Beyond Transfer Accuracy: Mechanism-Guided Controlled Adaptation for Low-Resource Languages

Khumaisa Nur'aini[1], **Ayu Purwarianti**[2], **Alham Fikri Aji**[1,3], **Derry Wijaya**[1,4]

[1]Monash University Indonesia, [2]Institut Teknologi Bandung, [3]MBZUAI, [4]Boston University

Correspondence: khumaisa.nuraini@monash.edu; derry.wijaya@monash.edu

Abstract

Existing circuit discovery methods rely on templated tasks with clean counterfactuals, limiting their use on diverse natural text. We adapt Contextual Decomposition for Transformers (CD-T) for unstructured settings via label-balanced activation means and task-directional relevance scoring, enabling counterfactual-free circuit discovery. We leverage the discovered circuits for Circuit-Targeted Supervised Fine-Tuning (CT-SFT), restricting parameter updates to task-relevant heads and LayerNorm. Experiments on NusaX cross-lingual sentiment transfer show that CT-SFT is highly competitive for low-resource adaptation. While non-circuit sparse updates and full fine-tuning sometimes match target accuracy through capacity recruitment, CT-SFT most consistently avoids catastrophic forgetting, preserving source-language and related-task performance. Extensions to XNLI support the source-retention and intervention findings on a harder task and two model families, showing that circuit-targeted adaptation provides a more controlled, intervention-supported alternative to global fine-tuning.

1 Introduction

Mechanistic interpretability aims to reverse-engineer neural networks into human-understandable mechanisms by identifying internal components and circuits that implement model behavior (Wang et al., 2023; Bereska and Gavves, 2024; Olah et al., 2020). In transformer language models, circuit discovery has localized behavior to sparse sets of heads, neurons, or edges (Wang et al., 2023; Conmy et al., 2023; Syed et al., 2024; Hanna et al., 2024; Hsu et al., 2025).

However, many circuit discovery methods are evaluated on structured or template-based tasks, where counterfactuals and task metrics are well-defined (Wang et al., 2023; Conmy et al., 2023; Hsu et al., 2025). Natural-text settings remain harder (Sharkey et al., 2025), as diverse inputs make faithful counterfactuals expensive or ambiguous. This limitation is critical for low-resource language adaptation, where labeled data is scarce (Joshi et al., 2020; Aji et al., 2022; Hu et al., 2020) and fine-tuning often risks instability or catastrophic forgetting (Somayajula et al., 2024; Du et al., 2023; Mosbach et al., 2021; Dodge et al., 2020).

This challenge is especially visible in Indonesian local languages, where resources remain limited despite high linguistic diversity (Aji et al., 2022; Cahyawijaya et al., 2023) and recent benchmarks show persistent gaps in model support (Adilazuarda et al., 2025; Aji and Cohn, 2025). NusaX (Winata et al., 2023) provides a parallel sentiment benchmark for Indonesian and ten local languages, making it a useful testbed for low-resource source-to-target transfer. This setting also highlights a limitation of standard transfer evaluation: target-language gains may coincide with changes in source behavior, related tasks, broader classification ability, or perturbation consistency, which are connected to cross-lingual forgetting and behavioral testing (Du et al., 2023; Vu et al., 2022; Ribeiro et al., 2020; Wu et al., 2021).

To address this, we investigate whether circuits discovered from natural source-language data can guide controlled adaptation. We adapt Contextual Decomposition for Transformers (CD-T) (Hsu et al., 2025) using label-balanced activation means and task-directional relevance scoring, enabling counterfactual-free discovery on natural text. We then introduce Circuit-Targeted Supervised Fine-Tuning (CT-SFT), which restricts parameter updates exclusively to task-relevant attention heads and LayerNorm.

Our main evaluation uses NusaX sentiment transfer (Indonesian to Acehnese, Buginese, Javanese, and Minangkabau) (Winata et al., 2023) with Llama-3.2-1B-Instruct (Grattafiori et al., 2024; Meta AI, 2024), comparing CT-SFT against full

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 40715–40742
October 24-29, 2026 ©2026 Association for Computational Linguistics

fine-tuning and sparse non-circuit controls while evaluating retention on IndoNLU (Wilie et al., 2020; Saputri et al., 2018) and SIB200 (Adelani et al., 2024); we further test scope using Llama-3.2-3B-Instruct for scalability and Llama-3.2-1B-Instruct with Qwen2.5-0.5B (Yang et al., 2024; Qwen Team, 2024) on XNLI (Conneau et al., 2018) for task and family breadth.

We ask four questions: (1) can counterfactual-free CD-T identify task-relevant circuit targets from natural text? (2) can circuit-targeted tuning support low-resource transfer, and what do non-circuit sparse controls reveal about adaptation? (3) does circuit-targeted tuning provide more controlled adaptation beyond target accuracy, including source retention, related- and less-related task retention, perturbation consistency, and interpretability? and (4) do the findings extend beyond the main NusaX Llama-1B setting, across model scale and a different multilingual classification task?

Contributions. Our contributions are as follows.
C1. We adapt CD-T to natural-text classification using label-balanced activation means and task-directional relevance scoring, enabling counterfactual-free discovery of task-relevant circuit targets.
C2. We introduce circuit-targeted supervised fine-tuning (CT-SFT), a controlled low-resource adaptation strategy that updates only discovered task-relevant heads.
C3. We show that circuit tuning is competitive for transfer, while non-circuit sparse updates can also improve target accuracy, suggesting both mechanism editing and capacity recruitment.
C4. We evaluate adaptation beyond target accuracy using source retention, related- and less-related task retention, perturbation consistency, and faithfulness/knockout intervention metrics, showing that CT-SFT gives the most balanced and interpretable profile.
C5. We test the scope of CT-SFT with Llama-3.2-3B on NusaX and both Llama-1B and Qwen2.5-0.5B on XNLI, covering scalability and task/language-family breadth.

2 Related Works

Mechanistic interpretability (MI) localizes computations inside transformers (Rai et al., 2024; Sharkey et al., 2025). Prior work has identified behavior-specific circuits using path patching (Wang et al., 2023), causal interventions, automated circuit discovery (Conmy et al., 2023), and attribution-based methods (Nanda, 2023; Syed et al., 2024; Hanna et al., 2024; Zhang et al., 2025b). Contextual Decomposition for Transformers (CD-T) offers an alternative by decomposing activations into relevant and irrelevant streams (Hsu et al., 2025).

Many circuit discovery methods are evaluated on template-based tasks, where counterfactuals and reference states are easier to define than in unstructured natural text (Sharkey et al., 2025). This gap is especially important for multilingual settings, as MI research has focused primarily on English (Yu et al., 2025). Recent work has begun to examine multilingual interpretability and cross-lingual circuit structure, finding both shared pathways and language-specific specialization (Resck et al., 2025; Zhang et al., 2025c).

Recent work links mechanistic structure to adaptation: circuits can be reused across tasks (Merullo et al., 2024), fine-tuning updates may concentrate in a small subset of heads (Wang et al., 2025), and circuit-guided tuning can update selected subgraphs/components instead of all parameters (Li et al., 2025; Zhang et al., 2025a).

In contrast, we adapt CD-T (Hsu et al., 2025) to unstructured, low-resource natural text and use the discovered circuits as targets for circuit-targeted supervised fine-tuning. This treats circuit discovery not only as a diagnosis but also as a controlled adaptation strategy for low-resource transfer with behavior preservation. By comparing circuit updates with non-circuit sparse controls, we analyze whether adaptation edits an existing task mechanism or recruits additional capacity.

3 Background: Contextual Decomposition for Transformers (CD-T)

Contextual Decomposition for Transformers (CD-T) (Hsu et al., 2025) identifies task-relevant components by decomposing each component activation into a baseline stream and an input-induced relevant stream. Let $a_x(s)$ be the activation of attention head s on input x, and let $\mu(s)$ be the precomputed mean activation for that head. CD-T defines the baseline processing as $\gamma_s = \mu(s)$ and the input-induced deviation as $\beta_s = a_x(s) - \mu(s)$. In the original CD-T setting, μ is estimated from faithful counterfactual examples, such as name or year

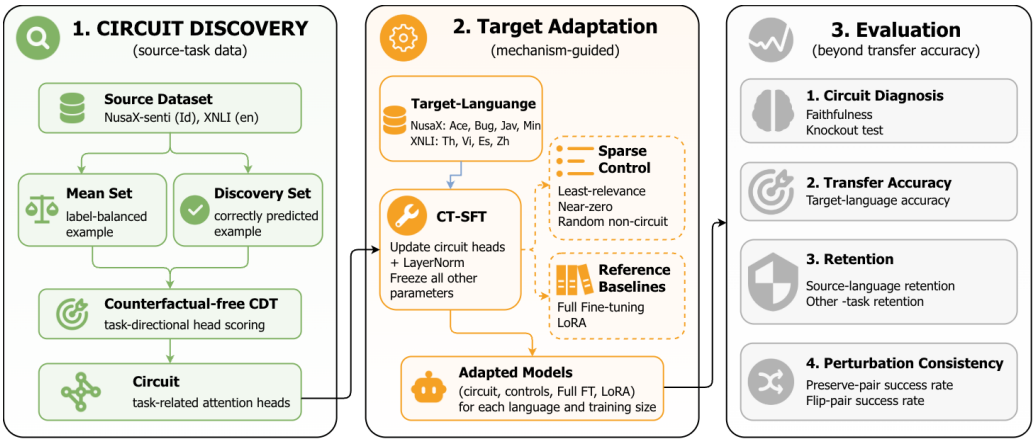


Figure 1: Overview of the controlled adaptation pipeline: counterfactual-free CD-T discovers a source-task circuit, CT-SFT adapts it to low-resource target languages, and evaluation measures circuit diagnosis, transfer, retention, and perturbation consistency.

swaps in templated tasks like IOI, greater-than, and docstring completion. The γ and β streams are then propagated through the network using contextual decomposition rules to estimate the contribution $\beta_{s\to t}(x)$ from head s to a target t.

CD-T ranks components using an unsigned relevance score based on the magnitude ratio between relevant and irrelevant streams:

$$R(s,t) = \frac{\|\beta_{s\to t}\|_1}{\|\gamma_{s\to t}\|_1}. \quad (1)$$

This score favors heads whose relevant contribution is large relative to the baseline stream. Starting from the output readout, CD-T constructs a circuit by iteratively selecting the top-scoring heads and treating them as new targets for the next backward expansion step.

While effective for controlled settings, this formulation has two limitations for natural-text classification. First, natural text does not provide obvious faithful counterfactuals for estimating μ. Second, the unsigned magnitude score can select heads with large but task-irrelevant or label-opposing contributions. The next section describes our counterfactual-free adaptation of CD-T for this setting.

4 Counterfactual-Free Circuit Discovery for Natural Text

Figure 1 summarizes our overall pipeline: we discover a source-language circuit using counterfactual-free CD-T, use the resulting circuit as the update target for CT-SFT, and evaluate the adapted models beyond target-transfer accuracy.

This section describes the discovery stage. We modify CD-T for natural-text classification in three steps. First, we replace counterfactual-based activation means with label-balanced means computed from natural source-language examples. Second, we rank heads by task-directional relevance, so that selected components are scored by whether they support the correct label rather than by magnitude alone. Third, we construct circuits through iterative backward expansion and apply score-based pruning to obtain a compact final circuit for adaptation.

4.1 Label-Balanced Precomputed Mean

A key dependency in CD-T is the precomputed mean μ, which determines what is treated as background (γ) versus signal (β). In templated tasks, μ can be estimated from faithful counterfactuals, but such counterfactuals are difficult to construct for unstructured natural text. We therefore estimate μ from a label-balanced set of in-distribution examples $\mathcal{M}$, using equal numbers of examples per class to reduce class-directional leakage into the baseline, defining $\mu(s) = \mathbb{E}_{x\sim\mathcal{M}}[a_x(s)]$ to approximate a neutral operating point for each head.

4.2 Task-Directional Relevance

Original CD-T ranks components using an unsigned magnitude ratio. Still, large contributions are not necessarily task-supporting: a head may be non-discriminative or even anti-aligned with the gold label. To address this, we score heads by whether their relevant contribution supports the correct label direction.

At the logit readout, let y_c be the correct label

and $\mathcal{Y}_{\text{other}} = \mathcal{Y} \setminus \{y_c\}$. We define:

$$R(s,t) = \beta_{s\to t}(y_c) - \frac{1}{|\mathcal{Y}_{\text{other}}|} \sum_{y \in \mathcal{Y}_{\text{other}}} \beta_{s\to t}(y). \tag{2}$$

For intermediate residual contributions, we define a task direction from the unembedding matrix:

$$\mathbf{v}_{\text{task}} = W_U(y_c) - \frac{1}{|\mathcal{Y}_{\text{other}}|} \sum_{y \in \mathcal{Y}_{\text{other}}} W_U(y), \tag{3}$$

and score the contribution by its projection onto this direction:

$$R(s,t) = \frac{\beta_{s\to t}(x) \cdot \mathbf{v}_{\text{task}}}{\|\mathbf{v}_{\text{task}}\|_2}. \tag{4}$$

Our task-directional score is related to direct logit attribution (DLA) (Wang et al., 2023; Elhage et al., 2021): both use the unembedding direction to estimate how an internal contribution supports the output decision. Unlike standard DLA, however, we apply this projection after CD-T decomposition, scoring the propagated relevant contribution $\beta_{s\to t}(x)$ rather than the raw residual-stream activation. Thus, the score measures the task alignment of a head's input-induced contribution. Positive scores indicate alignment with the correct-label direction, while negative scores indicate opposition. Figure 2 illustrates this distinction.

