Discursive Circuits: How Do Language Models Understand Discourse Relations?

Anonymous ACL submission

Abstract

Which components in transformer language models are responsible for discourse understanding? We hypothesize that sparse computational graphs, termed as *discursive circuits* control how models process discourse relations. Unlike simpler tasks, discourse relations involve longer spans and complex reasoning. To make circuit discovery feasible, we introduce a task called Completion under Discourse Relation (CuDR), where a model completes a discourse given a specified relation. To support this task, we construct a corpus of minimal contrastive pairs tailored for activation patching in circuit discovery. Experiments show that sparse circuits ($\approx 0.2\%$ of a full GPT-2 model) recover discourse understanding in the English PDTB-based CuDR task.

These circuits generalize well to unseen discourse frameworks such as RST and SDRT. Further analysis shows lower layers capture linguistic features such as lexical semantics and coreference, while upper layers encode discourse-level abstractions. Feature utility is consistent across frameworks (e.g., coreference supports Expansion-like relations).

1 Introduction

Discourse structure is essential for ensuring language models (LMs) to behave safely and ethically (Kim et al., 2025; Nakshatri et al., 2025). Yet, little is known about how discourse is internally processed by LMs, limiting our ability to guarantee that they are reliable and free from harmful outputs. Transformer circuit discovery (Zhang and Nanda, 2024) is a promising method that identifies sparse computational subgraphs causally responsible for specific behaviors. Unlike attention visualization (Jain and Wallace, 2019) or rationale generation (Wiegreffe and Marasovic, 2021), circuits provide mechanistic, intervention-based explanations that reveal which components causally drive the model's output. Existing circuit discovery methods focus on simple tasks, like numeric comparison (Hanna et al., 2023) which is well-suited for next-word prediction (e.g. "The year after 1731 is →"). In contrast, discourse relation involves longer contexts and more complex reasoning, making direct adaptation of existing methods infeasible.

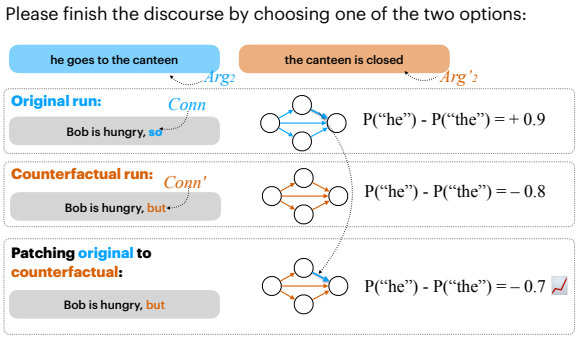

Figure 1: **Task Overview:** The CuDR task enables discovery of discursive circuits by contrasting model predictions under minimal changes to the discourse connectives. Activation patching reveals components causally responsible for shifting the model's prediction.

We contribute a key insight that bridging the linguistic structure of discourse and the requirements of circuit discovery, which offers a new path for mechanistic understanding of complex language tasks. On the discourse side, we hold the initial argument Arg_1 (e.g. "Bob is hungry", Figure 1) unchanged and introduce a counterfactual connective $Conn'$ (e.g., "but") that prompts the model to select an alternative continuation Arg_2' ("the canteen is closed"), which is only coherent under the counterfactual discourse relation. On the circuit discovery side, the method relies on minimal contrastive pairs, where inputs differ slightly but yield significantly different outputs. To identify influential model components, we patch activations (Nanda, 2023) from the original run into the counterfactual run and observe changes in prediction. The resulting discursive circuits are composed of

connections with significant causal influence.

To support this task, we construct a dataset spanning major discourse frameworks, including Penn Discourse Treebank (PDTB; Webber et al.,2019), Rhetorical Structure Theory (RST; Mann and Thompson,1987), and Segmented Discourse Representation Theory (SDRT; Asher and Lascarides,2003). Each instance contains an original annotation from the source corpus, along with a set of counterfactual connectives and their alternative completions. The three frameworks have 10 to 17 distinct discourse relations each, and together contribute a total of 27,754 instances.

Using our datasets, we discover discursive circuits in the GPT-2 medium model. For most discourse relations, the identified circuits achieve over around 90% faithfulness while involving only 0.2% of model connections. We show that circuits derived from PDTB generalize well to unseen discourse frameworks such as RST and SDRT, suggesting that language models may encode a shared representation of discourse relations. We also construct a novel circuit hierarchy adapted from PDTB's three-level taxonomy. To our knowledge, this is the first discourse hierarchy grounded in neural circuit components. Together, our circuits and hierarchy provide a new form of discourse representation, enabling direct cross-framework comparison and fine-grained decomposition into linguistic features. We discover similar utilities across different frameworks (e.g., coreference is prominent in all Expansion-like relations).

2 Circuit Discovery with CuDR

We propose a generic workflow to dissect a language model's discourse understanding via circuit discovery, which is compatible with any discourse framework. We introduce the Completion under Discourse Relation task (CuDR, pronounced "kooder"), where Arg_1 remains fixed, while the connective is swapped ($Conn \rightarrow Conn'$), requiring the model to shift its prediction from Arg_2 to Arg_2'.

2.1 Completion under Discourse Relation

CuDR creates a controlled environment to test a model's discursive behavior. By simply altering the discourse connective (from original (ori) to counterfactual (CF); Table 1), the model's continuation shifts sharply in response. For example, in the original discourse, a Contingency relation is expressed with the connective "so", leading to a completion that "he goes to the canteen". However, when the discourse relation is shifted to a counterfactual Comparison relation (signaled by "but"), the model should sharply change its prediction to an argument that negates the expectation of eating (i.e., "the canteen was closed"). Note that circuit discovery has been applied under various settings (Zhang and Nanda, 2024), we adopt such a setup to steer the model, because it captures the dynamic nature of discourse understanding.

Concretely, the original discourse consists of two arguments, Arg_1 and Arg_2, linked by a discourse relation R and connective $Conn$, formally denoted as $d_{\text{ori}} = (Arg_1, Conn, Arg_2, R)$. The counterfactual instance, $d_{\text{cf}} = (Arg_1, Conn', Arg_2', R')$, preserves Arg_1 but substitutes the continuation and relation ($R' \neq R$), forming a minimal contrastive pair required by activation patching.

Input: $d_{ori} = (Arg_1, Arg_2, R, Conn)$ $d_{cf} = (Arg_1, Arg_2', R', Conn')$
CuDR Task (Original): `Please finish the discourse by choosing one of the two options:` Arg_2, Arg_2' `To complete:` $Arg_1, Conn$
Correct answer: Arg_2, **Incorrect answer:** Arg_2'
Example: `Please finish the discourse by choosing one of the two options:` "he goes to the canteen" or "the canteen is closed" `To complete:` [Bob is hungry]$_{Arg1}$ [so]$_{Conn}$ ⇒ [he goes to the canteen]$_{Arg2}$
CuDR Task (Counterfactual): `Please finish the discourse by choosing one of the two options:` Arg_2 or Arg_2' `To complete:` $Arg_1, Conn'$
Correct answer: Arg_2', **Incorrect answer:** Arg_2
Example: `Please finish the discourse by choosing one of the two options:` "he goes to the canteen" or "the canteen is closed" `To complete:` [Bob is hungry]$_{Arg1}$ [but]$_{Conn'}$ ⇒ [the canteen is closed]$_{Arg2'}$

Table 1: **Formalization of the CuDR task:** the model must complete the discourse by either Arg_2 or the counterfactual Arg_2', based on which best fits as a continuation of Arg_1 following Conn or Conn' (best in color).

2.2 Circuit Discovery

Activation Patching. Transformer circuits are computational graphs that model the information flow from an input token, through residual flow among intermediate nodes (i.e., MLP layers and attention heads) to the output probability of the next token. To identify influential connections inside the circuits, we intervene in the model by replacing the activation of a counterfactual (corrupted) run

by the activation of a original (clean) run.

$$g(e) = L(x_{cf}|do(E = e_{ori})) - L(x_{cf}) \quad (1)$$

Concretely, we define the impact of introducing an intervening edge e (denoted by $g(e)$) as the difference in a metric L when patching the activation of edge e from the original run. Formally, $g(e)$ is computed as the difference between $L(x_{cf}|do(E = e_{ori}))$ where e is restored to its clean value, and $L(x_{cf})$, the metric value under the corrupted run.

Accelerate by Attribution Patching. To overcome the low speed for activation patching (Conmy et al., 2023), we adopt a first order Taylor approximation to Equation 1 and use the Edge Attribution Patching (EAP) method Nanda (2023); Syed et al. (2024). For an edge $e = (u, v)$, the change of metric $g(e)$ is:

$$g(e) \approx (z_u^{ori} - z_u^{cf})^\top \nabla_v L(x_{cf}), \quad (2)$$

where z_u^{ori} and z_u^{cf} denote the activation at node u in the original or counterfactual runs, and $\nabla_v L(x_{cf})$ is the gradient of metric L at node v. With the approximation, we can now calculate $g(e)$ for all edges by two forward passes and one backward pass, greatly enhancing efficiency (10^3 times faster in our practice).

