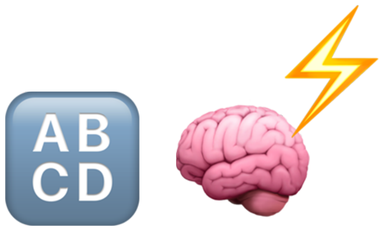
NUS | Computing
National University of Singapore

WING
Web Information Retrieval / Natural Language Processing Group

# Discursive Circuits:
## How Do Language Models Understand Discourse Relations?



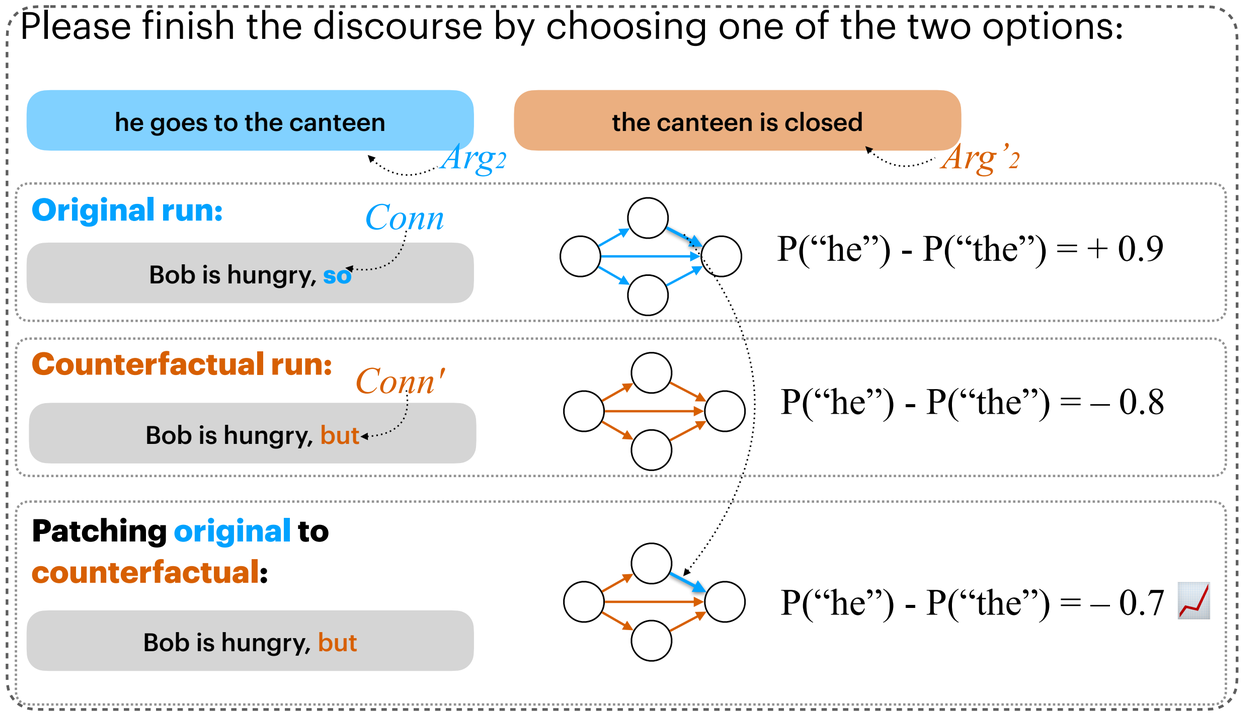

Yisong Miao, Min-Yen Kan

https://github.com/YisongMiao/Discursive-Circuits

Dataset and code @ Github

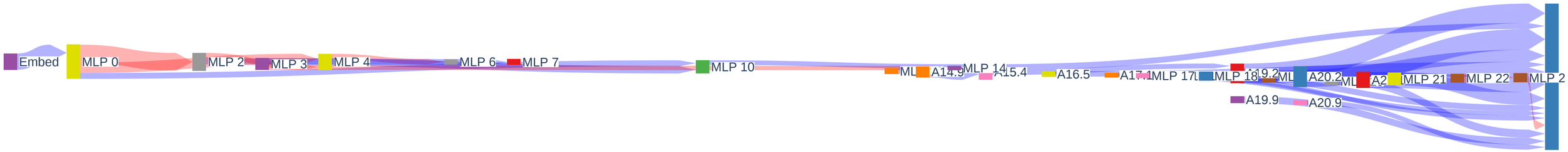


EMNLP 2025
Suzhou, China 中国苏州




**Want to mechanistically interpret how LMs understand discourse relations?**

**Our Completion under Discourse Relation (CuDR) task makes circuit discovery possible.**



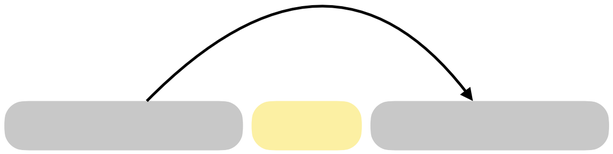
# What is Discourse?

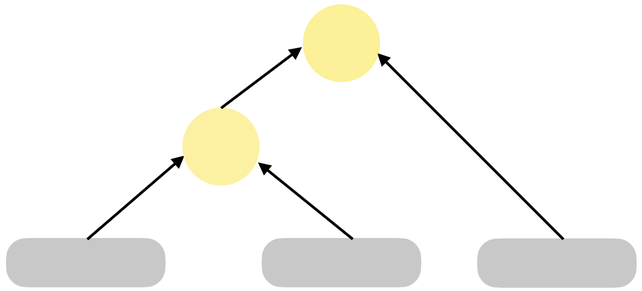
[Bob is hungry]$_{Arg1}$ [so]$_{Conn}$ [he goes to the canteen]$_{Arg2}$

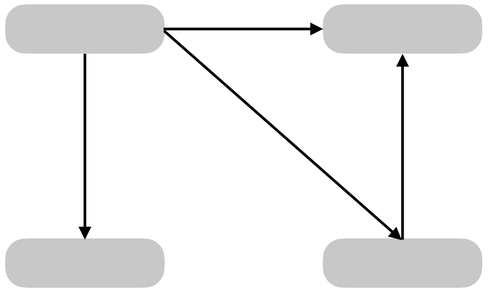
[Bob is hungry]$_{Arg1}$ [however]$_{Conn}$ [the canteen is closed]$_{Arg2}$

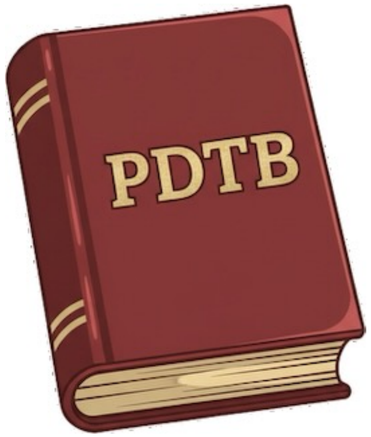
Penn Discourse Treebank (PDTB) style discourse representation.

# Background — Discourse



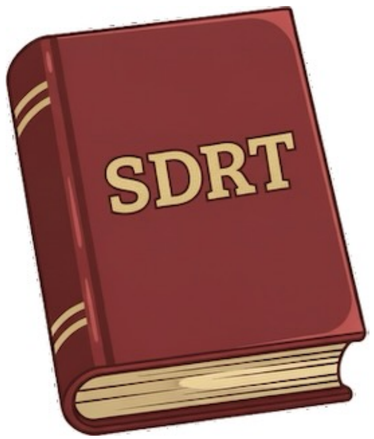






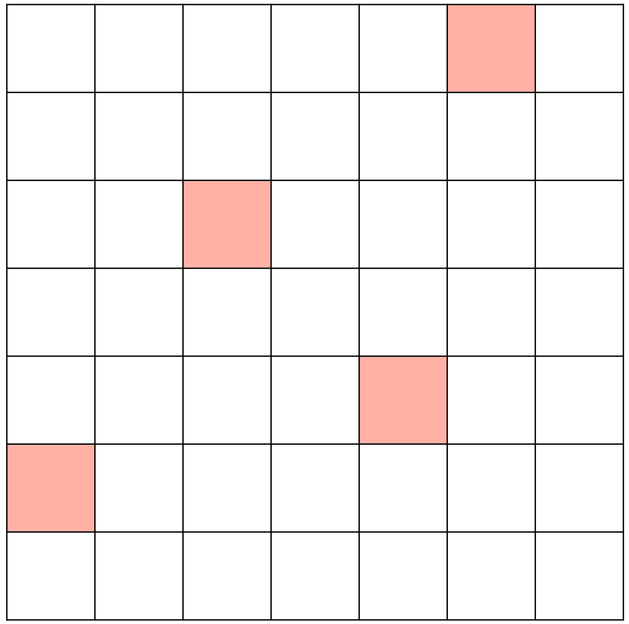

Penn Discourse Treebank


Rhetorical Structure Theory




Segmented Discourse Representation Theory

# Background — Interpretability



> “Here, Bob’s action (going to the canteen) is contingent upon his state (being hungry). His hunger provides the reason for his action.”
>
> — GPT5’s explanation.

***“Attention is NOT explanation!”***

(Jain and Wallace, NAACL 2019)

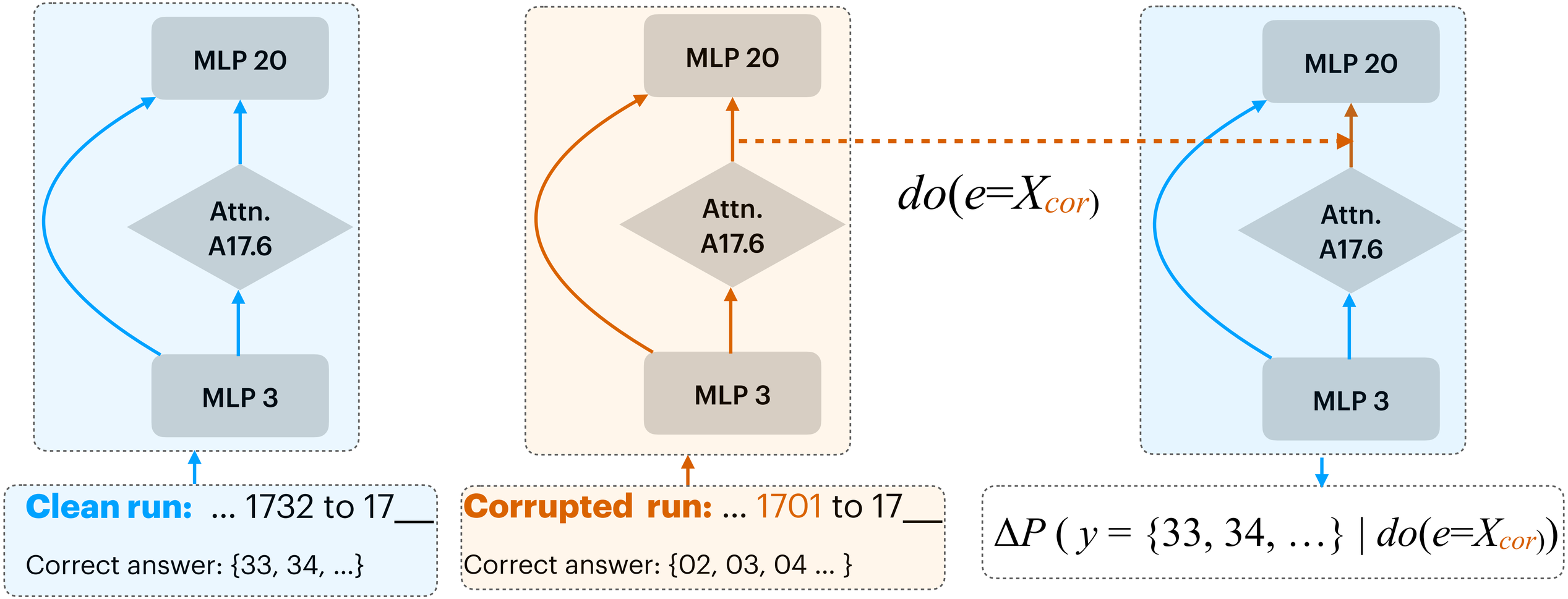
**Are these explanations faithful?**

Lyu, Apidianaki, & Callison-Burch, CL Journal 2024)

