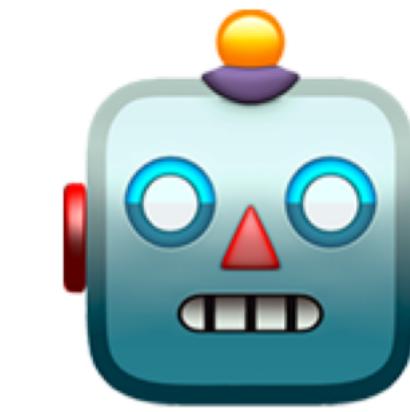
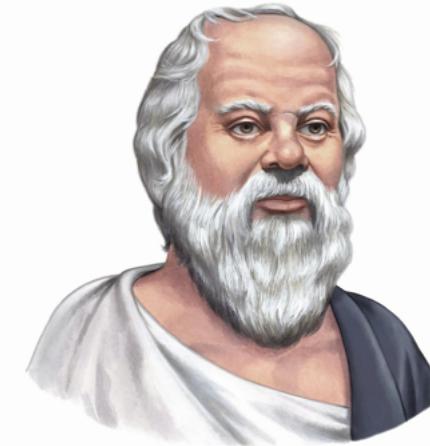


Discursive Socratic Questioning: Evaluating the Faithfulness of Language Models' Understanding of Discourse Relations



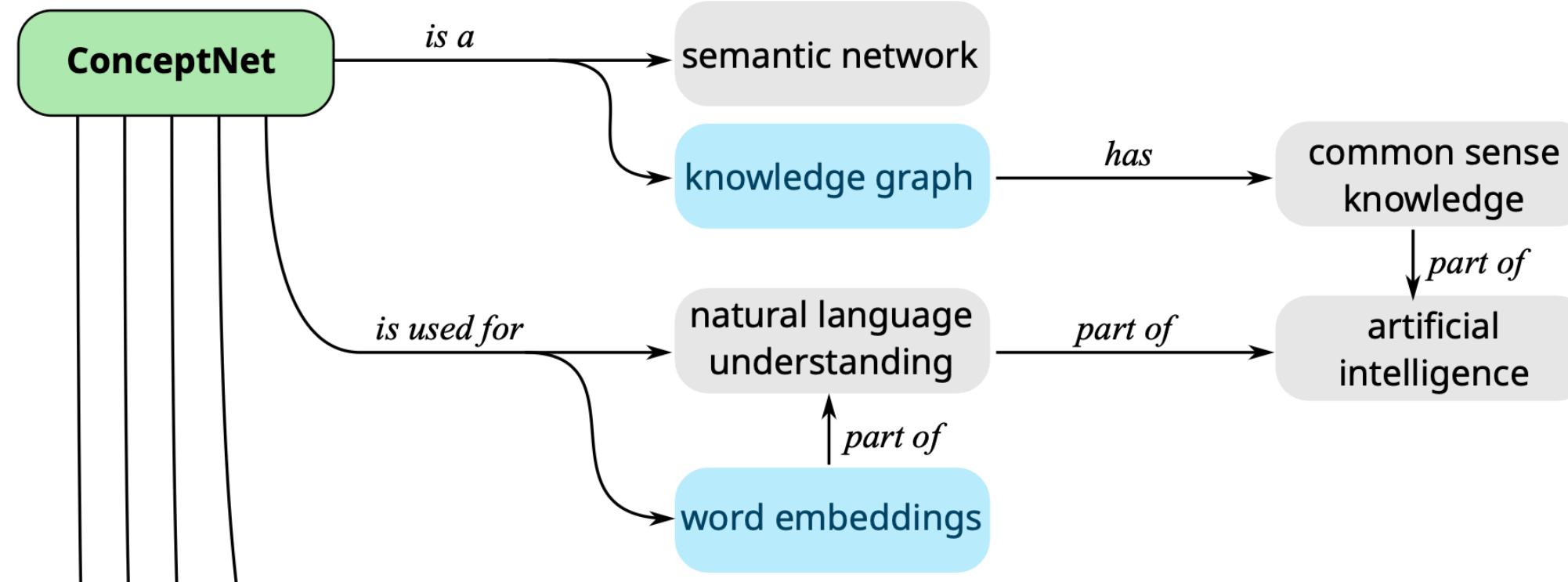
Yisong Miao, Hongfu Liu, Wenqiang Lei, Nancy F. Chen, Min-Yen Kan

<https://github.com/YisongMiao/DiSQ-Score>

Dataset and code @ Github

What is discourse semantics?

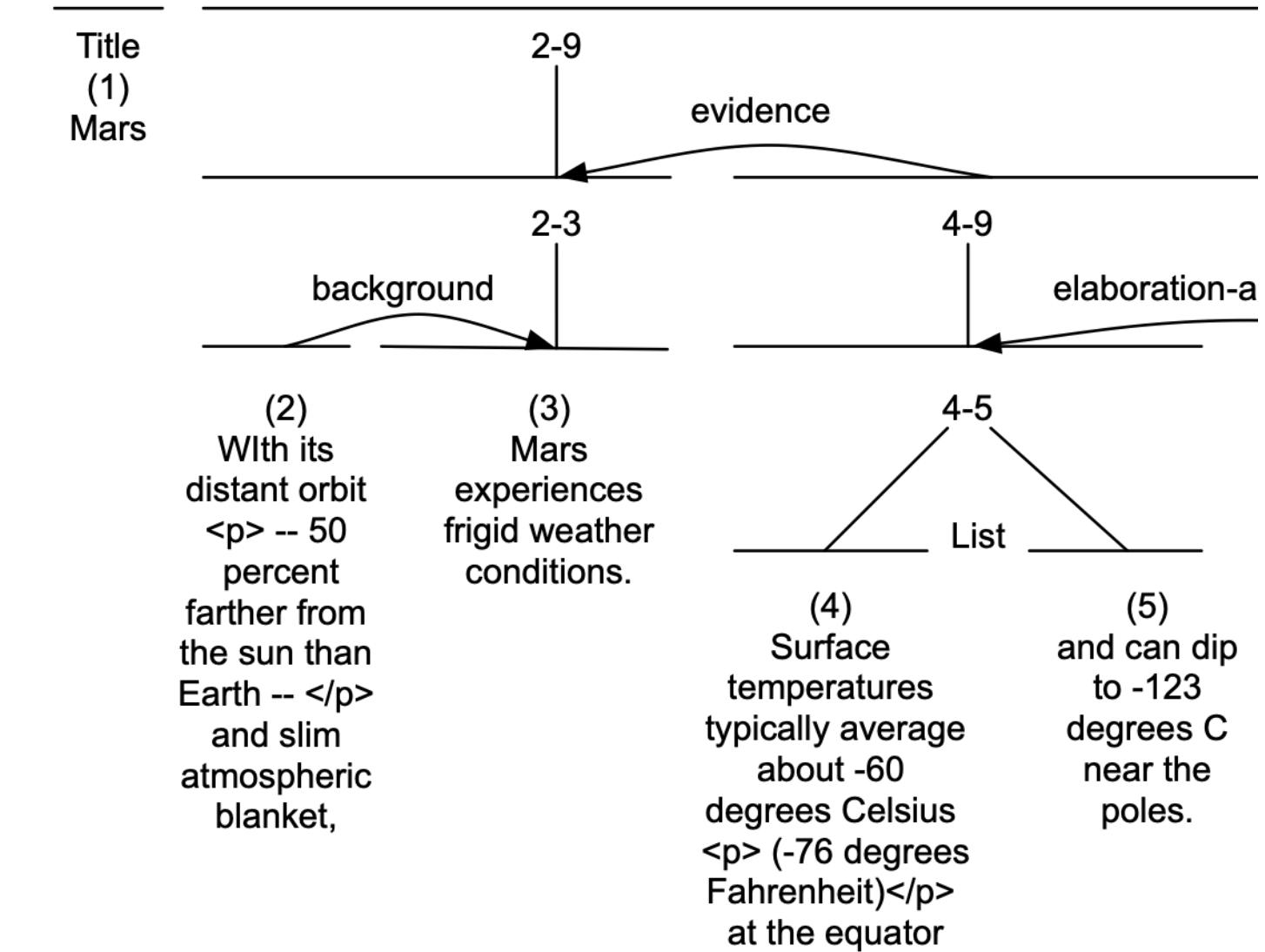
Lexical Semantics



Ontology

Relationships between words and phrases;
Non-contextual.

