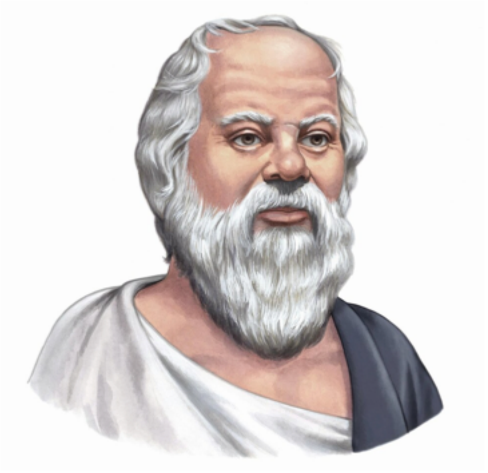
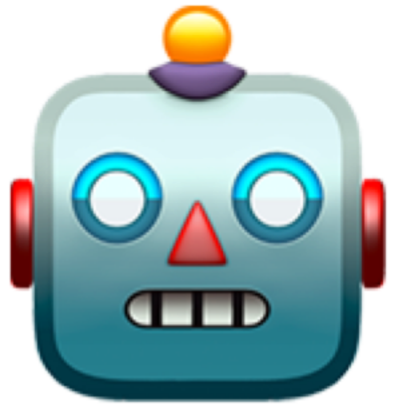
Discursive Socratic Questioning:

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

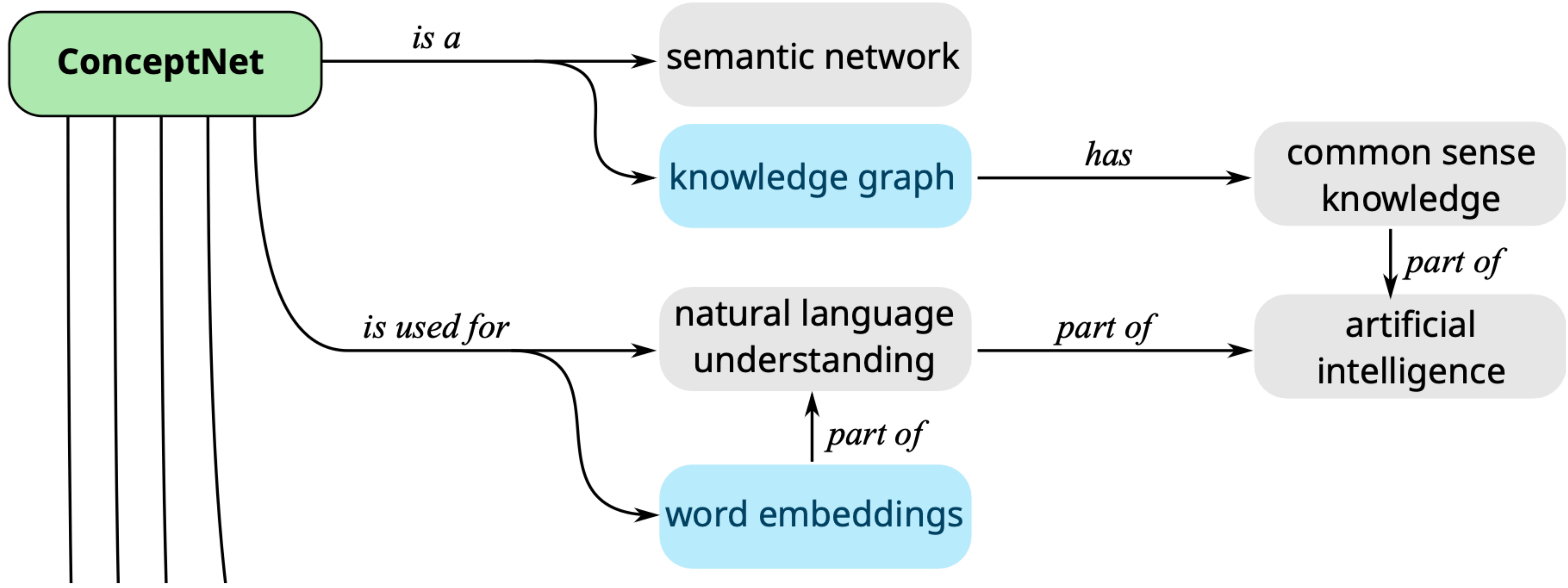
Dataset and code @ Github

ACL 2024

Bangkok, Thailand

What is discourse semantics?