Attribution Patching Using CuDR. We first input the model with the counterfactual (CF) input, and the model produces a CF output. Using the same CF input, we then perform activation patching from the original (Ori) to restore the model's prediction to the Ori output. In the CF run, the model receives x_{cf}, constructed from Arg_1 and a counterfactual discourse connective ($Conn'$). The correct prediction is the counterfactual completion (Arg_2'). In the ori run, the model receives x_{ori} as input, which consists of ($Arg_1, Conn$). The correct output is the original Arg_2. Attribution patching (Figure 2) works by replacing activations from the CF run with those from the Ori run. For example, to evaluate the edge between MLP 20 and Attention Head 21.9 (Attn. 21.9), we replace the activation flowing from MLP 20 into Attn. 21.9 with the corresponding activation from the Ori run and observe $g(e)$, which is the change in the model's output.

Construct Discursive Circuits. The discursive circuit for a given discourse relation is constructed by applying attribution patching to the CuDR task

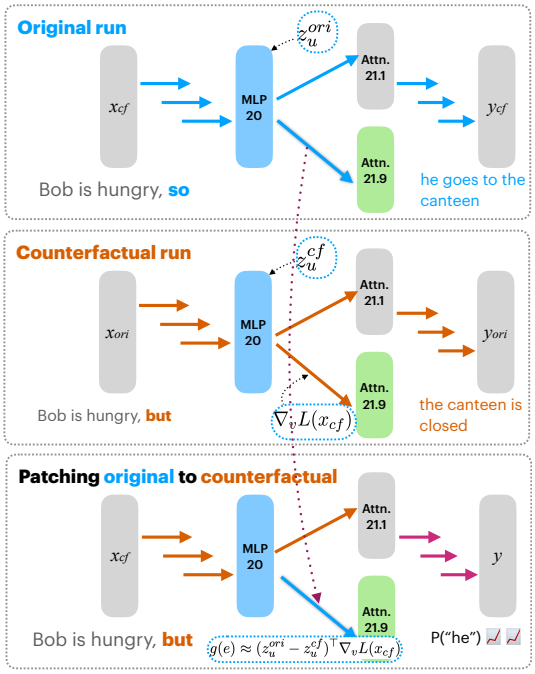

Figure 2: **Illustration of attribution patching with CuDR:** We steer the model's prediction from the counterfactual toward the original outcome. Activations from the original run are patched into the counterfactual run to influence the model's prediction.

over a set of samples for that relation. We compute the average $g(e)$ for each edge and select those with the highest absolute $g(e)$ values as the most important. In practice, the top 1000 such edges are sufficient to steer the model faithfully, similar to prior work (Hanna et al., 2024).

2.3 The CuDR Dataset

We construct an augmented dataset by prompting a large language model (LLM) with the original Arg_1 and a counterfactual $Conn'$, along with detailed instructions and discourse relation definitions (Appendix A.3). We employ GPT-4o-mini for its good instruction-following ability and lower cost.

Building on the taxonomy of counterfactual discourse relations proposed by Miao et al. (2024), our CuDR dataset adopts a PDTB3-based design (Table 2). For each discourse relation alongside its original connective, we construct five counterfactual discourse connectives. For example, the Comparison.Concession.Arg2-as-denier relation (e.g., "however", Row 1 in Table 2) is considered counterfactual to both a Contingency relation (signaled by "because") and an Instantiation relation ("for example"). We provide a complete list of connectives and their mappings in Appendix A.1.

We extend our dataset construction beyond

Discourse Relation	Ori Connective	CF Connective
Comparison.Concession.Arg2-as-denier	however	because for example
Comparison.Contrast	by comparison	specifically in other words
Contingency.Reason	because	so however
Contingency.Result	so	because by comparison
Expansion.Conjunction	and	however so
Expansion.Equivalence	in other words	however for example
Expansion.Instantiation.Arg2-as-instance	for example	because however
Expansion.Level-of-detail.Arg1-as-detail	in short	however so
Expansion.Level-of-detail.Arg2-as-detail	specifically	instead by comparison
Expansion.Substitution.Arg2-as-subst	instead	because in other words
Temporal.Asynchronous.Precedence	then	however previously
Temporal.Asynchronous.Succession	previously	so then
Temporal.Synchronous	while	so then

Table 2: **CuDR Dataset:** PDTB's discourse relations with corresponding original (Ori) connectives and counterfactual (CF) connectives (subset displayed for CF).

Discourse framework	# of DR	# of CuDR data
PDTB	13	11,843
GDTB	12	5,253
GUM-RST	17	6,805
SDRT	10	3,853
Total		27,754

Table 3: **CuDR Dataset Statistics:** Number of unique discourse relations and CuDR data across frameworks.

PDTB to include additional corpora: the GUM Discourse Treebank (GDTB; Liu et al. 2024b), a more up-to-date PDTB-style corpus, as well as GUM-RST (Zeldes, 2017) and SDRT (Asher and Lascarides, 2003). To enable the generation of counterfactual instances from non-PDTB corpora, we construct relation mappings from RST to PDTB (Table 7) and from SDRT to PDTB (Table 8 in Appendix A). For example, SDRT's Explanation relation is mapped to PDTB's Contingency.Cause.Reason, then its corresponding counterfactual relations Result ("so") and Contrast ("however"), are found in the PDTB-based taxonomy.

Table 3 summarizes the metadata per discourse framework. Each original and counterfactual discourse pair, $(d_{\text{ori}}, d_{\text{cf}})$, is treated as a single data instance in the CuDR dataset. For each discourse relation in each corpus, we sample up to 50 original instances. With five counterfactual connectives per relation, this yields up to 300 CuDR instances per relation. We discard minority relations with fewer than 20 instances, as well as low-quality instances where Arg_2 and Arg_2' are overly similar. We consider 300 instances per relation sufficient, as Yao et al. (2024) use a median of only 52. To validate the automated constructions, one author manually verified 40 CuDR samples and found them all valid as an indicative evaluation, with Arg_2' coherent with Arg_1 and $Conn'$. The language in Arg_2' tends to be straightforward, but it is desired because we want salient relations (Appendix A.3). Preliminary trials with open-source Llama-3.1-8B-Instruct (Grattafiori et al., 2024) to generate CuDR data was unsuccessful as it did not follow our task instruction.

3 Evaluate Discursive Circuits

We conduct our evaluation to answer following research questions (RQs):
RQ1: Do discursive circuits faithfully recover the full model's performance?
RQ2: Do discursive circuits generalize across different discourse frameworks and relation types?
RQ3: Are discursive circuits composed of components associated with specific linguistic features?

Implementation Detail. Following Hanna et al. (2024); Mondorf et al. (2025), we focus on a single model for in-depth analysis and adopt their choice of GPT-2 medium (Radford et al., 2019) for its manageable memory requirements. To identify circuits for specific discourse relations, we use a sample size of 32 for both circuit discovery and validation, and apply the standard practice of using the batch mean for node value patching (Miller et al., 2024). We repeat each experiment five times with different data samples and average the outcomes for stability. Before circuit discovery, we fine-tune the model on held-out CuDR data (half of the PDTB subset) to align it with our task setting and ensure it follows the intended instructions (Appendix B.1).
Baseline Circuits: We replicate the **Indirect Object Identification (IOI)** circuit (Wang et al., 2023) in our own model as a baseline circuit. In the IOI task, the model is given a prompt like "John and Mary went to a bar. Mary gave a beer to", and should predict "John". This circuit represents the model's general next-word prediction ability, without discourse-specific reasoning. Comparing against IOI allows us to test whether discursive circuits capture discourse-specific computation beyond standard language modeling.

Evaluation Metric. Our metric follows Miller et al. (2024) to calculate the **logit difference** between the correct and incorrect answers. Specif-

ically, we treat the original discourse's Arg_2 as correct and the counterfactual Arg_2' as incorrect, and compute $\Delta L = L(Arg_2) - L(Arg_2')$, where $L(\cdot)$ denotes the logit of the corresponding answer.

Normalized faithfulness: Since different discourse relations yield different raw scores, we report normalized faithfulness scores (Miller et al., 2024), which quantify the percentage of the full model's performance that a sparse circuit restores. Concretely, we compute $\frac{\Delta L_{\text{patch}}}{\Delta L_{\text{full}}}$, where ΔL_{patch} is the logit difference obtained by patching clean activations into a corrupted input, and ΔL_{full} is the logit-difference of the full model on clean input. In our CuDR task, faithfulness begins at a large negative value (since the unpatched model selects Arg_2'), increases as clean edges are patched, and reaches 100% when the full model is restored (which predicts Arg_2).