Discourse Semantics

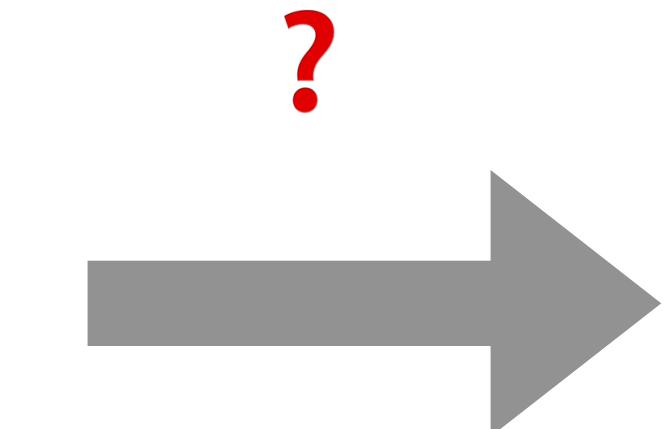
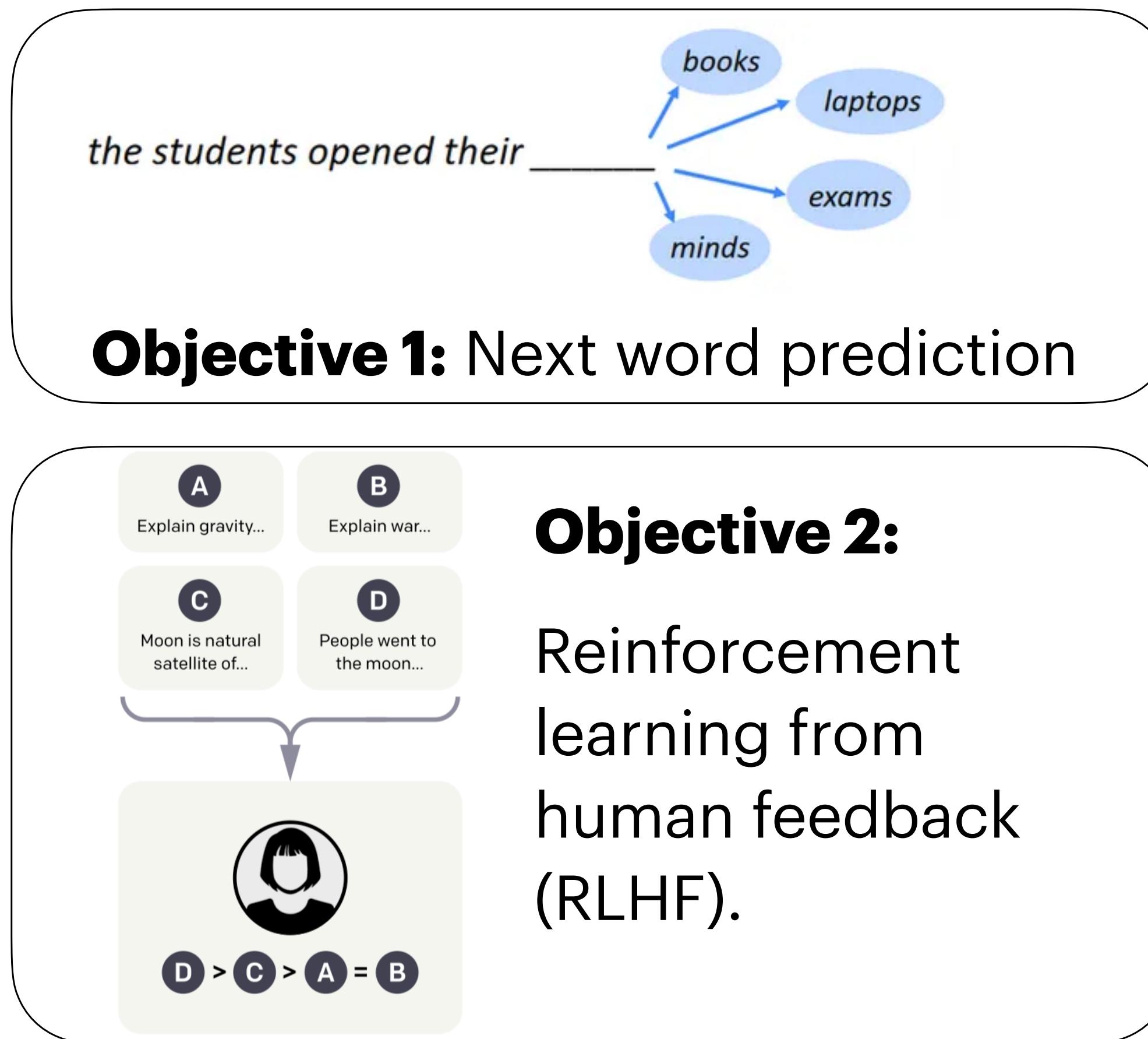


Treebanks

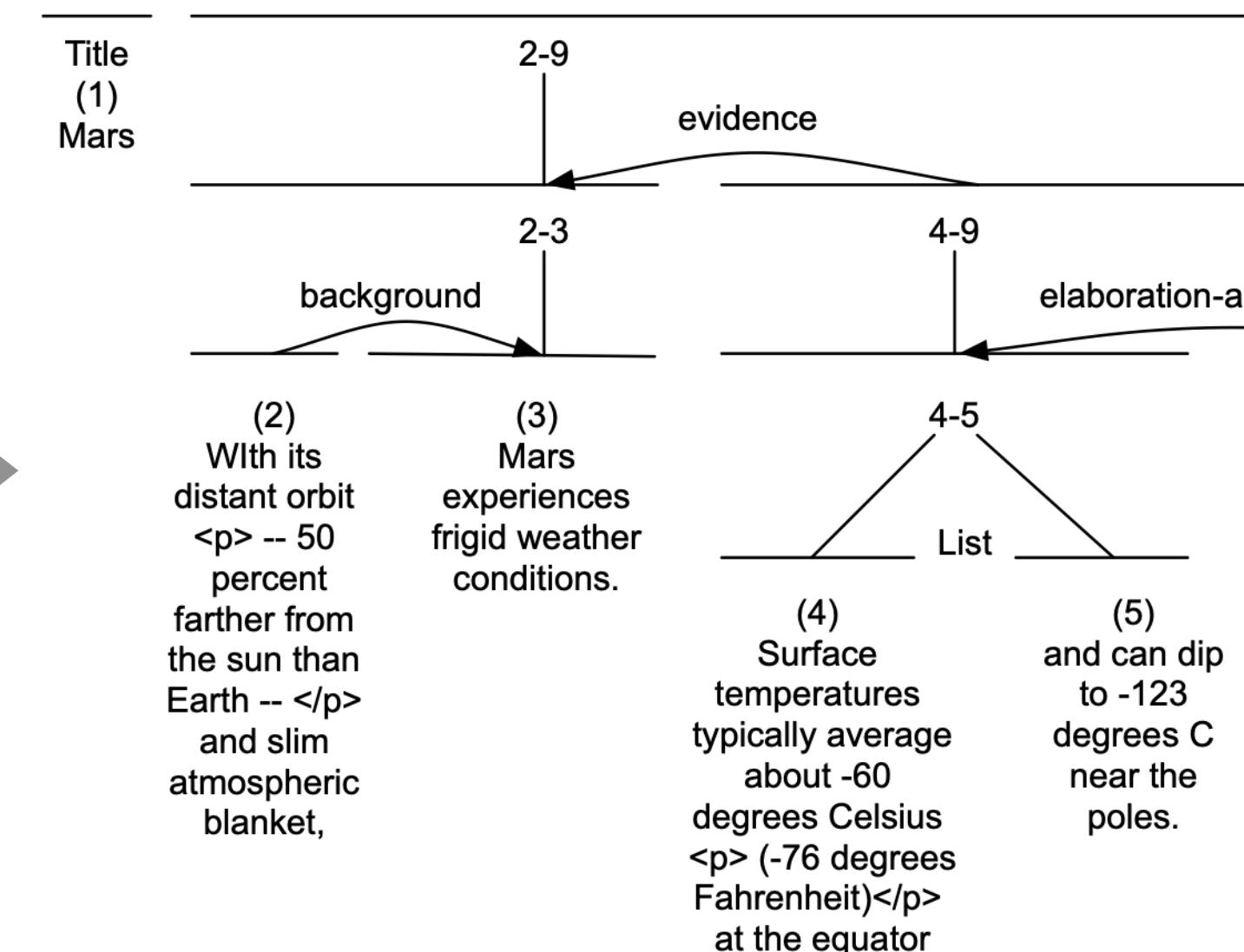
Larger nested units;
Depend on context.