# How to find the global flow of discourse in LMs?

# Background — Activation Patching

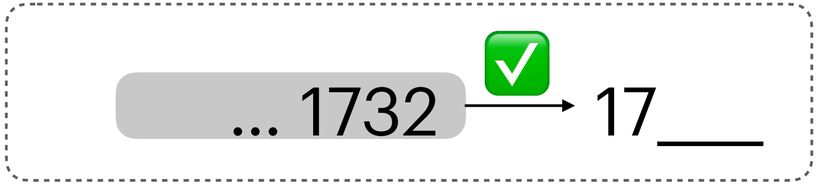
## How to estimate an edge's importance ?



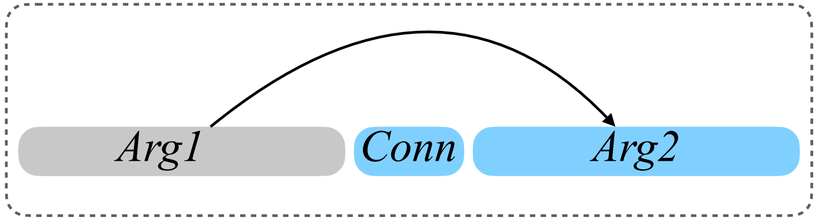

Reference: How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. Hanna, Liu, Variengien. NeurIPS 2023.

# Challenges of Discourse

**“Greater-Than” relation**



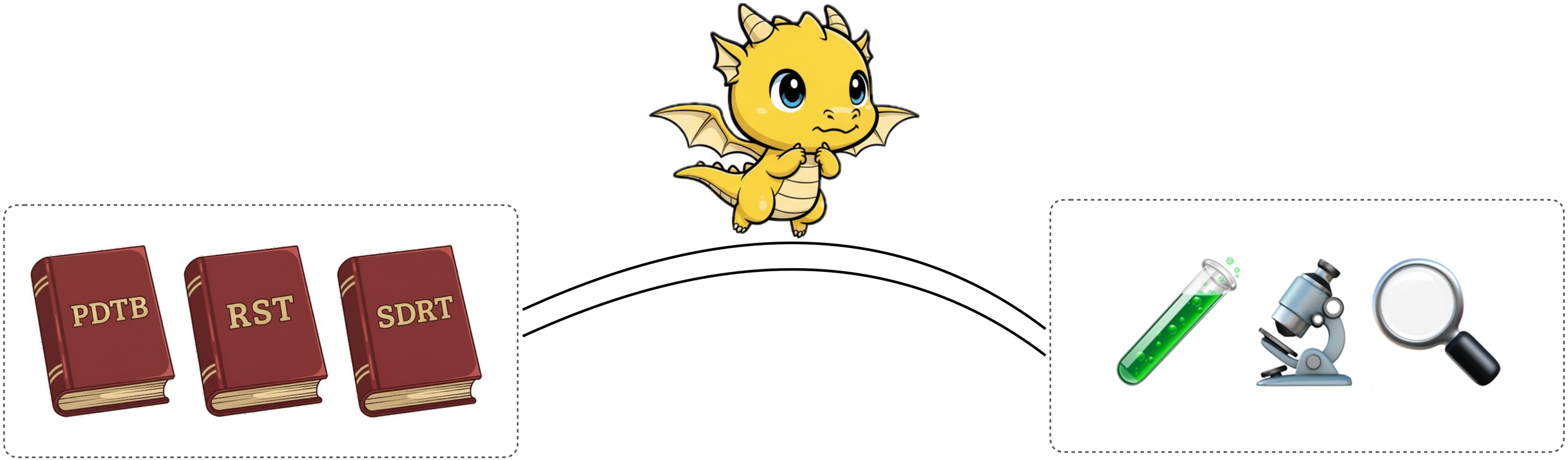

**Discourse relations**




**Challenge 1:** Next word prediction does not fit;

**Challenge 2:** No free lunch for counterfactuals;

# Thesis of This Paper




Discourse Relations

Mechanistic Interpretability

# Thesis of This Paper


Discourse Relations

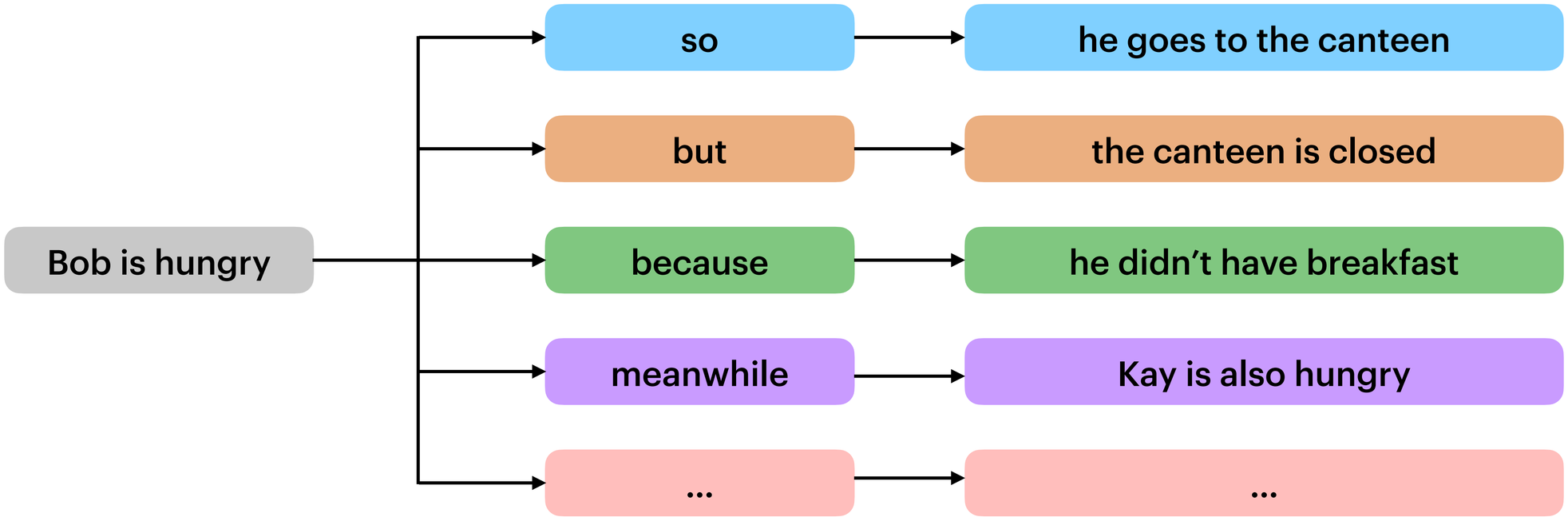
**Completion under Discourse Relation (CuDR)**

Mechanistic Interpretability

# Completion under Discourse Relation

## Our CuDR Task

Intuition: What **makes** us arrange the next sentence in this direction?

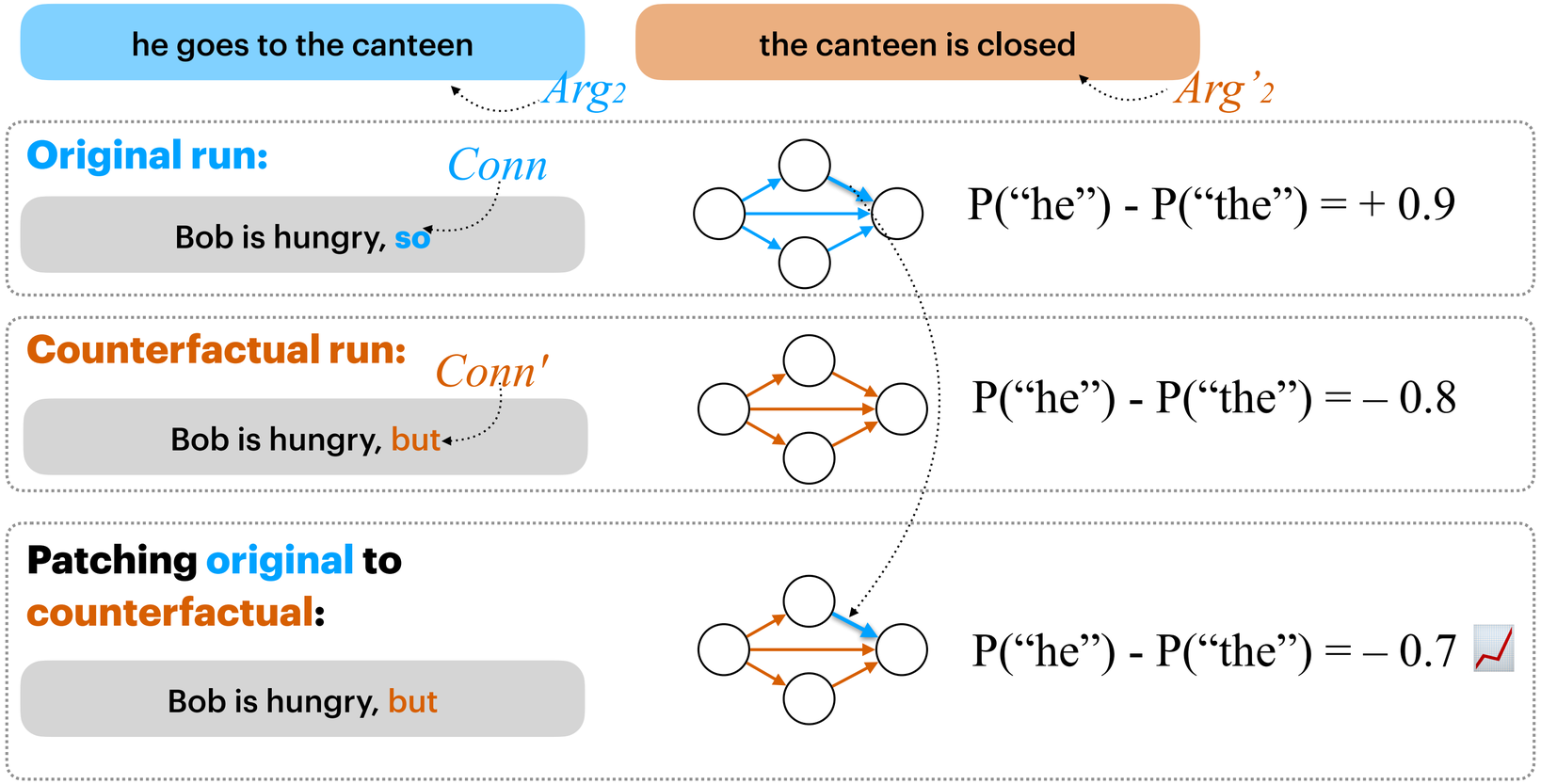



🤯 Infinite number of possible completions!