L1	L2	L3
Comparison (566)	Concession ✗	Arg2-as-denier
	/	Contrast
Contingency (564)	/	Reason
	/	Result
Expansion (200)	/	Conjunction
	/	Equivalence
	Instantiation ✗	Arg2-as-instance
	Level-of-detail (565) ✓	Arg1-as-detail
		Arg2-as-detail
	Substitution ✗	Arg2-as-subst
Temporal (405)	Asynchronous (575) ✓	Precedence
		Succession
	/	Synchronous

Table 4: **Discursive Circuits Hierarchy (L1–L3):** All "leaf node" relations are classified as L3. Only two circuits appear at the L2 level, each merging more than one L3 circuit. (Numbers) indicate edge counts. L3 circuit has 1,000 edges, and L0 circuit has 137 edges.

Hierarchical Discursive Circuits. With the learned circuits, we construct a new PDTB-style circuit hierarchy. To the best of our knowledge, this is the first discourse hierarchy derived from neural components. We first learn circuits for all 13 Level-3 (L3) relations and use the top 1,000 edges to merge them to form higher-level circuits. That is, L3 ∋ L2 ∋ L1 ∋ L0 (Table 4). Note that our circuit hierarchy differs from the PDTB taxonomy in two ways: (1) All "leaf node" relations are treated as L3 since they have no children to merge (e.g., Temporal.Synchronous) and circuit discovery operates on the finest-grain level; (2) Some L2 relations are removed (e.g., Concession ✗) as they contain only one valid L3 relation due to data scarcity, so merging would be meaningless. In the end, L2 circuits

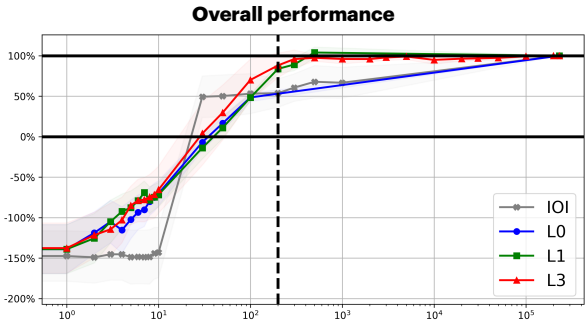

Figure 3: **RQ1: Overall Faithfulness of Discursive Circuits:** We report average faithfulness across 13 PDTB relations for circuits L3, L1, L0, and the IOI baseline. The Y-axis shows faithfulness (%), and the X-axis shows the number of patched edges (log scale). Shaded areas indicate standard deviation. L3 and L1 reach strong faithfulness at ≈ 200 edges (vertical dashed line).

contain over 500 edges, L1 circuits have 200–500+ edges, and the meta L0 circuit contains 137 edges.

3.1 Discursive Circuits are Faithful (RQ1)

We first validate the faithfulness of discursive circuits on the PDTB dataset. The average performance across 13 discourse relations (Figure 6) shows strong overall effectiveness. We omit L2 as it covers only a subset of relations. For both L3 and L1 circuits, strong faithfulness (≈ 90%) is achieved with only ≈ 200 edges. L3 outperforms L1 in the 10–200 edge range, likely due to its ability to capture more fine-grained information. Both L3 and L1 surpass L0 and IOI after 100 edges. This gap is likely due to L0's small size (137 edges). Even though IOI reasons over next objects, it still lacks of discourse skills, as it plateaus quickly around ≈ 50% faithfulness, showing the unique skills needed for discourse competence.

We then analyze the performance breakdown by relation types (Figure 4) and make the following observations: **(1) Finer-grained circuits are more effective than coarser ones.** There is a consistent trend across relation types: L3 > L2 ≈ L1 > L0 > IOI. However, fine-grained circuits also show greater variance (large red shades). L1 is more stable and has a lower variance. In practice, we recommend L1 as a balanced choice: while slightly less effective in only stages, it matches L3 after ≈ 300 edges and works for all lower-level relations. **(2) L2 does not necessarily outperform L1.** This is evident in the four relations that have L2 circuits, including Expansion.Details (8th and 9th subfigures in Figure 4, compared with Expansion L1's circuit) and Temporal.Asynchronous (12th and 13th,

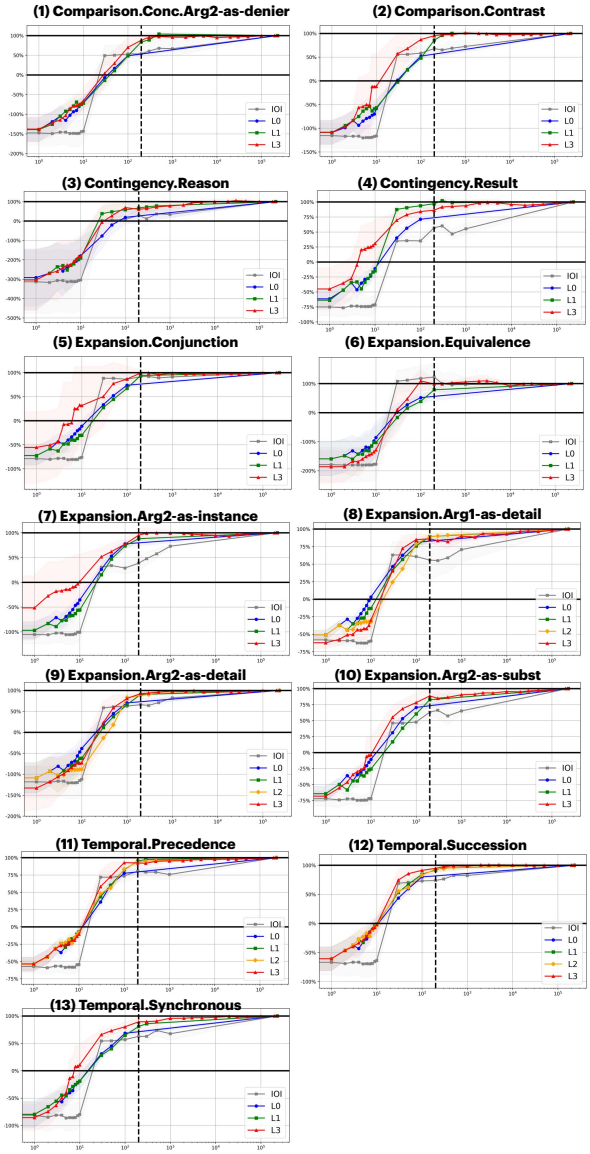

Figure 4: **RQ1** Faithfulness of Discursive Circuits by Discourse Relation (see indices 1–13).

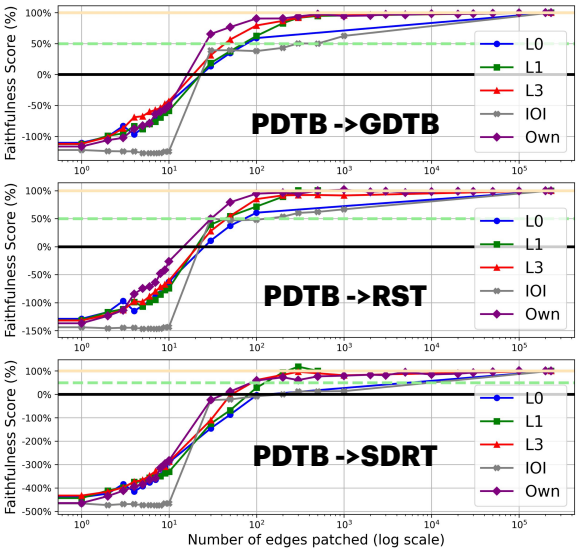

Figure 5: **RQ2 Cross-dataset generalization:** Performance by applying PDTB's circuits to other datasets.

compared with Temporal L1 circuit). This suggests that L2 and L1 operate at a similar level of abstraction, with comparable degrees of information loss. **(3) Discursive circuits reflect task difficulty.** Two Contingency relations (3rd and 4th) are exceptions where L1 matches or outperforms L3. Further inspection shows that these relations have lower absolute faithfulness scores, suggesting the model struggles with them. In such case, L3 may overfit, while L1 retains core patterns and generalizes better. IOI generally underperforms due to its lack of discourse specificity. However, in Conjunction (5th) and Equivalence (6th), it performs comparably or better than discursive circuits, suggesting these relations are easier to model. In contrast, larger gaps in Comparison (1st–2nd) and Contingency (3rd–4th) indicate greater complexity.