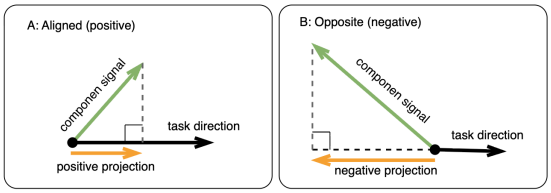


Figure 2: Illustration of task-directional relevance scoring. Panel (A) shows an aligned component with a positive projection, while Panel (B) shows an opposing component with a negative projection.

4.3 Circuit Construction and Pruning

Using the same iterative backward expansion procedure as in standard CD-T (Section 3), we construct our circuit by expanding backward from the output decision, guided by task-directional relevance. These selected heads are then treated as new targets for the next depth, and the candidate circuit is obtained by aggregating selected heads across depths. We then apply score-based pruning by removing the selected heads with the lowest CD-T relevance scores (Appendix A.4), yielding a smaller final circuit C for mechanism-guided tuning.

5 Mechanism-Guided Adaptation

5.1 Circuit-Targeted Supervised Fine-Tuning (CT-SFT) Setup

We use the discovered circuit as the update target for transfer: if the selected heads capture a task-relevant mechanism, then updating them should provide a more controlled route for adaptation than updating the full model. Given the final circuit $\mathcal{C}$ from Section 4.3, Circuit-Targeted Supervised Fine-Tuning (CT-SFT) restricts optimization to heads in $\mathcal{C}$ and LayerNorm. We freeze MLPs, embeddings, the unembedding matrix, and all unselected heads, using head-level gradient masking to localize learning to the discovered circuit. This updates 2.355% of Llama-1B and 2.410% of Llama-3B parameters; details are in Appendix A.5.

5.2 Baselines and Sparse Controls

Editing vs. recruitment. Table 1 summarizes the adaptation methods. Circuit tuning tests *mechanism editing*, since it updates heads identified as task-relevant. The sparse controls test whether similar low-resource transfer can also arise from *capacity recruitment*, where non-circuit or weakly task-related heads are repurposed for the target task. Random NC is sampled from the complement of the discovered circuit, so it presupposes circuit discovery; it is therefore a sparse-capacity control, not a standalone random selection rule.

Method	Heads	Role
Full FT	All params.	Global edit
Circuit	Circuit (C)	Mechanism-guided edit
Least-rel.	Negative-score	Negative recruitment
Near-zero	Near-zero-score	Indifferent recruitment
Random NC	Random heads outside C	Non-circuit recruitment control

Table 1: Adaptation methods and mechanistic roles.

6 Experimental Setup

6.1 Research Questions and Evaluation Map

We evaluate our method through four research questions (RQs) covering intervention behavior (Hanna et al., 2024; Conmy et al., 2023; Chen et al., 2025), transfer effectiveness, controlled adaptation beyond accuracy (Du et al., 2023; Ribeiro et al., 2020; Wu et al., 2021), and robustness across model scales and tasks. Table 2 maps these RQs to their corresponding evaluation protocols. Full definitions for intervention metrics and datasets are provided in Appendices A.6 and A.9.

RQ	Evaluation	Purpose
RQ1	Faithfulness, knockout	Identify task-relevant circuit targets
RQ2	Transfer, sparse controls	Test transfer and recruitment
RQ3	Retention, perturbations	Test controlled adaptation beyond accuracy
RQ4	Llama-3B, XNLI	Test scale, task, and language-family robustness

Table 2: Research questions and evaluation protocols.

6.2 Model

We use Llama-3.2-1B-Instruct (Grattafiori et al., 2024; Meta AI, 2024) as the main model for our reported experiments. To test scalability, we additionally run the NusaX experiment on Llama-3.2-3B-Instruct (Grattafiori et al., 2024; Meta AI, 2024). For the XNLI (Conneau et al., 2018) robustness experiment, we evaluate across two model families by using both Llama-3.2-1B-Instruct and Qwen2.5-0.5B (Yang et al., 2024; Qwen Team, 2024). Because these models perform near-chance on the three-way XNLI task before task adaptation, we first apply English-competence tuning before circuit discovery. Thus, the NusaX experiments use an instruction-tuned model with usable sentiment behavior out of the box, while the XNLI experiments first learn the English source task through competence tuning.

6.3 Datasets and Evaluation Setting

We utilize NusaX-Senti (Winata et al., 2023) for low-resource transfer (Indonesian source to Acehnese, Buginese, Javanese, and Minangkabau targets). We evaluate target transfer and Indonesian source retention using 400 examples per language. Behavior preservation is assessed via Indonesian IndoNLU EmoT (Wilie et al., 2020; Saputri et al., 2018) (related-task retention) and SIB200 topic classification (Adelani et al., 2024) (less-related retention), each using 400 examples. Perturbation consistency is evaluated using 200 label-preserving and 200 label-flipping pairs derived from NusaX-Senti, reported as pair-level success rates (Ribeiro et al., 2020).

Finally, we evaluate robustness on XNLI (Conneau et al., 2018). To match the NusaX evaluation budget, we construct XNLI subsets containing 500 training, 100 validation, and 400 test examples per language. We use English for competence tuning and circuit discovery, with Thai, Vietnamese, Spanish, and Chinese as target languages. We report XNLI target transfer and English source retention. Unlike the Indonesian benchmarks, XNLI starts from an English NLI competence-tuned checkpoint. Therefore, IndoNLU EmoT, SIB200, and Indonesian perturbation evaluations are not directly comparable in this setting. Further dataset details and examples are provided in Appendix A.9.

6.4 Training and Discovery Protocol

Each target language is fine-tuned with $n \in \{25, 50, 75, 100\}$ examples across four seeds. Circuit discovery uses a label-balanced mean set and a correctly predicted discovery set from the source language. Heads are selected using task-directional relevance, followed by score-based pruning to obtain the final circuit. All sparse controls are then matched to the resulting post-pruning head budget for each seed.

The Llama-3.2-3B-Instruct NusaX experiments follow this same protocol as the main Llama-3.2-1B-Instruct evaluation. For XNLI, we first competence-tune Llama-3.2-1B-Instruct and Qwen2.5-0.5B on 250 English XNLI training examples, then perform circuit discovery on the competence-tuned English checkpoints. These circuits are then used for target-language CT-SFT. Full implementation, discovery, training, and effective update-budget details are provided in Appendices A.2–A.5.

7 Results

7.1 RQ1: Counterfactual-free CD-T identifies task-relevant circuit targets in Llama

Key finding. *The discovered Llama circuits show higher faithfulness and stronger knockout effects than matched sparse controls.*

Table 3 reports validation-set intervention scores. Faithfulness measures retained label-token accuracy when only selected heads are preserved under mean ablation, while knockout measures the relative accuracy drop when selected heads are ablated. Across Indonesian and the target languages, Circuit achieves higher faithfulness and knockout than matched sparse controls, e.g., 53.91% vs. 4.30% knockout on Indonesian at 3B, indicating a stronger intervention target for CT-SFT. Sparse controls can retain non-zero faithfulness because transformer behavior may be distributed across redundant or self-repairing components (Wang et al., 2023; Rushing and Nanda, 2024); however, their weaker knockout effects suggest they are less task-specific than the

Method	Faithfulness (% full acc.)					Knockout (% acc. drop)				
	Ind	Ace	Bug	Jav	Min	Ind	Ace	Bug	Jav	Min
Llama-3.2-1B-Instruct, avg. 62.8/512 heads after pruning, 2.355% updated params										
Circuit	**80.99**	**88.36**	**102.55**	**82.50**	**90.95**	**52.82**	**37.93**	**17.35**	**43.33**	**34.91**
Least-relevant	39.08	59.91	82.14	56.67	59.91	3.87	0.43	-2.55	8.33	-3.88
Near-zero	41.20	60.78	85.71	56.25	61.64	1.76	6.47	0.00	5.42	3.45
Random NC	42.96	60.78	87.76	56.67	61.64	14.08	17.67	1.53	20.83	15.52
Llama-3.2-3B-Instruct, avg. 52.5/672 heads after pruning, 2.410% updated params										
Circuit	**73.44**	**101.63**	**99.44**	**89.50**	**90.69**	**53.91**	**23.37**	**9.44**	**30.00**	**32.84**
Least-relevant	34.77	68.48	80.56	60.00	61.76	31.25	6.52	4.44	13.50	13.73
Near-zero	37.50	70.11	80.56	61.50	60.78	9.77	1.63	3.89	6.50	7.35
Random NC	37.89	68.48	78.33	63.00	63.24	4.30	-0.54	5.56	4.50	1.47

Table 3: Intervention evaluation for Llama-3.2-1B and 3B-Instruct. Scores use validation label-token accuracy averaged over four seeds. Pruned circuits contain 62.8/512 heads on average for 1B and 52.5/672 for 3B, corresponding to effective CT-SFT update budgets of 2.355% and 2.410% of the parameters updated by Full FT, respectively. Sparse controls are matched to the per-seed head budget. Faithfulness is retained full-model accuracy when only selected heads remain active; knockout is the accuracy drop when selected heads are mean-ablated. Faithfulness can exceed 100% if ablating non-circuit heads removes noise.

discovered circuit.

As a sensitivity check, we also evaluate zero ablation. This intervention is harsher and less aligned with CD-T's mean/reference decomposition, so keep-only faithfulness is more variable. Still, Circuit has the strongest average knockout for both Llama-1B and Llama-3B, showing that the selected heads remain load-bearing under a second ablation baseline. Since high zero-ablation faithfulness can occur without strong knockout, we treat knockout as the more diagnostic signal in this check. Detailed results are provided in Appendix A.7.

7.2 RQ2: Circuit tuning supports transfer, but sparse adaptation also involves recruitment

Key finding. *Circuit tuning is competitive for low-resource transfer, but least-relevant and random non-circuit updates can also improve target accuracy, suggesting that sparse adaptation may involve both mechanism editing and capacity recruitment.*

Table 4 and Table 5 summarize target-language adaptation. Zero-shot results give the pre-adaptation baseline, while Full FT provides a high-plasticity reference that updates all parameters. Averaged across training sizes, Circuit is highly competitive among sparse methods; detailed trends are provided in Appendix B.3 and Figure 4. However, non-circuit sparse updates can sometimes match or exceed Circuit, showing that transfer accuracy alone is insufficient to identify the most controlled or interpretable update strategy.

Because Random NC is sampled from outside the discovered circuit, it presupposes circuit discovery and should be interpreted as a sparse-capacity control rather than a standalone random selection rule. Its strength is consistent with sparse-update and lottery-ticket-style findings, where restricted trainable subsets can still support adaptation while reducing interference (Frankle and Carbin, 2019; Mehta, 2019; Ansell et al., 2022). More broadly, it is compatible with intervention-based self-repair observations, which suggest that transformer behavior can be distributed and supported by compensatory components rather than localized to a single non-redundant circuit (Wang et al., 2023; Rushing and Nanda, 2024).

7.3 RQ3: Circuit tuning provides more controlled adaptation beyond target accuracy

Key finding. *Non-circuit sparse updates can be competitive on individual metrics, but Circuit provides the most balanced and interpretable profile across faithfulness, transfer, retention, and perturbation consistency.*

We use the preservation evaluations defined in Section 6.3 to test whether sparse transfer gains remain stable beyond target accuracy.

Source-language retention. On Indonesian source-task retention, Random non-circuit achieves the strongest average score (Table 4). One possible explanation is that Random non-circuit does not directly update the discovered source-task circuit, thereby leaving the original sentiment mechanism less disrupted. At the same time, the updated non-circuit heads can provide additional capacity for target-language adaptation. This supports the recruitment interpretation: non-circuit updates can