The big question

Do LLMs Understand Discourse?



?



Discourse understanding

Existing evaluations

Discourse relation: Contingency.Cause.Result

Arg1: When I want to buy, they run from you -- *they keep changing their prices.*

Arg2: *It's very frustrating.*

| | | |
|-----------|-----------------|--------------------------------------|
| Expansion | Conjunction | -- |
| | Disjunction | -- |
| | Equivalence | -- |
| | Instantiation | Arg1-as-instance Arg2-as-instance |
| | Level-of-detail | Arg1-as-detail Arg2-as-detail |
| | Substitution | Arg1-as-subst Arg2-as-subst |
| | Exception | Arg1-as-excpt Arg2-as-excpt |
| | Manner | Arg1-as-manner Arg2-as-manner |
| | | |
| | | |

The two layer relation taxonomy in PDTB 3.0.

- **Classification task** on a hierarchical taxonomy
- **Metrics:** Acc / F1.
- **Acc / F1 are not suitable for the evaluation for LLMs:**
 - Prompts carry randomness.
 - Only one-off predictions. Cannot measure the faithfulness of the prediction.

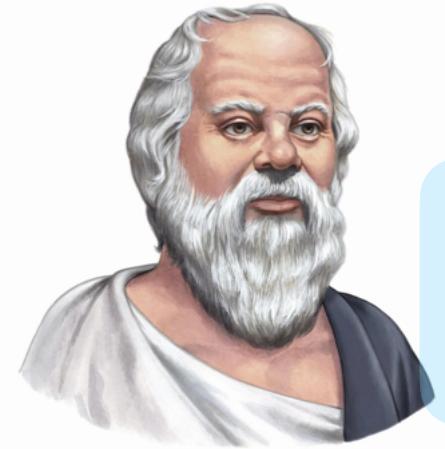
Source: [Discourse Annotation in the PDTB: The Next Generation](#)
@ ACL-ISO Workshop '18

| Method | Top | | Second | |
|---------------------------|-------|-------|--------|-------|
| | F1 | Acc | F1 | Acc |
| Random | 24.74 | 25.47 | 6.48 | 8.78 |
| Liu et al. (2020) | 63.39 | 69.06 | 35.25 | 58.13 |
| Jiang et al. (2022) | 65.76 | 72.52 | 41.74 | 61.16 |
| Long and Webber (2022) | 69.60 | 72.18 | 49.66 | 61.69 |
| Chan et al. (2023b) | 70.84 | 75.65 | 49.03 | 64.58 |
| ChatGPT _{Prompt} | 29.85 | 32.89 | 9.27 | 15.59 |
| ChatGPT _{PE} | 33.78 | 34.94 | 10.73 | 20.31 |
| ChatGPT _{ICL} | 36.11 | 44.18 | 16.20 | 24.54 |

F1 and Accuracy scores are reported in most papers.

Source: [ChatGPT Evaluation on Sentence Level Relations: A Focus on Temporal, Causal, and Discourse Relations](#) @EACL '24

Socratic Questioning



Why?

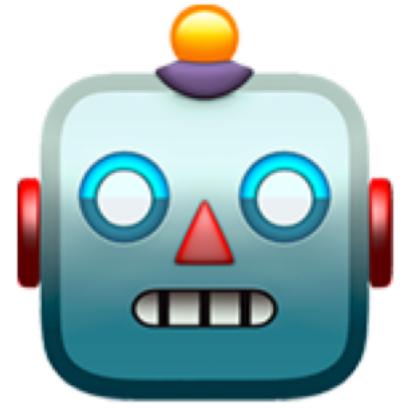
When will it rain?

When will you come back home?

I need to take an umbrella.

Because it's going to rain.

9PM.



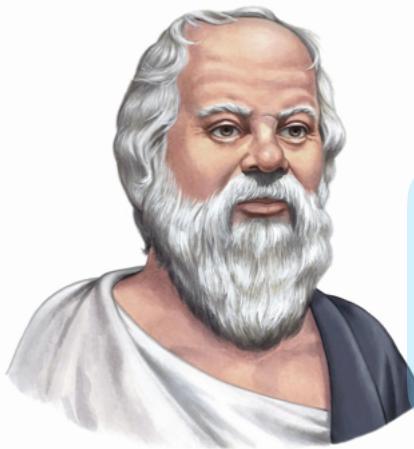
Socratic method is to ask a series of questions to challenge thoughts, clarify ideas and deepen understandings.

Socratic Questioning

Discourse relation: Contingency.Cause.Result

Arg1: When I want to buy, they run from you -- they keep changing their prices.

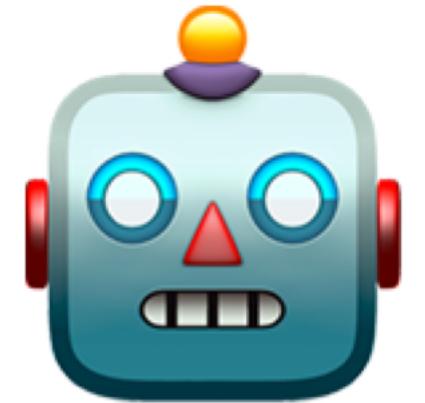
Arg2: It's very frustrating.



Why?

Can we comprehend them as other relations?

I think it's a Contingency discourse.

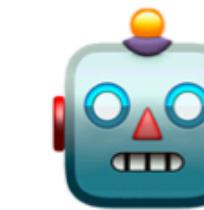
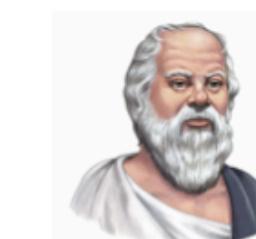


There is a cause-result event pair.

...

Socratic method is to ask a series of questions to challenge thoughts, clarify ideas and deepen understandings.

Discursive Socratic Questioning (DiSQ)



Is “they keep changing their prices” **a reason for** “it’s very frustrating”?

Ground-Truth Answer: True

Model’s Answer: True

Targeted Score = 1

Is “they keep changing their prices” **contrasted with** “it’s very frustrating”?

Ground-Truth Answer: False

Model’s Answer: True

Counterfactual Score = 0

Is “it’s very frustrating” **the result of** “they keep changing their prices”?

Ground-Truth Answer: True

Model’s Answer: False

Consistency Score = 0

DiSQ is composed of three scores to evaluate models’ faithfulness.

Discursive Socratic Questioning (DiSQ)

In this paper, we address:

- **What to ask?**
- **How to ask?**
- **How well do models answer?**

Preliminaries

What Counts as Discourse Understanding?