Lexical Semantics

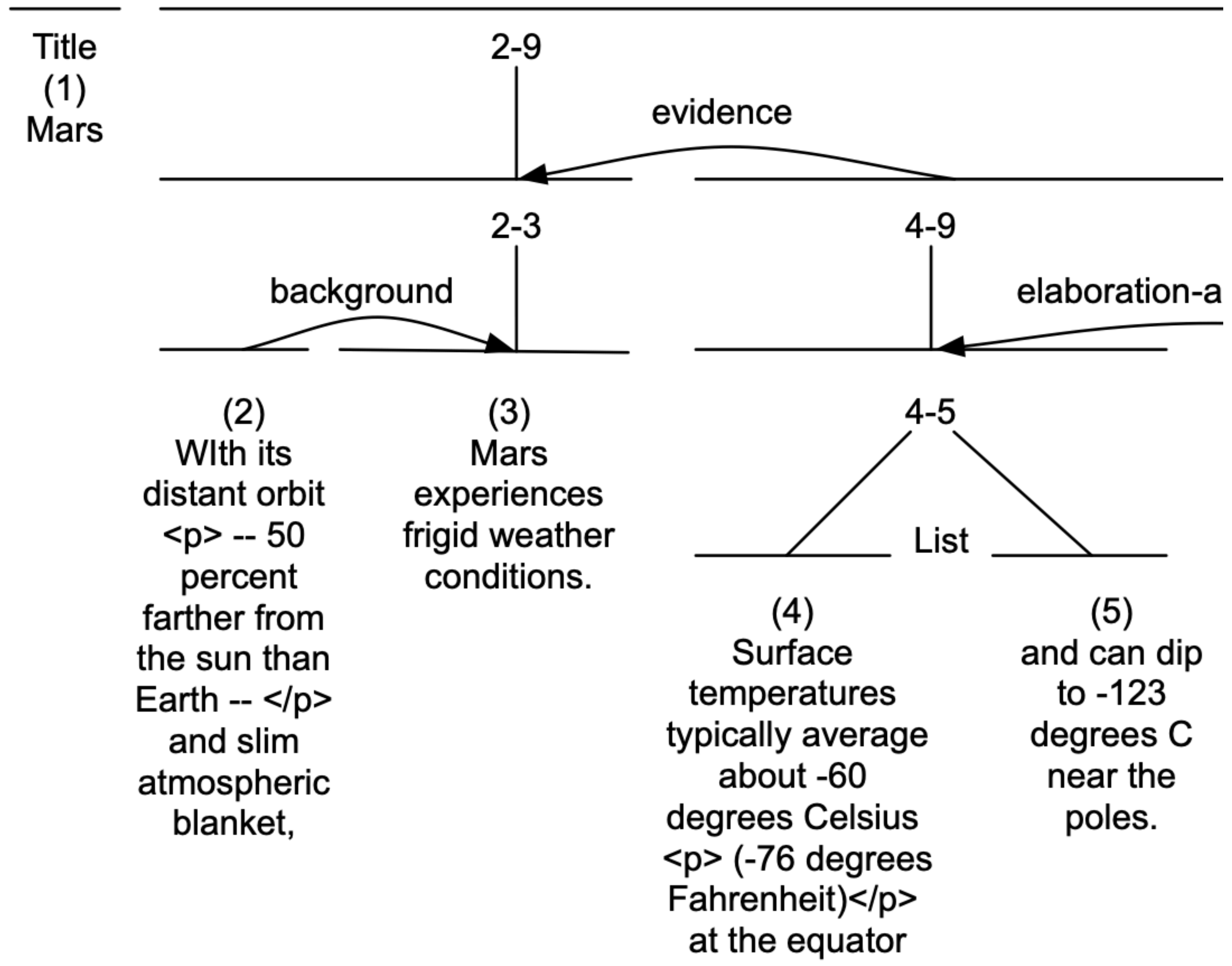


Ontology

Relationships between words and phrases;

Non-contextual.

Discourse Semantics



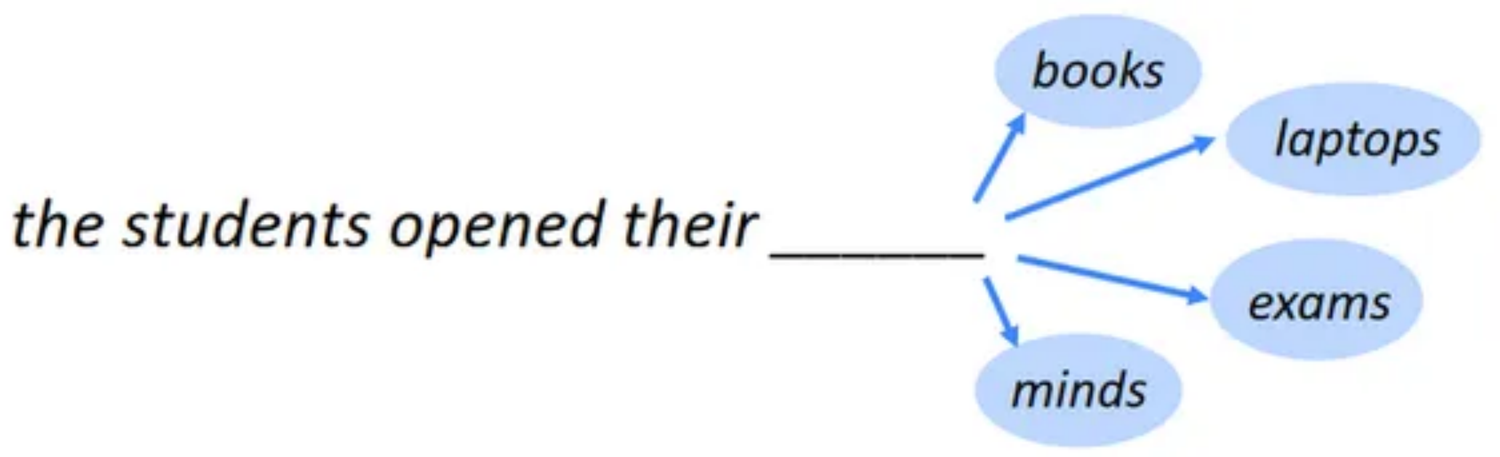
Treebanks

Larger nested units;