Method	Evaluation	Zero-shot	Ace	Bug	Jav	Min	Avg.
Zero-shot	Transfer	–	52.25	39.00	56.00	52.00	49.81
Circuit	Transfer	–	**50.00**	42.90	**63.50**	61.30	**54.40**
	Source	73.50	69.20	68.40	70.30	**70.30**	69.50
	Emotion	41.25	**40.10**	39.10	**41.40**	40.00	**40.20**
	Topic	51.25	**68.40**	**69.40**	**69.60**	69.60	**69.30**
	Pres. Pair	64.00	64.30	63.80	65.70	65.00	64.70
	Flip Pair	81.00	79.90	78.40	79.90	79.70	79.50
Least-relevant	Transfer	–	48.20 (-1.80)	42.60 (-0.30)	60.50 (-3.00)	**61.50 (+0.20)**	53.20 (-1.20)
	Source	73.50	70.20 (+1.00)	67.80 (-0.60)	**71.10 (+0.80)**	69.70 (-0.60)	69.70 (+0.20)
	Emotion	41.25	32.80 (-7.30)	33.00 (-6.10)	32.90 (-8.50)	32.70 (-7.30)	32.90 (-7.30)
	Topic	51.25	**68.40 (0.00)**	69.10 (-0.30)	69.10 (-0.50)	**70.20 (+0.60)**	69.20 (-0.10)
	Pres. Pair	64.00	64.20 (-0.10)	62.60 (-1.20)	**65.80 (+0.10)**	64.00 (-1.00)	64.10 (-0.60)
	Flip Pair	81.00	75.50 (-4.40)	67.70 (-10.70)	75.10 (-4.80)	76.80 (-2.90)	73.80 (-5.70)
Near-zero	Transfer	–	47.10 (-2.90)	39.90 (-3.00)	59.10 (-4.40)	55.80 (-5.50)	50.50 (-3.90)
	Source	73.50	66.20 (-3.00)	65.50 (-2.90)	65.90 (-4.40)	64.40 (-5.90)	65.50 (-4.00)
	Emotion	41.25	37.50 (-2.60)	**40.40 (+1.30)**	39.60 (-1.80)	**40.10 (+0.10)**	39.40 (-0.80)
	Topic	51.25	57.20 (-11.20)	59.80 (-9.60)	59.40 (-10.20)	58.50 (-11.10)	58.70 (-10.60)
	Pres. Pair	64.00	62.30 (-2.00)	62.10 (-1.70)	62.20 (-3.50)	61.30 (-3.70)	62.00 (-2.70)
	Flip Pair	81.00	**83.00 (+3.10)**	**82.80 (+4.40)**	**81.30 (+1.40)**	**80.50 (+0.80)**	**81.90 (+2.40)**
Random NC	Transfer	–	48.00 (-2.00)	**44.50 (+1.60)**	61.80 (-1.70)	60.30 (-1.00)	53.60 (-0.80)
	Source	73.50	**70.30 (+1.10)**	**69.70 (+1.30)**	70.90 (+0.60)	70.10 (-0.20)	**70.30 (+0.80)**
	Emotion	41.25	35.80 (-4.30)	35.10 (-4.00)	37.10 (-4.30)	36.30 (-3.70)	36.10 (-4.10)
	Topic	51.25	63.30 (-5.10)	66.10 (-3.30)	63.80 (-5.80)	64.00 (-5.60)	64.30 (-5.00)
	Pres. Pair	64.00	**65.00 (+0.70)**	**64.40 (+0.60)**	65.20 (-0.50)	**65.20 (+0.20)**	**64.90 (+0.20)**
	Flip Pair	81.00	80.80 (+0.90)	79.30 (+0.90)	78.10 (-1.80)	78.90 (-0.80)	79.30 (-0.20)

Table 4: Transfer and preservation of behavior after target-language adaptation. Scores are percentages averaged across training sizes and seeds; parentheses show differences from Circuit. Bold marks the best score, gray shading the lowest. Random NC denotes random non-circuit updates. Full FT is reported separately in Table 5; detailed training-size trends are shown in Appendix B.3.

preserve the original circuit while repurposing other heads to support transfer.

Other-task retention. The circuit is more stable than non-circuit updates on the two other-task retention evaluations (Table 4). This suggests that localizing updates to the discovered circuit does not strongly disrupt the broader classification logic used by related and less-related tasks. In contrast, non-circuit updates can support transfer through recruitment, but the recruited heads are not guaranteed to preserve other task-relevant computations. Since Least-relevant and Near-zero are defined only with respect to sentiment relevance, updating them may still disturb affective or topic-classification behavior, which may explain the larger drops on IndoNLU EmoT and SIB200.

Perturbation consistency. Perturbation results show that label-preserving and label-flipping consistency capture different behaviors (Table 4). Random non-circuit is slightly better than Circuit on label-preserving pairs, while Near-zero obtains the highest label-flipping score. This suggests that Near-zero heads are more plastic: because they are weakly tied to the original sentiment circuit, they can be repurposed more easily when explicit sentiment cues are flipped. However, this flexibility is not equivalent to a better sentiment mechanism, since Near-zero is also weakest on transfer, source retention, topic retention, and label-preserving perturbations. Thus, Near-zero is more responsive to label-changing edits, while Circuit provides a more balanced profile across transfer, retention, and consistency.

Reference baselines. Table 5 reports Full FT and LoRA as same-data/same-protocol reference baselines, not exhaustively tuned optima. All methods use the same target examples, seeds, training setup, and evaluation pipeline, so the comparison isolates behavior under a shared protocol rather than proving dominance over all possible FT or PEFT variants. Full FT slightly improves average transfer over Circuit but substantially weakens preservation. LoRA preserves source retention, emotion retention, and label-flipping consistency well, but is weaker on target transfer, SIB200 Topic retention, and label-preserving perturbations. Thus, LoRA is effective for sentiment-adjacent preservation, while

Circuit provides a stronger broader controlled-adaptation profile. A per-language Circuit–LoRA breakdown is provided in Appendix B.4.

Eval	Full FT	LoRA	Circuit	Δ_{FT}	Δ_{LoRA}
Trans.	55.30	50.32	54.41	+0.89	-4.09
Source	62.43	72.98	69.55	-7.11	+3.44
Emot.	6.09	41.07	40.16	-34.07	+0.92
Topic	12.17	48.51	69.29	-57.12	-20.77
Pres.	60.09	63.04	64.72	-4.62	-1.68
Flip	47.98	82.54	79.47	-31.48	+3.07

Table 5: NusaX reference baselines. Scores are averaged over seed-level averages; deltas are relative to Circuit.

LayerNorm and localization. LN-only remains close to the pre-adaptation baseline, especially on transfer and Topic retention (Table 6). In contrast, heads-only no-LN recovers most of CT-SFT's transfer and Topic gains. This indicates that LayerNorm alone is not a competitive adaptation path and cannot explain CT-SFT's main gains. The selected heads carry the primary adaptive capacity, while LayerNorm provides an auxiliary, stabilizing, or complementary update. Detailed per-language LayerNorm results and heads-only sparse-control comparisons are provided in Appendix B.1.

Eval.	Pre	LN	Heads	Heads+LN
Transfer	49.81	50.70	54.22	**54.41**
Source	73.50	**73.84**	71.94	69.55
Emotion	41.25	40.98	**41.30**	40.16
Topic	51.25	51.71	67.48	**69.29**
Pres. Pair	64.00	63.74	64.10	**64.72**
Flip Pair	81.00	82.66	**85.81**	79.47

Table 6: Aggregate Llama-1B LayerNorm ablation results, averaged over training sizes and seeds. Pre is the pre-adaptation model; LN updates only LayerNorm/RMSNorm; Heads updates selected attention heads; Heads+LN is CT-SFT. Higher is better.

7.4 RQ4: Does the finding extend beyond the main Llama-1B NusaX setting?

Key finding. *The core findings extend beyond the main Llama-1B NusaX setting: Circuit remains load-bearing under intervention and avoids the severe forgetting of Full FT, while non-circuit transfer gains recur, supporting the recruitment pattern from RQ2.*

Scalability on Llama-3.2-3B-Instruct. Tables 3 and 7 test whether the NusaX findings hold at 3B scale. The intervention results strengthen RQ1: Circuit achieves much higher faithfulness and knockout than matched controls, while Random NC has near-zero knockout. Thus, the discovered heads remain strong intervention targets in the larger model.

Method	Transfer	Source	Emotion	Topic
Zero-shot	30.12	56.00	34.50	20.00
Circuit	58.68	76.99	36.84	68.54
Least-rel.	59.66	76.70	34.91	58.11
Near-zero	58.01	76.25	30.88	45.34
Rand. NC	58.70	**77.11**	33.51	**70.30**
Full FT	**60.08**	71.45	1.01	4.99
LoRA	55.14	73.05	**38.16**	32.39

Table 7: Llama-3B NusaX results. Scores are averaged over four seeds after averaging across target languages and training sizes. Full per-language sparse-control and reference-baseline results are provided in Appendices C.1 and C.2.

The downstream results also support RQ3. Full FT gives the highest transfer but severely damages related and broader behavior, reducing Emotion and Topic retention to near-zero levels. LoRA is less destructive, but remains weak on Topic retention and target transfer. Circuit therefore gives a stronger controlled-adaptation profile: it remains competitive in transfer while avoiding the severe behavioral collapse of Full FT and the narrower Topic-retention profile of LoRA.

The sparse controls reinforce the recruitment pattern from RQ2. Random NC matches Circuit on transfer and source retention and slightly exceeds it on Topic retention, showing that non-circuit heads can sometimes provide useful adaptation capacity. However, Random NC has near-zero knockout and is not selected by source-task intervention evidence. Thus, circuit targeting is valuable not because it always dominates every sparse control, but because it gives a comparable profile from a principled, interpretable, knockout-supported update location. The 3B results also show that sparse controls are not interchangeable: relative to Circuit, Near-zero loses 23.2 points, and Least-relevant loses 10.4 points on SIB200 Topic retention.

Model	Zero	Circuit	LoRA	Full FT
Llama-1B	14.16	14.10	14.08	135.74
Llama-3B	9.60	9.55	9.54	16.10

Table 8: Indonesian Wikipedia perplexity after adaptation. Lower is better. Scores are averaged over target-language adaptation settings.

As a diagnostic for broad language-model drift, we evaluate Indonesian Wikipedia perplexity af-

ter adaptation (Table 8). LoRA remains close to Circuit in perplexity, so its weaker SIB200 Topic retention is not simply explained by broad language-model drift. Full FT, especially at 1B, shows much larger perplexity degradation.

XNLI task and language-family robustness. XNLI is used as a narrower robustness check for source retention and intervention robustness, not as a full replication of the NusaX multi-axis preservation suite. Table 9 summarizes the XNLI robustness check. Since XNLI starts from an English NLI competence-tuned checkpoint, the IndoNLU, SIB200, and Indonesian perturbation evaluations used for NusaX are not directly comparable. We use XNLI to test source retention and the RQ1 intervention finding on a harder task across two model families. Intervention results in Appendix C.3 and Table 27 show that Circuit has the strongest knockout effects across both models. Faithfulness is lower than in NusaX, suggesting that NLI behavior is more distributed: the selected heads are load-bearing under intervention, but not sufficient alone to reconstruct the full behavior.

Method	Qwen-0.5B		Llama-1B	
	Trans.	En	Trans.	En
Comp. eval	62.56	**74.38**	44.39	47.00
Circuit	63.52	71.99	**46.76**	**47.04**
Avg. NC	**64.49**	73.29	46.73	45.82
Full FT	39.48	41.98	34.05	34.05
LoRA	63.09	71.63	46.34	46.43

Table 9: XNLI transfer and English retention. Scores are averaged over target languages, training sizes, and four seeds; Avg. NC averages Least-relevant, Near-zero, and Random NC. Full results are in Appendix C.

Downstream, Full FT causes severe English forgetting, whereas sparse methods and LoRA remain closer to the competence-tuned checkpoint. On Qwen, Avg. non-circuit slightly exceeds Circuit on transfer, echoing the recruitment pattern from RQ2. Llama starts from weaker XNLI competence, so we treat it as a weak-competence stress test rather than a scale comparison. Overall, XNLI supports the preservation and intervention findings while showing that transfer can also benefit from non-circuit recruitment.

8 Discussion and Conclusion

This paper connects circuit discovery with controlled adaptation in natural-text classification. By adapting CD-T with label-balanced activation means and task-directional relevance scoring, we identify task-relevant attention-head circuits without manually constructed counterfactuals. These circuits show stronger faithfulness and knockout scores than matched sparse controls, making them useful intervention-supported targets for CT-SFT.

Our results suggest an editing–recruitment distinction. Circuit tuning updates heads identified by source-task intervention evidence, while non-circuit sparse updates can sometimes improve transfer by recruiting capacity outside the discovered circuit. Random NC is informative as a sparse-capacity control: it shows that restricted non-circuit subsets can support adaptation, but it is not a standalone selection rule because it is sampled from the complement of the discovered circuit. Its competitiveness is compatible with sparse-update findings and with broader evidence that transformer behavior can be distributed and partly compensatory. However, its weak knockout effects show that it is not the same kind of source-task intervention target as Circuit.

The baseline comparisons show why controlled adaptation matters. Full FT can improve transfer, but it causes severe behavioral drift. LoRA preserves sentiment-adjacent behavior better, but is weaker on broader Topic retention and label-preserving perturbations. The Llama-3B and XNLI extensions support the same framing: the selected circuit heads remain load-bearing under intervention, Full FT forgets strongly, and non-circuit gains recur as evidence of recruitment.

Overall, CT-SFT is not a method that consistently maximizes all metrics. Its value is that it provides a principled update location: a compact, knockout-supported set of task-relevant heads that can be adapted while limiting damage to existing behavior. When target accuracy alone matters, other sparse or PEFT methods may be competitive; when controlled low-resource adaptation matters, circuit targeting offers a more interpretable balance between transfer, preservation, and intervention evidence.

Limitations

This work has several limitations regarding model competence, evaluation scope, baseline coverage, and interpretability granularity.

Dependence on checkpoint competence. CT-SFT assumes that the model already has usable source-task behavior before circuit discov-

ery. As seen in the Llama-1B XNLI setting, if the competence-tuned checkpoint remains near chance, sparse adaptation cannot reliably construct the capability from scratch. CT-SFT should therefore be understood as an editing and routing technique that adapts an existing source-task capability, not as a method for creating that capability from scratch.

Scope of architectures and scales. Our experiments focus on relatively small decoder-only models: Llama-3.2-1B, Llama-3.2-3B, and Qwen2.5-0.5B. We do not evaluate encoder-only models or deployment-scale models. At larger scales, representations may become more distributed, which could change the redundancy and self-repair dynamics observed during circuit knockout.

Scope of tasks and evaluation axes. Our counterfactual-free CD-T adaptation relies on task-directional relevance scoring, which is most natural for tasks with discrete label tokens or compact output sets, such as classification, entity labeling, or controlled next-token prediction. Extending it to open-ended generation requires defining a suitable task direction over less constrained outputs. In addition, XNLI provides a task and language-family robustness check, but not a full replication of the NusaX multi-axis preservation suite. Thus, the strongest controlled-adaptation claims come from NusaX, while XNLI mainly supports source-retention and intervention findings.

