Discursive Circuits:
How Do Language Models Understand Discourse Relations?

Yisong Miao Min-Yen Kan

Want to mechanistically interpret how LMs understand discourse relations? We propose a new CuDR task to discover sparse computation graphs for discourse relations.

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

he goes to the canteen | the canteen is closed

Original run: Bob is hungry, so — P("he") - P("the") = + 0.9

Counterfactual run: Bob is hungry, but — P("he") - P("the") = – 0.8

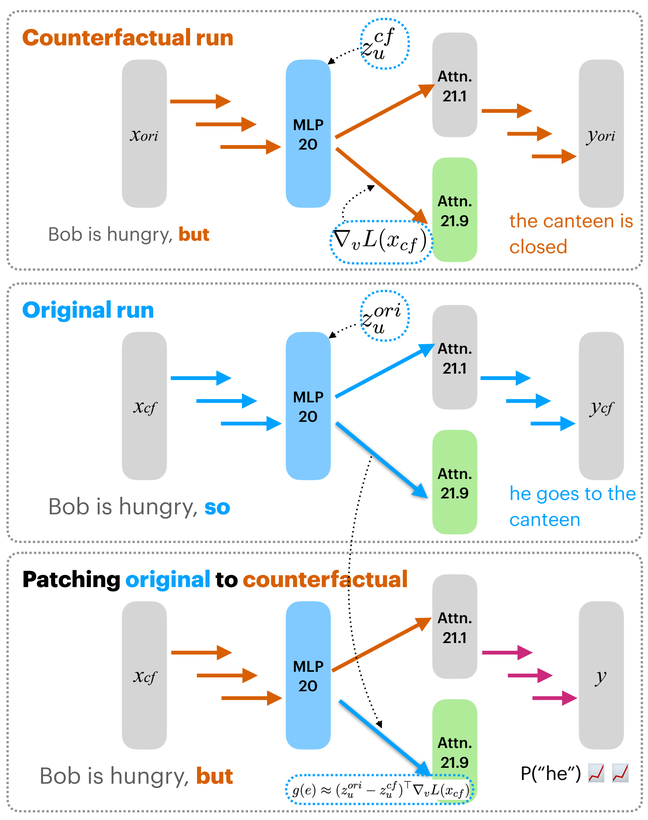
Patching original to counterfactual: Bob is hungry, but — P("he") - P("the") = – 0.7

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 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]$_{Conn}$ ⇒ [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]$_{Conn'}$ ⇒ [the canteen is closed]$_{Arg_2'}$

Task Formalization: Completion under Discourse Relation (CuDR)



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

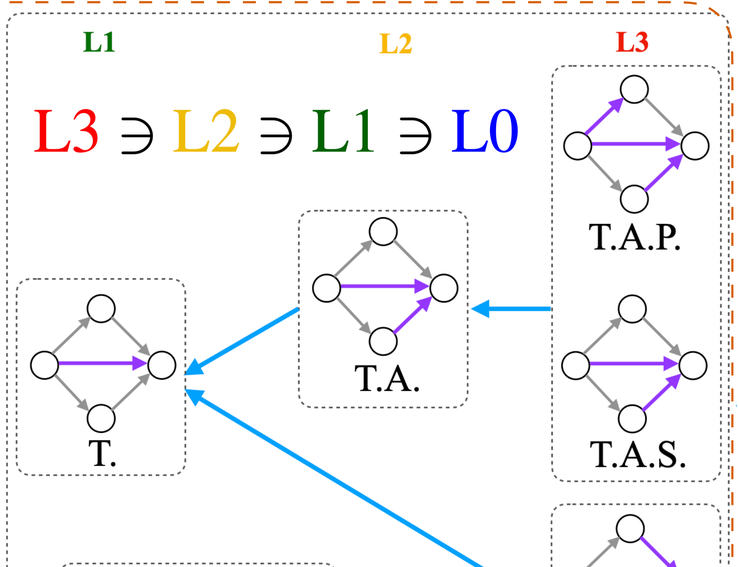
Estimates the causal effect of any edge.

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

Accelerate by Taylor approximation

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.)

L0 (137)