Discourse relation (R): Contingency.Cause.Result

Arg_1 : When I want to buy, they run from you – they keep changing their prices

Arg_2 : It's very frustrating

s_{11} : I want to buy;

s_{12} : they run from you;

s_{13} : they keep changing their prices

s_{21} : It's very frustrating

Salient signals: (s_{13}, s_{21}, r) , r is “the reason for”.

Targeted question: Is s_{13} the reason for s_{21} ?

Counterfactual question: Does s_{13} contrast against s_{21} ?

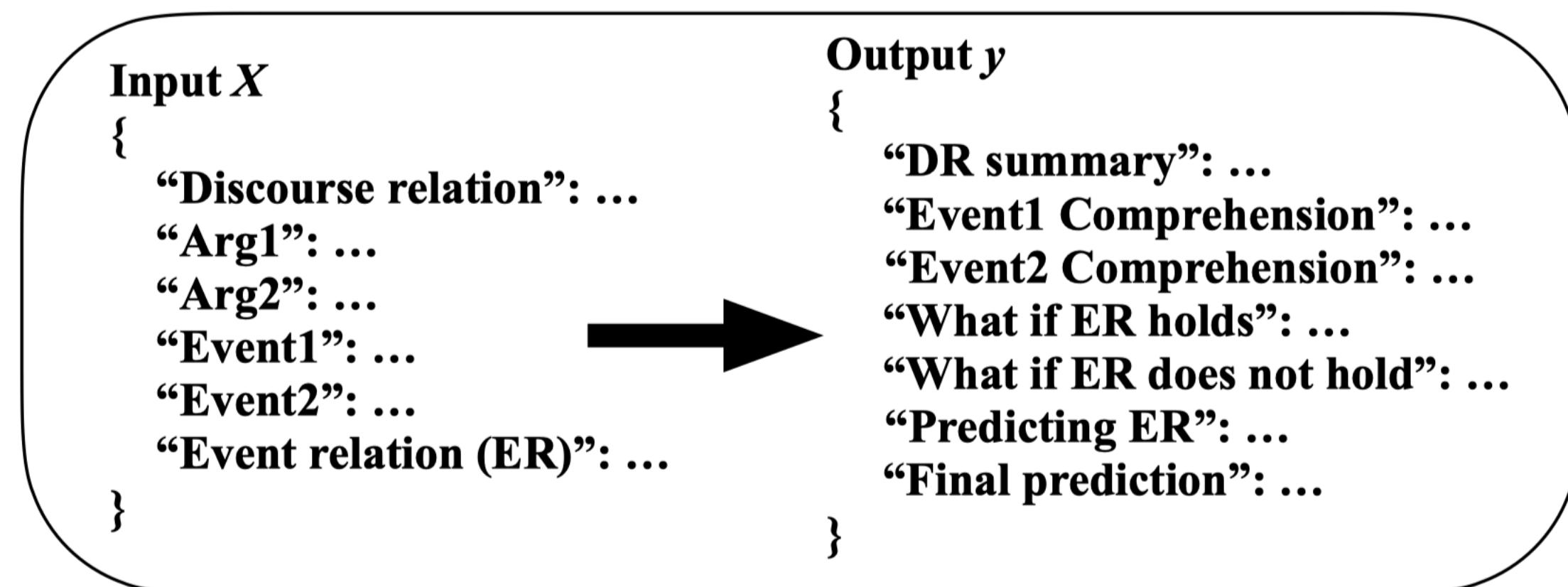
Converse question: Is s_{21} the result of s_{13} ?

We define salient signals as event pairs that highlight the discourse relation.

- **Responsiveness** to targeted questions.
- **Robustness** to counterfactual questions.
- **Consistency** across converse questions.

Preliminaries

Annotate Salient Signals



We employ in-context learning (ICL) as a semi-automatic approach:

- **Task:** Determine if event 1 and event 2 are salient signals for a discourse relation. JSON format in a chain-of-thought style.
- **Requirements:** One positive and one negative demonstration;
- **Hypothesis:** By informing a model the discourse relation, the model can do the reasoning for salient events.

Preliminaries

Annotation Outcome

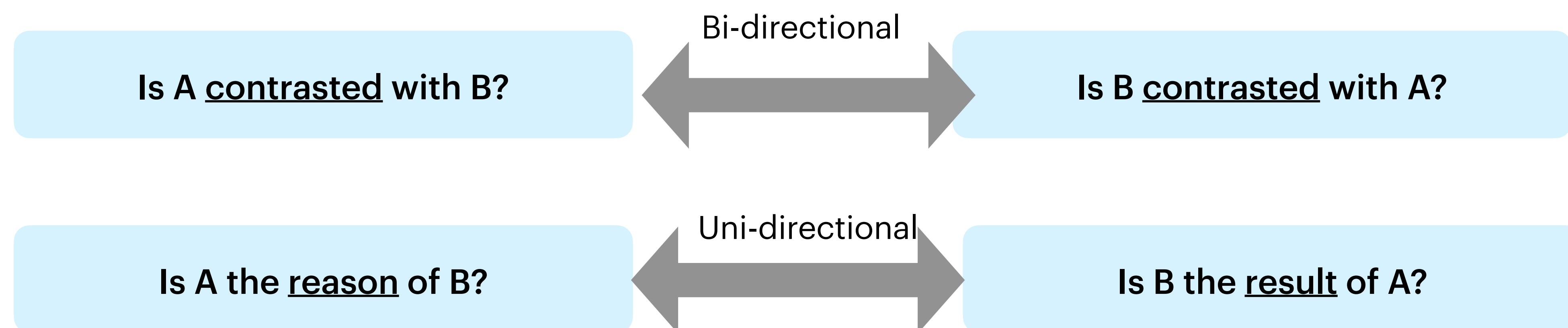
| Discourse relation (R) | Event relation (r) | Q Type | # of Q |
|----------------------------------|----------------------------------|--------|--------|
| Comparison.Concession | deny or contradict with | Bi- | 1,764 |
| Comparison.Congruence | contrast with | Bi- | 876 |
| Contingency.Reason | reason of | Uni- | 3,264 |
| Contingency.Result | result of | Uni- | 2,796 |
| Expansion.Conjunction | contribute to the same situation | Bi- | 4,596 |
| Expansion.Equivalence | equivalent to | Bi- | 420 |
| Expansion.Instantiation | example of | Uni- | 2,352 |
| Expansion.Level-of-detail | provide more detail about | Uni- | 3,888 |
| Expansion.Substitution | alternative to | Uni- | 216 |
| Temporal.Aynchronous | happen before/after | Uni- | 1,368 |
| Temporal.Synchronous | happen at the same time as | Bi- | 840 |
| Total | | | 22,380 |

| | A1&A2 | A1&ICL | A2&ICL |
|----------------------|-------|--------|--------|
| Agreements | 85.2% | 85.2% | 83.7% |
| Cohen's Kappa | 38.5% | 48.8% | 44.9% |
| Success Rate | / | 95.8% | 93.8% |

- **Annotation outcome:** Our dataset covers all eleven Level-2 relations in PDTB 3.0 taxonomy. We cover both uni- and bi-directional relations.
- **Human verification:** We employ humans to verify the correctness of event pairs determined as salient by the ICL method. The ICL between humans shows similar agreement levels to those between humans.

Discursive Socratic Questioning for Evaluation

| Type | Formalization | Expected Answer | Score |
|----------|---|------------------------|-----------|
| Targeted | $\mathcal{Q}_t = \{QG(s_1, s_2, r)\}$ | True | s_t |
| CF | $\mathcal{Q}_c = \{QG(s_1, s_2, r')\}$ | False | s_{cf} |
| Converse | $\tilde{\mathcal{Q}}_t = \{QG(s_2, s_1, \overleftarrow{r})\}$ | Equivalent to original | s_{con} |



Discursive Socratic Questioning for Evaluation

Discourse relation: Contingency.Cause.Result

Arg1: When I want to buy, they run from you -- they keep changing their prices.

Arg2: It's very frustrating.