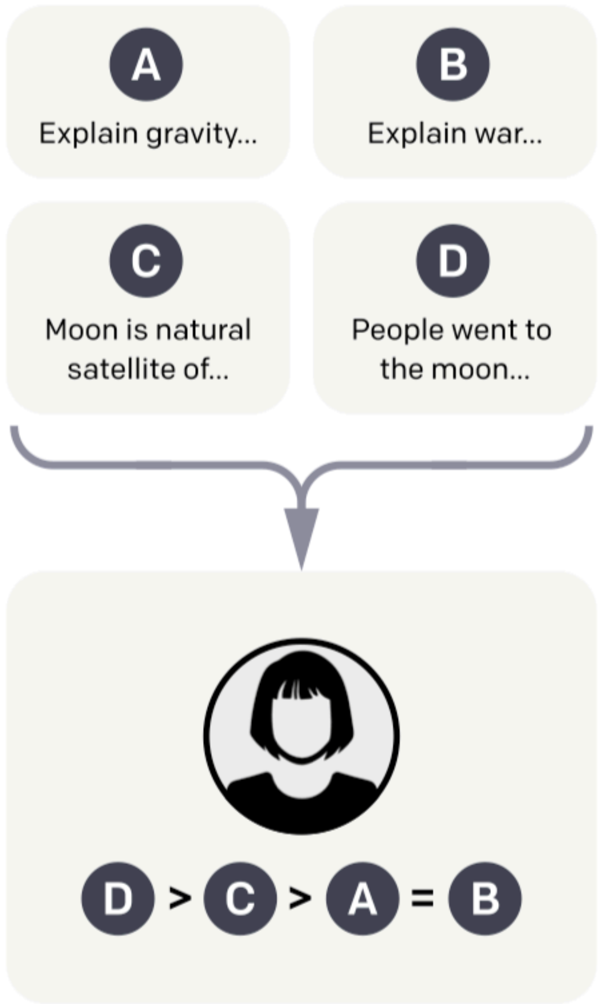
Depend on context.

Image credit1, credit2.

The big question — Do LLMs understand discourse?



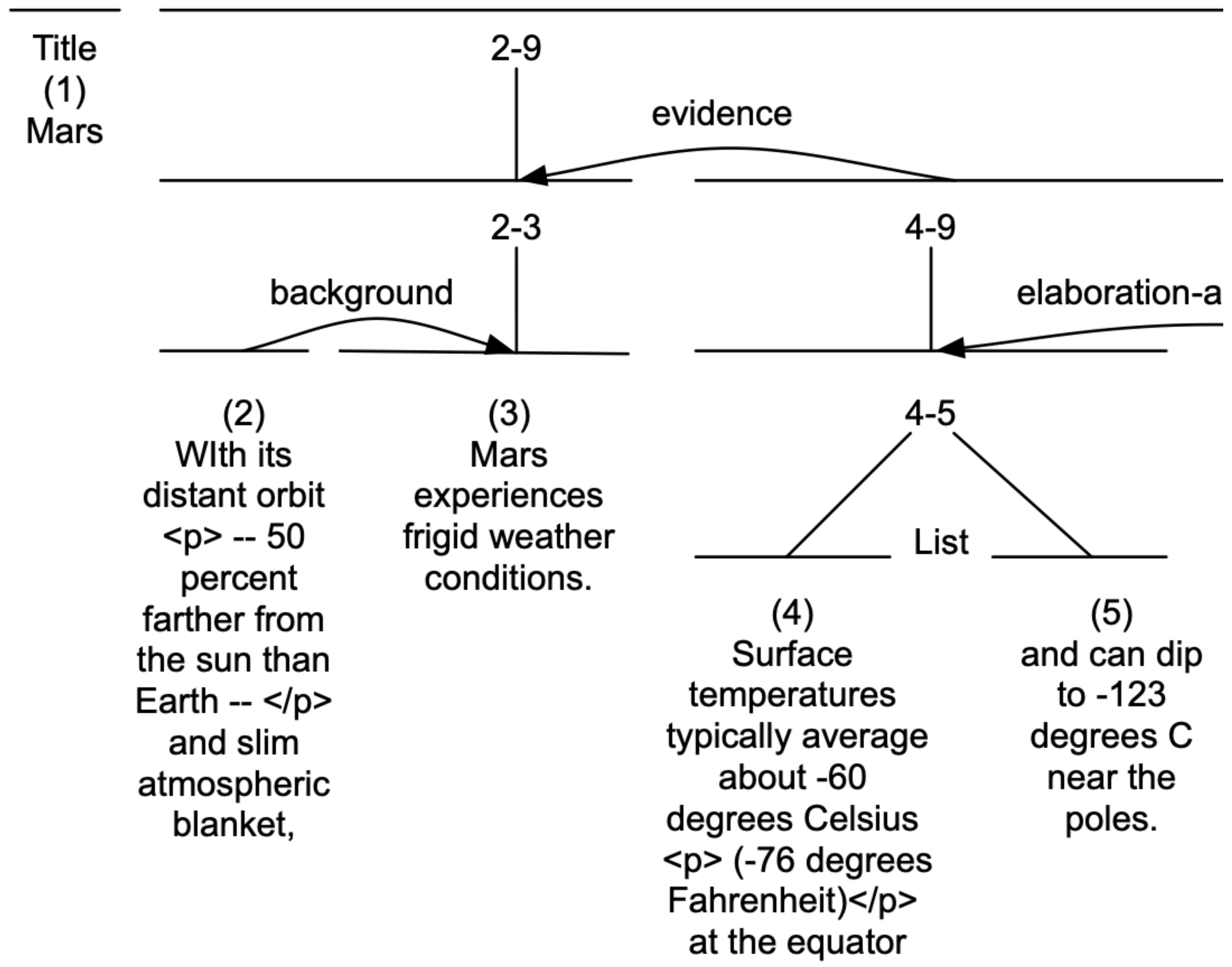
Objective 1: Next word prediction



Objective 2:

Reinforcement learning from human feedback (RLHF).

