For What Reason? Interpreting Models' Encoding of Causation and Antithesis

Abhidip Bhattacharyya[1], Shira Wein[2]
[1]University of Massachusetts Amherst, [2] University of South Florida

SIGdial 2026 Atlanta, GA, USA

Introduction

Background

- **Discourse relations** structure documents and support language understanding, influencing model performance and AI reliability.
- **We investigate** how instruction-tuned Transformers (LLaMA, Mistral) encode discourse using interpretability techniques.
- **Focus:** Compare representations of **causation** (cause-effect) and **antithesis** (contrast).

- **Causation:** John studied hard, so he passed the exam.
- **Antithesis**: John studied hard, yet he failed the exam

Methods

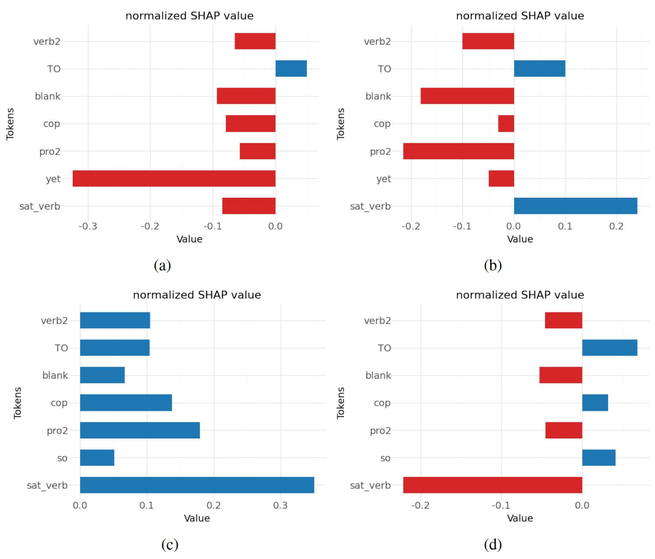
Data Collection

- **Paired dataset:** Causation–antithesis sentence pairs with matched word counts.
- **Data generation:** GPT-4o-generated pairs seeded with manual examples (130 train, 30 manually verified test).
- **Controlled design:** Introduce negative verbs so models cannot rely solely on discourse markers (*so*, *yet*).

- **Causation:** John studied hard, so he was able to pass the exam.
- **Antithesis**: John studied hard, yet he was unable to pass the exam

Results

RQ1: Which token-level features drive the model's decisions between antithesis and causation?



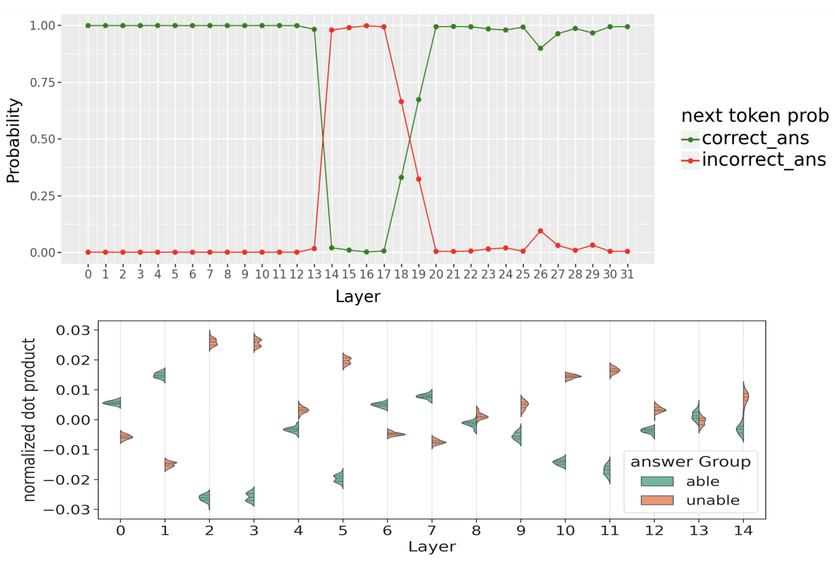
Method

- Computed **token-level SHAP values** to measure each token's contribution to the model's causation vs. antithesis prediction.
- Compared SHAP patterns across four conditions: **{so, yet} × {positive verb, negative verb}**.
- Performed **two-sample t-tests** to test whether token contributions differed significantly across conditions.

Key Findings

- Model primarily relies on the **discourse marker** and **verb polarity**.
- Their influence changes systematically with semantic context.
- We therefore patch activations at the **connector token** and **final token** in RQ2.

RQ2: Which layers play a pivotal role in identifying antithesis and causation?



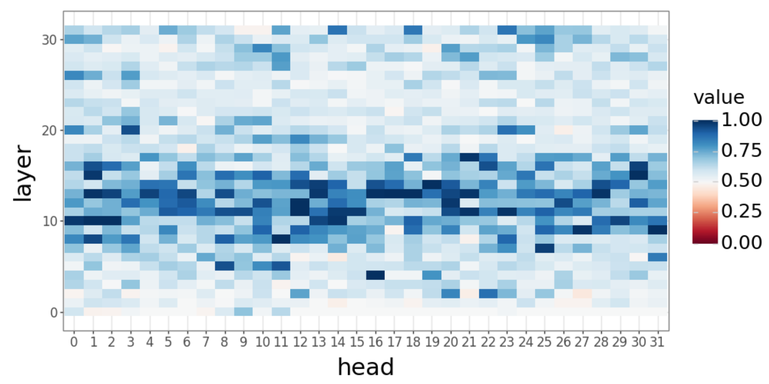
Method

- Perform **activation patching** at the discourse marker (*so/yet*) and the **final token**.
- Compare interventions at **attention, MLP, and block outputs** across layers.
- Use **logit attribution** to identify layers that contribute toward the correct prediction.

Key Findings

- **Attention and block outputs are primarily responsible** for distinguishing antithesis from causation.
- **Early–mid layers** encode local discourse-marker semantics, while **mid layers** integrate sentence-level meaning into the final decision.
- After this integration, **higher layers mainly propagate the decision**, making later interventions largely ineffective.

RQ3: What key concepts help the model distinguish between antithesis and causation?



Method

- Train logistic regression probes on attention-head representations.

Key Findings

- **Mid-layer attention heads** encode the strongest decision and semantic signals.
- The model explicitly represents **verb polarity** and **satellite-clause sentiment**, indicating these concepts support the final prediction.
- **Early heads** already encode predictive semantic information, while **later layers** contain consolidated decision representations.

RQ4: How does discursive information propagate through the model?

Method

- Apply **EAP-IG** to trace influential information pathways.
- Compare pathways across **shot settings** and **verb types**.

Key Findings

- Over half of the influential edges are shared across shot settings, indicating **stable information pathways**.

Conclusion

- **Token-level cues drive decisions:** Discourse markers (*so/yet*) and verb polarity provide key signals for distinguishing antithesis and causation.
- **Decision-making emerges in mid layers:** Attention and block representations in early-to-mid layers encode and integrate discursive meaning
- **Semantic concepts guide predictions:** The model represents verb and clause-level sentiment as important intermediate concepts.
- **Information follows stable pathways:** Discursive information propagates through consistent attention circuits, with prompting conditions influencing pathway stability.

Email:
- abhidipbhatt@umass.edu
- swein@amherst.edu