realistic tasks (Rai et al., 2024; Hanna et al., 2024). We build on this line of work methodologically: our pipeline integrates EAP-IG (Hanna et al., 2024) for circuit attribution.

**Mechanistic studies of uncertainty and verbal confidence.** Teplica et al. (2025) studies the mechanistic basis of actual uncertainty and argues that it is generated in largely the same parts of the network that support factuality. Relatedly, Orgad et al. (2025) show that LLMs internally encode more truthfulness information than they externally reveal. On verbal confidence specifically, Ji et al. (2025) characterize verbal uncertainty as a linear feature and intervene on it to reduce hallucinations. Kumaran et al. (2026) analyze how verbal confidence is computed and argue for a cached-retrieval account, while Stolfo et al. (2024) identify neurons that regulate next-token confidence more generally. Our work focuses on a failure mode: the tendency of LLMs to be *confidently wrong*. We identify the compact

circuit that drives this inflated verbalized confidence, validate its causal role, and show that targeted intervention on these circuits yields substantial calibration improvement.

## 7 Conclusion

We present a circuit-level mechanistic analysis of inflated verbalized confidence in LLMs. By capturing output-level verbal confidence as a differentiable internal proxy, we identify compact *Confidence Mover Circuits* that causally drive the model's overconfident behavior. These circuits are causally validated and substantially conserved across datasets, suggesting that verbalized overconfidence is supported by a stable model-internal mechanism rather than a benchmark-specific artifact. We further show that targeted inference-time intervention on only a small set of top-ranked components substantially improves calibration, establishing that the discovered circuits are not only interpretable but also actionable. Overall, our results suggest that verbalized overconfidence is rooted in identifiable internal circuit, and that circuit-level intervention offers a promising complement to post-hoc calibration methods.

## References

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WCRQFlji2q.

W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *Advances in Neural Information Processing Systems*, 36:76033–76060, 2023.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT.

Guande He, Peng Cui, Jianfei Chen, Wenbo Hu, and Jun Zhu. Investigating uncertainty calibration of aligned language models under the multiple-choice setting. *arXiv preprint arXiv:2310.11732*, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3769–3793, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Dharshan Kumaran, Arthur Conmy, Federico Barbero, Simon Osindero, Viorica Patraucean, and Petar Veličković. How do LLMs compute verbal confidence? In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=t7oQT3uX66.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1612. URL https://aclanthology.org/P19-1612/.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l0tg0jzsdL.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Neel Nanda. Attribution patching: Activation patching at industrial scale, 2023. URL https://www.neelnanda.io/mechanistic-interpretability/attribution-patching.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Maja Pavlovic. Understanding model calibration - a gentle introduction and visual exploration of calibration and the expected calibration error (ECE). In *The Fourth Blogpost Track at ICLR 2025*, 2025. URL https://openreview.net/forum?id=BxBeCjQd2y.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

KiJung Seo, Sehun Lim, and Taeuk Kim. ADVICE: Answer-dependent verbalized confidence estimation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23942–23962, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1098. URL https://aclanthology.org/2026.acl-long.1098/.

Alessandro Stolfo, Ben Peng Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0og7nmvDbe.

Carter Teplica, Yixin Liu, Arman Cohan, and Tim GJ Rudner. Sciurus: Shared circuits for interpretable uncertainty representations in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12451–12469, 2025.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7752–7764, 2024.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 609–616, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

## A Models and Datasets

**Models.** We experiment on two open-weight instruction-tuned model families with distinct architectures: **Qwen2.5-3B-Instruct** (Yang et al., 2024) (36 layers, 16 query heads, GQA with 2 KV heads) and **Llama-3.2-3B-Instruct** (Grattafiori et al., 2024) (28 layers, 24 query heads, GQA with 8 KV heads).

**Datasets.** We evaluate on three knowledge-intensive QA benchmarks spanning different question formats: **PopQA** (Mallen et al., 2023) (14,267 open-domain factoid questions with associated Wikipedia-popularity scores), **MMLU** (Hendrycks et al., 2021) (14,042 multiple-choice knowledge questions across 57 subjects), and **NQOpen** (Lee et al., 2019) (3,610 open-domain questions derived from Google search queries). PopQA and NQOpen test free-form factual recall; MMLU introduces a constrained multiple-choice format. For each model$\times$dataset combination, we run the two-step elicitation pipeline (§2), retain only incorrect predictions, and stratify into three buckets for circuit discovery.