?



Discourse understanding

Image credit1, credit2

Existing evaluations

- **Classification task** on a hierarchical taxonomy
- **Existing 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.

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
$\text{ChatGPT}_{\text{Prompt}}$	29.85	32.89	9.27	15.59
$\text{ChatGPT}_{\text{PE}}$	33.78	34.94	10.73	20.31
$\text{ChatGPT}_{\text{ICL}}$	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

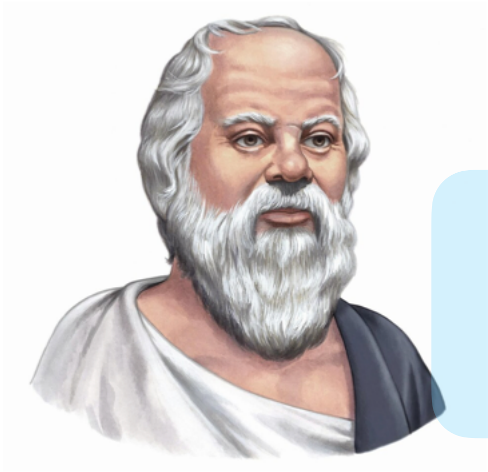

⚙️DiSQ
Faithfulness
Score

Socratic Questioning

Discourse relation: Contingency.Cause.Result

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

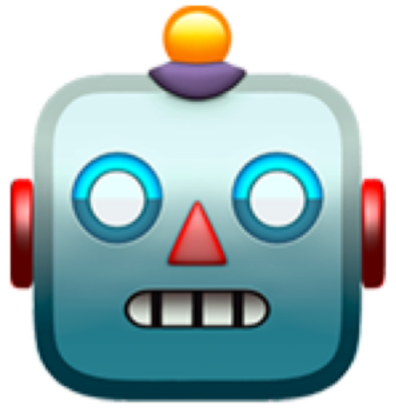
Arg2: *It's very frustrating*.



Why?

Can we comprehend them as other relations?

I think it's a Contingency discourse.

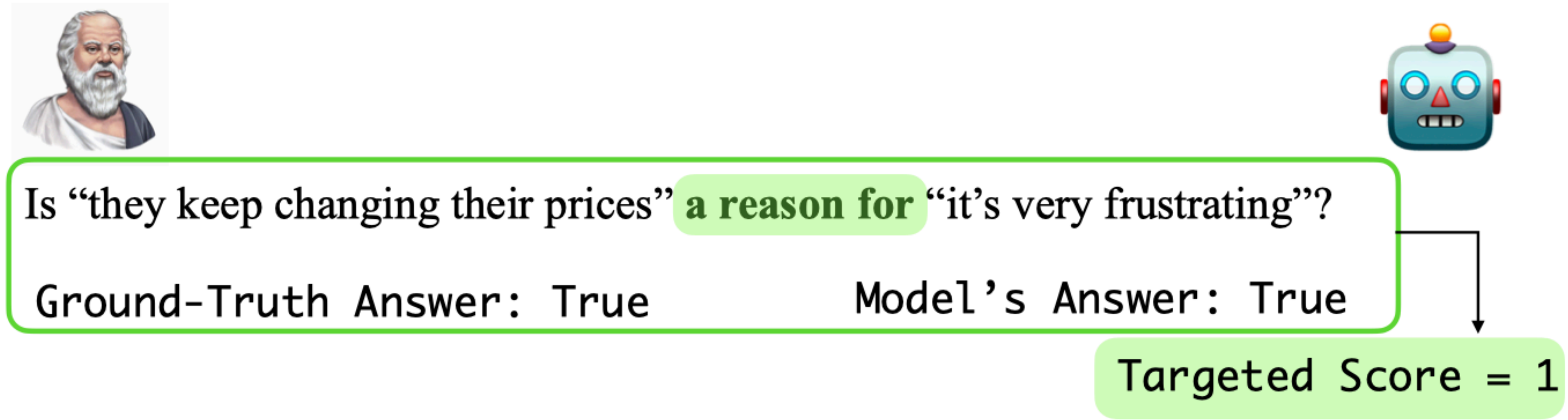


There is a cause-result event pair.

...