# Completion under Discourse Relation

## Change your mind with **minimal changes** in the input.

Please finish the discourse by choosing one of the two options:



# Completion under Discourse Relation

## Our CuDR Task

**Input:**
$d_{ori} = (Arg_1, Arg_2, R, Conn)$
$d_{cf} = (Arg_1, Arg_2', R', Conn')$

**Activation patching:**

**Requirement 1:** Minimal contrastive pairs;

**Requirement 2:** Significant change in output;

**CuDR Task (Original):**
Please finish the discourse by choosing one of the two options: $Arg_2$ or $Arg_2'$
To complete: $Arg_1, Conn$

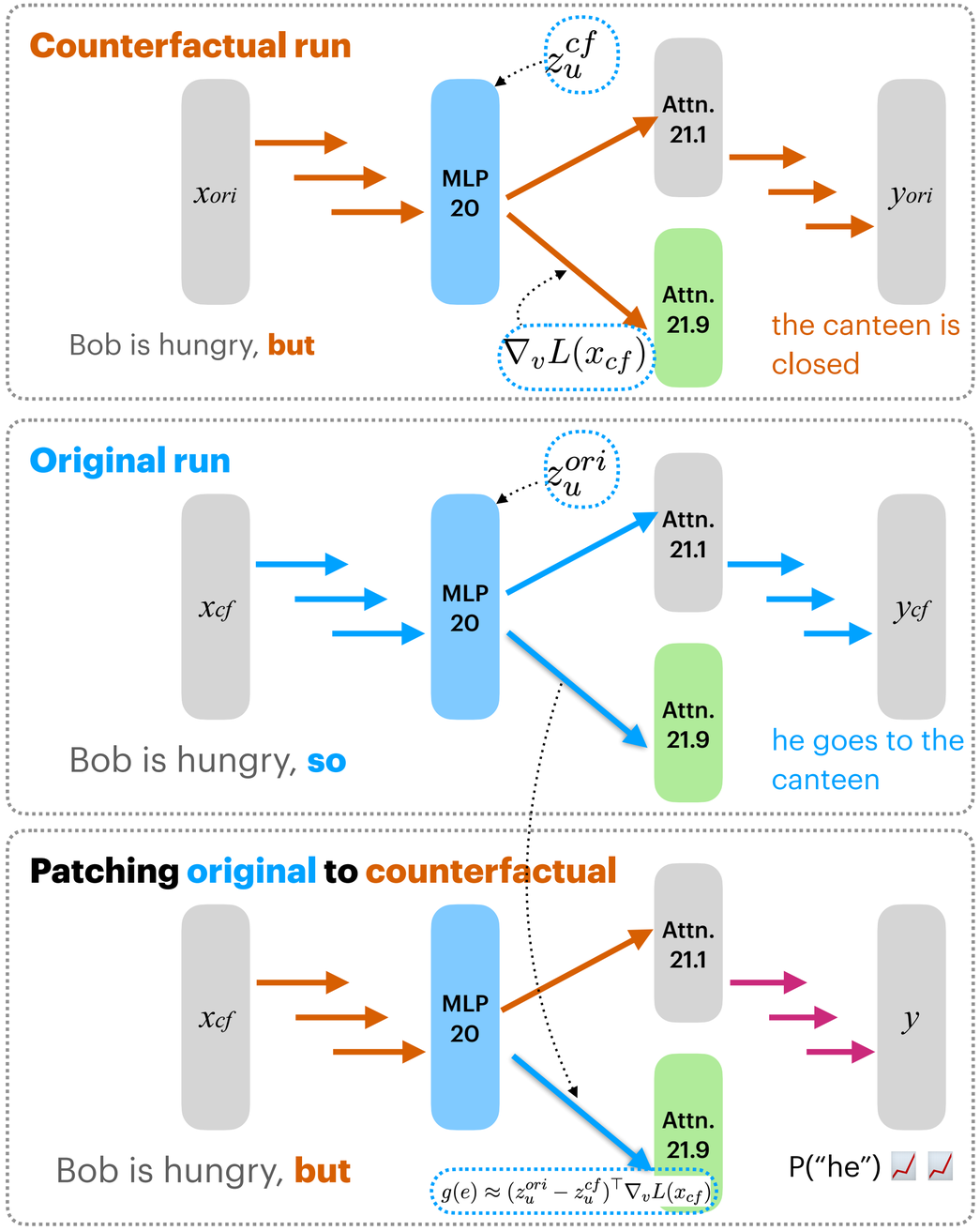
**Correct answer:** $Arg_2$, **Incorrect answer:** $Arg_2'$

**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]$_{Arg_1}$ [so]$_{Conn}$ $\Rightarrow$ [he goes to the canteen]$_{Arg_2}$

**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]$_{Arg_1}$ [but]$_{Conn'}$ $\Rightarrow$ [the canteen is closed]$_{Arg_2'}$




# Activation Patching with CuDR

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

**Equation (1)** estimates the causal effect of any edge in the residual flow.

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

**Equation (2)** accelerates the computation by a first order Taylor approximation.

**Discursive circuits** are composed of top $k$ important edges.


Reference: Attribution patching outperforms automated circuit discovery. Syed, Rager, Conmy. BlackboxNLP 2024.

Transformer circuit evaluation metrics are not robust. Miller, Chughtai, Saunders. COLM 2024.

# Evaluation

## Counterfactual CuDR data

CuDR Data Statistics

| 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 |

Original and Counterfactuals (CF)

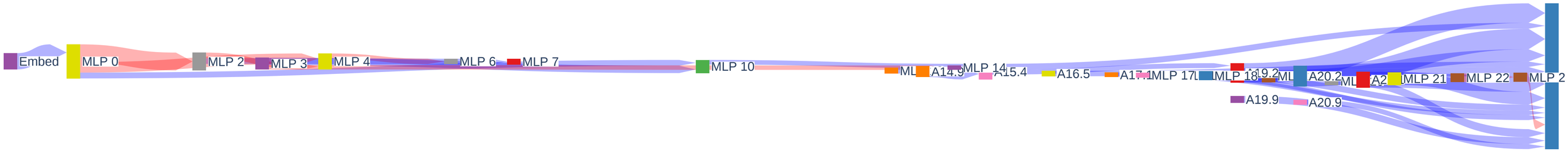
| Discourse Relation | Ori Connective | CF Connectives |
|---|---|---|
| Comp.Conc.Arg2-as-denier | however | because<br>for example<br>specifically<br>so<br>in other words |
| … | … | … |

**Creation of CuDR Data:**

- Prompt GPT 4o-mini with original $Arg_1$ and a counterfactual $Conn'$.
- Each original discourse relation has 5 counterfactual relations (details in paper).

Reference: Discursive Socratic Questioning: Evaluating the Faithfulness of Language Models' Understanding of Discourse Relations. Miao, Liu, Lei, Chen, Kan. ACL 2024.

# Discovering Discursive Circuit



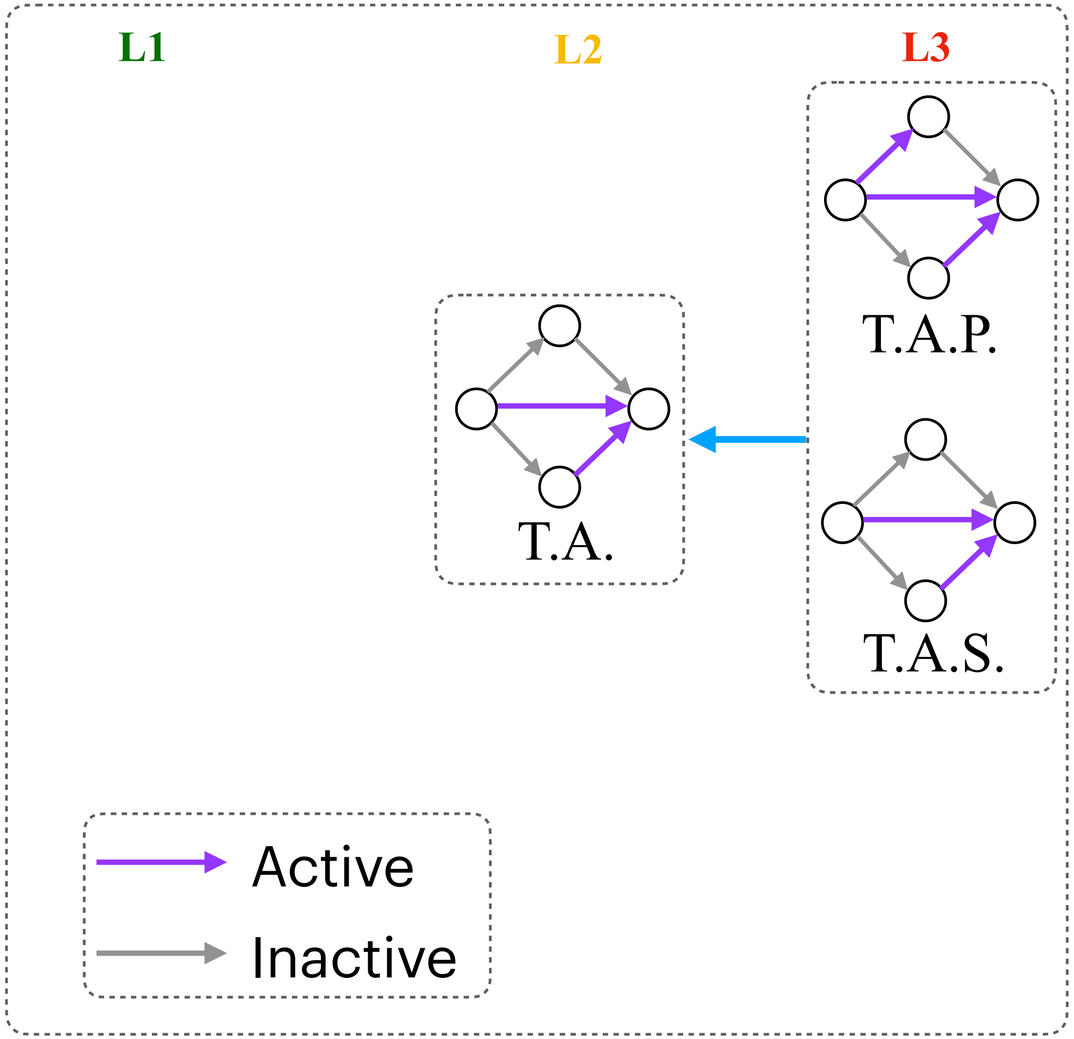

**Discursive Circuits for Contingency.Result relation.**

## Setup:

- Sample size: 32 CuDR data (a pair of (ori, cf) instance).
- Average over 5 randomly sampled runs.