## B LLM as a Computational Graph

We conceptualize the LLM as a Directed Acyclic Graph (DAG), denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, following the standard view in circuit-level mechanistic interpretability. In this graph, nodes $\mathcal{V}$ correspond to computational modules, and edges $\mathcal{E}$ represent information flow between modules through the residual stream. This formulation provides a convenient abstraction for identifying the subgraph that causally supports a target behavior.

**Nodes ($\mathcal{V}$).** To enable fine-grained circuit discovery, we decompose the standard Transformer blocks into the atomic linear projections of the models:

$$\mathcal{V} = \{W_Q^{(l,h)}, W_K^{(l,h)}, W_V^{(l,h)}, W_O^{(l,h)}, W_{MLP}^{(l)} \mid l \in [1, L], h \in [1, H]\} \tag{9}$$

where $L$ is the number of layers and $H$ is the number of heads per layer. $W_Q$, $W_K$, $W_V$ represent the query, key, and value projections for the $h$-th head in layer $l$, $W_O$ is the output projection, and $W_{MLP}$ represents the two-layer feed-forward network.

**Edges ($\mathcal{E}$).** An edge $e_{u \to v}$ exists if the output of module $u$ contributes to the input of module $v$ via the residual stream. Intuitively, edges capture how intermediate representations are transmitted and transformed across the network. Our goal is to identify the subset of edges whose activations are causally responsible for the target behavior of interest, namely inflated verbalized confidence.

**Behavioral metric.** Given an input $x$, let $f(x)$ denote the scalar behavior of interest. In this work, $f(x)$ is instantiated as the Target-Set Logit Difference (TSLD) defined in Section 2, which serves as a differentiable proxy for verbalized confidence. Circuit attribution is therefore performed with respect to changes in this scalar signal.

## C Edge Attribution Patching with Integrated Gradients (EAP-IG)

To identify the subgraph responsible for a target behavior, we require a method that attributes changes in the behavioral metric to specific internal edges. Standard methods like Activation Patching (Meng et al., 2022) are faithful but computationally expensive, while vanilla gradient attribution is efficient but can be unreliable in deep nonlinear networks due to gradient saturation. To balance fidelity and efficiency, we employ Edge Attribution Patching with Integrated Gradients (EAP-IG) (Hanna et al., 2024).

Vanilla EAP (Nanda, 2023) approximates the importance of an edge $e$ using the product of its activation $x_e$ and the gradient of the scalar objective $\mathcal{L}$ with respect to that activation:

$$\phi_{EAP} = x_e \cdot \frac{\partial \mathcal{L}}{\partial x_e} \tag{10}$$

where, in our setting, $\mathcal{L}$ is the confidence-related signal induced by TSLD. While requiring only one backward pass, this method can fail under saturation effect. When a neuron is fully activated, the local gradient is near zero, even if that neuron is critically important for the output, leading to incomplete or fragmented circuits.

To overcome saturation, EAP-IG (Hanna et al., 2024) accumulates gradients along a linear path between a corrupted baseline input ($I_{corrupt}$) and the clean target input ($I_{clean}$). Let $x_e$ denote the activation along edge $e$. A path is defined as $x_e(\gamma) = x_e^{corrupt} + \gamma(x_e^{clean} - x_e^{corrupt})$ for $\gamma \in [0, 1]$. The attribution score $\phi(e)$ for an edge $e$ is calculated as:

$$\phi(e) = (x_e^{clean} - x_e^{corrupt}) \times \int_{\gamma=0}^{1} \frac{\partial \mathcal{L}(x(\gamma))}{\partial x_e(\gamma)} d\gamma \tag{11}$$

By accumulating gradients along the interpolation path rather than relying on a single local derivative, EAP-IG yields more reliable edge attributions under saturation. In practice, this integral is approximated using a Riemann sum with $m$ steps:

$$\phi(e) \approx (x_e^{clean} - x_e^{corrupt}) \times \frac{1}{m} \sum_{k=1}^{m} \frac{\partial \mathcal{L}(x(\frac{k}{m}))}{\partial x_e} \tag{12}$$