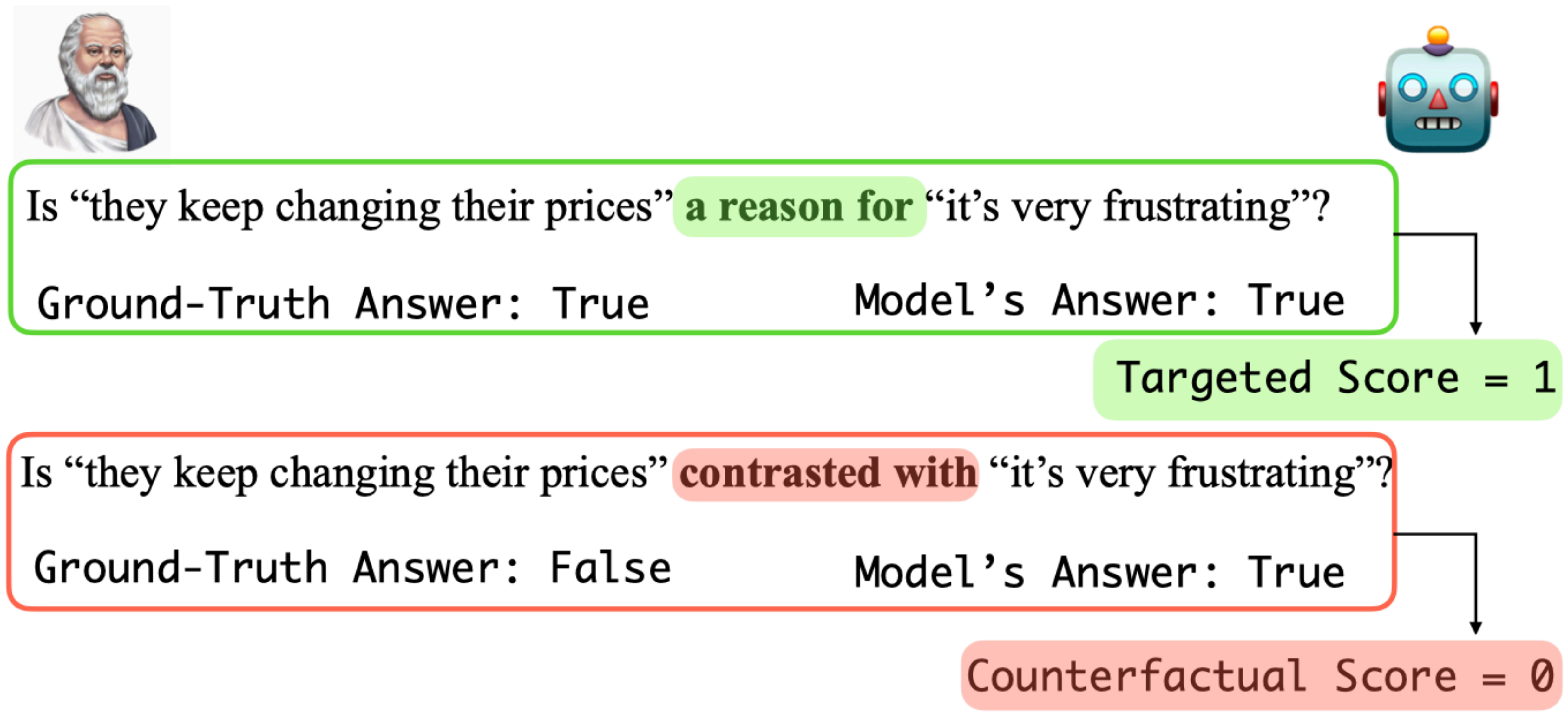
Socratic method is to ask a series of questions to challenge thoughts, clarify ideas and deepen understandings.

Discursive Socratic Questioning (DiSQ)



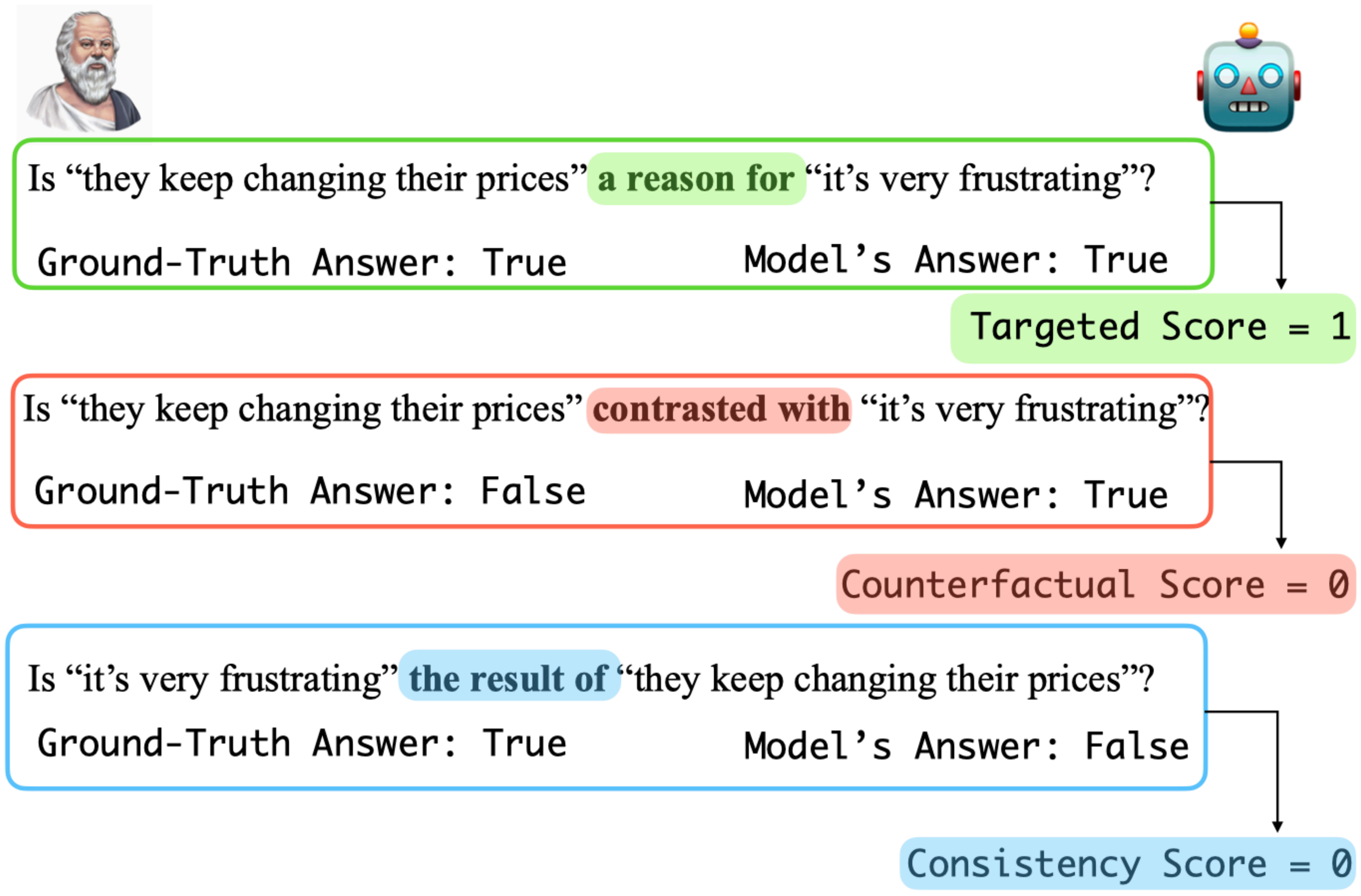
DiSQ is composed of three scores to evaluate models' faithfulness.