Algorithm 1 DiSQ interrogates a language model.

```

1: Input: Discourse  $d$  and its corresponding questions  $\mathcal{Q}$ .
2:  $\mathcal{H} = \{\emptyset\}$                                  $\triangleright$  The history is initialized.
3: Stage 1: Targeted and Counterfactual QA
4: for  $q_i$  in  $\mathcal{Q}_t$  and  $\mathcal{Q}_c$  do
5:    $a_i = LM(q = q_i, c = d)$        $\triangleright$  The model performs
     QA. The context  $c$  is the discourse  $d$ .
6:    $\mathcal{H} \leftarrow (q_i, a_i)$            $\triangleright$  The history is updated.
7: end for
8: Stage 2: Converse QA
9: for  $(q_i, a_i)$  in  $\mathcal{H}$  do
10:    $\tilde{q} = Lookup(q, \{\tilde{\mathcal{Q}}_c, \tilde{\mathcal{Q}}_t\})$   $\triangleright$  Look up the converse
    question in converse question sets.
11:    $\tilde{a}_i = LM(q = \tilde{q}_i, c = d, (q_i, a_i) \in \mathcal{H})$   $\triangleright$  The model
    executes QA on the converse question,  $\tilde{q}_i$ , optionally
    utilizing the previous response  $(q_i, a_i)$  as supplemental
    context.
12:    $\mathcal{H} \leftarrow (\tilde{q}_i, \tilde{a}_i)$            $\triangleright$  The history is updated.
13: end for
14: Output:  $\mathcal{H}$ 
```

Discourse relation: Contingency
Targeted question:
Is A the result of with B?

Counterfactual question:
Is A contrasted with B?
Is A the example with B?
Is A an alternative of B?
...

Converse question:
(Given you answered A is the result
of B.) Is B the reason of A?

$$s_t = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[a_i = \text{True}], q_i \in \{\mathcal{Q}_t, \tilde{\mathcal{Q}}_t\} \quad (1)$$

$$s_{cf} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[a_i = \text{False}], q_i \in \{\mathcal{Q}_c, \tilde{\mathcal{Q}}_c\} \quad (2)$$

$$s_{con} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[a_i = \tilde{a}_i], q_i \in \mathcal{Q}, \tilde{q}_i \in \tilde{\mathcal{Q}} \quad (3)$$

$$s_{disq} = s_t \times s_{cf} \times s_{con}$$

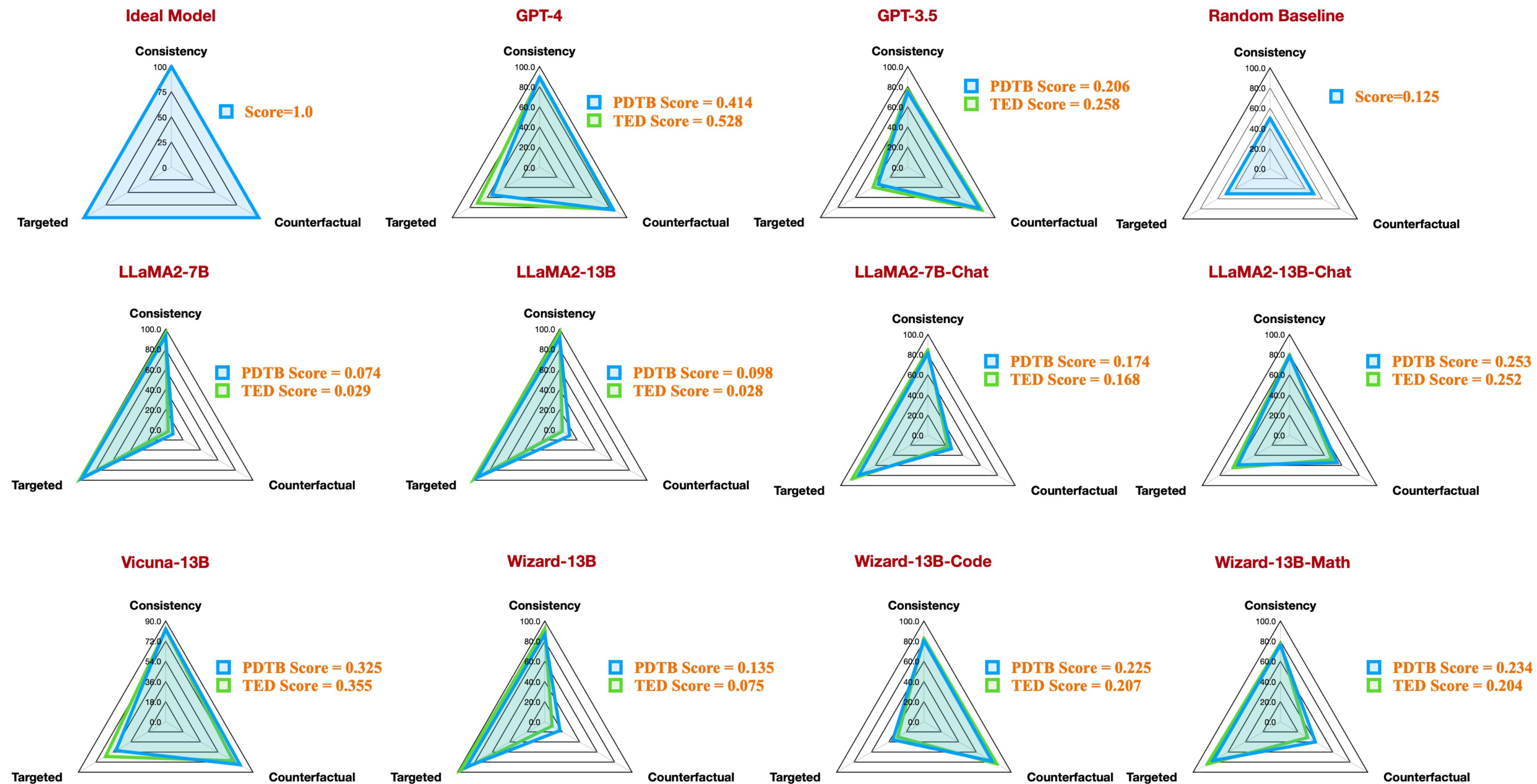
DiSQ Score is the multiplication of the three scores because we believe they are equally important (0.6, 0.6, 0.6) is better than (0.9, 0.9, 0).

Evaluation

Experiment Setup

- **Datasets:** **PDTB** (WSJ News) and **TED-MDB** corpus (also in PDTB discourse style). TED has 448 instances and 8,378 questions, about half the size of PDTB.
-  **Closed-source models:** **GPT-3.5 / GPT-4**.
-  **Open-source models:** **LLaMA-2** (with or without chat fine-tuning). **Vicuna** (further fine-tuned a LLaMA based on user interaction). **WizardLM** (complex instruction).
- **Zero-shot evaluation:** To mitigate the randomness from few-shot example selection, we adopt a zero-shot approach. We experiment with 4 different templates and select the best, to marginalize the impact of the templates.

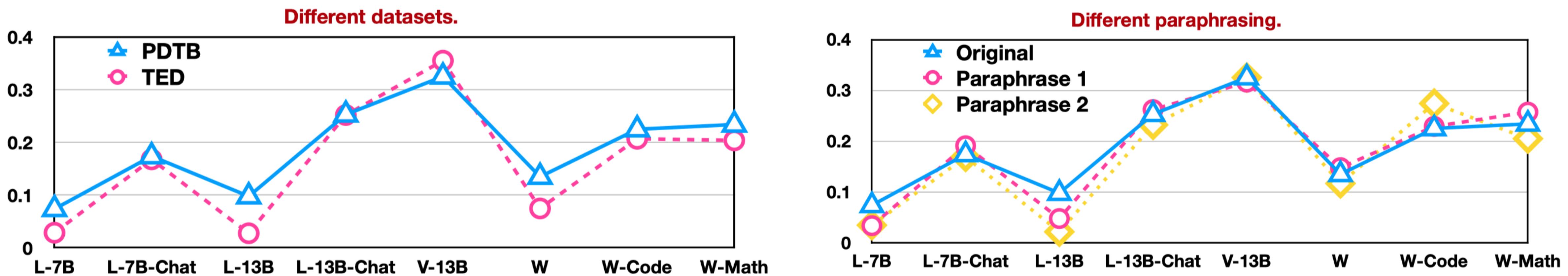
Evaluation



Findings: (1) Gap between Close and Open source models; (2) Improvements from further fine-tuning; (3) Consistency between two datasets.