# D Raw Verbal Confidence Distributions

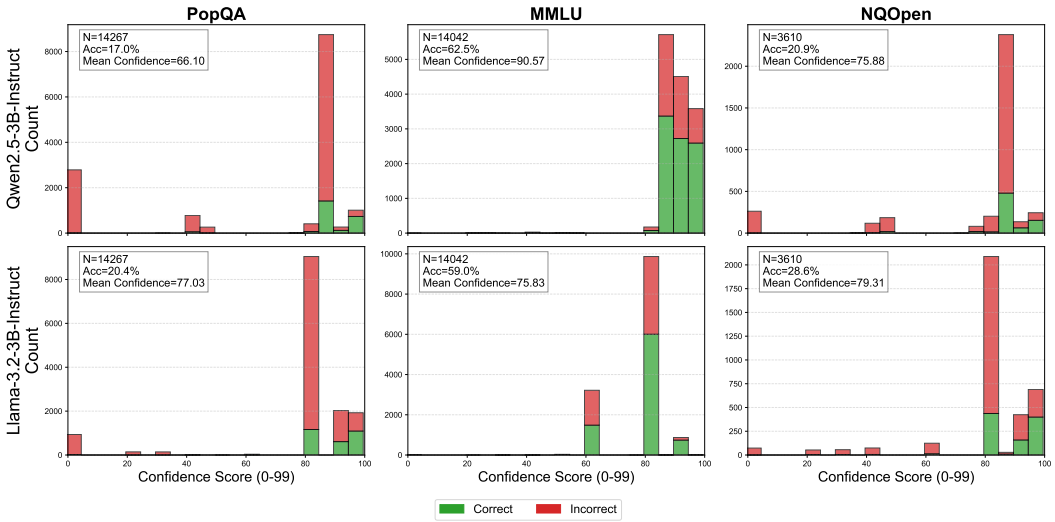



Figure 6: **Raw verbalized confidence distributions** across all six model×dataset configurations. Green: correct predictions; red: incorrect predictions.

Figure 6 shows the distribution of verbalized confidence scores across all six model×dataset configurations, separated by answer correctness. These distributions vividly illustrate the core phenomenon studied in this paper: the models are not merely wrong, but *confidently wrong*, systematically assigning high confidence scores that poorly discriminate between correct and incorrect answers.

# E Robustness Controls for Circuit Discovery

## E.1 Answer-Substitution Control

The truth-injection counterfactual replaces the model's incorrect Step 1 answer with the ground-truth answer. Although this intervention changes the factual status of the committed answer, it may also alter semantic content, surface form, token identity, and sequence length.

To test whether these factors account for the recovered circuit, we repeat EAP-IG circuit discovery on Qwen2.5-3B-Instruct/PopQA using four Step 1 answer substitutions:

1. the ground-truth answer used in the main experiment;
2. a same-meaning paraphrase, implemented as a different valid PopQA gold alias;
3. a type-matched plausible-wrong answer, sampled from another answer associated with the same relation type; and
4. a length-matched distractor with comparable token length but unrelated semantic content.

For example, when the question is "What is Henry Feilden's occupation?" and the model answers "architect," the four substitutions are "politician," "political leader," "priest," and "board game," respectively.

We evaluate each condition using the mean $\Delta_{\text{TSLD}}$ and the overlap between its top-10 attributed components and those recovered under the original ground-truth substitution.

As shown in Table 4, all four substitutions produce confidence decreases of similar magnitude and recover substantially overlapping component sets. The same-meaning alias result shows that the circuit is not tied to the exact surface form of the ground-truth answer. The

| Substituted Step 1 answer | Mean $\Delta_{\text{TSLD}}$ | Top-10 overlap |
|---|---|---|
| Ground truth | $-8.50$ | 10/10 |
| Same-meaning gold alias | $-8.24$ | 9/10 |
| Type-matched plausible-wrong answer | $-7.71$ | 10/10 |
| Length-matched distractor | $-8.68$ | 7/10 |

Table 4: **Answer-substitution controls on Qwen2.5-3B-Instruct/PopQA.** Mean $\Delta_{\text{TSLD}}$ measures the change relative to the original incorrect-answer context, with more negative values indicating a larger confidence decrease. Top-10 overlap is measured against the circuit recovered using ground-truth substitution.

length-matched control indicates that answer length alone does not explain the attribution pattern. More importantly, the plausible-wrong condition produces a confidence shift and component ranking comparable to those obtained with the ground truth. Therefore, the discovered circuit cannot be interpreted as selectively detecting whether the substituted answer is factually correct.