# Circuit Hierarchy

## A new representation of discourse.



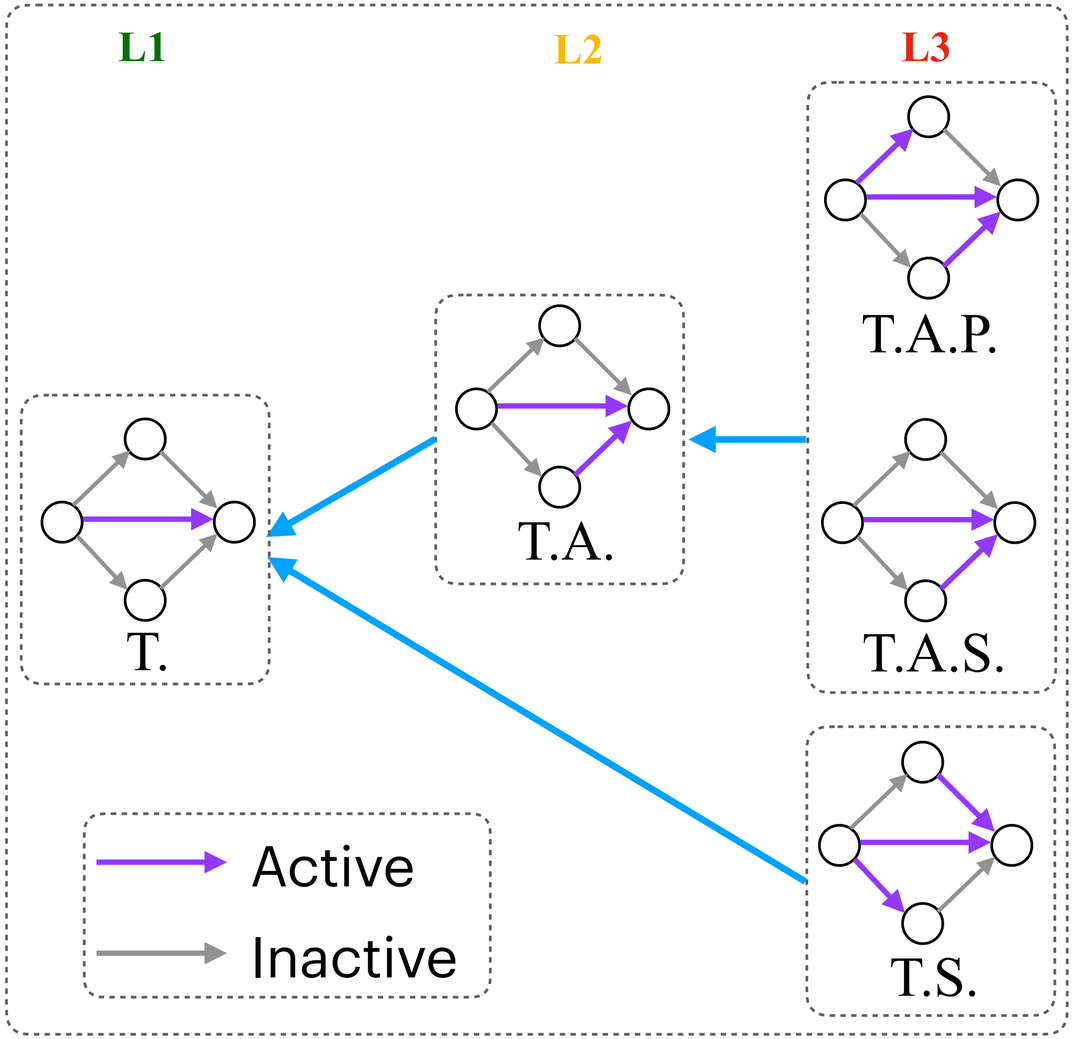

L0 (137)

| L1 | L2 | L3 (1000) |
|---|---|---|
| **Comparison (568)** | 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 (T.A.P.) |
| | | Succession (T.A.S.) |
| | / | Synchronous (T.S.) |

$$L3 \ni L2 \ni L1 \ni L0$$

# Circuit Hierarchy

## A new representation of discourse.

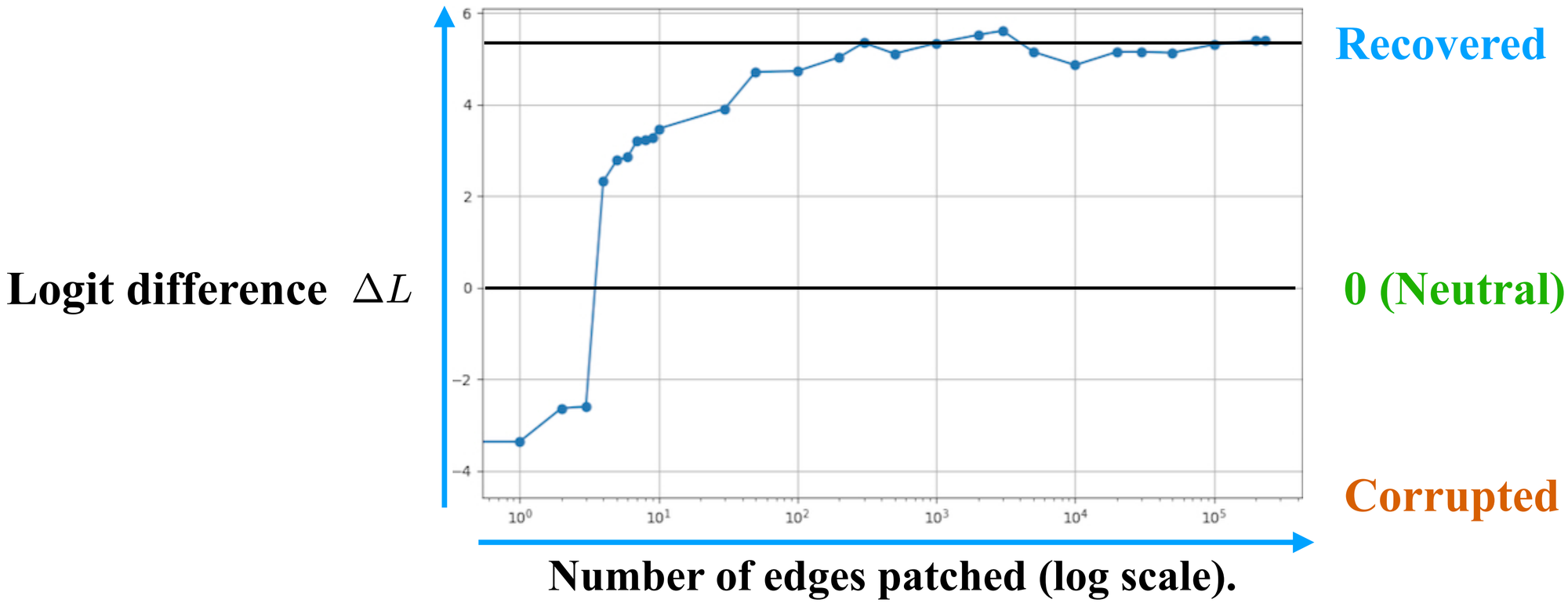



L0 (137)

| L1 | L2 | L3 (1000) |
|---|---|---|
| **Comparison (568)** | 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 (T.A.P.) |
| | | Succession (T.A.S.) |
| | / | Synchronous (T.S.) |

$$L3 \ni L2 \ni L1 \ni L0$$

# Evaluation Metric: Faithfulness Score

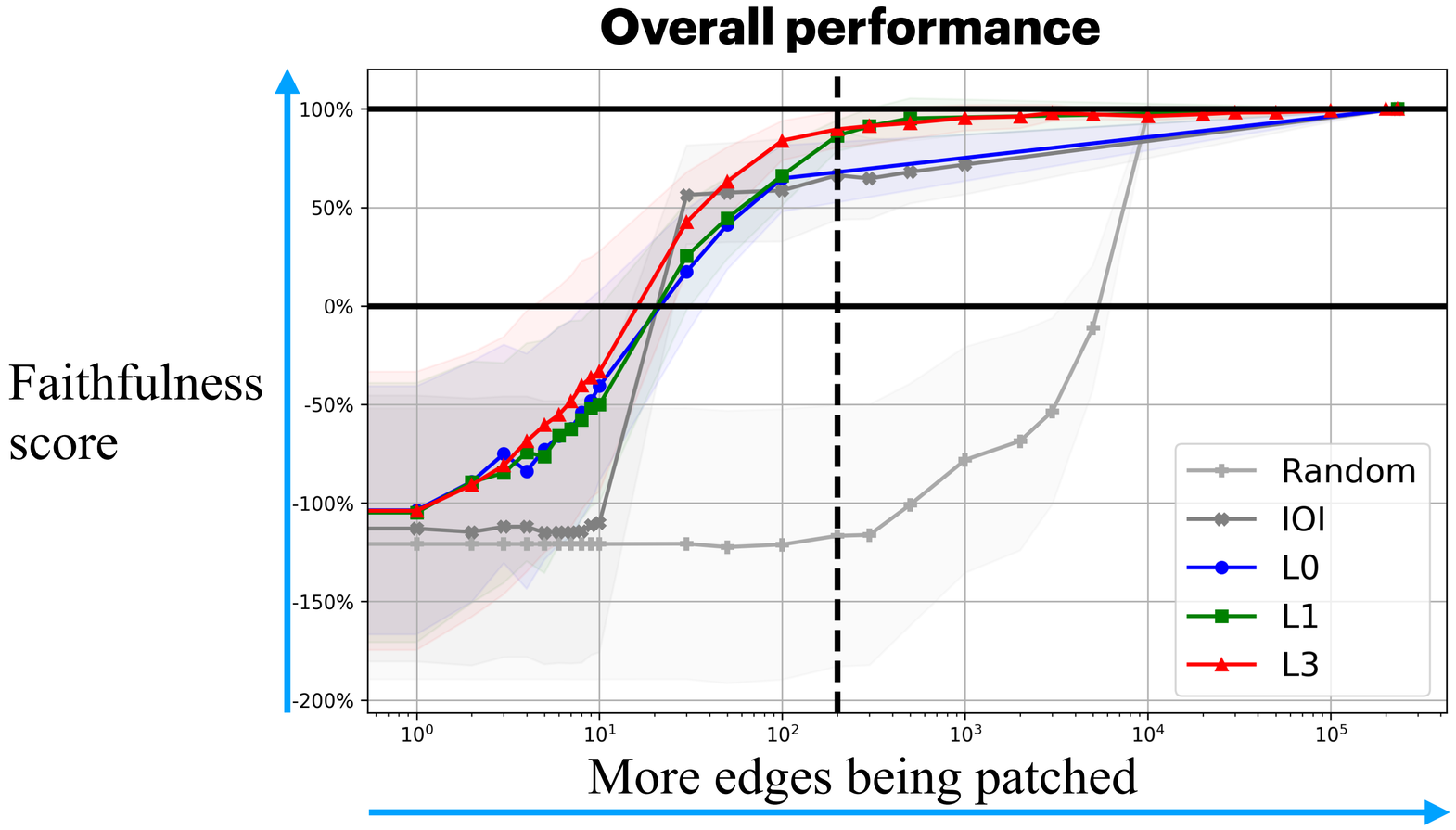



**Faithfulness score:**

- $\Delta L = L(Arg_2) - L(Arg_2')$ : logit difference between the two answers.
- $\frac{\Delta L_{\text{patch}}}{\Delta L_{\text{full}}}$ : normalize to the full model's performance.