Evaluation

RQ2: Consistency of DiSQ Scores



Operations: We compare DiSQ Scores in terms of different datasets (left) and different paraphrasing of the questions;

Findings: DiSQ Scores are consistent in terms of the variations.

Evaluation

RQ3: Impact of Discourse Relations on DiSQ Scores

| Models | Overall | Comp.Conce | Comp.Contra | Cont.Reason | Cont.Result | Exp.Conj | Exp.Equiv | Exp.Inst | Exp.Detail | Exp.Subst | Temp.Async | Temp.Sync |
|--------|----------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| PDTB | 1. Random Basline | 0.125 | 0.125 | 0.125 | 0.125 | 0.125 | 0.125 | 0.125 | 0.125 | 0.125 | 0.125 | 0.125 |
| | 2A. LLaMA2-7B | 0.074 | 0.029 | 0.083 | 0.094 | 0.095 | 0.076 | 0.056 | 0.087 | 0.067 | 0.156 | 0.035 |
| | 3A. LLaMA2-7B-Chat | 0.174 | 0.231 | 0.431 | 0.131 | 0.174 | 0.213 | 0.104 | 0.120 | 0.150 | 0.199 | 0.108 |
| | 4A. LLaMA2-13B | 0.098 | 0.037 | 0.100 | 0.082 | 0.097 | 0.127 | 0.101 | 0.113 | 0.107 | 0.086 | 0.084 |
| | 5A. LLaMA2-13B-Chat | 0.253 | 0.193 | 0.477 | 0.129 | 0.172 | 0.288 | 0.157 | 0.326 | 0.373 | 0.291 | 0.195 |
| | 6A. Vicuna-13B | 0.325 | 0.087 | 0.513 | 0.200 | 0.353 | 0.369 | 0.000 | 0.334 | 0.462 | 0.195 | 0.511 |
| | 7A. Wizard | 0.135 | 0.221 | 0.256 | 0.067 | 0.107 | 0.170 | 0.072 | 0.167 | 0.128 | 0.108 | 0.097 |
| | 8A. Wizard-Code | 0.225 | 0.032 | 0.268 | 0.175 | 0.287 | 0.121 | 0.008 | 0.283 | 0.329 | 0.174 | 0.545 |
| | 9A. Wizard-Math | 0.234 | 0.132 | 0.264 | 0.241 | 0.286 | 0.192 | 0.046 | 0.240 | 0.323 | 0.201 | 0.240 |
| | 10A. GPT-3.5 | 0.206 | 0.151 | 0.278 | 0.082 | 0.161 | 0.246 | 0.067 | 0.257 | 0.262 | 0.232 | 0.388 |
| TED | 11A. GPT-4 | 0.414 | 0.053 | 0.567 | 0.119 | 0.351 | 0.610 | 0.192 | 0.659 | 0.481 | 0.422 | 0.692 |
| | 2B. LLaMA2-7B | 0.029 | 0.011 | 0.021 | 0.053 | 0.047 | 0.024 | 0.037 | 0.027 | 0.040 | 0.037 | 0.018 |
| | 3B. LLaMA2-7B-Chat | 0.168 | 0.182 | 0.315 | 0.139 | 0.160 | 0.184 | 0.105 | 0.089 | 0.124 | 0.099 | 0.167 |
| | 4B. LLaMA2-13B | 0.028 | 0.001 | 0.044 | 0.020 | 0.017 | 0.034 | 0.029 | 0.053 | 0.030 | 0.000 | 0.027 |
| | 5B. LLaMA2-13B-Chat | 0.252 | 0.196 | 0.492 | 0.141 | 0.175 | 0.260 | 0.211 | 0.376 | 0.369 | 0.284 | 0.203 |
| | 6B. Vicuna-13B | 0.355 | 0.098 | 0.551 | 0.241 | 0.387 | 0.386 | 0.064 | 0.509 | 0.508 | 0.219 | 0.448 |
| | 7B. Wizard | 0.075 | 0.100 | 0.159 | 0.028 | 0.075 | 0.063 | 0.063 | 0.120 | 0.100 | 0.063 | 0.062 |
| | 8B. Wizard-Code | 0.207 | 0.015 | 0.330 | 0.114 | 0.269 | 0.163 | 0.000 | 0.360 | 0.254 | 0.096 | 0.616 |
| | 9B. Wizard-Math | 0.204 | 0.126 | 0.240 | 0.228 | 0.258 | 0.165 | 0.062 | 0.302 | 0.303 | 0.208 | 0.224 |
| | 10B. GPT-3.5 | 0.258 | 0.159 | 0.562 | 0.132 | 0.181 | 0.240 | 0.326 | 0.350 | 0.500 | 0.089 | 0.346 |
| | 11B. GPT-4 | 0.528 | 0.061 | 0.688 | 0.238 | 0.481 | 0.652 | 0.593 | 0.652 | 0.403 | 0.314 | 0.812 |

Findings: (1) Persistent challenges on minority classes; (2) Stronger open source models are on par with GPT-3.5

Evaluation

RQ3: Impact of Discourse Relations on DiSQ Scores

| | <i>Exp.Inst.Arg2-as-instance</i> | <i>Exp.Detail.Arg1-as-detail</i> | <i>Exp.Detail.Arg2-as-detail</i> | <i>Exp.Subst.Arg2-as-subst</i> | <i>Cont.Result</i> | <i>Cont.Reason</i> | <i>Comp.Conc.Arg1-as-denier</i> | <i>Comp.Conc.Arg2-as-denier</i> | <i>Temp.Async.Prec</i> | <i>Temp.Async.Succ</i> |
|------------------------|----------------------------------|----------------------------------|----------------------------------|--------------------------------|--------------------|--------------------|---------------------------------|---------------------------------|------------------------|------------------------|
| LLaMA2-7B | 0.087 | 0.07 | 0.066 | 0.156 | 0.095 | 0.094 | 0.005 | 0.032 | 0.037 | 0.009 |
| LLaMA2-7B-Chat | 0.12 | 0.067 | 0.158 | 0.199 | 0.174 | 0.131 | 0.149 | 0.239 | 0.116 | 0.025 |
| LLaMA2-13B | 0.113 | 0.116 | 0.107 | 0.086 | 0.097 | 0.082 | 0.037 | 0.037 | 0.085 | 0.076 |
| LLaMA2-13B-Chat | 0.326 | 0.289 | 0.383 | 0.291 | 0.172 | 0.129 | 0.155 | 0.197 | 0.203 | 0.122 |
| Vicuna-13B | 0.334 | 0.273 | 0.487 | 0.195 | 0.353 | 0.2 | 0.048 | 0.091 | 0.53 | 0.354 |
| Wizard | 0.167 | 0.132 | 0.128 | 0.108 | 0.107 | 0.067 | 0.22 | 0.22 | 0.102 | 0.043 |
| Wizard-Code | 0.283 | 0.269 | 0.335 | 0.174 | 0.287 | 0.175 | 0.053 | 0.03 | 0.558 | 0.417 |
| Wizard-Math | 0.24 | 0.405 | 0.314 | 0.201 | 0.286 | 0.241 | 0.161 | 0.128 | 0.248 | 0.143 |

Arg1 is the detail.

Arg2 is the detail.

Findings: There are task difficulty asymmetries in converse relations.