Discursive Socratic Questioning (DiSQ)



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

Discursive Socratic Questioning (DiSQ)



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?**

Annotate Salient Signal

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}?

Event pair as the salient signal.

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": …
}
```

In-context learning (ICL) for annotation.

Annotation Outcome

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

Question statistics for PDTB dataset.

Human verification of our annotation.

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

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

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}

Is A <u>contrasted</u> with B?

Bi-directional

Is B <u>contrasted</u> with A?

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}

Is A <u>contrasted</u> with B? $\xleftrightarrow{\text{Bi-directional}}$ Is B <u>contrasted</u> with A?

Is A the <u>reason</u> of B? $\xleftrightarrow{\text{Uni-directional}}$ Is B the <u>result</u> of A?

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.

Input: Discourse d and its corresponding questions $\mathcal{Q}$.
$\mathcal{H} = \{\varnothing\}$ ▷ The history is initialized.
Stage 1: Targeted and Counterfactual QA
for q_i in $\mathcal{Q}_t$ and $\mathcal{Q}_c$ **do**
  $a_i = \mathrm{LM}(q = q_i, c = d)$ ▷ The model performs QA. The context c is the discourse d.
  $\mathcal{H} \leftarrow (q_i, a_i)$ ▷ The history is updated.
end for
Stage 2: Converse QA
for (q_i, a_i) in $\mathcal{H}$ **do**
  $\tilde{q} = Lookup(q, \{\tilde{\mathcal{Q}}_c, \tilde{\mathcal{Q}}_t)\}$ ▷ Look up the converse question in converse question sets.