3.2 Discursive Circuits Generalize to New Datasets and New Relations (RQ2)

Do discursive circuits generalize across different discourse frameworks? We extend the CuDR task to other frameworks by applying circuits obtained from PDTB to GDTB (same framework, different genre), as well as to RST and SDRT (different frameworks). We follow the same mapping (Appendix A.2) for cross-framework transfer; for example, Explanation (SDRT) is mapped to Contingency.Cause.Reason (PDTB). Figure 5 shows the generalization performance, with each line representing the average performance across all relations in the dataset. **PDTB circuits generalize well to other datasets.** We set an "upper bound" using the Own circuits (learned via CuDR task in-dataset, e.g. SDRT's Explanation). PDTB's L3 circuits close the gap with Own using only $\approx$ 200 edges, despite initially lagging due to dataset-specific features. Across the three generalization targets, the trend is consistent: Own > L3 > L1 $\approx$ L0 > IOI. L1 and L0 are weaker in the first 100 edges, likely because both abstractions lose fine-grained information (L2 is skipped due to limited coverage). SDRT is the hardest to generalize to, with only 50% faithfulness after 100 patched edges, highlighting the gap between the datasets.

Do circuits learned for one discourse relation generalize to others? We study all 13 PDTB L3 relations by applying each circuit to the other 12, using the top 200 edges per circuit (enough for strong

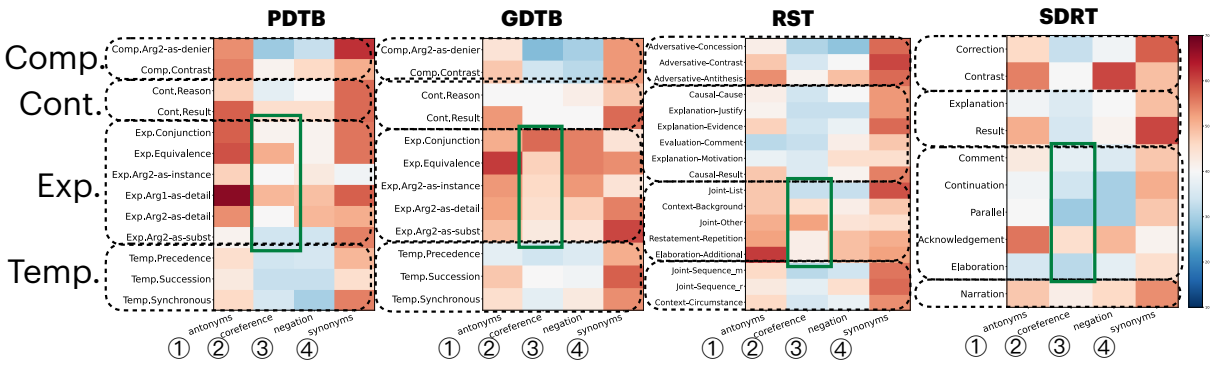

Figure 6: **RQ3** Overlap of discursive circuits with circuits for linguistic features: antonymy, synonymy, coreference, and negation. Similar pattern is shared across frameworks, (e.g. coreference signal in Expansion relations).

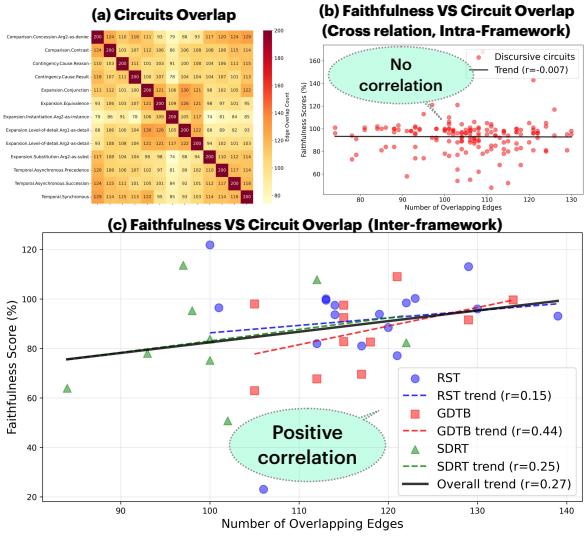

Figure 7: **RQ2 Cross-relation Generalization:** (a) The overlap among PDTB's relation circuits; (b) Intra-framework generalization in PDTB; (c) Inter-framework generalization from PDTB.

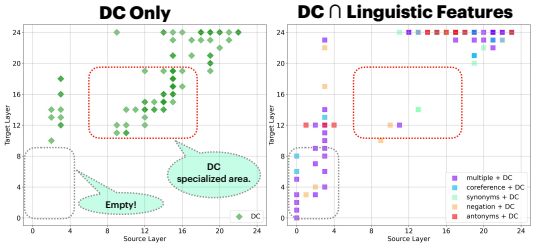

Figure 8: **RQ3 Layer-wise Edge Analysis:** Source (X-axis) and target (Y-axis) layers of edges in discursive and linguistic circuits. DC-only edges emerge independently in higher layers and are absent in lower layers.

faithfulness): **(1)** Figure 7a shows the edge overlap among these circuits. While the diagonals are darker, indicating greater overlap between similar relations, the overall overlap remains consistently high (80–120 out of 200 edges). **(2)** Figure 7b shows no correlation between overlap and faithfulness ($r = -0.007$). This is counterintuitive, as one might expect more overlap to imply better generalization. The narrow overlap range (80–120) likely limits the variation. Recently, Hanna et al. (2024) also reports faithfulness does not necessarily require high overlap. **(3)** Cross-framework results (Figure 7c) reveal a positive correlation between overlap and performance, e.g., PDTB → GDTB yields $r = 0.44$. In summary, higher circuit overlap *does not* imply better intra-framework faithfulness, but *does* support inter-framework transfer.

3.3 Discursive Circuits Overlap with Linguistic Features' Circuits (RQ3)

Are discursive circuits composed of sub-circuits linked to linguistic features? Inspired by the eRST and RST Signaling Corpus (Zeldes et al., 2025; Das and Taboada, 2018), we discover circuits for four key features, ① antonymy, ② coreference, ③ negation, and ④ synonymy, as a preliminary and non-exhaustive study, using similar activation prompts (Appendix B.3). Figure 6a shows that **the utility of linguistic features per L1 relation is consistent across datasets**. Lexical features ① antonymy and ④ synonymy are broadly used in all relations, which is consistent across frameworks. ② coreference is most active in Expansion relations (the ② is darkest in Expansion-like rows, highlighted by the **green boxes**), where continuity relies on entity reference. SDRT shows less reliance on coreference, likely due to shorter texts. This suggests that LMs encode discourse relations with similar linguistic cues across frameworks.

Figure 8 shows the layer-wise distribution of discursive circuits (DC) and linguistic circuits by source and target node layers (Top 200 edges). DC-only edges are absent in lower layers (noted as

"empty"). A distinct region (source: 8–16, target: 10–20) contains DC-only edges, with very limited overlap with linguistic features. This suggests lower layers in discursive circuits capture shared linguistic features, while discursive abstraction emerges in higher layers.

Error case 1: [yay!!!!!]$_{Arg_1}$, (because) [I don't care who wins now]$_{Arg_2}$
Error case 2: [I'll give clay in return]$_{Arg_1}$, (because) [think clay is in abundance this game]$_{Arg_2}$
PDTB's missing edges: Resid Start→MLP0, A19.9→A21.1, MLP3→MLP7, MLP7→MLP11

Table 5: **Case Study:** PDTB circuit ✗; SDRT circuit ✓

We further examine the cases where SDRT's Own circuits succeed but PDTB's L3 circuits (both using the first 30 edges). Table 5 shows a subset of representative errors. Case 1 involves an interjection ("yay!"), and Case 2 features an ellipsis of the subject "I" in Arg_2, both are rare phenomena in PDTB. Our method pinpoints missing elements in PDTB that SDRT captures, such as early edges (Resid Start→MLP 0, aiding connective reasoning) and late edges (e.g., 19.9→21.1, shared only with the coreference feature among the four features).

4 Related Works

Discourse Modeling and Evaluation. Discourse modeling has been studied under three major frameworks: PDTB (Webber et al., 2019; Prasad et al., 2008), RST (Mann and Thompson, 1987; Zeldes, 2017; Zeldes et al., 2025), and SDRT (Asher and Lascarides, 2003). Recent studies seek to unify these frameworks, with advances in discourse relation prediction (Zhao et al., 2023; Wu et al., 2023a; Anuranjana, 2023; Chan et al., 2023; Rong and Mo, 2024; Liu and Strube, 2023; Long et al., 2024), discourse parsing (Li et al., 2024a,b; Thompson et al., 2024; Pastor et al., 2025; Liu et al., 2025), and annotation (Yung et al., 2024; Pyatkin et al., 2023; Ruby et al., 2025; Saeed et al., 2025). Fu (2022) outline early plans for unification, and the DISRPT benchmark (Braud et al., 2024) enables cross-framework evaluation with data annotated under all three schemes. Liu et al. (2024b) propose automatic RST-to-PDTB transformation via sense mapping. Liu and Zeldes (2023); Eichin et al. (2025) examine generalization across domains and languages. While linguistically insightful, these approaches overlook mechanistic interpretability.