The results instead support a more precise interpretation: the CMC writes verbalized confidence conditioned on the model's committed answer, while remaining largely insensitive to whether that answer is factually correct. That is, the model is confident in what it said, not in what is true.

### E.2 Elicitation-Prompt Control

The main experiments use a fixed Step 2 prompt to elicit verbalized confidence. To test whether the recovered components depend on this exact wording, we repeat circuit discovery on Qwen2.5-3B-Instruct/PopQA using three semantically related prompt variants.

| Step 2 confidence prompt | Mean $\Delta_{\text{TSLD}}$ | Top-5 overlap | Top-10 overlap |
|---|---|---|---|
| Original prompt | $-7.07$ | 5/5 | 10/10 |
| "How certain are you that it is correct?" | $-10.63$ | 4/5 | 8/10 |
| "On a scale from 0 to 100, how sure are you?" | $-3.75$ | 3/5 | 7/10 |
| "How likely is it that your answer is accurate?" | $-3.35$ | 4/5 | 6/10 |

Table 5: **Confidence-prompt robustness on Qwen2.5-3B-Instruct/PopQA.** The circuit remains partially conserved under three semantically related Step 2 prompt variants.

The top-ranked components remain conserved across all prompt variants, suggesting that the discovered circuits are properties of the models rather than artifacts of a single elicitation template.

### E.3 TSLD Set Sensitivity

We vary the high- and low-confidence sets used in Eq. 1 and measure the Jaccard similarity between the resulting top-10 components and those discovered with the default sets (Table 6).

| High/low target sets | Top-10 Jaccard vs. default |
|---|---|
| $\mathcal{H} = \{70, 75, 80, 85, 90, 99\}$; $\mathcal{L} = \{0, 10, 15, 20, 25, 30\}$ | 1.00 |
| $\mathcal{H} = \{80, 85, 90, 99\}$; $\mathcal{L} = \{0, 10, 20\}$ | 1.00 |
| $\mathcal{H} = \{60, 65, 70, 75, 80, 85, 90, 95, 99\}$; $\mathcal{L} = \{0, 5, 10, 15, 20, 25, 30, 35, 40\}$ | 0.82 |
| $\mathcal{H} = \{80, 90, 99\}$; $\mathcal{L} = \{0, 15, 30\}$ | 0.82 |
| $\mathcal{H} = \{75, 80, 85, 90, 95, 99\}$; $\mathcal{L} = \{5, 10, 15, 20, 25, 35\}$ | 0.82 |

Table 6: **Sensitivity of Qwen2.5-3B-Instruct/PopQA circuit discovery to TSLD target sets.** The recovered top-10 components remain stable across narrower, wider, coarser, and shifted target sets.

### E.4 Bucket-2 Contrast

Running the same EAP-IG discovery on Bucket 2 (the opposite-sign contrast in which truth injection increases confidence) yields an attention-heavy circuit whose top components (L26H3, L28H4, M13, M19, L35H3, and others) share none of the Bucket 1 top-10 (Jaccard 0.00). A distinct circuit is expected for an opposite-sign phenomenon; the disjointness confirms that the Bucket 1 circuit is specific to overconfident collapse rather than a by-product of the stratification.

## F Calibration Metrics

We evaluate recalibration using two standard metrics.

**Expected Calibration Error (ECE)** (Pavlovic, 2025) measures the average gap between predicted confidence and observed accuracy. Predictions are partitioned into $B$ equal-width bins by confidence score, and ECE is computed as the weighted average of per-bin calibration gaps:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} \big|\mathrm{acc}(b) - \mathrm{conf}(b)\big|, \tag{13}$$

where $n_b$ is the number of samples in bin $b$, $N$ is the total number of samples, $\mathrm{acc}(b)$ is the empirical accuracy within the bin, and $\mathrm{conf}(b)$ is the mean predicted confidence. A perfectly calibrated model has $\mathrm{ECE} = 0$. We use $B = 10$ bins throughout.