# RQ1: Faithfulness

## Evaluation




**Baseline:**

- **Random** circuits;
- **Indirect object identification (IOI)** circuits: *"Bob and Kay went to the bar, Kay gave a drink to ___"* (Answer: "Bob").

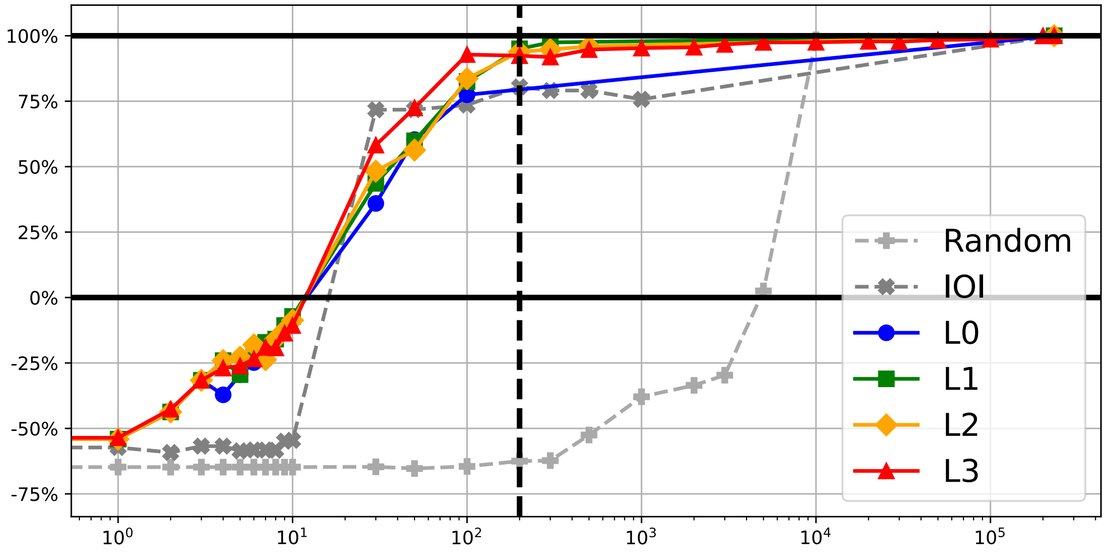
**Faithfulness** is recovered with around 200 edges (<0.2% of the full model).

* Shadows denote variance.

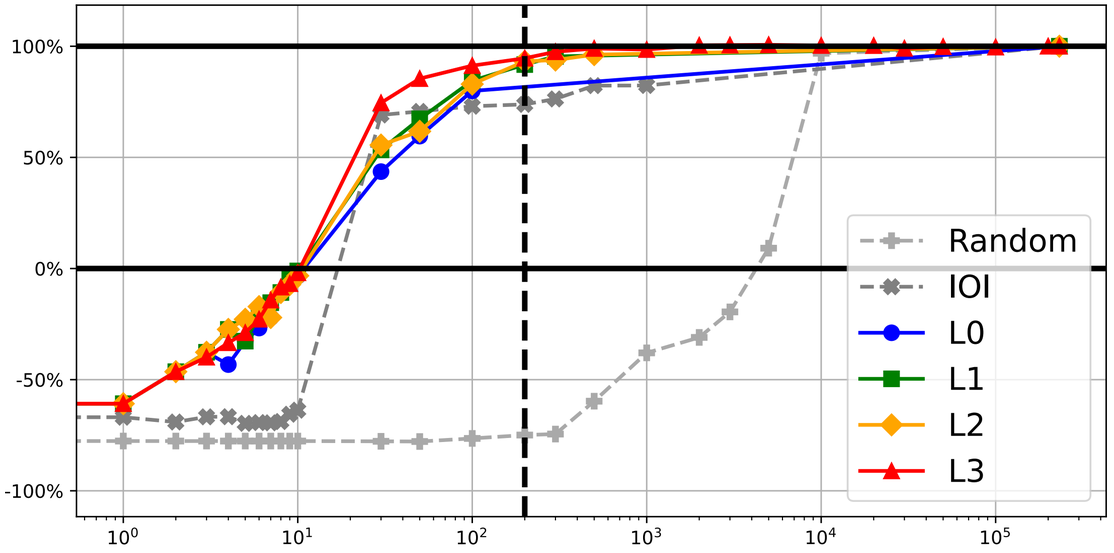
# RQ1: Faithfulness

## Evaluation

### (11) Temporal.Precedence




### (12) Temporal.Succession



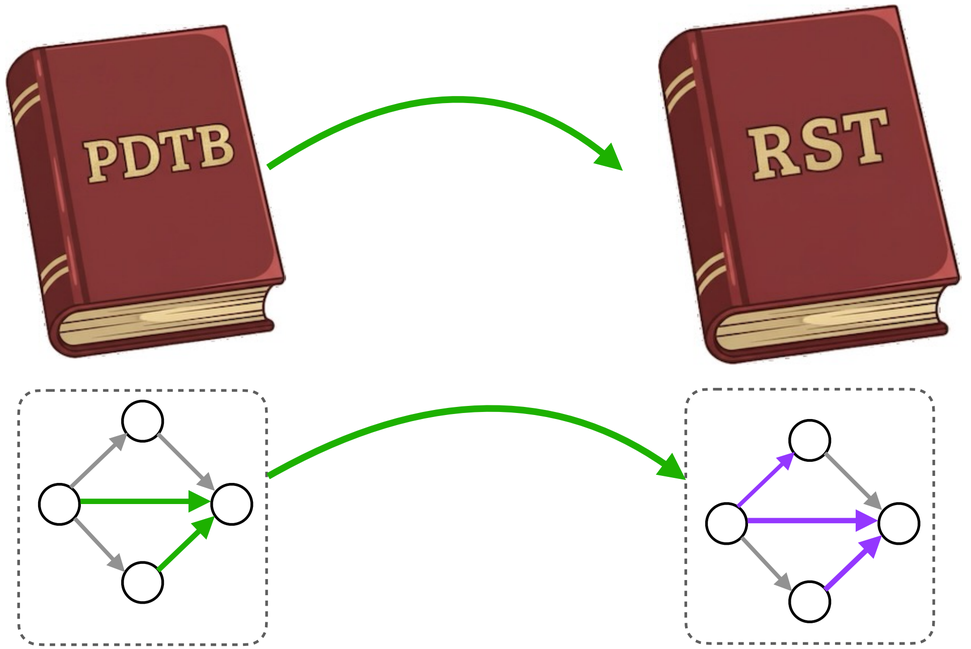

L3 > L2 ≈ L1 > L0 > IOI >> Random

*Full results in the paper.

# RQ2: Generalization

## Evaluation

Given discursive circuits are exact ...



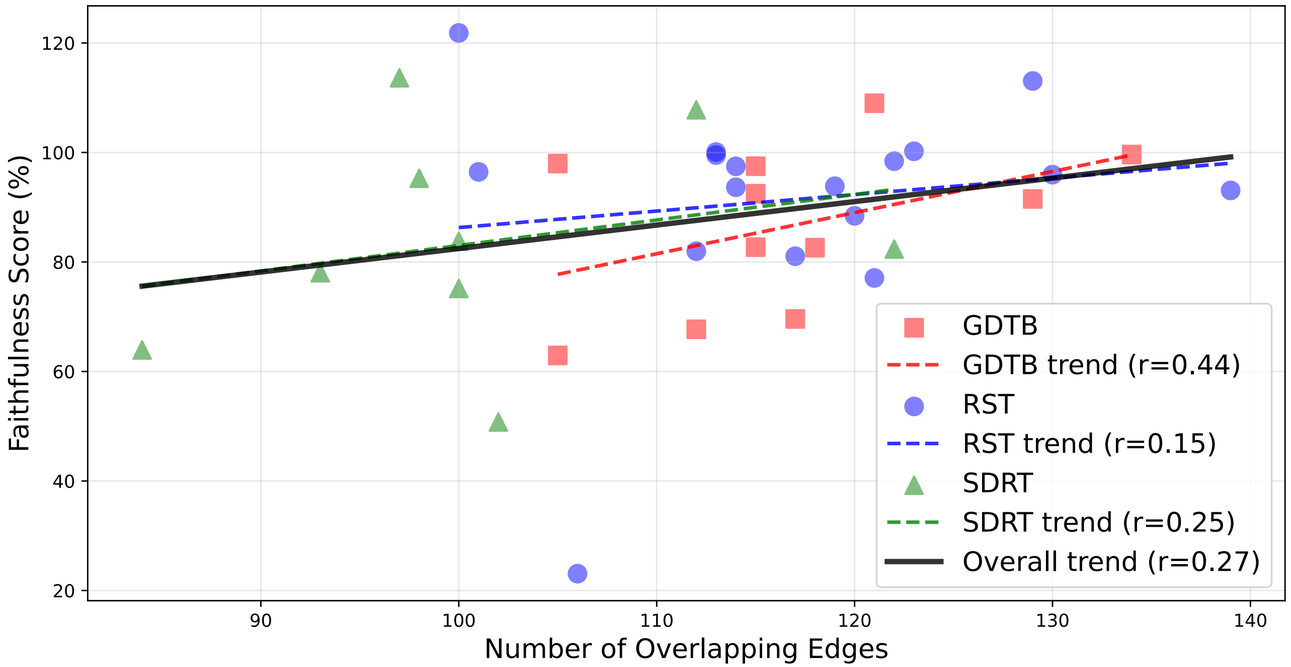

Comparison.Contrast

Adversative.Contrast

*See our paper for the full mapping among frameworks.