Question answering has also been explored as a bridge across frameworks. Fu (2025) link Questions Under Discussion (QUD) (Wu et al., 2023b; Ko et al., 2023) to PDTB, RST, and SDRT. Miao et al. (2024) propose a QA-based evaluation, though their prompts offer limited insight into model internals. LLMs have been used to synthesize discourse data (Yung et al., 2025; Cai, 2025), mainly to augment low-resource relations (Omura et al., 2024). In contrast, our CUDR dataset targets interpretability rather than data expansion.

Mechanistic Interpretability. Unlike visualizations (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019) or textual explanations (Lyu et al., 2024; Zhu et al., 2024), mechanistic interpretability identifies components in a model that drive predictions. Circuits, as global computation graphs, can be identified through activation patching (Conmy et al., 2023; Miller et al., 2024; Syed et al., 2024). We do not adopt sparse autoencoders (SAEs) (Huben et al., 2024; Makelov et al., 2024), as our goal is to understand discourse processing at a global model rather than isolate local activity. Circuit discovery has mostly been applied to simplistic tasks, such as indirect object identification (IOI) (Wang et al., 2023), numerical comparison (Hanna et al., 2023), subject-verb agreement (SVA) (Ferrando and Costa-jussà, 2024), MCQ (Lieberum et al., 2023), knowledge acquisition (Yao et al., 2024; Ou et al., 2025; Hanna et al., 2024), colored objects (Merullo et al., 2024), and context-free grammars (Mondorf et al., 2025). No existing work addresses complex discourse phenomena.

5 Conclusion and Future Work

In this work, we introduce discursive circuits, the first mechanistic interpretation of how discourse understanding is realized within language models. To make circuit discovery feasible, we propose a novel CUDR task that enables activation patching, along with a collection of CUDR datasets for PDTB, RST, and SDRT discourse frameworks. Our identified discursive circuits are shown to be faithful in restoring the full model's performance and exhibit strong cross-framework generalization. Discursive circuits provide a new lens for mechanistically representing discourse, enabling the construction of a circuit hierarchy that supports direct comparison of discourse relations both within and across frameworks. We already observe promising evidence of shared linguistic features utility across them. In future work, we aim to extend CUDR to multiple languages and adapt it for a broader range of tasks.

Limitations

Our work also has the following limitations: (1) We only study English-based corpora. It would be promising to extend circuit discovery to multiple languages and explore whether a unified circuit space exists across different languages, similar to the universal discourse label set explored by Eichin et al. (2025). This is feasible, as we can construct the CUDR dataset for other languages as well. (2) We follow Hanna et al. (2023, 2024); Mondorf et al. (2025) in focusing on one single transformer-based language model to enable more in-depth analysis. While it would be interesting to extend our method to other model architectures such as multi-layer perceptrons (MLPs) (Fusco et al., 2023) or LSTMs (Sundermeyer et al., 2012), we limit our scope to transformers due to their predominant use today and because activation patching is not directly compatible with MLPs or LSTMs. (3) We do not compare discourse processing in language models with that in the human brain (Case and Oetama-Paul, 2015; Perfetti and Frishkoff, 2008). For example, Eviatar and Just (2006) report that discourse processing triggers specific brain activations observable via fMRI. While intriguing, this is beyond the scope of our study.

Ethical Statement and Potential Risks

Our research on discourse relations does not pose direct ethical risks. However, as with all mechanistic interpretability studies, the identified circuits could be used to influence model behavior in specific capacities, such as modifying numerical reasoning (Hanna et al., 2023) or, in our case, discourse processing and generation. By making the model's reasoning about discourse relations more transparent, our work has the potential to aid in detecting and mitigating biases in scenarios where discourse structure plays a role.

Declaration of AI Tool Usage

We used AI tools at the following stages of this research: (1) GPT-4o-mini (via API) was used to generate the counterfactual instances for our CUDR dataset; prompt details are provided in Appendix A; (2) Cursor AI was used during coding, primarily for debugging assistance; (3) ChatGPT-4o (via web interface) was employed only for grammatical checking of the manuscript. All research ideas, analyses, and findings were developed and written independently by the authors.

References

Kaveri Anuranjana. 2023. DiscoFlan: Instruction fine-tuning and refined text generation for discourse relation label classification. In *Proceedings of the 3rd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2023)*, pages 22–28, Toronto, Canada. The Association for Computational Linguistics.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.

Xinyi Cai. 2025. Fine-grained evaluation for implicit discourse relation recognition. *Preprint*, arXiv:2503.05326.

Susan S Case and Angela J Oetama-Paul. 2015. Brain biology and gendered discourse. *Applied Psychology*, 64(2):338–378.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Wong, and Simon See. 2023. DiscoPrompt: Path prediction prompt tuning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 35–57, Toronto, Canada. Association for Computational Linguistics.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Debopam Das and Maite Taboada. 2018. Rst signalling corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52(1):149–184.

Florian Eichin, Yang Janet Liu, Barbara Plank, and Michael A. Hedderich. 2025. Probing llms for multilingual discourse generalization through a unified label set. *Preprint*, arXiv:2503.10515.

Zohar Eviatar and Marcel Adam Just. 2006. Brain correlates of discourse processing: An fmri investigation of irony and conventional metaphor comprehension. *Neuropsychologia*, 44(12):2348–2359.

Javier Ferrando and Marta R. Costa-jussà. 2024. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125, Miami, Florida, USA. Association for Computational Linguistics.

Yingxue Fu. 2022. Towards unification of discourse annotation frameworks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 132–142, Dublin, Ireland. Association for Computational Linguistics.

Yingxue Fu. 2025. A survey of QUD models for discourse processing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1722–1732, Albuquerque, New Mexico. Association for Computational Linguistics.

Francesco Fusco, Damian Pascual, Peter Staar, and Diego Antognini. 2023. pNLP-mixer: an efficient all-MLP architecture for language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 53–60, Toronto, Canada. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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*.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Zae Myung Kim, Anand Ramachandran, Farideh Tavazoee, Joo-Kyung Kim, Oleg Rokhlenko, and Dongyeop Kang. 2025. Align to structure: Aligning large language models with structural information. *Preprint*, arXiv:2504.03622.

Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srinivas, Greg Durrett, and Junyi Jessy Li. 2023. Discourse analysis via questions and answers: Parsing dependency structures of questions under discussion. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11181–11195, Toronto, Canada. Association for Computational Linguistics.

Chuyuan Li, Chloé Braud, Maxime Amblard, and Giuseppe Carenini. 2024a. Discourse relation prediction and discourse parsing in dialogues with minimal supervision. In *Proceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024)*, pages 161–176, St. Julians, Malta. Association for Computational Linguistics.

Chuyuan Li, Yuwei Yin, and Giuseppe Carenini. 2024b. Dialogue discourse parsing as generation: A sequence-to-sequence LLM-based approach. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–14, Kyoto, Japan. Association for Computational Linguistics.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *Preprint*, arXiv:2307.09458.

Shannan Liu, Peifeng Li, Yaxin Fan, and Qiaoming Zhu. 2025. Enhancing multi-party dialogue discourse parsing with explanation generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1531–1544, Abu Dhabi, UAE. Association for Computational Linguistics.

Wei Liu and Michael Strube. 2023. Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024a. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *The Twelfth International Conference on Learning Representations*.

Yang Janet Liu, Tatsuya Aoyama, Wesley Scivetti, Yilun Zhu, Shabnam Behzad, Lauren Elizabeth Levine, Jessica Lin, Devika Tiwari, and Amir Zeldes. 2024b. GDTB: Genre diverse data for English shallow discourse parsing across modalities, text types, and domains. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12287–12303, Miami, Florida, USA. Association for Computational Linguistics.

Yang Janet Liu and Amir Zeldes. 2023. Why can't discourse parsing generalize? a thorough investigation of the impact of data diversity. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3112–3130, Dubrovnik, Croatia. Association for Computational Linguistics.

Wanqiu Long, Siddharth N, and Bonnie Webber. 2024. Multi-label classification for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8437–8451, Bangkok, Thailand. Association for Computational Linguistics.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Aleksandar Makelov, Georg Lange, and Neel Nanda. 2024. Towards principled evaluations of sparse autoencoders for interpretability and control. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

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*.

Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy Chen, and Min-Yen Kan. 2024. Discursive socratic questioning: Evaluating the faithfulness of language models' understanding of discourse relations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6277–6295, Bangkok, Thailand. Association for Computational Linguistics.

Joseph Miller, Bilal Chughtai, and William Saunders. 2024. Transformer circuit evaluation metrics are not robust. In *First Conference on Language Modeling*.

Philipp Mondorf, Sondre Wold, and Barbara Plank. 2025. Circuit compositions: Exploring modular structures in transformer-based language models. *Preprint*, arXiv:2410.01434.

Nishanth Nakshatri, Nikhil Mehta, Siyi Liu, Sihao Chen, Daniel J. Hopkins, Dan Roth, and Dan Goldwasser. 2025. Talking point based ideological discourse analysis in news events. *Preprint*, arXiv:2504.07400.

Neel Nanda. 2023. Attribution patching: Activation patching at industrial scale. Accessed: 2025-04-12.

Kazumasa Omura, Fei Cheng, and Sadao Kurohashi. 2024. An empirical study of synthetic data generation for implicit discourse relation recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1073–1085, Torino, Italia. ELRA and ICCL.

Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun Chen. 2025. How do llms acquire new knowledge? a knowledge circuits perspective on continual pre-training. *Preprint*, arXiv:2502.11196.

Martial Pastor, Nelleke Oostdijk, Patricia Martin-Rodilla, and Javier Parapar. 2025. Enhancing discourse parsing for local structures from social media with LLM-generated data. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8739–8748, Abu Dhabi, UAE. Association for Computational Linguistics.

Charles A Perfetti and Gwen A Frishkoff. 2008. The neural bases of text and discourse processing. *Handbook of the neuroscience of language*, 2:165–174.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Valentina Pyatkin, Frances Yung, Merel C. J. Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. Design choices for crowdsourcing implicit discourse relations: Revealing the biases introduced by task design. *Transactions of the Association for Computational Linguistics*, 11:1014–1032.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Yuetong Rong and Yijun Mo. 2024. NCPrompt: NSP-based prompt learning and contrastive learning for implicit discourse relation recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1159–1169, Miami, Florida, USA. Association for Computational Linguistics.

Ahmed Ruby, Christian Hardmeier, and Sara Stymne. 2025. Multimodal extraction and recognition of Arabic implicit discourse relations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5415–5429, Abu Dhabi, UAE. Association for Computational Linguistics.

Muhammed Saeed, Peter Bourgonje, and Vera Demberg. 2025. Implicit discourse relation classification for Nigerian Pidgin. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2561–2574, Abu Dhabi, UAE. Association for Computational Linguistics.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Interspeech*, volume 2012, pages 194–197.

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.

Kate Thompson, Akshay Chaturvedi, Julie Hunter, and Nicholas Asher. 2024. Llamipa: An incremental

discourse parser. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6418–6430, Miami, Florida, USA. Association for Computational Linguistics.

Kevin Ro 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 *The Eleventh International Conference on Learning Representations*.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable nlp.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Hongyi Wu, Hao Zhou, Man Lan, Yuanbin Wu, and Yadong Zhang. 2023a. Connective prediction for implicit discourse relation recognition via knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5908–5923, Toronto, Canada. Association for Computational Linguistics.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023b. QUDeval: The evaluation of questions under discussion discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363, Singapore. Association for Computational Linguistics.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. Prompting implicit discourse relation annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julians, Malta. Association for Computational Linguistics.

Frances Yung, Varsha Suresh, Zaynab Reza, Mansoor Ahmad, and Vera Demberg. 2025. Synthetic data augmentation for cross-domain implicit discourse relation recognition. *Preprint*, arXiv:2503.20588.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, 51(1):23–72.

Fred Zhang and Neel Nanda. 2024. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*.

Haodong Zhao, Ruifang He, Mengnan Xiao, and Jing Xu. 2023. Infusing hierarchical guidance into prompt tuning: A parameter-efficient framework for multi-level implicit discourse relation recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6477–6492, Toronto, Canada. Association for Computational Linguistics.

Zining Zhu, Hanjie Chen, Xi Ye, Qing Lyu, Chenhao Tan, Ana Marasovic, and Sarah Wiegreffe. 2024. Explanation in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 19–25, Mexico City, Mexico. Association for Computational Linguistics.

A CuDR Dataset Details

A.1 Counterfactual Connectives

To create counterfactual instances in the CuDR dataset, we rely on the taxonomy by Miao et al. (2024), which defines each discourse relation along with five irrelevant counterfactual relations. Due to space constraints, Table 2 in Section 2 lists only a subset of the counterfactual connectives. The complete set of five counterfactual connectives is provided in Table 6.

Discourse Relation	Ori Connective	CF Connectives
Comparison.Concession.Arg2-as-denier	however	because for example specifically so in other words
Comparison.Contrast	by comparison	specifically in other words because for example so
Contingency.Reason	because	so however by comparison for example in other words
Contingency.Result	so	because by comparison for example however in other words
Expansion.Conjunction	and	however so because by comparison instead
Expansion.Equivalence	in other words	however for example because so by comparison
Expansion.Instantiation.Arg2-as-instance	for example	because however by comparison so in other words
Expansion.Level-of-detail.Arg1-as-detail	in short	however so by comparison in other words instead
Expansion.Level-of-detail.Arg2-as-detail	specifically	instead by comparison however so in other words
Expansion.Substitution.Arg2-as-subst	instead	because in other words so for example specifically
Temporal.Asynchronous.Precedence	then	however previously by comparison for example because
Temporal.Asynchronous.Succession	previously	so then by comparison however for example
Temporal.Synchronous	while	so then by comparison however for example

Table 6: **CuDR Dataset Details (Full Counterfactual Connectives):** PDTB discourse relations with their original (Ori) connective and the corresponding set of five counterfactual (CF) connectives.

A.2 Aligning Discourse Frameworks

We refer to cross-framework relation mapping both to prepare counterfactual CuDR data for frameworks beyond PDTB (Section 2.3) and to perform cross-framework transfer (Section 3.2). The mapping between PDTB and the GUM Discourse Treebank (GDTB) (Liu et al., 2024b) is straightforward, as GDTB adopts the PDTB relation taxonomy. For the GUM Rhetorical Structure Theory (GUM-RST) dataset (Zeldes, 2017), we closely examine the annotation guidelines and the mapping approach used by Liu et al. (2024b). Based on this, we define a mapping shown in Table 7, which includes 17 RST relations, excluding those with insufficient data. This mapping offers broad coverage, aligning the 17 RST relations with 9 distinct PDTB relations. For the Segmented Discourse Representation Theory (SDRT) dataset (Asher and Lascarides, 2003), we also examine the relation definitions and construct the mapping presented in Table 8. This results in 10 distinct SDRT relations mapped to 8 PDTB relations.

RST Label	Mapped PDTB Label
joint-list_m	Expansion.Conjunction
joint-sequence_m	Temporal.Asynchronous.Precedence
elaboration-additional_r	Expansion.Level-of-detail.Arg2-as-detail
context-circumstance_r	Temporal.Synchronous
adversative-concession_r	Comparison.Concession.Arg2-as-denier
causal-cause_r	Contingency.Cause.Reason
causal-result_r	Contingency.Cause.Result
adversative-contrast_m	Comparison.Contrast
explanation-justify_r	Contingency.Cause.Reason
context-background_r	Expansion.Conjunction
joint-other_m	Expansion.Conjunction
adversative-antithesis_r	Comparison.Contrast
explanation-evidence_r	Contingency.Cause.Reason
evaluation-comment_r	Contingency.Cause.Reason
explanation-motivation_r	Contingency.Cause.Reason
restatement-repetition_m	Expansion.Equivalence
joint-sequence_r	Temporal.Asynchronous.Precedence

Table 7: **RST to PDTB Mapping:** Mapping of RST discourse labels to PDTB labels for the CuDR dataset.

SDRT Label	Mapped PDTB Label
Acknowledgement	Expansion.Equivalence
Comment	Expansion.Conjunction
Continuation	Expansion.Conjunction
Contrast	Comparison.Contrast
Correction	Comparison.Concession.Arg2-as-denier
Elaboration	Expansion.Level-of-detail.Arg2-as-detail
Explanation	Contingency.Cause.Reason
Narration	Temporal.Asynchronous.Precedence
Parallel	Expansion.Conjunction
Result	Contingency.Cause.Result

Table 8: **SDRT to PDTB Mapping:** Mapping of SDRT discourse labels to PDTB labels for the CuDR dataset.