Evaluation

RQ4: Impact of Linguistic Features

Discourse relation: Contingency.Cause.Result

Arg1: When I want to buy, they run from you -- *they keep changing their prices.*

Arg2: *It's very frustrating.*

"so"

Previous context

Discourse relation: Contingency.Cause.Result

Arg1: When I want to buy, they run from you -- *they keep changing their prices.*

Arg2: *It's very frustrating.*

Subsequent context

| | Models | Overall | Exp. | Cont. | Comp. | Temp. |
|----------|------------------------|---------|-------|-------|-------|-------|
| w/o Conn | LLaMA2-13B-Chat | 0.253 | 0.322 | 0.147 | 0.293 | 0.149 |
| | Vicuna-13B | 0.325 | 0.374 | 0.262 | 0.220 | 0.357 |
| | Wizard-Code | 0.225 | 0.220 | 0.223 | 0.101 | 0.382 |
| w/ Conn | LLaMA2-13B-Chat | 0.273 | 0.328 | 0.187 | 0.329 | 0.149 |
| | Vicuna-13B | 0.396 | 0.418 | 0.382 | 0.289 | 0.401 |
| | Wizard-Code | 0.264 | 0.249 | 0.275 | 0.192 | 0.376 |

| | Models | Overall | Exp. | Cont. | Comp. | Temp. |
|-------------|------------------------|---------|-------|-------|-------|-------|
| w/o Context | LLaMA2-13B-Chat | 0.253 | 0.322 | 0.147 | 0.293 | 0.149 |
| | Vicuna-13B | 0.325 | 0.374 | 0.262 | 0.220 | 0.357 |
| | Wizard-Code | 0.225 | 0.22 | 0.223 | 0.101 | 0.382 |
| w/ Context | LLaMA2-13B-Chat | 0.311 | 0.402 | 0.231 | 0.186 | 0.169 |
| | Vicuna-13B | 0.369 | 0.424 | 0.333 | 0.192 | 0.380 |
| | Wizard-Code | 0.253 | 0.245 | 0.273 | 0.152 | 0.331 |

DiSQ Scores +0.04 with the help of discourse connective.

DiSQ Scores +0.05 with the help of the context (previous and subsequent one paragraph).

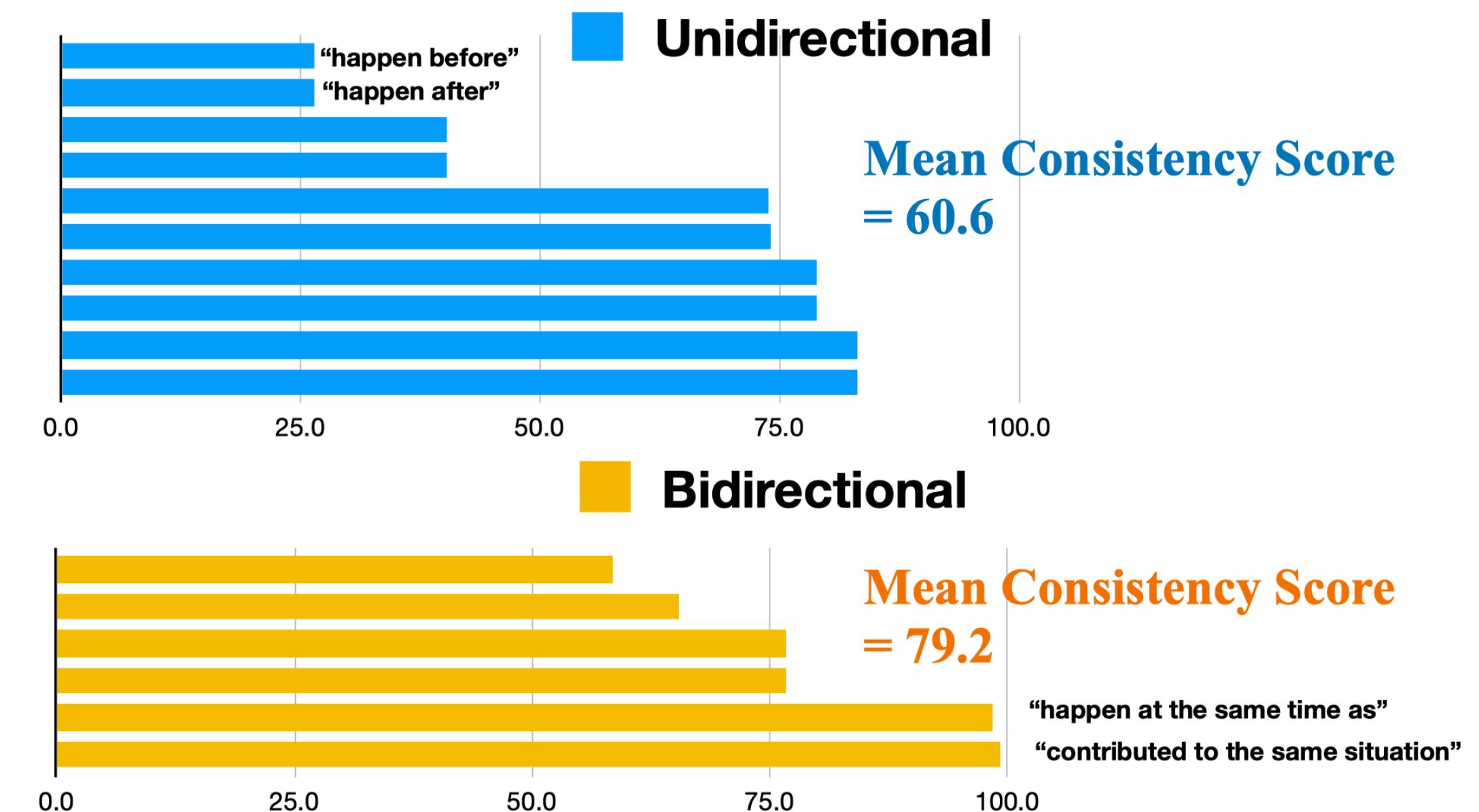
Evaluation

RQ4: Impact of Linguistic Features

| | w/o history | w/ history |
|------------------------|-------------|------------|
| LLaMA2-13B-Chat | 78.6 | 70.1 |
| Vicuna-13B | 82.8 | 88.7 |
| Wizard-Code | 81.6 | 99.8 |

Models' consistency score with the help of previous QA history.