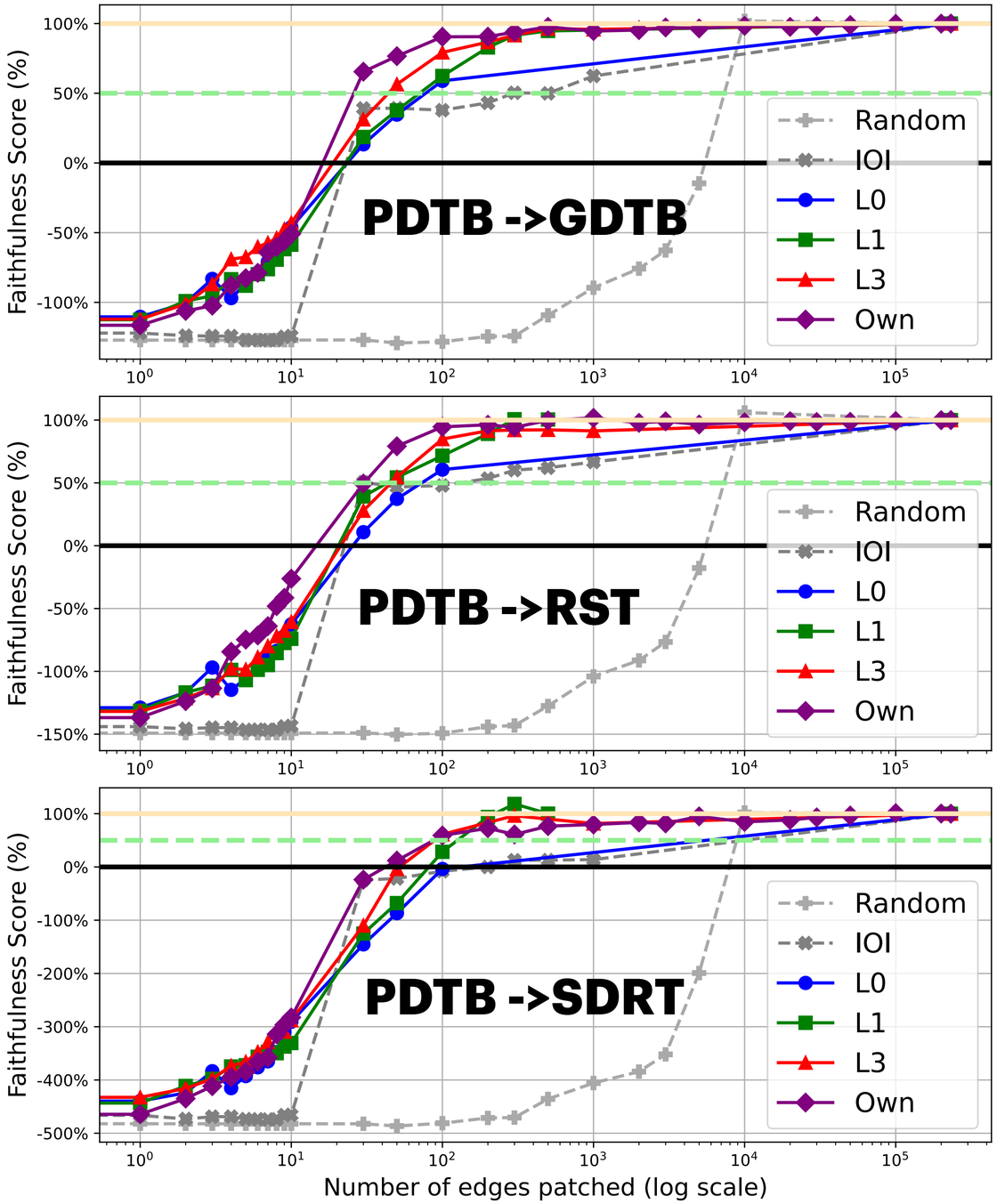
# RQ2: Generalization

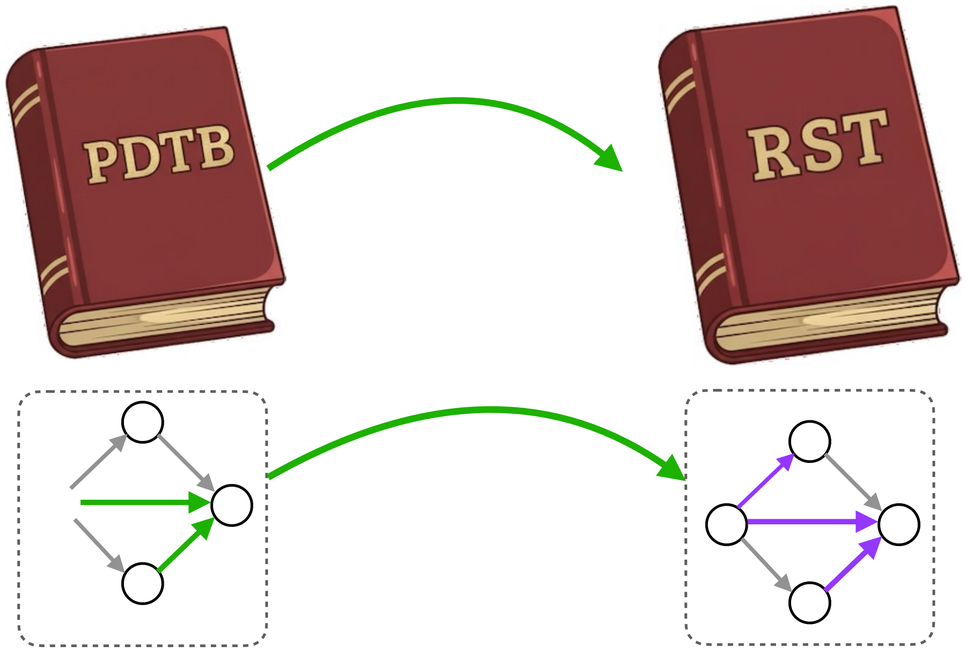
## Evaluation




**Takeaway:**

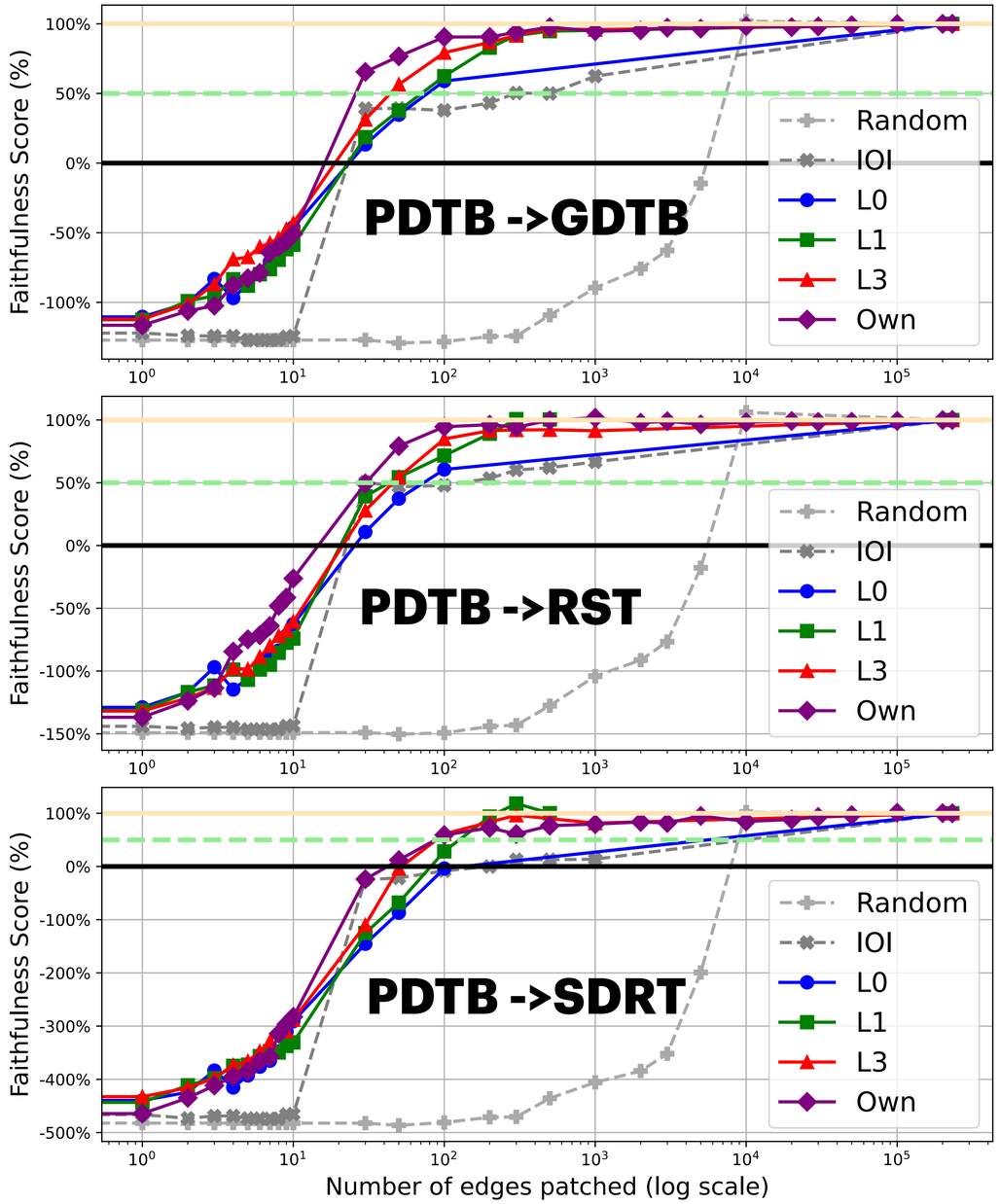
- Circuits generalize well cross-framework.
- Faithfulness ∝ Overlap