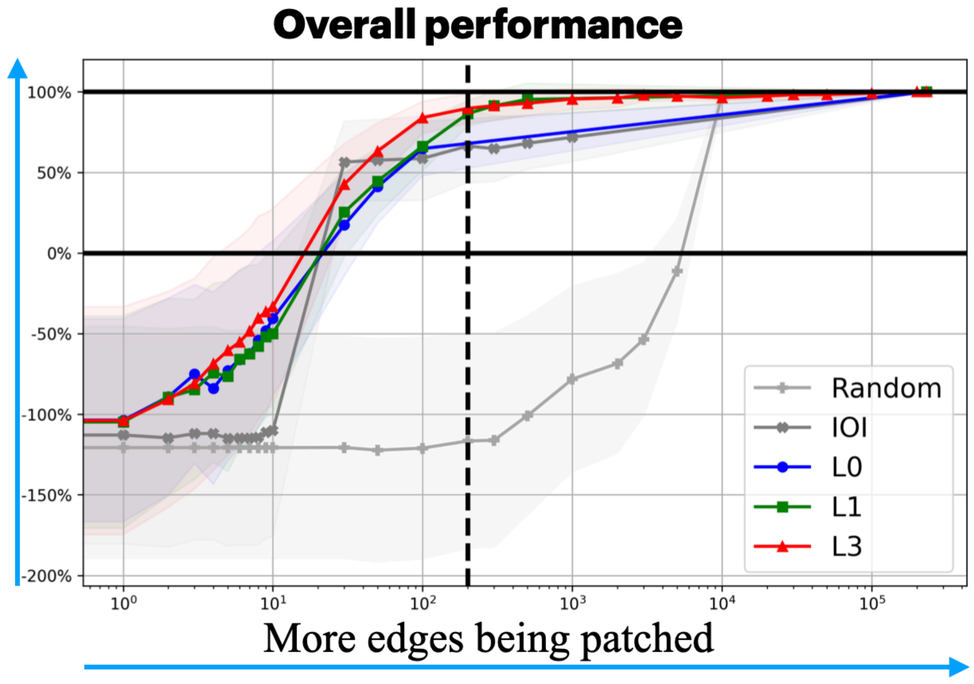
Discursive circuits generate a new hierarchy of discourse using model's internal representations.

Discursive circuits are discovered by edge attribution patching with CuDR data.

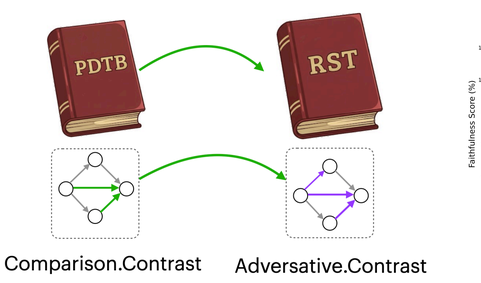
Evaluated metric: **Logit difference** $\Delta L = L(Arg_2) - L(Arg_2')$

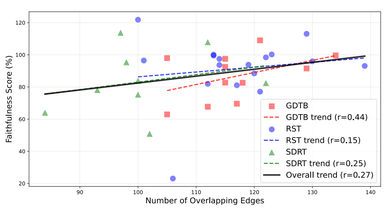
$\frac{\Delta L_{\text{patch}}}{\Delta L_{\text{full}}}$ **Normalized faithfulness score** (to the full model)

Datasets: **PDTB, GDTB, RST, SDRT** (GPT-4o-mini was used to generate counterfactual CuDR data).

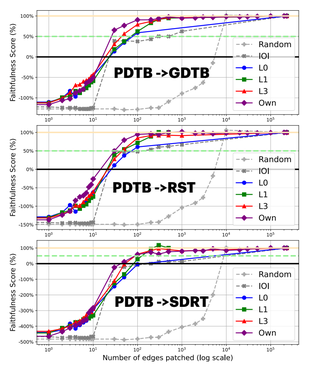


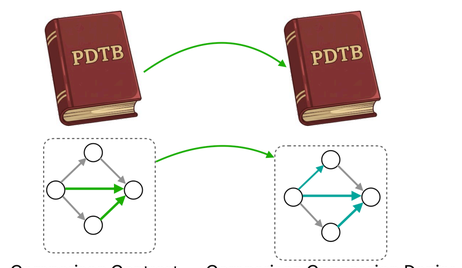
Overall performance (RQ1): **(1)** Discursive circuits are sparse. Faithfulness is recovered with around 200 edges (<0.2% of the full model). **(2)** L1~L3 circuits outperform baselines significantly.

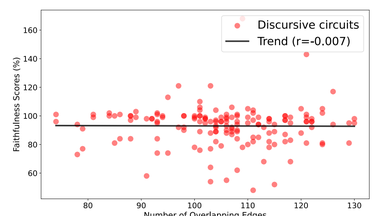




Faithfulness ∝ Overlap

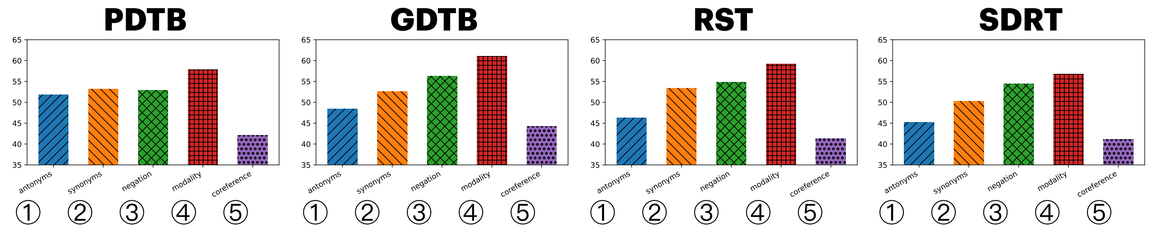






Overlap ⊭ Faithfulness

Generalization (RQ2): Discursive circuits show strong generalization across inter- and intra-discourse frameworks.



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

Composition of linguistic features (RQ3): Discursive circuits exhibit consistent utilization of linguistic features.

NUS | Computing
National University of Singapore
WING
Web Information Retrieval / Natural Language Processing Group
EMNLP 2025
Suzhou, China | 中国苏州