**Brier Score** (Glenn et al., 1950) provides a complementary assessment by measuring the mean squared error between the predicted confidence (normalized to $[0, 1]$) and the binary correctness indicator:

$$\mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} \big(c_i - y_i\big)^2, \tag{14}$$

where $c_i \in [0, 1]$ is the predicted confidence for sample $i$ and $y_i \in \{0, 1\}$ indicates whether the answer is correct. Unlike ECE, which captures only calibration, the Brier score jointly penalizes miscalibration and poor discrimination (inability to separate correct from incorrect predictions), providing a more holistic measure of confidence quality. Lower values indicate better performance for both metrics.

## G Reliability Curves

Reliability curves provide a bin-level view of calibration. They plot *empirical accuracy* against *mean predicted confidence* over confidence bins, with the diagonal corresponding to perfect calibration: if the model reports confidence $p$, it should be correct approximately $p$ fraction of the time. Curves below the diagonal indicate *overconfidence*, meaning that the model is less accurate than its stated confidence suggests, while curves above the diagonal indicate *underconfidence*. Compared with aggregate metrics such as ECE and Brier score, reliability curves provide a more fine-grained and interpretable view of calibration by showing where miscalibration occurs across the confidence range and whether it takes the form of overconfidence or underconfidence. The shaded area between the curve and the diagonal visualizes the total calibration gap, which ECE summarizes as a single scalar. Intuitively, the closer the curve lies to the diagonal, the better calibrated the model is.

Figures 7 and 5 show reliability curves for Qwen2.5-3B-Instruct and Llama-3.2-3B-Instruct respectively, across all three datasets under baseline (before intervention), mean ablation, and steering at $\alpha \in \{0.2, \ldots, 0.7\}$.

## H Full Steering Strength Sweep from 0.1 to 1

Figure 8 shows ECE (left axis) and Brier score (right axis) as a function of steering strength $\alpha$. Across all six model$\times$dataset settings, steering exhibits a clear dose-response pattern: both

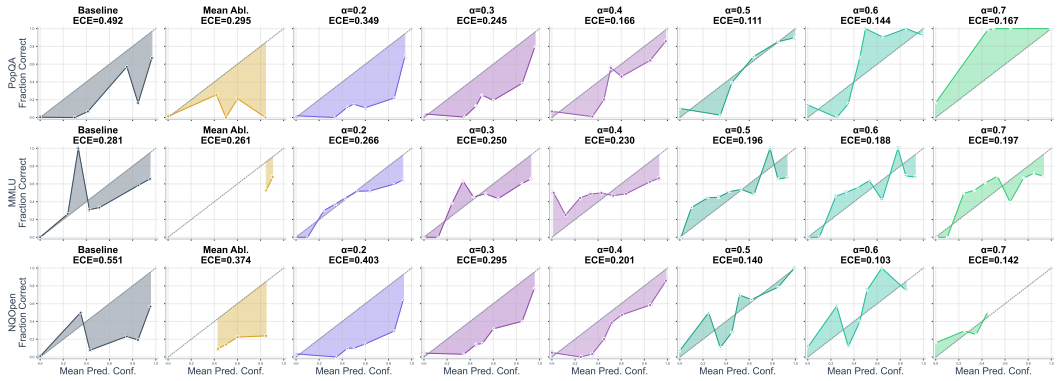



Figure 7: **Reliability curves for Qwen2.5-3B-Instruct** across PopQA (top), MMLU (middle), and NQOpen (bottom) under baseline (before intervention), mean ablation, and steering at $\alpha \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$.