A.3 Details for CuDR Dataset Construction

To construct the counterfactual argument Arg_2', we ensure it is coherent with both the original argument Arg_1 and the counterfactual discourse relation, along with its connective $Conn'$. **Input:** We generate the dataset by prompting the GPT-4o-mini model via API, chosen for its balance of instruction-following ability and efficiency. Each prompt includes Arg_1, $Conn'$, and a `CF_dr_description` field defining the discourse relation. For example, `Contingency.Cause.Reason` is described as "Arg_2 is the reason for Arg_1: when Arg_1 gives the effect, and Arg_2 provides the reason, explanation, or justification", adapted from the PDTB annotation guidelines (Webber et al., 2019). **Requirements:** We ask the model to complete a structured JSON template. To maintain quality and discourage shallow completions, we explicitly instruct the model *not* to repeat $Conn'$ verbatim, and instead to use relation-specific language patterns. We also request that Arg_2' match the length of Arg_1, improving stylistic and structural consistency. **Output and Postprocessing:** The model is prompted independently for each CuDR data instance, and its output is saved as a plain text file. These files are subsequently parsed into usable JSON format using a custom loader. The final prompt template, with inserted variables such as Arg_1 and $Conn'$, is shown below:

```
You are an expert in discourse semantics. In discourse
      theory, arg1 and arg2 are two arguments connected by a
       relation (a connective word).
I am going to give you an original discourse argument (*
      original_arg1*) and a counterfactual relation (*CF_dr
      *). Your task is to generate a new counterfactual
      argument (*counterfactual_arg2*) that aligns with *
      original_arg1* while reflecting the given
      counterfactual relation.

**Requirements:**
1. *counterfactual_arg2* must be **coherent** with *
      original_arg1* and appropriately reflect the given
      counterfactual relation (by writing after
      counterfactual_connective).
2. The length of *counterfactual_arg2* should be around {
      original_arg2_length} words.
3. Make the relation between *counterfactual_arg2* and *
      original_arg1* easy to understand and as salient as
      possible.
4. Do not repeat the connective word in your *
      counterfactual_arg2*. Instead, try to use negation or
      contrastive signal (for comparison counterfactuals),
      specific causal events of result or reason (for
      contingency counterfactual), specific examples like
      entities and concrete details (for expansion
      counterfactuals).

Complete the following dictionary and only return the
      dictionary as your output:
{
    "original_arg1": "{original_arg1}",
    "counterfactual_relation": "{CF_dr}", which means {
         CF_dr_description},
    "counterfactual_connective": "{conn_CF}",
    "counterfactual_arg2": TO BE COMPLETED
}
```

Manual Verification One author manually verified the quality of our CuDR data samples. We randomly sampled 10 instances from each discourse framework and present subsets of CuDR examples from the PDTB (Table 9), GDTB (Table 11), RST (Table 10), and SDRT (Table 12) datasets. Although each framework uses different terminology, we adopt a unified notation of Arg_1 and Arg_2 throughout. **Across the 40 samples, we find all to be valid:** the generated Arg_2' is coherent with the original Arg_1 and aligns well with the intended counterfactual connective $Conn'$. For example, in the first PDTB sample, the original Arg_1 is "Robert S. Ehrlich resigned as chairman, president and chief executive", which is linked by a denying relation (signaled by "however") to "Mr. Ehrlich will continue as a director and a consultant". Under the counterfactual connective $Conn'$ "so", our generated Arg_2' becomes "the company faced significant leadership challenges afterward", directly expressing the consequence of Mr. Ehrlich's resignation and appropriately realizing the intended relation. Beyond PDTB, our CuDR construction performs well across other frameworks. For instance, although SDRT often contains shorter text spans, the generated Arg_2' still effectively reflects the intended $Conn'$. In Sample 2 from Table 12, "others settle for less" clearly presents a contrasting scenario, demonstrating that the model can express discourse relations concisely.

However, we do find our generated data to be straightforward in their expression. In all samples we examined, rare words are seldom used, and the model tends to prefer simple sentence structures. For example, Sample 3 in SDRT (Table 12) has an original Arg_1 as "yep saturday's looking promising", and continues with an Arg_2' expression "the weather forecast predicts sunshine", using the counterfactual connective "because". This is a valid instance, but discussing the weather is relatively expected and less surprising. Sample 3 in PDTB (Table 9) has an Arg_1 as "Much is being done in Colombia to fight the drug cartel mafia", and it assigns Arg_2' as "the government recognizes that drug trafficking severely undermines national security and social stability". While this is a valid continuation aligning with the counterfactual connective "because", it lacks specific knowledge about the drug situation in Colombia. In contrast, the original Arg_2 is "luxurious homes and ranches have been raided by the military authorities, and sophisticated and powerful communica-

tions equipment have been seized", which offers more phrase variation and concrete detail.

This straightforward style is expected, as we explicitly prompt the model to "Make the relation between counterfactual Arg_2' and original Arg_1 easy to understand and as salient as possible." We adopt this approach to make the CuDR task as sharply steering as possible. In future work, we can explore the CuDR task under more complex texts and ambiguous scenarios.

CuDR data samples for PDTB framework:
Sample 1: Arg_1: Robert S. Ehrlich resigned as chairman, president and chief executive Arg_2: Mr. Ehrlich will continue as a director and a consultant Original relation: Comparison.Concession.Arg2-as-denier (however) $Conn'$: so Arg_2': the company faced significant leadership challenges afterward.
Sample 2: Arg_1: Shortly after Texas Air took control of Eastern, some Machinists union supervisors received a 20% pay raise Arg_2: the pilots argued that this triggered a pay raise for them Original relation: Contingency.Cause.Result (so) $Conn'$: but Arg_2': most other employees were not granted any wage increase.
Sample 3: Arg_1: Much is being done in Colombia to fight the drug cartel mafia Arg_2: luxurious homes and ranches have been raided by the military authorities, and sophisticated and powerful communications equipment have been seized Original relation: Expansion.Instantiation.Arg2-as-instance (for example) $Conn'$: because Arg_2': the government recognizes that drug trafficking severely undermines national security and social stability.

Table 9: CuDR data samples for PDTB framework: counterfactual Arg_2' being coherent with original Arg_1 and counterfactual $Conn'$.

CuDR data samples for RST framework:
Sample 1: Arg_1: that cultural behaviors are not genetically inherited from generation to generation Arg_2: must be passed down from older members of a society to younger members Original relation: adversative-antithesis (however) $Conn'$: specifically Arg_2': they are learned through social interactions and environmental influences
Sample 2: Arg_1: I came up with an individual story called Thad 's World Destruction and , she wanted to illustrate it Arg_2: that 's the way we ended up doing it Original relation: causal-result (so) $Conn'$: but Arg_2': she thought it was too dark for children
Sample 3: Arg_1: fisherman first noticed the people Arg_2: a warship was deployed to retrieve them Original relation: joint-sequence (then) $Conn'$: because Arg_2': he heard their laughter nearby

Table 10: CuDR data samples for RST framework: counterfactual Arg_2' being coherent with original Arg_1 and counterfactual $Conn'$.

CuDR data samples for GDTB framework:
Sample 1: Arg_1: Due to its remarkable biodiversity, with over a third of the local plant species found nowhere else, Socotra has been designated a UNESCO World Heritage Site Arg_2: With over 40,000 inhabitants, though, it's not just a nature reserve Original relation: Comparison.Concession.Arg2-as-denier (however) $Conn'$: so Arg_2': many conservation efforts are now focused on preserving its unique ecosystems.
Sample 2: Arg_1: So this place was so cool we could have spent hours in here Arg_2: The best thing that I thought about this bookstore was that they mixed in new copies of books with used copies Original relation: Contingency.Cause.Result (so) $Conn'$: but Arg_2': the uncomfortable seating made it difficult to stay for long, despite the incredible atmosphere surrounding us.
Sample 3: Arg_1: There are flights from Sana'a via Al Mukalla Arg_2: Yemenia Airlines offers one flight per week on Thursday morning Original relation: Expansion.Instantiation.Arg2-as-instance (for example) $Conn'$: because Arg_2': the airport reopened after extensive renovations

Table 11: CuDR data samples for GDTB framework: counterfactual Arg_2' being coherent with original Arg_1 and counterfactual $Conn'$.