Baseline and sparse-control scope. Our baseline comparisons isolate update location under a shared training protocol, rather than establishing dominance over all tuned baselines. Full FT and LoRA use the same data, seeds, epochs, learning rate, batch size, prompts, and evaluation pipeline as CT-SFT, but we do not perform extensive method-specific hyperparameter search. Random NC is also a strong sparse-capacity control in some settings, limiting any claim that circuit targeting is necessary for target accuracy alone. Our main claim is narrower: circuit targeting provides an intervention-supported update location with a more interpretable transfer–preservation tradeoff.

Intervention metrics. Faithfulness and knockout evaluate internal intervention behavior using label-token accuracy from a single forward pass, not autoregressively generated predictions. These metrics are useful for controlled circuit comparison but are not identical to downstream generation performance. Their absolute values may also depend on the intervention baseline, reference activations, and intervention sites.

Circuit granularity and interpretability. Our circuit granularity is restricted to attention heads. CT-SFT updates selected heads and LayerNorm, while MLPs, embeddings, the unembedding matrix, and unselected heads are frozen. This localization makes the update easier to analyze, but may miss task-relevant computation in MLPs, residual interactions, or finer-grained edges. The discovered circuits should therefore be interpreted as useful intervention-supported update targets, not exhaustive explanations of model behavior.

Use of AI Assistants

We used AI assistants as writing and coding aids, including for improving clarity and grammar, restructuring text, drafting LaTeX/table snippets, debugging scripts, and refactoring small code fragments. All scientific content, experimental design, implementation decisions, analyses, interpretations, and reported results were produced and verified by the authors. AI assistants were not used to generate datasets, train models, or produce experimental results.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Muhammad Farid Adilazuarda, Musa Izzanardi Wijanarko, Lucky Susanto, Khumaisa Nur'aini, Derry Tanti Wijaya, and Alham Fikri Aji. 2025. NusaAksara: A multimodal and multilingual benchmark for preserving Indonesian indigenous scripts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28371–28401, Vienna, Austria. Association for Computational Linguistics.

Alham Fikri Aji and Trevor Cohn. 2025. LORAXBENCH: A multitask, multilingual benchmark suite for 20 Indonesian languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17421–17446, Suzhou, China. Association for Computational Linguistics.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of ACL*.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety – a review. *Transactions on Machine Learning Research*.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, and 29 others. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Hang Chen, Jiaying Zhu, Xinyu Yang, and Wenya Wang. 2025. Rethinking circuit completeness in language models: And, or, and adder gates. In *Advances in Neural Information Processing Systems*.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Wei Du, Thien Huu Nguyen, David Rosenberg, David Grangier, Andrew McCallum, Hongyu Lin, and Alexander M. Rush. 2023. Measuring the instability of fine-tuning in neural language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1728–1747, Toronto, Canada. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 538 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. 2024. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Aliyah R. Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Odisho, Peter R. Carroll, and Bin Yu. 2025. Efficient automated circuit discovery in transformers using contextual decomposition. In *International Conference on Learning Representations (ICLR)*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Yueyan Li, Wenhao Gao, Caixia Yuan, and Xiaojie Wang. 2025. Fine-tuning is subgraph search: A new lens on learning dynamics. *arXiv preprint*, arXiv:2502.06106.

Rahul Mehta. 2019. Sparse transfer learning via winning lottery tickets. *arXiv preprint arXiv:1905.07785*.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Meta AI. 2024. meta-llama/llama-3.2-1b-instruct. `https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct`.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Neel Nanda. 2023. Attribution patching: Activation patching at industrial scale. Blog post.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and interpretability of multilingual large language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20465–20497, Suzhou, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Cody Rushing and Neel Nanda. 2024. Explorations of self-repair in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42836–42855. PMLR.

Mei Silviana Saputri, Rahmad Mahendra, and Mirna Adriani. 2018. Emotion classification on indonesian twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, and 10 others. 2025. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*.

Sai Ashish Somayajula, Youwei Liang, Li Zhang, Abhishek Singh, and Pengtao Xie. 2024. Generalizable and stable finetuning of pretrained language models. In *Proceedings of NAACL*.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416, Miami, Florida, US. Association for Computational Linguistics.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300. Association for Computational Linguistics.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. In *International Conference on Learning Representations (ICLR)*.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. 2025. Towards understanding fine-tuning mechanisms of llms via circuit analysis. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, X. Li, Zhi Yuan Lim, S. Soleman, R. Mahendra, Pascale Fung, Syafri Bahar, and A. Purwarianti. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S. Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan

Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Haeun Yu, Seogyeong Jeong, Siddhesh Pawar, Jisu Shin, Jiho Jin, Junho Myung, Alice Oh, and Isabelle Augenstein. 2025. Entangled in representations: Mechanistic investigation of cultural biases in large language models. *arXiv preprint arXiv:2508.08879*.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. 2025a. Exploring the translation mechanism of large language models. In *NeurIPS 2025*.

Lin Zhang, Wenshuo Dong, Zhuoran Zhang, Shu Yang, Lijie Hu, Ninghao Liu, Pan Zhou, and Di Wang. 2025b. Eap-gp: Mitigating saturation effect in gradient-based automated circuit identification. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025c. The same but different: Structural similarities and differences in multilingual language modeling. In *International Conference on Learning Representations (ICLR)*.

A Experimental Details

A.1 Overview

This appendix provides the implementation, training, evaluation, circuit discovery, and update budget details for the experiments in the main paper. We separate the details into four parts for readability. Appendix A.2 summarizes the model families and compute setup. Appendix A.3 describes the adaptation and evaluation protocol shared across CT-SFT, sparse controls, Full FT, and LoRA. Appendix A.4 describes circuit discovery, pruning, and the CT-SFT configuration. Appendix A.5 reports the effective updated-parameter budget for CT-SFT and the baselines.

Circuit discovery, circuit pruning, CT-SFT, and sparse-control evaluations were implemented in TransformerLens[1]. LoRA baselines were implemented using Hugging Face Transformers and PEFT. For each target language, all sparse-control methods are matched to the same post-pruning circuit head budget for each seed.

A.2 Model and Compute Setup

Component	Setting
Models	Qwen2.5-0.5B, Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct
Approx. model sizes	0.5B, 1B, and 3B parameters
Main model	Llama-3.2-1B-Instruct
Scalability model	Llama-3.2-3B-Instruct
XNLI robustness models	Qwen2.5-0.5B and Llama-3.2-1B
Compute infrastructure	SLURM-managed GPU cluster
GPU setup	Single GPU per run
GPU types	NVIDIA A40
Memory consideration	Circuit discovery requires activation caching and is more memory-intensive than standard fine-tuning

Table 10: Model sizes and computational setup.

Table 10 summarizes the model families and compute environment used in our experiments. The main experiments use Llama-3.2-1B-Instruct, while Llama-3.2-3B-Instruct is used for the NusaX scalability check. For XNLI, we evaluate both Llama-3.2-1B-Instruct and Qwen2.5-0.5B to test whether the source-retention and intervention patterns hold across model families. Each run uses a single GPU on a SLURM-managed cluster. Circuit discovery is more memory-intensive than standard fine-tuning because it requires activation caching for CD-T scoring and intervention evaluation.

[1] https://transformerlensorg.github.io/TransformerLens/

A.3 Adaptation and Evaluation Protocol

Table 11 summarizes the shared adaptation and evaluation protocol. For NusaX, Indonesian is used as the source language for circuit discovery, and Acehnese, Buginese, Javanese, and Minangkabau are used as target languages. Each target language is adapted with $n \in \{25, 50, 75, 100\}$ training examples across four seeds. All adaptation methods use the same target-language examples, prompt format, optimizer, learning rate, batch size, number of epochs, maximum sequence length, and evaluation pipeline. This shared protocol is intended to isolate the effect of update location under a common training setup, rather than to perform method-specific hyperparameter search.

Component	Setting
Implementation framework	TransformerLens
Base model	Llama-3.2-1B-Instruct
Source language	Indonesian
Target languages	Acehnese, Buginese, Javanese, Minangkabau
Target training size	$n \in \{25, 50, 75, 100\}$
Seeds	$\{31, 777, 2025, 12345\}$
Optimizer	AdamW
Learning rate	5×10^{-5}
Training epochs	5
Batch size	16
Maximum sequence length	128
Training prompt style	Instruct format
Output generation	Maximum 3 new tokens
XNLI competence tuning size	250 English examples

Table 11: Training and evaluation setup.

A.4 Circuit Discovery, Pruning, and CT-SFT Configuration

Circuit discovery uses a label-balanced mean set and a separate correctly predicted discovery set from the source language. Heads are selected with task-directional relevance and expanded backward from the output decision. Figure 3 illustrates this iterative construction procedure: starting from the logit readout, the top-scoring heads are selected and then treated as new targets for the next backward

expansion step. The final candidate circuit is the union of selected heads across depths.

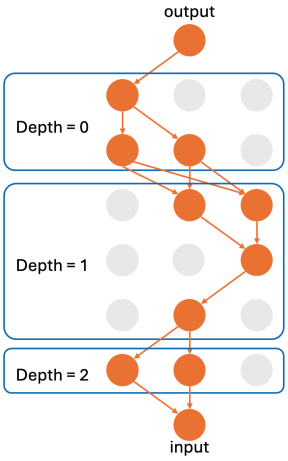


Figure 3: Iterative circuit construction. Beginning at the logit readout, we select the top-scoring heads, recursively expand backward by treating the selected heads as new targets, and form the final candidate circuit by taking the union of selected heads across depths.

After constructing the initial CD-T circuit, we prune the selected heads using a multi-pass faithfulness-guided greedy procedure. In each pass, candidate heads are considered in increasing order of absolute CD-T score. A head is removed if its removal improves the pruning objective while keeping the degradation of faithfulness within a fixed tolerance. We repeat this procedure for up to five passes, stopping early when a full pass removes no additional heads. The pruning objective is:

$$\text{Score}(C) = 0.7 \cdot \text{Faith.}(C) + 0.3 \cdot \text{Knockout}(C).$$

Here, faithfulness measures the accuracy retained when only the circuit heads are kept and all non-circuit heads are mean-ablated. Knockout measures the drop in accuracy when the circuit heads are mean-ablated while all other heads remain active. We weight faithfulness higher than knockout to prioritize the standalone task competence of the discovered circuit, while using knockout as a secondary constraint to prune redundant heads. These weights were selected empirically based on preliminary experiments on the Indonesian validation set, where weighting knockout equally or higher led to overly aggressive pruning and degraded the circuit's standalone competence. Each accepted pruning step is constrained to have at most a 2-point faithfulness drop. The resulting pruned circuit is used for CT-SFT, and all sparse controls are matched to the final post-pruning head budget for each seed.

Table 12 summarizes the circuit discovery, pruning, and CT-SFT configuration.

Component	Setting
Mean set	75 label-balanced Indonesian examples
Discovery set	75 correctly predicted Indonesian examples
Scoring	Task-directional relevance
Circuit construction	Iterative backward expansion from output decision
Maximum discovery depth	Depths 0–2, corresponding to three backward expansion rounds
Loaded circuit depth for CT-SFT	Pruned cumulative circuit from depths 0–2
Selection ratio	$p = 0.10$ per depth before pruning
Pruning objective	$0.7\times$ faithfulness $+0.3\times$ knockout
Faithfulness metric	Accuracy retained when non-circuit heads are mean-ablated
Knockout metric	Accuracy drop when circuit heads are mean-ablated
Pruning order	Lowest absolute CD-T score first
Pruning acceptance	Accept removal if objective improves and faithfulness drop is at most 2 points
Final circuit	Pruned circuit used for CT-SFT
CT-SFT trainable parameters	Selected circuit heads and LayerNorm
Frozen parameters	MLPs, embeddings, unembedding, and unselected heads
Sparse control budget	Matched to final post-pruning circuit size per seed

Table 12: Circuit discovery and CT-SFT setup.

A.5 Effective Updated Parameter Budget

Table 13 reports the effective updated-parameter budget for CT-SFT and the main baselines. For CT-SFT and matched sparse controls, these counts are not obtained by simply checking `requires_grad`, because the implementation applies head-level gradient masking inside the training function. Accordingly, we report the effective update budget implied by the mask: selected attention-head Q/K/V/O slices plus LayerNorm. Full FT updates all model parameters, while LoRA updates only the q/v adapter parameters.

A.6 Faithfulness and Knockout Metrics

We evaluate discovered circuits using label-token intervention accuracy rather than generated accuracy. For each validation example, we append the

Model	Method	Updated components	Eff. updated params	% Full FT
Llama-1B	CT-SFT	62.75 heads avg + LN	29.10M	2.355%
Llama-1B	Matched sparse controls	62.75 heads avg + LN	29.10M	2.355%
Llama-1B	LoRA q/v $r = 8$	q/v all layers	0.852M	0.069%
Llama-1B	Full FT	all parameters	1,235.81M	100%
Llama-3B	CT-SFT	52.50 heads avg + LN	77.44M	2.410%
Llama-3B	Matched sparse controls	52.50 heads avg + LN	77.44M	2.410%
Llama-3B	LoRA q/v $r = 8$	q/v all layers	2.294M	0.071%
Llama-3B	Full FT	all parameters	3,212.75M	100%

Table 13: Effective updated-parameter counts for CT-SFT and baselines. CT-SFT and matched sparse controls update selected attention-head Q/K/V/O slices plus LayerNorm under a head-level gradient mask. LoRA updates q/v adapters only, and Full FT updates all parameters.

gold label y to the prompt x and run a single forward pass. We then read the logits at the position immediately before the gold label token. The model prediction is the highest-logit label among the task-specific label set:

$$\hat{y} = \arg\max_{y' \in \mathcal{Y}} z_{y'}(x),$$

where $\mathcal{Y}$ is the task-specific label set and $z_{y'}(x)$ denotes the next-token logit assigned to label y'. For NusaX, $\mathcal{Y} = \{\texttt{negative}, \texttt{neutral}, \texttt{positive}\}$; for XNLI, $\mathcal{Y} = \{\texttt{entailment}, \texttt{neutral}, \texttt{contradiction}\}$.