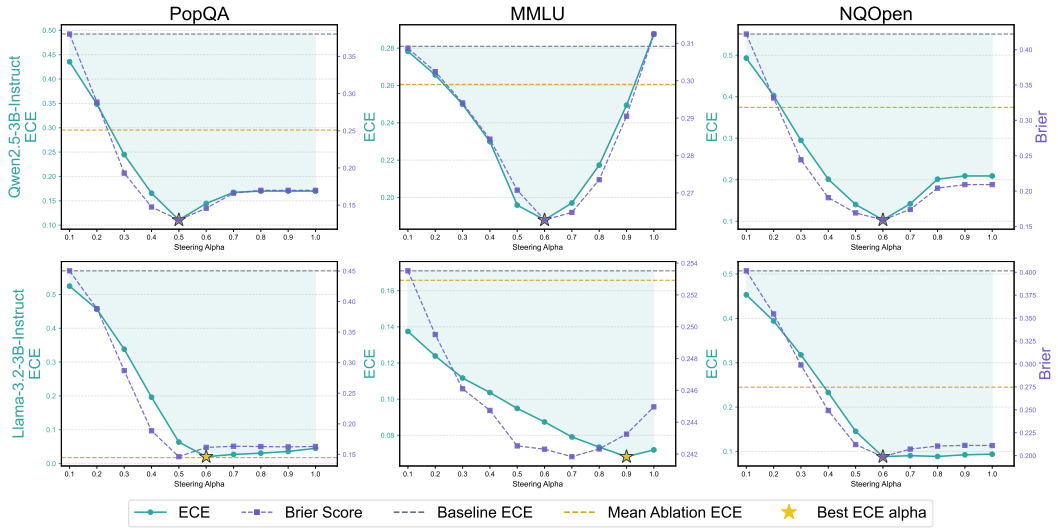



Figure 8: **Steering $\alpha$ sweep across all six configurations.** ECE (teal, left axis) and Brier score (purple, right axis) as a function of steering strength $\alpha$. Dashed grey: baseline ECE; dashed yellow: mean ablation ECE; star: optimal $\alpha$.

ECE and Brier score improve substantially as $\alpha$ increases from small values, indicating that progressively stronger suppression of the overconfidence direction yields better calibration. However, this improvement is not monotonic. In most settings, performance reaches an optimum at a moderate intervention strength and then plateaus or degrades as $\alpha$ becomes too large. Overall, the sweep confirms that activation steering provides a robust and tunable recalibration mechanism. It also shows that the intervention is not simply "the stronger, the better": calibration improves most when the overconfidence direction is attenuated to an intermediate degree rather than fully suppressed.

## I Comparison with Post-Hoc Calibration Baselines

We compare circuit-level intervention against four standard output-level calibrators: temperature scaling (Guo et al., 2017), Platt scaling (Platt et al., 1999), isotonic regression (Zadrozny & Elkan, 2002), and histogram binning (Zadrozny & Elkan, 2001). Each calibrator is fit on a labeled 50% split of the verbalized confidence scores and evaluated on the disjoint 50% split; the Raw ECE column is computed on that evaluation split and therefore deviates slightly

from the full-dataset baselines in Tables 3, 11, and 12. The mean-ablation and best-steering columns restate our full-dataset results from those tables. The comparison is included to position circuit-level recalibration as a viable, mechanistic complement to output-level post-hoc calibration methods, rather than to claim state-of-the-art calibration performance

| Model / Dataset | Raw ECE | Temp. | Platt | Isotonic | Histogram | Mean abl. (ours) | Best steer (ours) |
|---|---|---|---|---|---|---|---|
| Llama-3.2-3B / PopQA | 0.568 | 0.302 | 0.008 | 0.075 | 0.147 | 0.018 | 0.020 |
| Qwen2.5-3B / NQOpen | 0.555 | 0.304 | 0.016 | 0.054 | 0.033 | 0.374 | 0.103 |
| Llama-3.2-3B / MMLU | 0.176 | 0.021 | 0.026 | 0.078 | 0.232 | 0.166 | 0.074 |
| Llama-3.2-3B / NQOpen | 0.507 | 0.245 | 0.036 | 0.057 | 0.191 | 0.245 | 0.088 |

Table 7: **Comparison with post-hoc calibrators** (ECE, 10 bins; lower is better). Calibrators are fit on a labeled 50% split and evaluated on the disjoint half; Raw ECE is computed on the evaluation split. Mean ablation and best steering restate our full-dataset results from Tables 3, 11, and 12.