  $\tilde{a}_i = \mathrm{LM}(q = \tilde{q}_i, c = d, (q_i, a_i) \in \mathcal{H})$ ▷ The model executes QA on the converse question, $\tilde{q}_i$, optionally utilizing the previous response (q_i, a_i) as supplemental context.
  $\mathcal{H} \leftarrow (\tilde{q}_i, \tilde{a}_i)$ ▷ The history is updated.
end for
Output: $\mathcal{H}$

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

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

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: $\mathcal{H} = \{\varnothing\}$ ▷ 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: $a_i = \mathrm{LM}(q = q_i, c = d)$ ▷ The model performs QA. The context c is the discourse d.
6: $\mathcal{H} \leftarrow (q_i, a_i)$ ▷ The history is updated.
7: **end for**
8: **Stage 2:** Converse QA
9: **for** (q_i, a_i) in $\mathcal{H}$ **do**
10: $\tilde{q} = Lookup(q, \{\tilde{\mathcal{Q}}_c, \tilde{\mathcal{Q}}_t\})$ ▷ Look up the converse question in converse question sets.
11: $\tilde{a}_i = \mathrm{LM}(q = \tilde{q}_i, c = d, (q_i, a_i) \in \mathcal{H})$ ▷ The model executes QA on the converse question, $\tilde{q}_i$, optionally utilizing the previous response (q_i, a_i) as supplemental context.
12: $\mathcal{H} \leftarrow (\tilde{q}_i, \tilde{a}_i)$ ▷ The history is updated.
13: **end for**
14: **Output:** $\mathcal{H}$

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

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.

Input: Discourse d and its corresponding questions $\mathcal{Q}$.
$\mathcal{H} = \{\varnothing\}$ ▷ The history is initialized.
Stage 1: Targeted and Counterfactual QA
for q_i in $\mathcal{Q}_t$ and $\mathcal{Q}_c$ **do**
    $a_i = \mathrm{LM}(q = q_i, c = d)$ ▷ The model performs QA. The context c is the discourse d.
    $\mathcal{H} \leftarrow (q_i, a_i)$ ▷ The history is updated.