Let C be a selected set of attention heads. We use mean ablation at the attention-head output z: ablated heads are replaced by their mean activations computed over a label-balanced source-language mean set. Let f_{full} denote the unablated model, $f_{\text{keep } C}$ denote the intervention where only heads in C remain active and all non-selected heads are mean-ablated, and $f_{\text{ablate } C}$ denote the intervention where heads in C are mean-ablated while all other heads remain active.

Faithfulness is defined as the percentage of full-model label-token accuracy retained when only the selected heads are preserved:

$$\text{Faithfulness}(C) = 100 \times \frac{\text{Acc}(f_{\text{keep } C})}{\text{Acc}(f_{\text{full}})}.$$

Knockout is defined as the relative drop in full-model label-token accuracy when the selected heads are ablated:

$$\text{Knockout}(C) = 100 \times \frac{\text{Acc}(f_{\text{full}}) - \text{Acc}(f_{\text{ablate } C})}{\text{Acc}(f_{\text{full}})}.$$

Both metrics are computed on the validation split. The mean set is sampled from the source-language training split using a label-balanced sampling procedure: Indonesian for NusaX and English for XNLI.

A.7 Zero-Ablation Sensitivity

As a sensitivity check, we repeat the faithfulness and knockout evaluations using zero ablation rather than mean ablation. In this setting, ablated attention-head outputs are replaced with zero vectors rather than source-language mean activations. This intervention is harsher than mean ablation and is less aligned with CD-T, which decomposes activations around a mean/reference activation. Therefore, we use zero ablation only as a robustness check, not as the primary circuit-discovery or pruning criterion.

Model	Method	Avg. Faith.	Avg. Knockout
Llama-1B	Circuit	**56.73**	**34.55**
Llama-1B	Least-rel.	47.20	8.32
Llama-1B	Near-zero	43.94	3.81
Llama-1B	Random NC	54.62	19.30
Llama-3B	Circuit	63.95	**16.04**
Llama-3B	Least-rel.	52.00	9.04
Llama-3B	Near-zero	**69.90**	5.59
Llama-3B	Random NC	52.00	-1.80

Table 14: Aggregate zero-ablation sensitivity results. Faithfulness is retained full-model label-token accuracy when only the selected heads remain active and all other heads are zero-ablated. Knockout is the relative accuracy drop when the selected heads are zero-ablated. Scores are averaged across source and target languages.

Table 14 reports the aggregate zero-ablation results. Circuit remains strongest by average knockout in both Llama-1B and Llama-3B. This supports the conclusion that the selected heads are load-bearing under a second ablation baseline. Faithfulness is more variable under zero ablation: for example, Near-zero has high Llama-3B keep-only faithfulness but much lower knockout than Circuit.

Table 15 gives the full per-language breakdown. The detailed results show the same qualitative pattern as the aggregate table. For Llama-1B, Circuit

Model	Method	Ind F/K	Ace F/K	Bug F/K	Jav F/K	Min F/K	Avg. F/K
Llama-1B	Circuit	**52.48/49.62**	**58.20/34.92**	59.70/**15.32**	53.35/**37.92**	**59.90/34.93**	**56.73/34.55**
Llama-1B	Least-rel.	40.50/12.32	46.98/7.33	50.52/0.52	44.58/16.23	53.42/5.18	47.20/8.32
Llama-1B	Near-zero	42.27/0.70	43.98/7.75	49.48/0.50	40.83/5.83	43.12/4.30	43.94/3.81
Llama-1B	Random NC	43.33/29.55	55.17/18.55	**67.85**/6.15	**54.58**/25.00	52.15/17.23	54.62/19.30
Llama-3B	Circuit	**57.42/30.85**	70.67/**11.43**	59.45/**3.33**	65.50/**16.00**	66.67/**18.62**	63.95/**16.04**
Llama-3B	Least-rel.	38.27/18.38	54.90/3.23	58.35/2.75	55.50/11.50	52.95/9.32	52.00/9.04
Llama-3B	Near-zero	46.50/7.80	**75.00**/1.10	**93.35**/2.75	**67.50**/6.50	**67.12**/9.80	**69.90**/5.59
Llama-3B	Random NC	39.08/-5.85	57.08/-1.65	53.90/0.00	57.00/-3.50	52.95/1.98	52.00/-1.80

Table 15: Per-language zero-ablation sensitivity results. Each cell reports faithfulness/knockout. Faithfulness is the retained full-model label-token accuracy under the keep-only-zero ablation, and knockout is the relative accuracy drop after zero-ablating the selected heads. Bold marks the best value within each model block and column.

has the strongest average faithfulness and knockout, while Random NC is competitive in keep-only faithfulness but has substantially weaker knockout. For Llama-3B, Near-zero obtains the highest keep-only faithfulness on several target languages, but its knockout remains much smaller than Circuit's. This indicates that zero-ablation faithfulness can reflect redundancy or removal of interfering non-selected heads. In contrast, knockout more directly measures whether the selected heads are necessary for the task behavior.

A.8 Evaluation Prompts

All downstream accuracy evaluations are generation-based. Intervention metrics use label-token accuracy, as defined in Appendix A.6. The model generates a short label string using greedy decoding, and the generated output is parsed with the same parser for CT-SFT, sparse controls, Full FT, and LoRA. NusaX transfer and source-language retention use at most three generated tokens, while fixed local evaluations such as IndoNLU EmoT and SIB200 use at most five generated tokens.

NusaX sentiment classification. For controlled perturbation evaluation and for NusaX transfer/source evaluation when using the instruct prompt style, we use:

> Classify the sentiment of the following text.
> Reply with exactly one word only: positive, negative, or neutral.
> Do not explain.
>
> Text: {text}
> Sentiment:

The valid labels are `positive`, `negative`, and `neutral`. In all NusaX evaluations, the generated text is parsed into one of these labels.

IndoNLU EmoT emotion detection. For IndoNLU EmoT, the evaluation uses the prompt stored in the fixed evaluation file. The model is evaluated with the same generation-based parser and the label set:

> Classify the emotion of the following Indonesian text.
> Reply with exactly one label only: anger, fear, happy, love, sadness.
> Do not explain.
>
> Text: {text}
> Emotion:

SIB200 topic classification. For SIB200 topic classification, the evaluation uses the prompt stored in the fixed evaluation file. The valid labels are:

> Classify the topic of the following text.
> Reply with exactly one label only: science/technology, travel, politics, sports, health, entertainment, geography.
> Do not explain.
>
> Text: {text}
> Topic:

Controlled perturbation evaluation. For label-preserving and label-flipping perturbation pairs, we use the NusaX sentiment prompt above for both the original and perturbed sentences. Label-preserving consistency is counted as successful when the prediction remains unchanged across the pair. Label-flipping consistency is counted as successful when the original prediction matches the original label and the perturbed prediction matches the intended flipped label.

A.9 Evaluation Datasets and Examples

This section summarizes the evaluation datasets used in our experiments and provides examples

Evaluation	Dataset	Language(s)	Size	Label space / criterion
Target transfer	NusaX-Senti	Acehnese, Buginese, Javanese, Minangkabau	400 per language	negative, neutral, positive
Source-language retention	NusaX-Senti	Indonesian	400	negative, neutral, positive
Behavior preservation	IndoNLU EmoT	Indonesian	400	anger, fear, happy, love, sadness
Behavior preservation	SIB200	Indonesian	400	topic labels
Controlled perturbation consistency	NusaX-Senti-derived preserving pairs	Indonesian	200 pairs	prediction should remain unchanged; 70 negative, 70 positive, 60 neutral
Controlled perturbation consistency	NusaX-Senti-derived flipping pairs	Indonesian	200 pairs	prediction should flip in the intended direction; 100 negative→positive, 100 positive→negative

Table 16: Evaluation datasets and controlled perturbation sets. All dataset-level evaluations are filtered to a maximum tokenized length of 128 tokens after prompt formatting, including the prompt and label text. We use label-balanced 400-example subsets for dataset-level evaluations, matching the NusaX-Senti test-set size. Controlled perturbation sets are derived from Indonesian NusaX-Senti test examples and evaluated at the pair level.

from the actual evaluation sets. Table 16 gives an overview of all evaluation settings, including the dataset, language, size, and label space or success criterion. We group the evaluations into target transfer, behavior preservation, and controlled perturbation consistency. Target transfer and source-language retention use NusaX-Senti, while IndoNLU EmoT and SIB200 are used to evaluate whether adaptation preserves related and less-related classification behavior. The perturbation sets are derived from Indonesian sentiment examples and are evaluated at the pair level.

Evaluation set construction. All evaluation examples were obtained from the official Hugging Face dataset releases. Before evaluation, we applied the same prompt template used by the model and filtered out examples whose tokenized length exceeded 128 tokens, including the prompt and label text. For each evaluation dataset, we then constructed label-balanced subsets of 400 examples, matching the size of the NusaX-Senti test set used in the main transfer experiments. This filtering and balancing procedure was applied consistently to NusaX-Senti, IndoNLU EmoT, and SIB200.

For controlled perturbation evaluation, we constructed Indonesian sentiment pairs from the NusaX-Senti test set. The label-flipping set contains 200 pairs: 100 negative-to-positive and 100 positive-to-negative. The label-preserving set also contains 200 pairs: 70 negative, 70 positive, and 60 neutral pairs. Label-preserving pairs are used to test prediction invariance, while label-flipping pairs are used to test whether predictions change in the intended sentiment direction.

NusaX-Senti. Table 17 shows a parallel NusaX-Senti example across Indonesian and the four target languages used in our transfer experiments.

Lang.	Example	Label
Indonesian	*Kecewa saya dengan provider ini, sampai kesal, marah, emosi, saya keluar, gara gara, sinyal yang tidak benar.*	negative
Acehnese	*Keucewa lon keu provider nyoe, luwat, beungeh, emosi, lon teubiet, gara-gara sinyal nyang hana pah.*	negative
Buginese	*Masara nyawa ka mitai iyewe provider e, lettu' ka mapella, macai, mappiddara, iya massu, nasaba', sinyal e de na tuju.*	negative
Javanese	*Gela aku karo provider iki, nganti sebel, nesu, emosi, aku metu, gara gara, sinyal sing ora bener.*	negative
Minangkabau	*Kecewa awak samo provider ko, sampai kesal, berang, emosi, awak kalua, dek karano, sinyal nan indak batua.*	negative

Table 17: Actual NusaX-Senti examples from the Indonesian source language and four target languages used for transfer evaluation.

IndoNLU EmoT. Table 18 shows examples from the IndoNLU EmoT evaluation set. This task is used to evaluate related-task retention because emotion detection and sentiment classification both involve affective language understanding. Unlike

NusaX-Senti, however, IndoNLU EmoT requires the model to distinguish finer-grained emotional categories such as anger, fear, happiness, love, and sadness rather than only positive, negative, and neutral sentiment. We therefore use it as a related but non-identical preservation test: a model that adapts to target-language sentiment while preserving IndoNLU EmoT performance is less likely to have damaged its broader affective-language representations.

Example	Label
bikin orang kesal saja	anger
Ya.....Allah..smoga 2019 tdk ada lagi 2 manusia super hero di DPR	fear
gp sudah hitam manis laku pula dari pada cuma ngaku2in pacar orang yeh	happy
Aku suka caramu menyukaiku :')	love
Buka puasa cuman minum sama makan gorengan, trus tengah malam makan bakso cabe rawitnya 3 sendok makan lebih, abis itu perut langsung melilit	sadness

Table 18: Actual examples from the IndoNLU EmoT evaluation set.

SIB200 topic classification. Table 19 shows examples from the Indonesian SIB200 topic classification evaluation set. This task is used to evaluate the preservation of less-related behavior beyond sentiment and affective classification. Unlike NusaX-Senti and IndoNLU EmoT, SIB200 requires the model to assign texts to broad semantic topics such as science/technology, sports, health, politics, and travel. We use it as a stricter preservation test: strong SIB200 retention suggests that target-language sentiment adaptation has not substantially damaged more general Indonesian classification behavior. This helps distinguish methods that preserve only sentiment-adjacent behavior from methods that maintain broader task competence after adaptation.

Controlled perturbation examples. Table 20 shows examples from the Indonesian controlled perturbation sets. Label-preserving pairs keep the original sentiment label unchanged and are counted as successful when the model prediction remains unchanged. Label-flipping pairs intentionally alter sentiment-bearing words and are counted as successful when the original prediction matches the original label and the perturbed prediction matches the intended flipped label.

Example	Label
Ia melakukan pembuatan bel pintu dengan teknologi WiFi, katanya.	science/ technology
Baik golf maupun rugbi akan kembali ke Olimpiade.	sports
Ia kemudian dipindahkan ke Rumah Sakit Addenbrooke di Cambridge.	health
Ini berlawanan dengan laporan sebelumnya, yang menyatakan bahwa membatalkan pemilu ulang berarti menentang konstitusi.	politics
Turkish Airlines terbang ke 39 tujuan di 30 negara Afrika pada 2014.	travel

Table 19: Actual examples from the Indonesian SIB200 topic classification evaluation set.