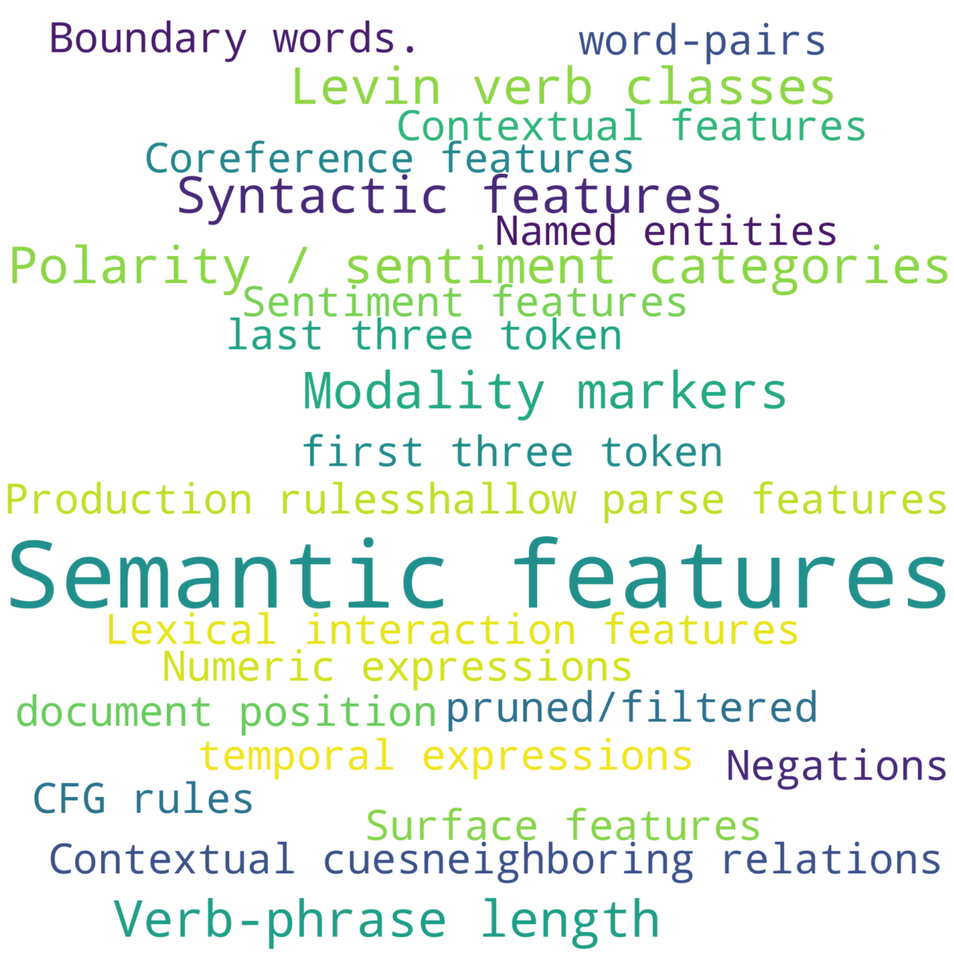

Comparison.Contrast

Adversative.Contrast



# RQ3: Composition of linguistic features

## Evaluation

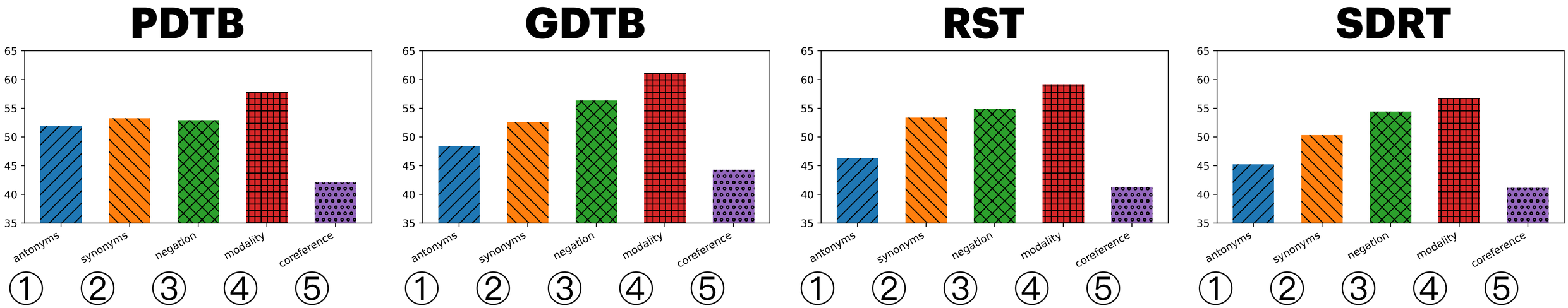




Reference: Automatic sense prediction for implicit discourse relations in text. Pitler, Louis, Nenkova. ACL 2009.

# RQ3: Composition of linguistic features

## Evaluation



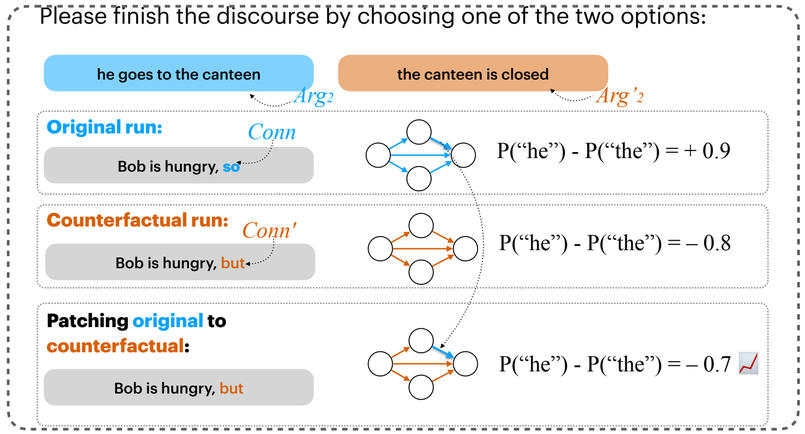

We consider (1) antonyms, (2) synonyms; (3) negation; (4) modality; (5) coreference.

- Utility of linguistic features: **Circuit overlaps.**
- Consistent trends of utilities among 1-5.

# Conclusion

## A bridge between discourse and interpretability.

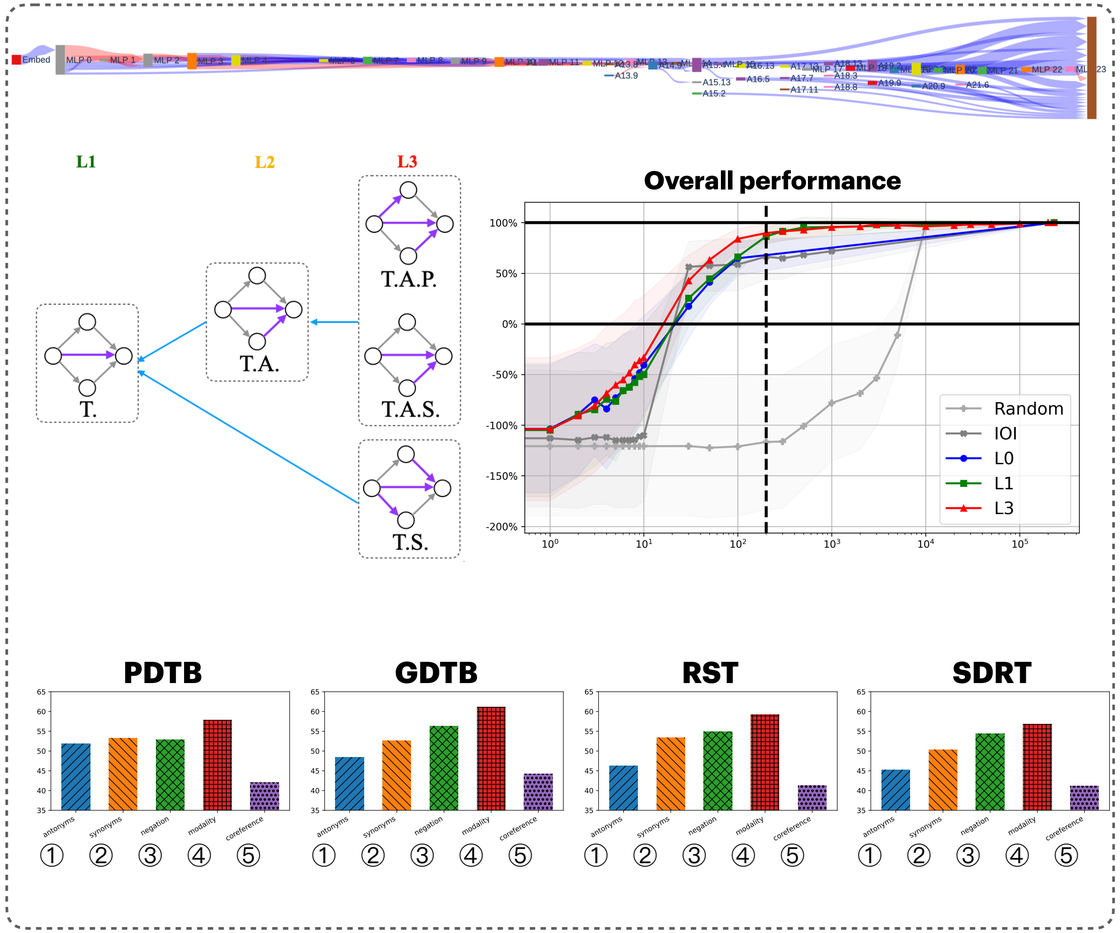
### New task.




### New datasets.

| 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 |

### New representations.



# Talk Feedback

https://forms.gle/ZPV4Wc4AKYXx6Cx89

# References


[1] Emily Pitler, Annie Louis, and Ani Nenkova. 2009. *Automatic sense prediction for implicit discourse relations in text.* ACL 2009.

[2] 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.* ACL 2024.

[3] Joseph Miller, Bilal Chughtai, and William Saunders. 2024. *Transformer circuit evaluation metrics are not robust.* COLM 2024.

[4] Aaquib Syed, Can Rager, and Arthur Conmy. 2024. *Attribution patching outperforms automated circuit discovery.* BlackboxNLP 2024.

[5] Philipp Mondorf, Sondre Wold, and Barbara Plank. 2025. *Circuit compositions: Exploring modular structures in transformer-based language models.* ACL 2025.

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

[7] 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.

[8] Yang Janet Liu, Tatsuya Aoyama, Wesley Scivetti, Yilun Zhu, Shabnam Behzad, Lauren Elizabeth Levine, Jessica Lin, Devika Tiwari, and Amir Zeldes. 2024. *GDTB: Genre diverse data for English shallow discourse parsing across modalities, text types, and domains.* EMNLP 2024.

[9] Amir Zeldes. 2017. *The gum corpus: Creating multilayer resources in the classroom.* Language Resources and Evaluation.

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

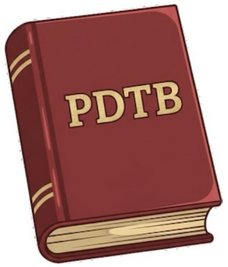
[11] Yingxue Fu. 2022. *Towards unification of discourse annotation frameworks.* ACL 2022 SRW.

[12] Florian Eichin, Yang Janet Liu, Barbara Plank, Michael A. Hedderich. 2025. Probing LLMs for Multilingual Discourse Generalization Through a Unified Label Set. ACL 2025.

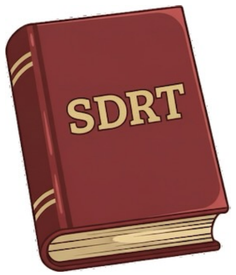
# Acknowledgements

We thank our anonymous reviewers for their time spent on reviewing our paper and their detailed feedback, which greatly helped us refine our work. We also thank several colleagues at National University of Singapore (NUS) for research discussions and proofreading of our drafts, especially Barid Xi Ai, Shumin Deng, Yajing Yang, Tongyao Zhu, Mahardika Krisna Ihsani, Xuan Long Do, and Xinyuan Lu. We appreciate Joseph Miller, Bilal Chughtai, and William Saunders for open sourcing their software and Neel Nanda's blog that guided the first author into mechanistic interpretability. We would also like to acknowledge a grant from National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-GC-2022-005).