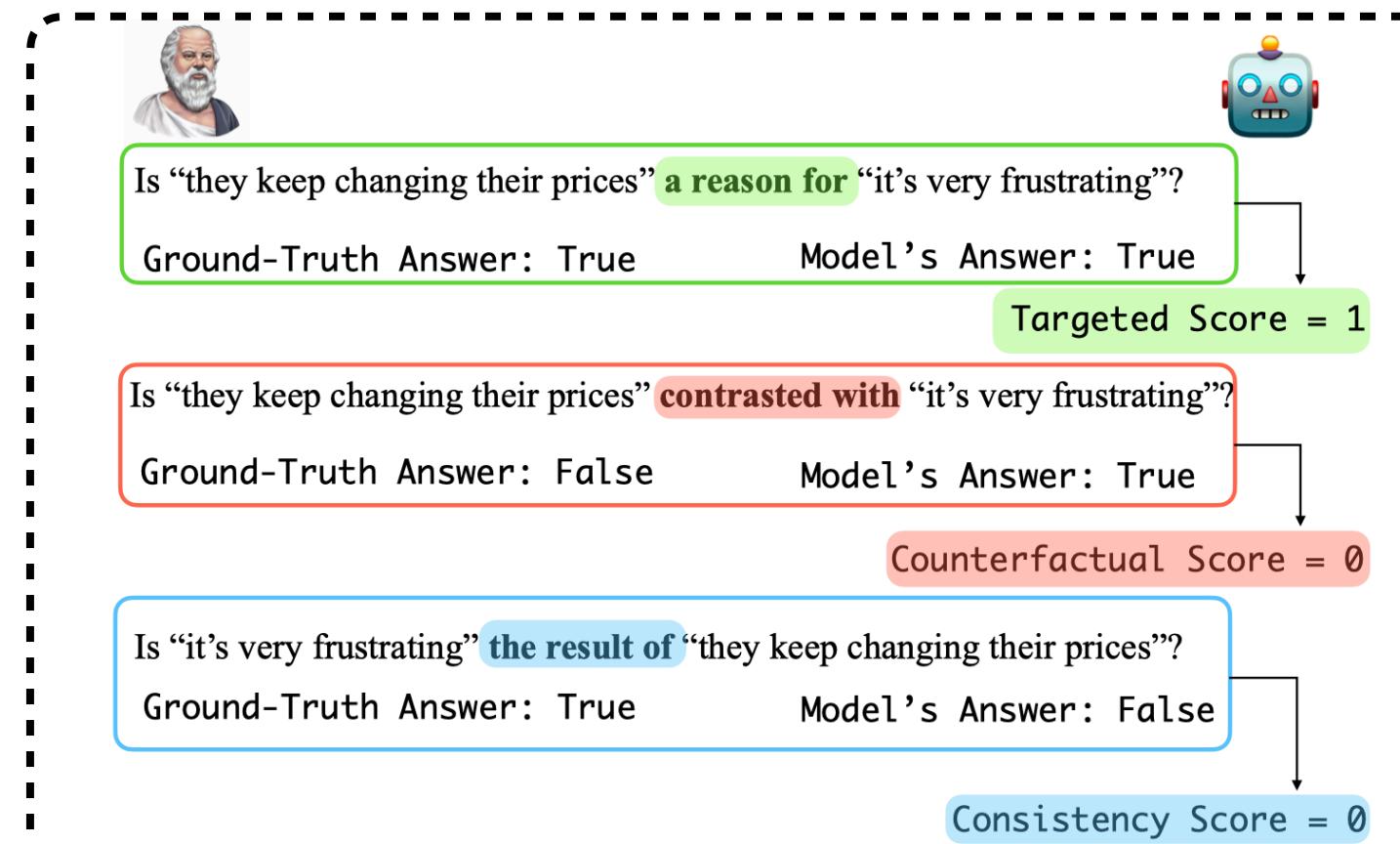
- Wizard code is nearly perfect.
- LLaMA2-13B-Chat has lower performance.



LLaMA2's consistency scores per question.

Conjecture: LLaMA2 model can only pay attention to verbatim keywords, and cannot do the real reasoning given previous QA.

Conclusion



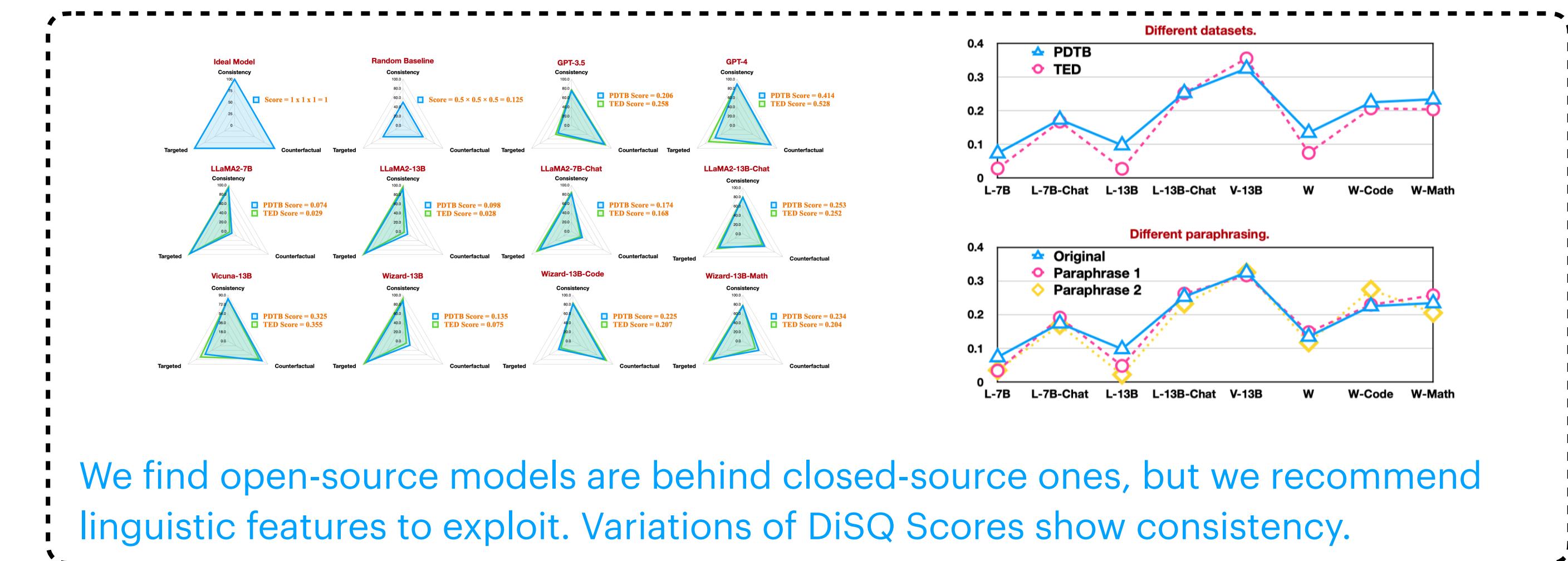
Discourse relation (R): Contingency.Cause.Result
Arg1: When I want to buy, they run from you – they keep changing their prices
Arg2: It’s very frustrating
s₁₁: I want to buy;
s₁₂: they run from you;
s₁₃: they keep changing their prices
S21: It’s very frustrating
Salient signals: (*s₁₃, s₂₁, r*), *r* is “the reason for”.
Targeted question: Is *s₁₃* the reason for *s₂₁*?
Counterfactual question: Does *s₁₃* contrast against *s₂₁*?
Converse question: Is *s₂₁* the result of *s₁₃*?

DiSQ is a new formalization using QA to evaluate models’ faithfulness in understanding discourse.

| Discourse relation (R) | Event relation (r) | Q Type | # of Q |
|---------------------------|----------------------------------|--------|--------|
| Comparison.Concession | deny or contradict with | Bi- | 1,764 |
| Comparison.Contrast | contrast with | Bi- | 876 |
| Contingency.Reason | reason of | Uni- | 3,264 |
| Contingency.Result | result of | Uni- | 2,796 |
| Expansion.Conjunction | contribute to the same situation | Bi- | 4,596 |
| Expansion.Equivalence | equivalent to | Bi- | 420 |
| Expansion.Instantiation | example of | Uni- | 2,352 |
| Expansion.Level-of-detail | provide more detail about | Uni- | 3,888 |
| Expansion.Substitution | alternative to | Uni- | 216 |
| Temporal.Asynchronous | happen before/after | Uni- | 1,368 |
| Temporal.Synchronous | happen at the same time as | Bi- | 840 |
| Total | | | 22,380 |

Input X
{
"Discourse relation": ...
"Arg1": ...
"Arg2": ...
"Event1": ...
"Event2": ...
"Event relation (ER)": ...
}
→
Output y
{
"DR summary": ...
"Event1 Comprehension": ...
"Event2 Comprehension": ...
"What if ER holds": ...
"What if ER does not hold": ...
"Predicting ER": ...
"Final prediction": ...
}

We employ in-context learning as semi-automatic annotation for salient discourse signals.



We find open-source models are behind closed-source ones, but we recommend linguistic features to exploit. Variations of DiSQ Scores show consistency.

Acknowledgements

We thank several group members from the Web Information Retrieval / Natural Language Processing Group (WING) at NUS for their research discussions and proofreading of our drafts, especially Xiao Xu, Vicor Li, Taha Aksu, Tongyao Zhu, Guanzhen Li, Zekai Li, and Yanxia Qin.

We also thank our anonymous reviewers for their time spent on our review and their detailed and insightful feedback, which greatly helped us refine our work.