A.10 LoRA Baseline

We include LoRA as a standard reference for parameter-efficient fine-tuning. Unlike CT-SFT, LoRA does not use circuit information and is not intended as a mechanism-localized intervention. We report it as an additional adaptation baseline, rather than as part of the matched sparse-control comparison.

Training setup. For each target language and training size, LoRA is trained on the same target-language examples used by the other adaptation methods. Specifically, for transfer from Indonesian to a target language L, we train LoRA on $n \in \{25, 50, 75, 100\}$ examples from L, using the same random seeds, data subsampling strategy, prompt format, batch size, number of epochs, and learning rate as the other fine-tuning baselines.

We use the Hugging Face PEFT implementation with LoRA applied to the query and value projection modules, `q_proj` and `v_proj`. The LoRA rank is $r = 8$, the scaling factor is $\alpha = 16$, and dropout is set to 0.05. We train for 5 epochs using AdamW with learning rate 5×10^{-5}, batch size 16, and maximum sequence length 128. During training, the base model parameters are frozen and only the LoRA adapter parameters are updated.

Evaluation protocol. After training, we merge the LoRA adapter into the base model using the PEFT `merge_and_unload()` procedure. The merged model is then evaluated using the same generation-based evaluation pipeline as CT-SFT, Full FT, and the sparse controls. This ensures that LoRA uses the same prompts, decoding procedure, parsing functions, and held-out evaluation examples as the other methods.

We evaluate LoRA on target-language NusaX

Type	Original	Perturbed	Expected
Preserving	*Karena lapar, saya pesan indomie.*	*Karena lapar, aku pesan indomie.*	neu. → neu.
Preserving	*Belanja di Giant diskon 20% dengan kartu kredit bca.*	*Belanja di Giant diskon 20% dengan kartu kredit bni.*	neu. → neu.
Preserving	*Pagi ini lupa lepas helm abang gojek setelah turun dari motor gara-gara mengejar bus*	*Siang ini lupa lepas helm abang maxim setelah turun dari motor gara-gara mengejar travel*	neu. → neu.
Preserving	*Salah satu mal dengan konsep yang cukup unik. Banyak pilihan restoran sehingga cocok untuk menghabiskan waktu bersama teman-teman.*	*Salah satu mal dengan konsep yang cukup unik. Banyak pilihan kafe sehingga cocok untuk menghabiskan waktu bersama sahabat-sahabat.*	pos. → pos.
Preserving	*Parkiran susah bener sampai berputar 3x. Saran: Parkir di dekat hotel saja ,tlalujalan kaki. Restoran di area perumahan.*	*Parkiran susah bener sampai berputar 3x. Saran: Parkir di dekat villa saja terus jalan kaki. Kafe di area perumahan.*	neg. → neg.
Flipping	*Saya tidak kecewa dengan produk apple*	*Saya sangat kecewa dengan produk apple*	pos. → neg.
Flipping	*Harga selangit porsi seiprit tidak janji lagi saya datang untuk yang kedua kali. Masakannya tidak beda jauh sama ema saya dong!*	*Harga bersahabat porsi melimpah, pasti saya datang untuk yang kedua kali. Masakannya jauh lebih enak dari ema saya dong!*	neg. → pos.
Flipping	*Macet di mana-mana kalau lagi liburan*	*Lancar di mana-mana walaupun lagi liburan*	neg. → pos.
Flipping	*Rumah makan nusa indah sekarang mengecewakan ah. Kotor, orang didiamkan saja merokok di ruangan ber ac, makanan lama datangnya.*	*Rumah makan nusa indah sekarang memuaskan ah. Bersih, orang dilarang tegas merokok di ruangan ber ac, makanan sangat cepat datangnya.*	neg. → pos.
Flipping	*Ini adalah sop stengkel terenak di aceh tamiang.*	*Ini adalah sop stengkel terburuk di aceh tamiang.*	pos. → neg.

Table 20: Actual examples from the Indonesian controlled perturbation sets. Label-preserving pairs test prediction invariance, while label-flipping pairs test whether predictions change in the intended sentiment direction.

sentiment accuracy, Indonesian source-language retention, IndoNLU EmoT retention, SIB200 topic retention, and Indonesian controlled perturbation consistency using label-preserving and label-flipping pairs. The perturbation metrics are reported as pair-level success rates. Temporary adapter and merged-model artifacts are deleted after evaluation; only the evaluation outputs and summary files are retained.

Role of the baseline. LoRA provides a practical PEFT reference, but it does not test the same mechanistic hypothesis as CT-SFT. CT-SFT asks whether updating an intervention-supported source-task circuit gives a controlled adaptation target, while LoRA applies low-rank updates to predefined projection modules independent of circuit discovery. Accordingly, LoRA is used to contextualize CT-SFT, while the main analysis focuses on circuit updates and matched non-circuit sparse controls.

A.11 Artifacts and Licenses

We use publicly available research artifacts: pretrained language models, public benchmark datasets, and open-source software libraries. We use these artifacts for research evaluation only and in accordance with their intended research use. Dataset and model artifacts are cited in the main paper, and implementation libraries are listed in Appendix A.

B Additional Llama-1B NusaX Results

This appendix provides additional analyses for the main Llama-3.2-1B NusaX setting. The main paper reports the core transfer, preservation, and intervention results, while this appendix gives more detailed views of the same experimental setting.

We organize the analyses around five complementary questions. First, we examine whether CT-SFT's gains are driven primarily by the selected attention heads or by LayerNorm. Second, we summarize how Circuit and matched sparse controls trade off the six downstream evaluation axes. Third, we show how target-language transfer varies with the number of target-language training examples. Fourth, we compare Circuit with LoRA in more detail to separate mechanism-guided adaptation from a same-data PEFT reference. Finally, we examine fixed-task retention trends on SIB200 Topic and IndoNLU EmoT to show how target-language adaptation affects broader Indonesian behavior across training sizes.

B.1 LayerNorm Ablation Details

The main text reports the aggregate LayerNorm ablation in Table 6. Here, we provide two additional views. First, Table 21 breaks the result down by source and target language. Second, Table 22 compares heads-only no-LN updates across Circuit and matched sparse controls. Together, these results

Method	Ind/Source	Ace	Bug	Jav	Min	Target Avg.
Pre-adapt.	73.50	52.25	39.00	56.00	52.00	49.81
LN-only	**73.84**	**53.83**	39.45	57.28	52.23	50.70
Heads-only no-LN	71.94	50.22	42.75	**63.86**	60.05	54.22
CT-SFT heads+LN	69.55	50.02	**42.86**	63.48	**61.27**	**54.41**

Table 21: Per-language Llama-1B LayerNorm ablation results. Ind/Source is Indonesian source-language retention; Target Avg. is the average of Acehnese, Buginese, Javanese, and Minangkabau target transfer. LN-only updates only LayerNorm/RMSNorm parameters, Heads-only no-LN updates only selected attention heads, and CT-SFT heads+LN is the original CT-SFT setting.

support the interpretation that LayerNorm alone is not the primary driver of CT-SFT's transfer and Topic gains.

The per-language results show the same pattern as the aggregate table. LN-only best preserves Indonesian source performance but provides little target-language adaptation. In contrast, heads-only no-LN and CT-SFT heads+LN produce much larger target-language gains, especially for Buginese, Javanese, and Minangkabau. This indicates that the selected attention heads carry most of the adaptation capacity, while LayerNorm provides an auxiliary stabilizing or complementary update.

When LayerNorm is frozen for all sparse methods, Circuit remains strongest on the balanced controlled-adaptation profile: it gives the best transfer, source retention, Emotion retention, and label-preserving consistency. Near-zero remains strongest on label-flipping consistency, consistent with the plasticity interpretation discussed in the main text. Random NC remains competitive, but unlike Circuit, it is not selected by source-task intervention evidence.

Eval.	Circuit	Least-rel.	Near-zero	Rand. NC
Transfer	**54.22**	52.66	51.48	53.44
Source	**71.94**	71.88	70.13	71.82
Emotion	**41.30**	34.62	40.57	36.71
Topic	67.48	**69.01**	56.50	64.39
Pres. Pair	**64.10**	62.93	62.55	63.09
Flip Pair	85.81	82.45	**86.13**	85.45

Table 22: Aggregate Llama-1B heads-only no-LN sparse-control results. All methods update only attention heads and freeze LayerNorm/RMSNorm. Scores are averaged over four target-language training sizes and four seeds. Higher is better.

B.2 Sparse-Control Rank Summary

Table 23 summarizes the metric-wise ranks of the sparse methods in the main Llama-1B NusaX experiment. We include this table as a descriptive overview of the downstream trade-offs across transfer, source retention, related-task retention, less-related-topic retention, and perturbation consistency. Circuit has the best average rank, but we do not use this summary alone as evidence that circuit targeting is practically necessary. Rather, it complements the intervention results in Table 3 and the preservation analyses in the main text.

Metric	Circuit	Least-rel.	Near-zero	Rand. NC
Transfer	**54.40 (1)**	53.20 (3)	50.50 (4)	53.60 (2)
Source	69.50 (3)	69.70 (2)	65.50 (4)	**70.30 (1)**
Emotion	**40.20 (1)**	32.90 (4)	39.40 (2)	36.10 (3)
Topic	**69.30 (1)**	69.20 (2)	58.70 (4)	64.30 (3)
Pres. Pair	64.70 (2)	64.10 (3)	62.00 (4)	**64.90 (1)**
Flip Pair	79.50 (2)	73.80 (4)	**81.90 (1)**	79.30 (3)
Avg. rank	**1.67**	3.00	3.17	2.17

Table 23: Metric-wise sparse-control ranks on the main Llama-1B NusaX setting. Numbers in parentheses indicate rank among sparse methods for each metric. Average rank is reported only as a descriptive summary and is not used as the primary evidence for practical necessity.

B.3 Target Transfer Trends by Training Size

Figure 4 shows target-language transfer accuracy across the four low-resource training sizes, $n \in \{25, 50, 75, 100\}$, for each target language. This complements the aggregate transfer scores in Table 4 by showing whether each method's behavior is stable across data sizes rather than driven by a single training budget.

Overall, Circuit remains competitive across training sizes and target languages, while non-circuit sparse controls can also improve with additional target-language data. The trends support the interpretation in RQ2: target transfer can arise both from updating the discovered source-task circuit and from recruiting non-circuit capacity. However, because transfer accuracy alone does not distinguish controlled editing from uncontrolled recruitment, these curves should be interpreted alongside the retention, perturbation, and intervention results

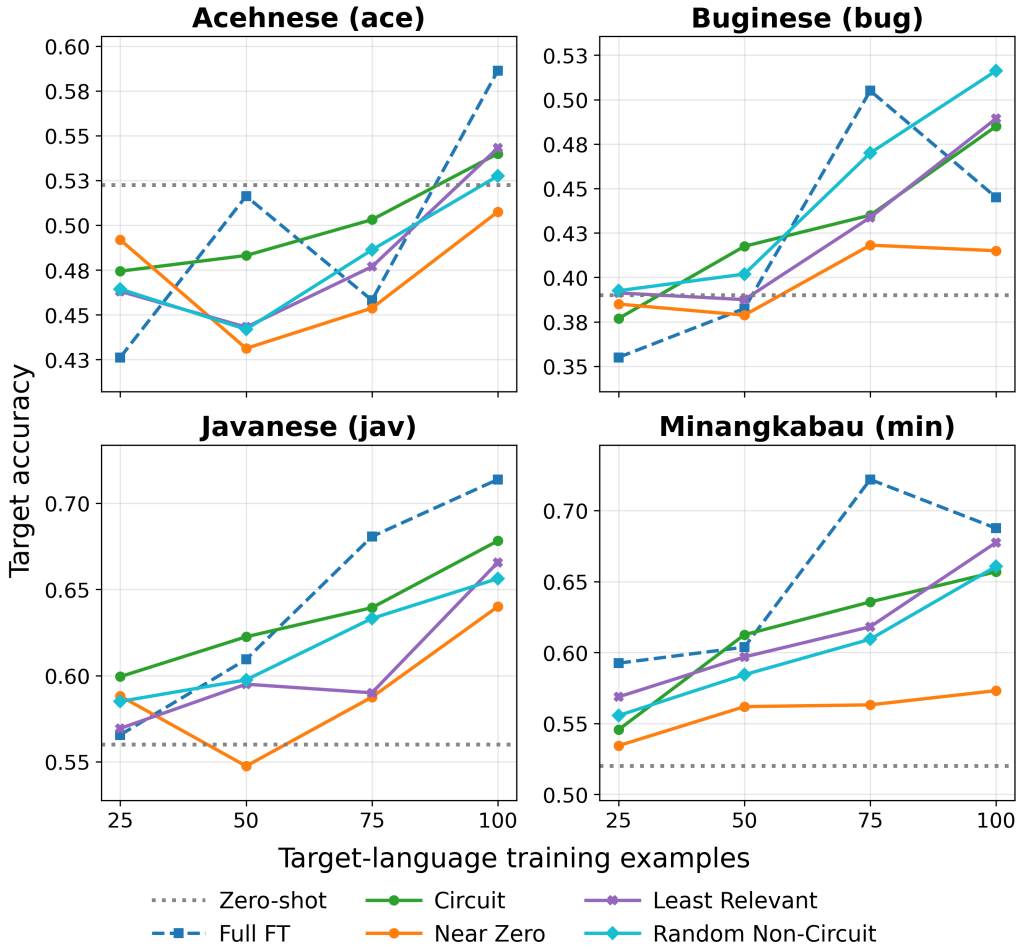


Figure 4: Target-language transfer across training sizes for Llama-3.2-1B-Instruct. Each panel corresponds to one target language, with zero-shot and Full FT included as reference baselines.

reported in the main text.