B Implementation Details

B.1 Model fine-tuning

The CuDR task imposes two key requirements: **(1) Instruction following:** the model must adhere to the task format by choosing between Arg_2 and Arg_2'; and **(2) Discourse comprehension:** it must interpret the discourse relation to select the continuation that matches the given connective. These requirements prove challenging for the widely used GPT-2 model (Conmy et al., 2023; Yao et al., 2024). To address (1), we pretrain GPT-2 on a next sentence prediction (NSP) task using randomly mismatched Arg_2' from PDTB. Without this step, the

CuDR data samples for SDRT framework:
Sample 1: Arg_1: the deal mechanism 's a bit clunky Arg_2: the key is to make sure you've checked the right colour box :D Original relation: Contrast (by comparison) $Conn'$: specifically Arg_2': it often requires multiple steps and lengthy approvals to finalize transactions
Sample 2: Arg_1: you drive a hard bargain Arg_2: that price is too good Original relation: Explanation (because) $Conn'$: by comparison Arg_2': others settle for less
Sample 3: Arg_1: yep saturday 's looking promising Arg_2: saturday evening good for me too Original relation: Parallel (and) $Conn'$: because Arg_2': the weather forecast predicts sunshine

Table 12: CuDR data samples for SDRT framework: counterfactual Arg_2' being coherent with original Arg_1 and counterfactual $Conn'$, while the arguments are shorter than PDTB.

	Accuracy		Logit Diff	
	Ori	**CF**	**Ori**	**CF**
Random Model	0.50	0.50	0.00	0.00
Ideal Model	1.00	1.00	+	+
GPT$_{\text{NSP}}$	0.46	0.63	-1.26	2.51
GPT$_{\text{CuDR}}$	**0.80**	**0.79**	**11.89**	**11.15**

Table 13: **Performance on the CuDR task:** Accuracy and logit difference are reported for each model under both original (Ori) and counterfactual (CF) scenarios.

model often generates irrelevant outputs. However, GPT$_{\text{NSP}}$ performs poorly on the actual CuDR task, with near random accuracy (0.46 and 0.63; see Table 13). To address (2), we further pretrain it on strictly held-out set of PDTB data, resulting in GPT$_{\text{CuDR}}$, which achieves 0.8 accuracy and a significantly larger logit margin. This ensures the model is sensitive to discourse relation, making it suitable for activation patching with CuDR.

These results also reflects the quality of our dataset. GPT$_{\text{NSP}}$ performs better on counterfactual instances than original ones (0.63 vs. 0.46 accuracy), suggesting that the counterfactual data is not only valid but also easier to interpret. The final GPT$_{\text{CuDR}}$ achieves balanced performance across both Ori and CF directions.

B.2 Computation Resource

All experiments are conducted on a server with four NVIDIA L40 GPUs (48GB RAM each). To accelerate circuit discovery, we use the implementation by Miller et al. (2024) [1] for the Edge Attribution Patching (EAP) method (Syed et al., 2024; Nanda, 2023), which completes discovery for a single discourse relation in about one minute using a sample size of 32 on a single GPU. This is substantially faster than the Automatic Circuit DisCovery (ACDC) method (Conmy et al., 2023)[2], which takes over 24 hours for the same task.

B.3 Details for Circuits Analysis Experiments

Antonymy **Input:** The sky was *bright*, far from, **Output:** dark **Input:** His explanation was *clear*, unlike, **Output:** confusing
Coreference **Input:** *John* went to the store because, **Output:** He **Input:** *Lisa* loves painting, and **Output:** She
Negation **Input:** The answer was expected, though arrival was **Output:** delayed **Input:** He expected an easy task, but it was **Output:** not
Synonymy **Input:** The road was *narrow*, and the alley even, **Output:** slim **Input:** The musician composed a *tune*, a catchy, **Output:** melody

Table 14: Data samples for discovering circuits for linguistic features, including antonymy, coreference, negation, and synonymy. If an anchor word exists (e.g. "John"), it was in italic form.

To identify circuits responsible for linguistic features (Zeldes et al., 2025; Das and Taboada, 2018), we adopt a simplified next-word prediction setting, where the model predicts a word tied to a specific linguistic feature. This setup follows tasks like subject–verb agreement (SVA) (Ferrando and Costa-jussà, 2024) and world knowledge (Yao et al., 2024). Following standard practice, we apply activation patching. The clean input is a context–target pair, while the corrupted input has the same context but a different (incorrect) target word. Activation patching identifies key edges that steer the model from the incorrect to the correct prediction. For example, for coreference, a clean input like "Lisa loves painting" should yield "she"; similarly, "John went to the store because" should produce he" (Table 14). We patch activations from the clean input into the corrupted one to restore the correct output and identify important edges to compose the corresponding circuits.

[1] https://github.com/UFO-101/auto-circuit

[2] https://github.com/ArthurConmy/Automatic-Circuit-Discovery

B.4 Discursive Circuits Help Uncover Underspecified Bias

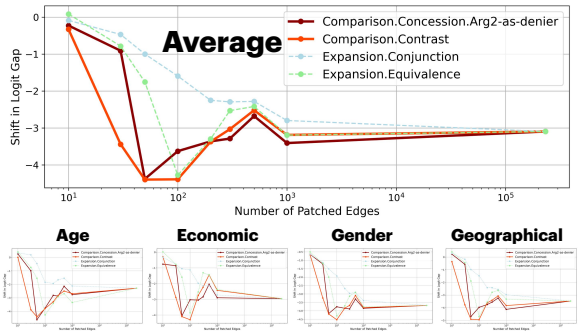

Figure 9: **Impact of discursive circuits on biased completions.** A sharper decrease in answer logit gap (Y-axis) w.r.t. patched edges (X-axis) indicates stronger circuit influence. The upper plot shows average effects.

The main body of the paper focuses on the CUDR task itself. To illustrate the utility of the identified discursive circuits, we present one possible use case where these circuits help reveal potential ethical biases in LLMs. We consider scenarios where the model predicts a next sentence given an underspecified discourse relation (i.e., without an explicit connective). For instance, "Girls like math" is followed by "Boys like sports". It is unclear whether the model interprets the two as equivalent or contrasting. Discursive circuits can uncover if the models generates the prediction for the correct reason. To test whether the model relies on a given discursive circuit (e.g., Contrast), we destroy the activation in that circuit by patching in values from an unrelated sentence, and observe whether the output shifts toward completing that unrelated context. Thus, a stronger reliance results in a sharper shift. We select four representative social biases (Liu et al., 2024a) and create 100 discourse instances with underspecified discourse relations. Using GPT-4o-mini, we prompt the model to generate short and simple cases that are coherent but intentionally underspecified in their discourse relation. For example, "[A young artist painted bold lines across the canvas]$_{Arg_1}$, [A senior man updated the date in his weather journal]$_{Arg_2}$" is a case for age bias. Figure 10 shows output shifts under four possible biases. We find that comparison circuits produce the steepest drops (50 edges to reach bottom), indicating stronger influence. Equivalence circuits follow but require more edges (100 edges to reach bottom), while Conjunction circuits show minimal impact. This provides mechanistic evidence that the model may exhibit a bias toward contrastive interpretations.

B.5 Samples of Discursive Circuits Visualization

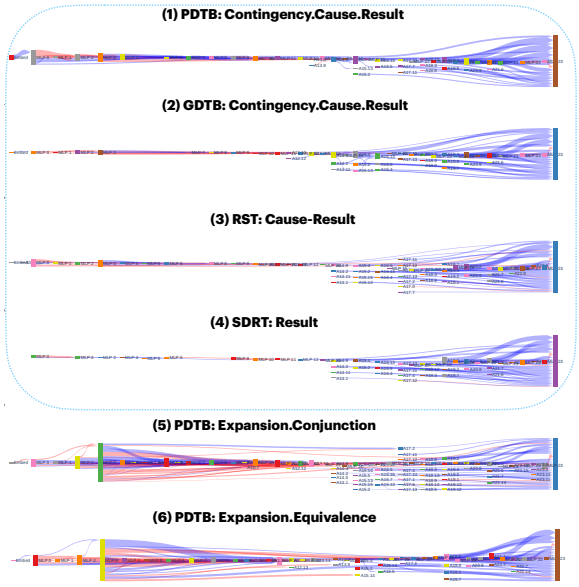

Figure 10: Examples of discursive circuits. Residual flows progress from left (residual start) to right (residual end).

We present representative samples of discursive circuits across different frameworks and thank the visualization tool by Miller et al. (2024). The left side marks the start of the residual flow from the embedding layer, continuing through 24 layers to the residual end. Each edge represents a connection between modular blocks (either MLPs or attention heads) in the transformer. The 1st to 4th samples (highlighted by the blue dot lines) correspond to contingency-like relations across the PDTB, GDTB, RST, and SDRT datasets. These circuits show a consistent pattern: a narrow, focused flow at the start that begins to build specialized representations from Layer 14 onward, dispersing toward the residual end. This aligns with our findings in Section 3.3, where discourse-specific information emerges in higher layers. In contrast, PDTB's Expansion.Conjunction and Expansion.Equivalence (5th and 6th) are more straightforward relations (Section 3.1). Their circuits resemble an "H" shape, with dense processing at both the beginning and end. Overall, these visualizations highlight both the consistency and divergence of circuit structure across different discourse relations.