## J Model Scale Generalization

| Rank | PopQA | NQOpen |
|---|---|---|
| 1 | m27 | m27 |
| 2 | m21 | m21 |
| 3 | m23 | m26 |
| 4 | m26 | m23 |
| 5 | m20 | m20 |
| 6 | m19 | m19 |
| 7 | m25 | m25 |
| 8 | m18 | m24 |
| 9 | m24 | m22 |
| 10 | a20.h1 | m18 |

Table 8: **Top-10 components by attribution score for Qwen2.5-7B-Instruct.** 9 of the top-10 components are shared across PopQA and NQOpen.

To test whether the discovered mechanism persists at larger scale within the same model family, we repeat circuit discovery and the steering sweep on **Qwen2.5-7B-Instruct** for PopQA and NQOpen. Table 8 lists the top-10 components: 9 of the top-10 are shared across the two datasets, and the circuit again consists of mid-to-late MLP blocks, the same locus as at 3B. Table 9 reports the steering sweep: at $\alpha = 0.6$, ECE falls from 0.647 to 0.068 on PopQA and from 0.626 to 0.109 on NQOpen, with the same dose–response pattern as at 3B, while mean ablation is effective on PopQA (0.155) but not on NQOpen (0.615), mirroring the brittleness observed at 3B. This cross-scale experimental result further complements the cross-family Qwen2.5-3B vs Llama-3.2-3B result.

| Dataset | Baseline | $\alpha = 0.2$ | $\alpha = 0.3$ | $\alpha = 0.4$ | $\alpha = 0.5$ | $\alpha = 0.6$ | Mean ablation |
|---|---|---|---|---|---|---|---|
| PopQA | 0.647 | 0.419 | 0.312 | 0.203 | 0.121 | **0.068** | 0.155 |
| NQOpen | 0.626 | 0.528 | 0.450 | 0.347 | 0.217 | **0.109** | 0.615 |

Table 9: **Intervention results (ECE) for Qwen2.5-7B-Instruct.** Steering sweep over $\alpha$ and mean ablation; best ECE per dataset in **bold**.

| Bucket | Criterion | Role |
|---|---|---|
| 1 | $\Delta_{\text{TSLD}} \leq -\tau$ | Overconfident collapse; circuit discovery target |
| 2 | $\Delta_{\text{TSLD}} \geq +\tau$ | Truth recognition; excluded |
| 3 | $\Delta_{\text{TSLD}} \approx 0$ | Noise; excluded |

Table 10: Three-bucket stratification.

| Model | Method | $\alpha$ | ECE | Brier | ECE Impr. | Brier Impr. |
|---|---|---|---|---|---|---|
| Qwen2.5-3B-Instruct | baseline | – | 0.281 | 0.310 | – | – |
| | mean ablation | – | 0.261 | 0.298 | 7.3% | 3.8% |
| | steering | 0.3 | 0.250 | 0.294 | 11.1% | 5.2% |
| | steering | 0.4 | 0.230 | 0.284 | 18.2% | 8.3% |
| | steering | 0.5 | 0.196 | 0.271 | 30.3% | 12.7% |
| | steering | 0.6 | **0.188** | **0.263** | **33.1%** | **15.3%** |
| | steering | 0.7 | 0.197 | 0.265 | 29.9% | 14.6% |
| | steering | 0.8 | 0.217 | 0.274 | 22.7% | 11.8% |
| Llama-3.2-3B-Instruct | baseline | – | 0.171 | 0.265 | – | – |
| | mean ablation | – | 0.166 | 0.269 | 3.1% | −1.8% |
| | steering | 0.3 | 0.112 | 0.246 | 34.7% | 7.0% |
| | steering | 0.4 | 0.104 | 0.245 | 39.4% | 7.5% |
| | steering | 0.5 | 0.095 | 0.243 | 44.5% | 8.3% |
| | steering | 0.6 | 0.087 | 0.242 | 48.9% | 8.4% |
| | steering | 0.7 | 0.079 | **0.242** | 53.7% | **8.6%** |
| | steering | 0.8 | **0.074** | 0.242 | **57.0%** | 8.4% |

Table 11: **Intervention results on MMLU.** Best ECE and Brier per model in **bold**.