B.4 Circuit and LoRA Details

Table 24 provides the per-target-language breakdown for Circuit and LoRA in the main Llama-1B NusaX setting. This complements the aggregate reference-baseline comparison in Table 5 by showing whether the aggregate trends are consistent across Acehnese, Buginese, Javanese, and Minangkabau. LoRA is included as a same-data and same-protocol PEFT reference: it uses the same target-language examples, training sizes, random seeds, epochs, learning rate, batch size, prompt format, decoding setup, parsing function, and held-out evaluation pipeline as the other adaptation methods. However, unlike CT-SFT, LoRA does not use circuit information or select update locations based on source-task intervention evidence.

The per-language results show a consistent trade-off between sentiment-adjacent preservation and broader controlled adaptation. LoRA preserves Indonesian source performance, IndoNLU Emotion retention, and label-flipping consistency better than Circuit. This suggests that LoRA is a strong preservation-oriented baseline for behavior that closely aligns with the original sentiment objective. In particular, its stronger label-flipping consistency indicates that low-rank updates can remain responsive to explicit sentiment-changing cues. However, this advantage does not extend uniformly to all preservation axes.

Circuit is stronger on average target transfer, SIB200 Topic retention, and label-preserving perturbation consistency. The Topic gap is especially consistent across all target languages, indicating that LoRA's preservation is narrower than Circuit's broader controlled-adaptation profile. This is important because SIB200 is less directly tied to sentiment than IndoNLU Emotion, so Topic retention provides a stricter test of whether adaptation disrupts broader classification behavior. Similarly, stronger label-preserving consistency suggests that

Method	Evaluation	Ace	Bug	Jav	Min	Avg.
Circuit	Transfer	50.00	**42.90**	**63.50**	**61.30**	**54.40**
	Source	69.20	68.40	70.30	70.30	69.50
	Emotion	40.10	39.10	**41.40**	40.00	40.20
	Topic	**68.40**	**69.40**	**69.60**	**69.60**	**69.30**
	Pres. Pair	**64.30**	**63.80**	**65.70**	**65.00**	**64.70**
	Flip Pair	79.90	78.40	79.90	79.70	79.50
LoRA	Transfer	**53.28 (+3.28)**	39.20 (-3.70)	56.58 (-6.92)	52.22 (-9.08)	50.32 (-4.08)
	Source	**72.95 (+3.75)**	**73.03 (+4.63)**	**72.97 (+2.67)**	**72.98 (+2.68)**	**72.98 (+3.48)**
	Emotion	**41.06 (+0.96)**	**40.95 (+1.85)**	41.20 (-0.20)	**41.08 (+1.08)**	**41.07 (+0.87)**
	Topic	48.27 (-20.13)	47.81 (-21.59)	49.25 (-20.35)	48.72 (-20.88)	48.51 (-20.79)
	Pres. Pair	63.06 (-1.24)	62.94 (-0.86)	63.03 (-2.67)	63.12 (-1.88)	63.04 (-1.66)
	Flip Pair	**82.38 (+2.48)**	**82.47 (+4.07)**	**82.62 (+2.72)**	**82.69 (+2.99)**	**82.54 (+3.04)**

Table 24: Circuit and LoRA results by target language, averaged across training sizes and seeds. Parentheses indicate the signed percentage-point difference between LoRA and Circuit. Bold marks the better score between Circuit and LoRA for each evaluation and language.

Circuit better maintains prediction invariance under meaning-preserving perturbations, rather than only responding flexibly to explicit sentiment flips.

These results support the interpretation in the main text: LoRA is a strong preservation-oriented PEFT reference, especially for behavior close to the sentiment objective, but it does not provide the same broader preservation profile as Circuit. Circuit is not uniformly better on every metric, and LoRA is clearly competitive on source retention, Emotion retention, and label-flipping consistency. However, Circuit better preserves less-related Topic behavior and label-preserving consistency while remaining competitive for target transfer. Thus, the difference between Circuit and LoRA is not simply about target-language accuracy; it reflects different preservation profiles under the same data and training protocol.

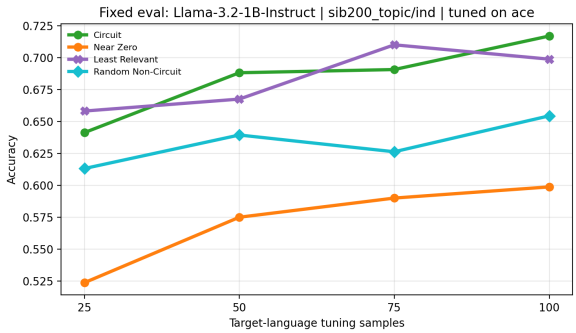
This distinction also clarifies the role of LoRA in our evaluation. LoRA is a useful practical baseline, but it does not test the same mechanistic hypothesis as CT-SFT. CT-SFT asks whether a source-task circuit identified by intervention evidence can serve as a controlled update target. LoRA instead applies low-rank updates to predefined projection modules independent of circuit discovery. The per-language comparison therefore supports the main claim that mechanism-guided updates offer a different kind of control: not universal dominance on every score, but a more interpretable tradeoff between transfer, broad behavior preservation, and perturbation consistency.

B.5 Fixed-Task Retention by Target Language and Training Size

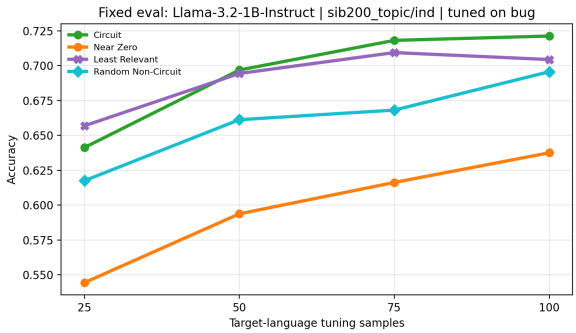
Figures 5 and 6 show fixed-task retention across target-language training sizes for the main Llama-1B NusaX setting. These plots complement the aggregate preservation results in Table 4 by showing whether retention behavior is stable across $n \in \{25, 50, 75, 100\}$, rather than being driven by a single target-data budget.

For each target language, the model is adapted to that target language and then evaluated on fixed Indonesian retention tasks. SIB200 Topic retention tests broader, less-related classification behavior, while IndoNLU EmoT tests affective behavior that is closer to sentiment classification. We omit Indonesian tuning results because the main transfer setting adapts from Indonesian to local target languages, and the goal here is to measure how target-language adaptation affects fixed Indonesian behavior.

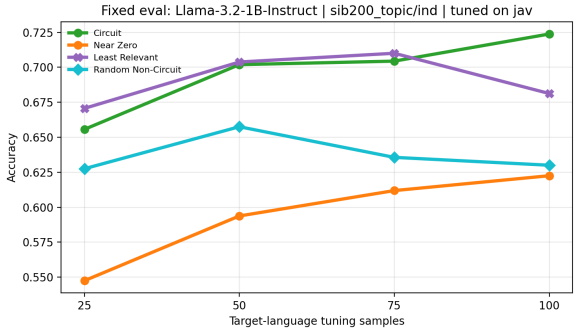
The trends show that retention behavior differs substantially across evaluation axes. Topic retention is the stricter preservation test: methods that are competitive for target transfer can still diverge strongly in their ability to preserve less-related topic classification. Emotion retention is more sentiment-adjacent, so it is generally less diagnostic of broad behavioral preservation than SIB200 Topic. These plots therefore support the main text's argument that controlled adaptation should be evaluated beyond target-language accuracy alone.



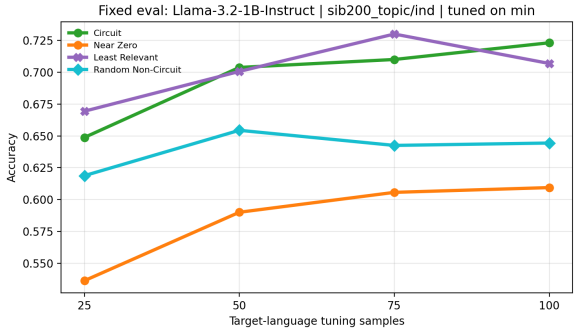
(a) Tuned on Acehnese



(b) Tuned on Buginese

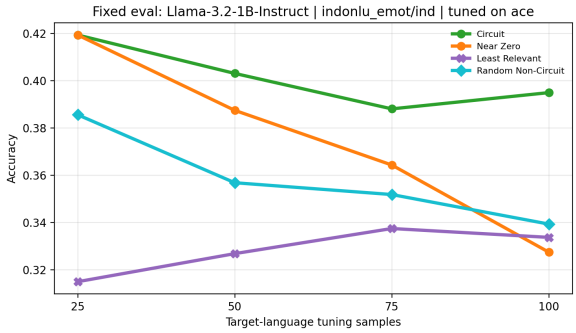


(c) Tuned on Javanese

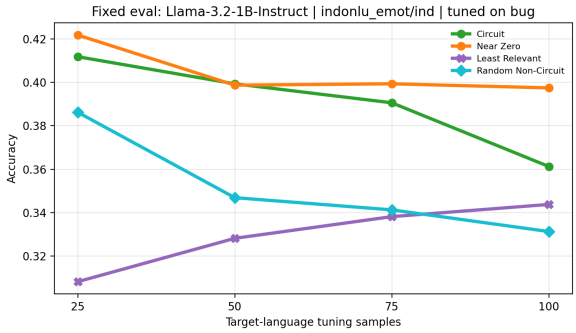


(d) Tuned on Minangkabau

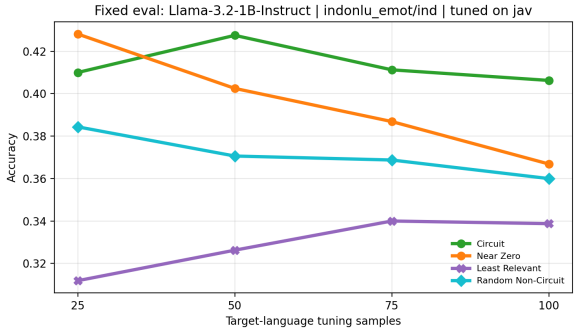
Figure 5: SIB200 topic retention by target-language training size. Each panel shows Indonesian SIB200 topic accuracy after adapting the model to a single target language. Indonesian tuning results are omitted.



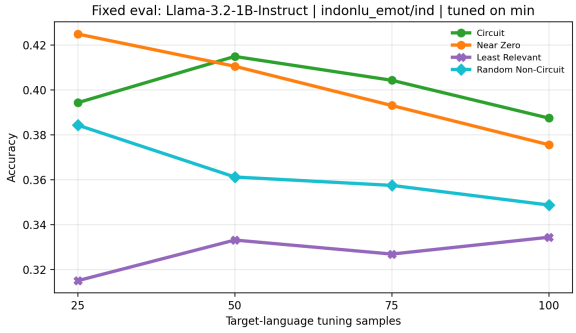
(a) Tuned on Acehnese



(b) Tuned on Buginese



(c) Tuned on Javanese



(d) Tuned on Minangkabau

Figure 6: IndoNLU EmoT retention by target-language training size. Each panel shows Indonesian emotion-detection accuracy after adapting the model to one target language. Indonesian tuning results are omitted.

C Additional Scalability Results

This appendix presents additional scalability and robustness results beyond those in the main Llama-3.2-1B NusaX setting. We include two kinds of evidence. First, we report detailed Llama-3.2-3B-Instruct results to examine whether the controlled-adaptation patterns observed at 1B also appear at a larger model scale. Second, we report XNLI results for Qwen2.5-0.5B and Llama-3.2-1B to test whether the intervention and source-retention patterns extend beyond NusaX sentiment classification.

The Llama-3B results should be read as a scalability check rather than a complete retuning of all baselines. As in the main experiments, sparse controls are matched to the post-pruning Circuit head budget, and Full FT and LoRA are same-data/same-protocol reference baselines. The XNLI experiments are narrower: they evaluate cross-lingual transfer, English source retention, and intervention behavior after English competence tuning, but they do not replicate the full NusaX multi-axis preservation suite.

The results show that sparse methods remain competitive at 3B scale, but the preservation profile differs by update location.

C.1 Llama-3B Sparse-Control Results

Table 25 reports the full per-language sparse-control comparison for Llama-3.2-3B-Instruct. This table complements the aggregate scalability results in Table 7 by showing whether the average trends are consistent across Acehnese, Buginese, Javanese, and Minangkabau. The comparison includes target-language transfer, Indonesian source retention, IndoNLU Emotion retention, and SIB200 Topic retention.

The results show that sparse methods remain competitive at 3B scale, but the preservation profile differs by update location. Least-relevant gives the highest average transfer, and Random NC is competitive on source and Topic retention. However, Near-zero suffers a large Topic-retention drop, and Least-relevant also loses substantial Topic retention relative to Circuit. Thus, the 3B results support the main text's conclusion that sparse updates are not interchangeable: matching the update budget alone does not guarantee the same preservation behavior.