# Image Credits








**Baby dragons** are generated with Google Gemini's image generation models.

**Discourse books** are also generated with Google Gemini's image generation models.

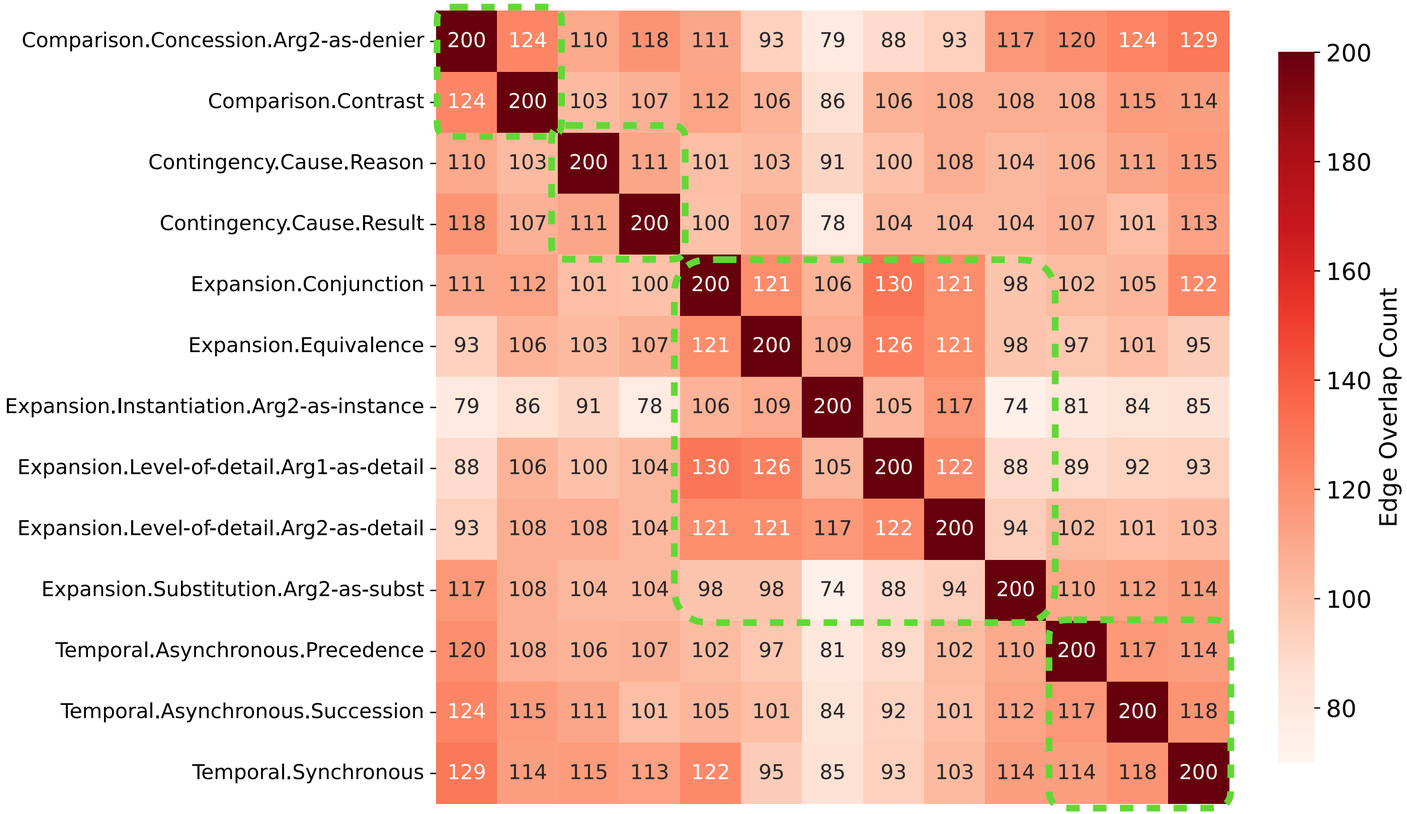
**Discursive circuits** are generated with Joseph Miller, Bilal Chughtai, and William Saunders's software auto-circuit.

**QR codes** are generated with QRCode Monkey.

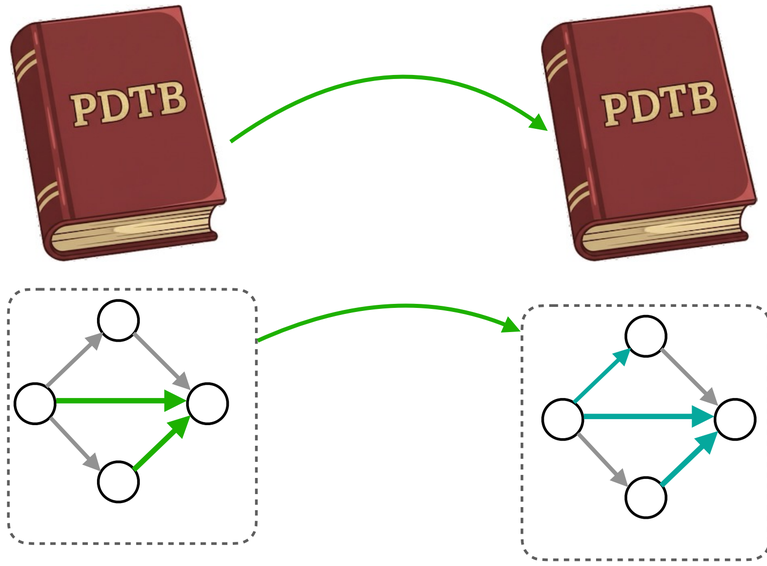
# Supplementary Slides

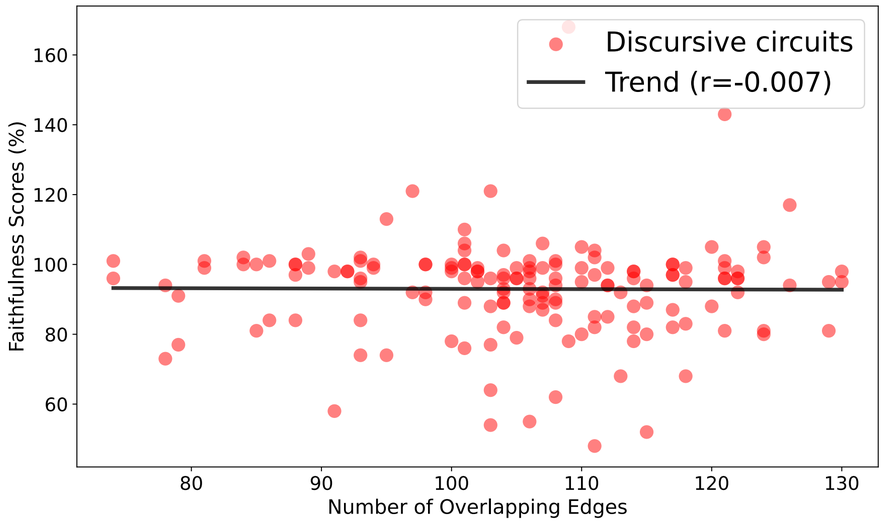
# Evaluation

## RQ2: Generalization



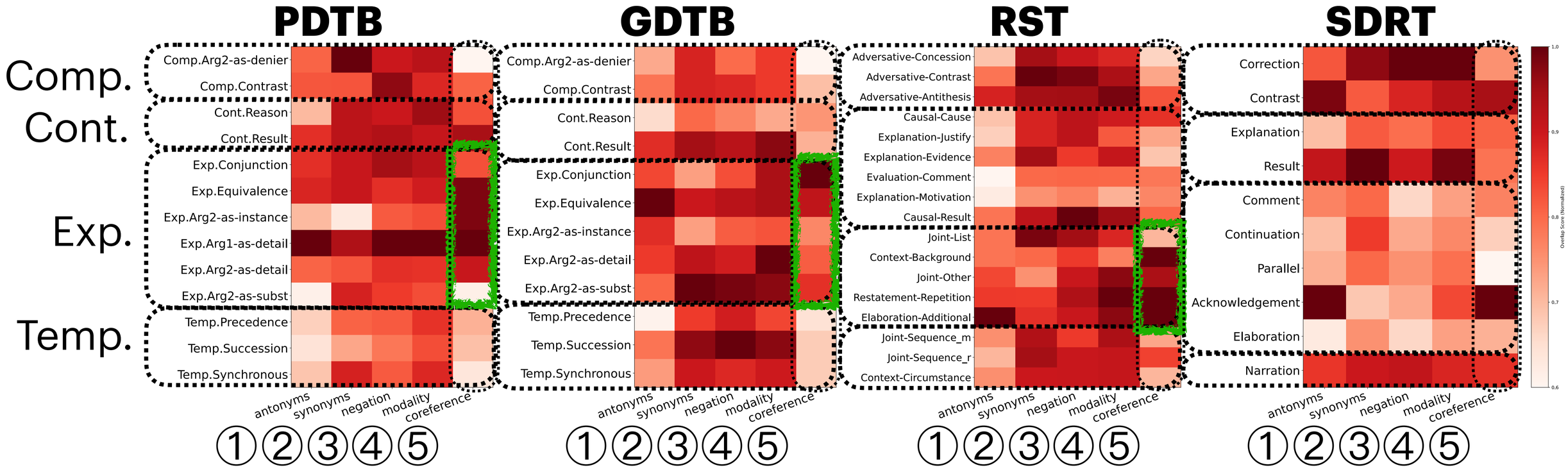

denotes L1 clusters



# Evaluation

## RQ3: Composition of linguistic features



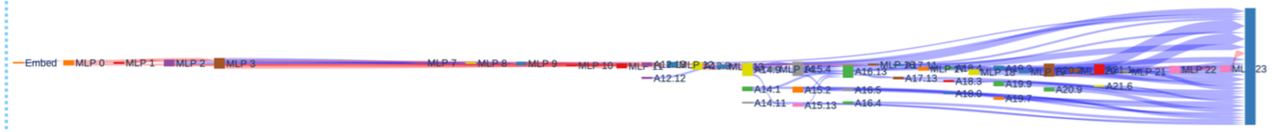

We consider (1) antonyms, (2) synonyms; (3) negation; (4) modality; (5) coreference.

- Column-wise normalization (all columns have a **dark cell** )
- Green boundaries denote coreference-heavy zone.

# Examples of Discursive Circuits

**(1) PDTB: Contingency.Cause.Result**

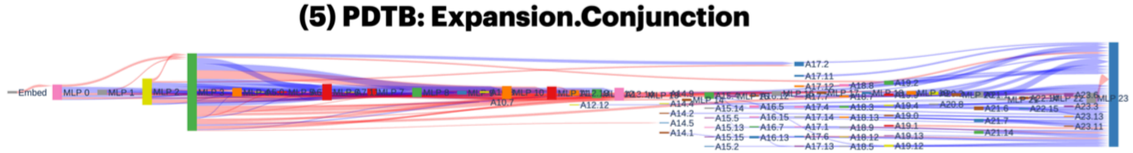
**(2) GDTB: Contingency.Cause.Result**



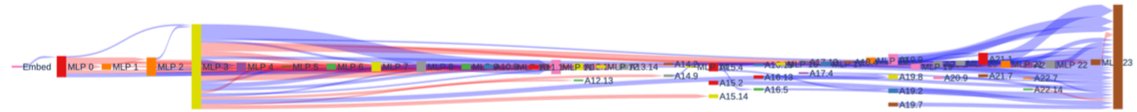
**(3) RST: Cause-Result**

**(4) SDRT: Result**

**(5) PDTB: Expansion.Conjunction**



**(6) PDTB: Expansion.Equivalence**



# Taylor Expansion Details

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

$$= L(z_u^{ori}) - L(z_u^{cf}) \tag{2}$$

$$\approx L(z_u^{cf}) + \nabla_{z_u} L(x_{cf})^\top (z_u^{ori} - z_u^{cf}) - L(z_u^{cf}) \tag{3}$$

$$= (z_u^{ori} - z_u^{cf})^\top \nabla_{z_u} L(x_{cf}) \tag{4}$$