| Model | Method | $\alpha$ | ECE | Brier | ECE Impr. | Brier Impr. |
|---|---|---|---|---|---|---|
| Qwen2.5-3B-Instruct | baseline | – | 0.551 | 0.485 | – | – |
| | mean ablation | – | 0.374 | 0.304 | 32.1% | 37.3% |
| | steering | 0.3 | 0.295 | 0.244 | 46.5% | 49.6% |
| | steering | 0.4 | 0.201 | 0.191 | 63.5% | 60.7% |
| | steering | 0.5 | 0.141 | 0.169 | 74.5% | 65.1% |
| | steering | 0.6 | **0.103** | **0.159** | **81.2%** | **67.2%** |
| | steering | 0.7 | 0.142 | 0.174 | 74.2% | 64.1% |
| | steering | 0.8 | 0.201 | 0.204 | 63.4% | 57.9% |
| Llama-3.2-3B-Instruct | baseline | – | 0.507 | 0.453 | – | – |
| | mean ablation | – | 0.245 | 0.273 | 51.7% | 39.8% |
| | steering | 0.3 | 0.318 | 0.299 | 37.4% | 34.1% |
| | steering | 0.4 | 0.233 | 0.249 | 54.1% | 45.0% |
| | steering | 0.5 | 0.145 | 0.212 | 71.3% | 53.2% |
| | steering | 0.6 | **0.088** | **0.199** | **82.6%** | **56.1%** |
| | steering | 0.7 | 0.091 | 0.207 | 82.2% | 54.3% |
| | steering | 0.8 | 0.089 | 0.210 | 82.5% | 53.6% |

Table 12: **Intervention results on NQOpen.** Best ECE and Brier per model in **bold**.

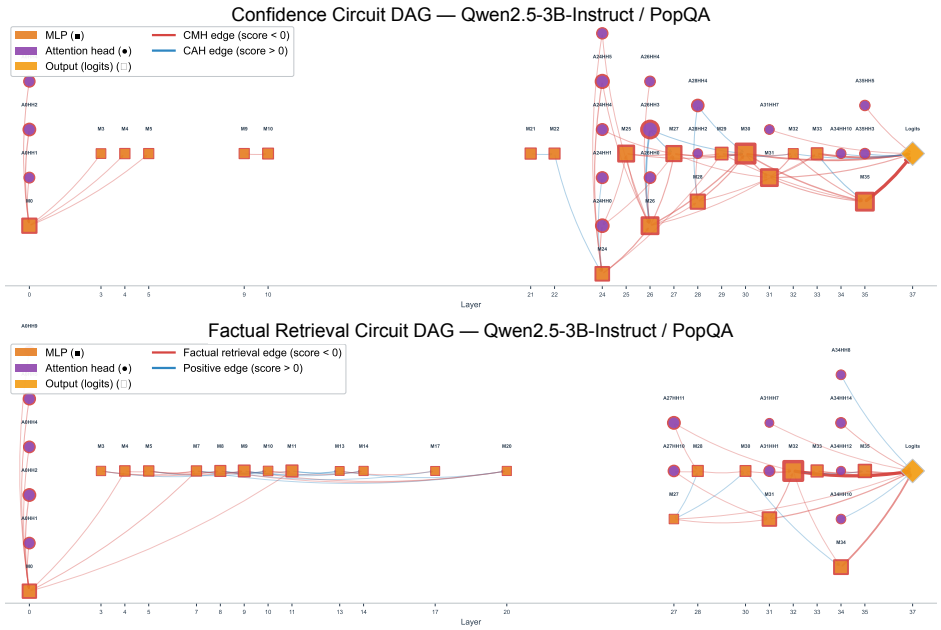



Figure 9: Verbal confidence circuit vs. factual retrieval circuit (Qwen2.5-3B-Instruct, PopQA). Both circuits are discovered on the same Bucket 1 records with the same EAP-IG machinery; the factual retrieval circuit uses an answer-logit objective in place of TSLD. The two circuits are structurally distinct: the Jaccard overlap of their top-10 (top-20) components is 0.18 (0.21), and the cosine similarity between their signed edge-attribution scores on the induced subgraphs is 0.004 (0.055).