Method	Evaluation	Pre-adapt.	Ace	Bug	Jav	Min	Avg.
Zero-shot	Transfer	–	29.00	19.00	36.00	36.50	30.12
Circuit	Transfer	–	55.98	44.03	68.31	66.38	58.68
	Source	56.00	**77.27**	76.02	77.31	77.36	76.99
	Emotion	34.50	**36.34**	**35.33**	**38.09**	**37.58**	**36.84**
	Topic	20.00	68.78	66.64	69.09	69.64	68.54
Least-relevant	Transfer	–	**57.19 (+1.20)**	**44.86 (+0.83)**	69.08 (+0.77)	**67.53 (+1.16)**	**59.66 (+0.99)**
	Source	56.00	76.86 (-0.41)	**76.16 (+0.14)**	76.70 (-0.61)	77.08 (-0.28)	76.70 (-0.29)
	Emotion	34.50	34.94 (-1.41)	33.39 (-1.94)	35.80 (-2.30)	35.50 (-2.08)	34.91 (-1.93)
	Topic	20.00	57.09 (-11.69)	57.23 (-9.41)	58.45 (-10.64)	59.66 (-9.98)	58.11 (-10.43)
Near-zero	Transfer	–	54.95 (-1.03)	44.27 (+0.23)	66.78 (-1.53)	66.05 (-0.33)	58.01 (-0.66)
	Source	56.00	76.09 (-1.17)	75.61 (-0.41)	76.72 (-0.59)	76.56 (-0.80)	76.25 (-0.74)
	Emotion	34.50	30.61 (-5.73)	30.22 (-5.11)	31.42 (-6.67)	31.28 (-6.30)	30.88 (-5.95)
	Topic	20.00	45.70 (-23.08)	43.59 (-23.05)	45.97 (-23.12)	46.11 (-23.53)	45.34 (-23.20)
Random NC	Transfer	–	55.55 (-0.44)	42.72 (-1.31)	**69.28 (+0.97)**	67.25 (+0.88)	58.70 (+0.02)
	Source	56.00	77.09 (-0.17)	76.14 (+0.12)	**77.38 (+0.06)**	**77.81 (+0.45)**	**77.11 (+0.12)**
	Emotion	34.50	33.12 (-3.22)	32.22 (-3.11)	34.31 (-3.78)	34.39 (-3.19)	33.51 (-3.32)
	Topic	20.00	**70.31 (+1.53)**	**69.39 (+2.75)**	**70.30 (+1.20)**	**71.19 (+1.55)**	**70.30 (+1.76)**

Table 25: Transfer and behavior preservation after target-language adaptation for Llama-3.2-3B-Instruct. The Pre-adapt. column reports zero-shot reference performance before adaptation for behavior-preservation evaluations. For Transfer, the zero-shot target-language accuracies are shown in the first row. Adapted scores are percentages averaged across training sizes and seeds; parentheses show differences from Circuit. Bold marks the best score per column, and gray shading marks the lowest score. Random NC denotes random non-circuit updates. Emotion denotes IndoNLU emotion retention, and Topic denotes SIB200 topic-classification retention.

C.2 Llama-3B Reference Baselines

Table 26 compares Circuit with Full FT and LoRA for Llama-3.2-3B-Instruct. These baselines are included as same-data and same-protocol references: they use the same target-language examples, training sizes, seeds, optimizer settings, prompt format, and evaluation pipeline as Circuit. They are not intended to represent exhaustively tuned method-specific optima.

The reference-baseline comparison shows a clear tradeoff. Full FT yields the highest average target transfer, but sharply degrades broader preservation metrics, especially IndoNLU Emotion and SIB200 Topic. LoRA preserves Emotion better than Circuit, but it substantially underperforms on Topic retention and is weaker on average transfer. Circuit therefore provides a more balanced scalability profile: it is not the best method on every metric, but it better preserves broader Indonesian classification behavior while remaining competitive for target-language adaptation.

Method	Evaluation	Ace	Bug	Jav	Min	Avg.
Circuit	Transfer	55.98	**44.03**	**68.31**	66.38	58.68
	Source	**77.27**	**76.02**	**77.31**	77.36	**76.99**
	Emotion	36.34	35.33	38.09	37.58	36.84
	Topic	**68.78**	**66.64**	**69.09**	**69.64**	**68.54**
Full FT	Transfer	**60.20 (+4.22)**	43.66 (-0.38)	67.31 (-1.00)	**69.16 (+2.78)**	**60.08 (+1.41)**
	Source	72.14 (-5.12)	57.41 (-18.61)	76.36 (-0.95)	**79.91 (+2.55)**	71.45 (-5.54)
	Emotion	0.67 (-35.67)	0.56 (-34.77)	1.98 (-36.11)	0.83 (-36.75)	1.01 (-35.82)
	Topic	4.12 (-64.66)	7.73 (-58.91)	1.34 (-67.75)	6.77 (-62.88)	4.99 (-63.55)
LoRA	Transfer	50.78 (-5.20)	42.81 (-1.22)	62.03 (-6.28)	64.95 (-1.42)	55.14 (-3.53)
	Source	73.25 (-4.02)	72.69 (-3.33)	73.17 (-4.14)	73.11 (-4.25)	73.05 (-3.93)
	Emotion	**38.17 (+1.83)**	**37.69 (+2.36)**	**38.52 (+0.42)**	**38.28 (+0.70)**	**38.16 (+1.33)**
	Topic	32.48 (-36.30)	30.38 (-36.27)	33.92 (-35.17)	32.80 (-36.84)	32.39 (-36.14)

Table 26: Llama-3.2-3B-Instruct comparison against reference baselines after target-language adaptation. Scores are percentages averaged across training sizes and seeds; parentheses show differences from Circuit. Bold marks the best score per column, and gray shading marks the lowest score. Emotion denotes IndoNLU emotion retention, and Topic denotes SIB200 topic-classification retention.

C.3 XNLI Intervention Robustness

Table 27 reports XNLI intervention results for Qwen2.5-0.5B and Llama-3.2-1B. Unlike the NusaX experiments, XNLI requires an English competence-tuning stage before circuit discovery and target-language adaptation. We therefore use XNLI as a narrower robustness check for intervention behavior rather than as a full replication of the NusaX controlled-adaptation suite.

Method	Faithfulness (% full acc.)					Knockout (% acc. drop)				
	En	Es	Th	Vi	Zh	En	Es	Th	Vi	Zh
Qwen2.5-0.5B, $n_{\text{circuit}} = 250$										
Circuit	**52.48**	**57.14**	67.04	**57.68**	52.19	**43.74**	**40.04**	**29.91**	**42.68**	**40.45**
Least-relevant	44.49	55.75	68.95	51.09	**53.40**	9.60	1.22	9.70	9.64	14.73
Near-zero	45.21	56.84	67.17	51.52	50.24	19.87	11.34	8.02	13.63	21.15
Random NC	44.74	53.01	**69.35**	52.03	50.44	33.42	29.42	27.79	32.15	29.03
Llama-3.2-1B, $n_{\text{circuit}} = 250$										
Circuit	**78.00**	67.99	69.07	67.64	**69.21**	**50.13**	**33.14**	23.48	**36.89**	**37.42**
Least-relevant	71.24	**73.61**	**70.51**	71.20	66.55	22.52	28.72	25.80	21.84	27.24
Near-zero	72.01	72.03	66.37	71.36	68.35	12.68	23.76	18.21	18.76	31.26
Random NC	68.33	69.90	64.68	**72.87**	63.93	21.38	28.37	**31.37**	20.69	32.92

Table 27: XNLI intervention evaluation with $n_{\text{circuit}} = 250$. Scores are validation label-token accuracies averaged across four seeds. Faithfulness is the percentage of full-model accuracy retained when only selected heads are preserved; knockout is the relative accuracy drop after selected-head ablation. Random NC is a matched, non-circuit-sparse control.

The intervention results show that Circuit generally has strong knockout effects across model families,

especially compared with Least-relevant and Near-zero controls. This means that ablating the selected Circuit heads produces a larger relative accuracy drop than ablating most matched sparse-control heads. Faithfulness is more mixed, which is expected because keep-only interventions are stricter and can be sensitive to distributed redundancy. Together, the XNLI intervention results support the claim that the selected heads remain informative intervention targets beyond the main NusaX setting, while also showing that faithfulness and knockout capture different aspects of circuit behavior.

C.4 XNLI Transfer and Source Retention

Tables 28 and 29 report XNLI cross-lingual transfer and English source retention after target-language adaptation. The target languages are Thai, Vietnamese, Spanish, and Chinese, and all scores are averaged across training sizes and seeds. Parenthesized values show the signed difference from Circuit for the same target language and evaluation axis.

These results should be interpreted differently from the NusaX preservation results. For XNLI, the goal is to check whether circuit-guided sparse updates preserve English competence while adapting to new target languages. We do not evaluate the broader NusaX-style preservation suite here, because the English competence-tuning stage would make it difficult to separate the effects of competence tuning from the effects of CT-SFT. The XNLI tables therefore provide evidence about transfer and source retention, not about broad behavioral preservation across unrelated Indonesian tasks.

Method	Eval	th	vi	es	zh	Avg.
Comp.-tuned ckpt (En retention ceiling: 74.38)		50.81	64.25	65.88	69.31	62.56
Circuit	Transfer	57.59	63.16	65.25	68.06	63.52
	Source (En)	70.25	71.81	72.14	73.77	71.99
Least-relevant	Transfer	59.05 (+1.46)	63.31 (+0.15)	66.89 (+1.64)	68.14 (+0.08)	64.35 (+0.83)
	Source (En)	73.02 (+2.77)	**73.80** (+1.99)	74.34 (+2.20)	73.69 (-0.08)	**73.71** (+1.72)
Near-zero	Transfer	**59.30** (+1.71)	63.41 (+0.25)	**67.28** (+2.03)	68.39 (+0.33)	**64.59** (+1.07)
	Source (En)	**73.30** (+3.05)	73.08 (+1.27)	73.80 (+1.66)	**74.05** (+0.28)	73.55 (+1.56)
Random NC	Transfer	58.64 (+1.05)	**63.92** (+0.76)	67.00 (+1.75)	**68.61** (+0.55)	64.54 (+1.02)
	Source (En)	70.14 (-0.11)	72.30 (+0.49)	**74.39** (+2.25)	73.58 (-0.19)	72.60 (+0.61)
Full FT	Transfer	36.70 (-20.89)	38.09 (-25.07)	42.70 (-22.55)	40.44 (-27.62)	39.48 (-24.04)
	Source (En)	40.31 (-29.94)	39.12 (-32.69)	45.36 (-26.78)	43.11 (-30.66)	41.98 (-30.01)
LoRA	Transfer	58.75 (+1.16)	62.58 (-0.58)	64.30 (-0.95)	66.75 (-1.31)	63.09 (-0.43)
	Source (En)	69.50 (-0.75)	70.77 (-1.04)	73.27 (+1.13)	73.00 (-0.77)	71.63 (-0.36)

Table 28: XNLI cross-lingual transfer and English retention for Qwen2.5-0.5B with 250-sample English competence tuning. Scores are percentages averaged over four seeds and $n \in \{25, 50, 75, 100\}$. Parentheses show signed differences from Circuit for the same column. **Bold** marks the best score per column among matched sparse methods.

Table 28 shows the Qwen2.5-0.5B results. We next report the corresponding Llama-3.2-1B results under the same XNLI protocol. Overall, the XNLI results support the robustness of the intervention evidence while also showing that Circuit is not uniformly the highest-transfer sparse method. Matched sparse controls can sometimes slightly outperform Circuit on transfer or source retention, especially when the evaluation is limited to the original XNLI label space. However, Circuit remains competitive while being tied to source-task intervention evidence, which is the main distinction from unguided sparse controls.

Method	Eval	th	vi	es	zh	Avg.
Comp.-tuned ckpt (En retention ceiling: 47.00)		41.69	45.31	45.50	45.06	44.39
Circuit	Transfer	**45.75**	**45.53**	48.89	46.87	46.76
	Source (En)	**45.67**	**48.33**	47.19	46.97	**47.04**
Least-relevant	Transfer	44.47 (-1.28)	44.95 (-0.58)	50.27 (+1.38)	**47.14** (+0.27)	46.71 (-0.05)
	Source (En)	41.36 (-4.31)	46.59 (-1.74)	46.64 (-0.55)	46.73 (-0.24)	45.33 (-1.71)
Near-zero	Transfer	44.17 (-1.58)	44.48 (-1.05)	**50.41** (+1.52)	46.92 (+0.05)	46.50 (-0.26)
	Source (En)	41.38 (-4.29)	46.91 (-1.42)	47.62 (+0.43)	46.86 (-0.11)	45.69 (-1.35)
Random NC	Transfer	45.45 (-0.30)	45.47 (-0.06)	50.03 (+1.14)	47.00 (+0.13)	**46.99** (+0.23)
	Source (En)	43.03 (-2.64)	47.53 (-0.80)	**48.00** (+0.81)	**47.17** (+0.20)	46.43 (-0.61)
Full FT	Transfer	33.42 (-12.33)	34.12 (-11.41)	33.91 (-14.98)	34.73 (-12.14)	34.05 (-12.71)
	Source (En)	34.02 (-11.65)	34.05 (-14.28)	33.84 (-13.35)	34.30 (-12.67)	34.05 (-12.99)
LoRA	Transfer	42.95 (-2.80)	45.69 (+0.16)	50.03 (+1.14)	46.70 (-0.17)	46.34 (-0.42)
	Source (En)	43.61 (-2.06)	47.92 (-0.41)	47.28 (+0.09)	46.92 (-0.05)	46.43 (-0.61)

Table 29: XNLI cross-lingual transfer and English retention for Llama-3.2-1B with 250-sample English competence tuning. Scores are percentages averaged over four seeds and $n \in \{25, 50, 75, 100\}$. Parentheses show signed differences from Circuit for the same column. **Bold** marks the best score per column among matched sparse methods.