end for
Stage 2: Converse QA
for (q_i, a_i) in $\mathcal{H}$ **do**
    $\tilde{q} = Lookup(q, \{\tilde{\mathcal{Q}}_c, \tilde{\mathcal{Q}}_t\})$ ▷ Look up the converse question in converse question sets.
    $\tilde{a}_i = \mathrm{LM}(q = \tilde{q}_i, c = d, (q_i, a_i) \in \mathcal{H})$ ▷ The model executes QA on the converse question, $\tilde{q}_i$, optionally utilizing the previous response (q_i, a_i) as supplemental context.
    $\mathcal{H} \leftarrow (\tilde{q}_i, \tilde{a}_i)$ ▷ The history is updated.
end for
Output: $\mathcal{H}$

Discourse relation: Contingency
Targeted question:
Is A <u>the result of</u> 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?

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

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

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

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

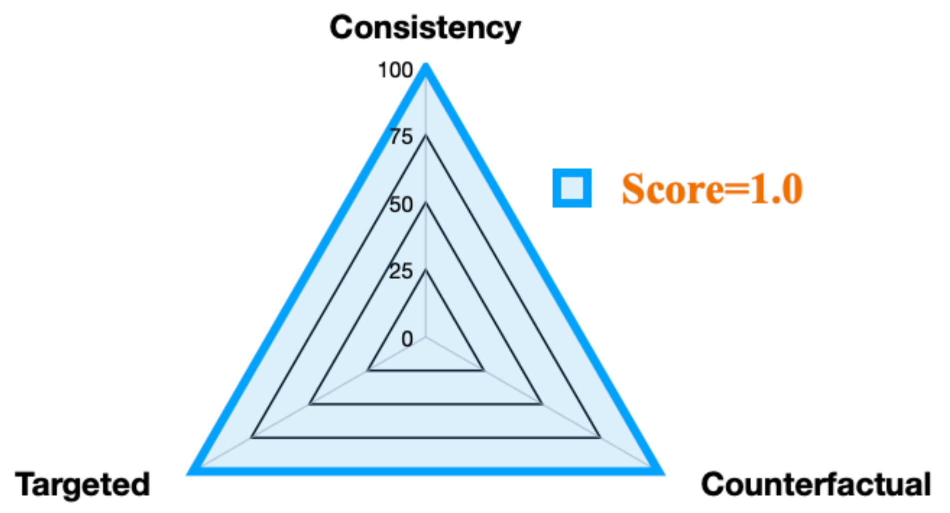
DiSQ Score is the multiplication of the three scores because we believe they are equally important (0.6, 0.6, 0.6) is better than (0.9, 0.9, 0).

Experiment setup

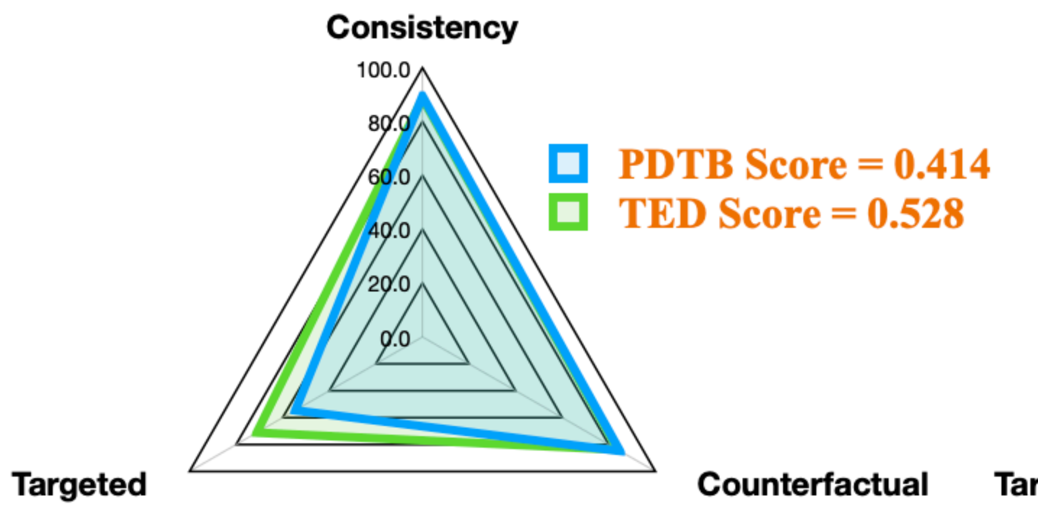
- **Datasets: PDTB** (WSJ News) and **TED-MDB** corpus (also in PDTB discourse style). TED has 448 instances and 8,378 questions, about half the size of PDTB.
- 🚪🔒 **Closed-source models: GPT-3.5 / GPT-4**.
- 🚪🔓 **Open-source models: LLaMA-2** (with or without chat fine-tuning). **Vicuna** (further fine-tuned 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.

Overall Performance

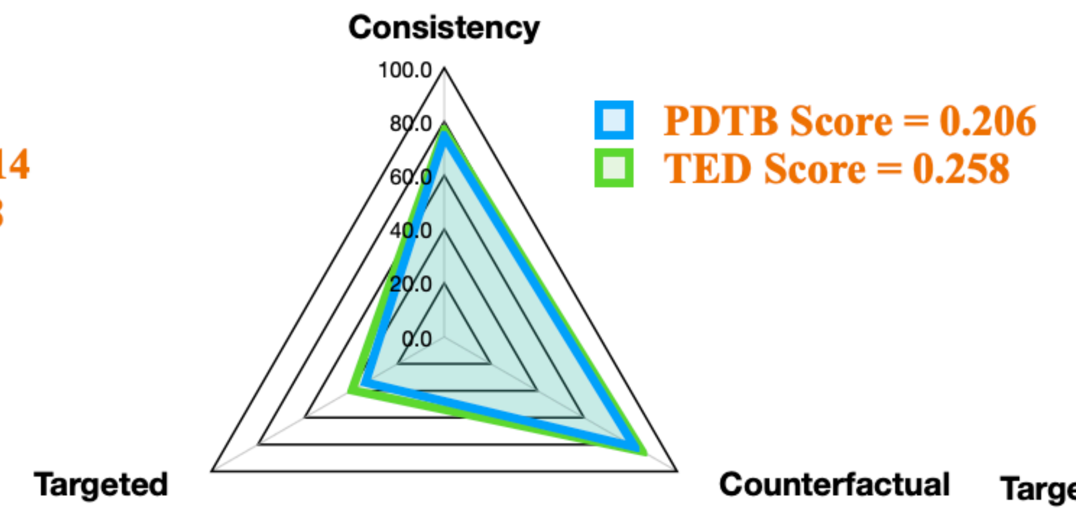
Ideal Model



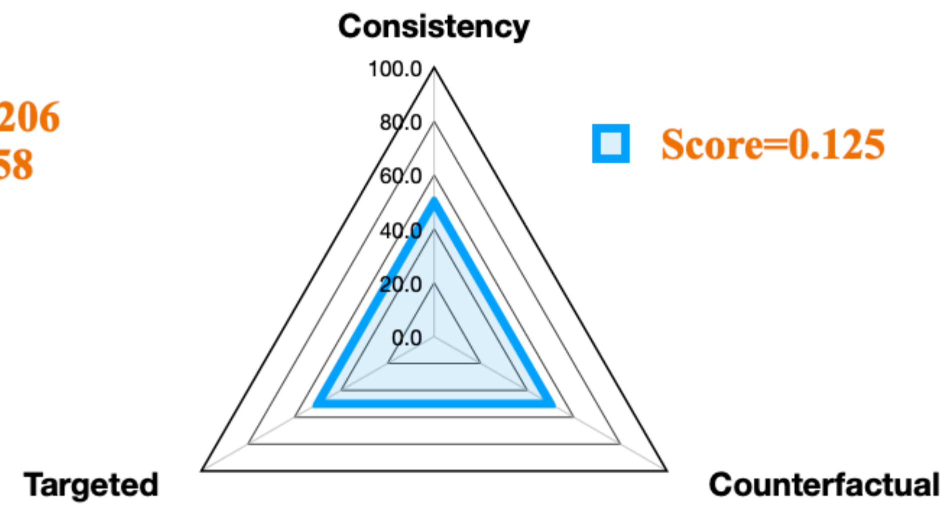
GPT-4



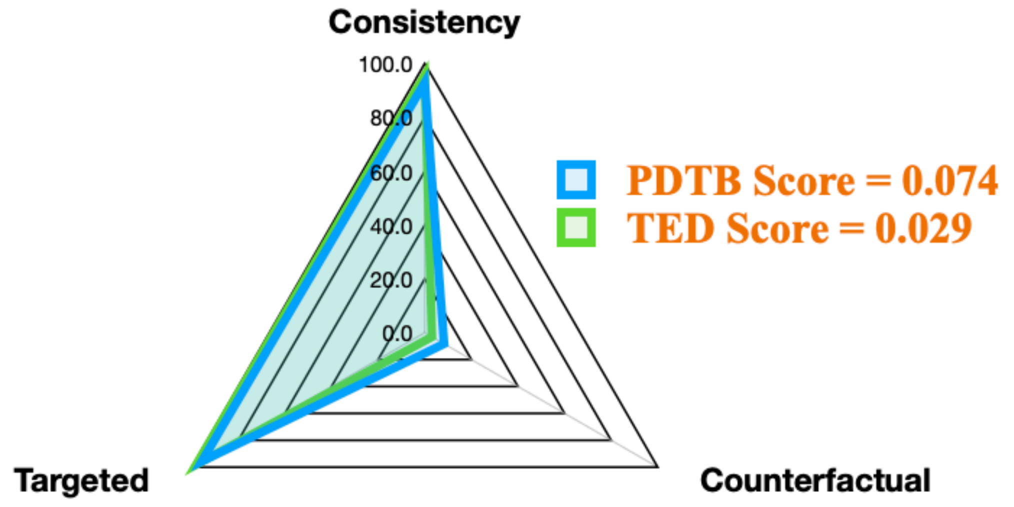
GPT-3.5



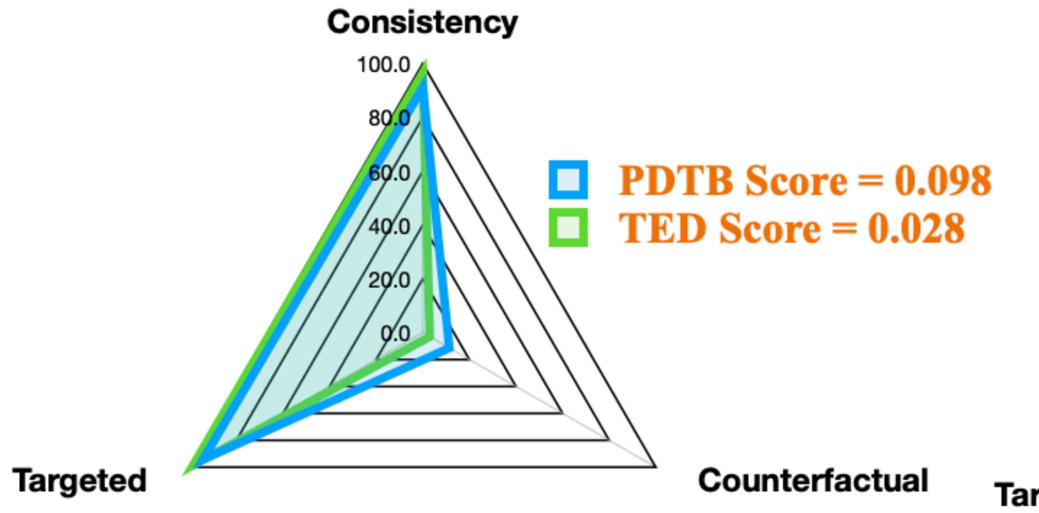
Random Baseline



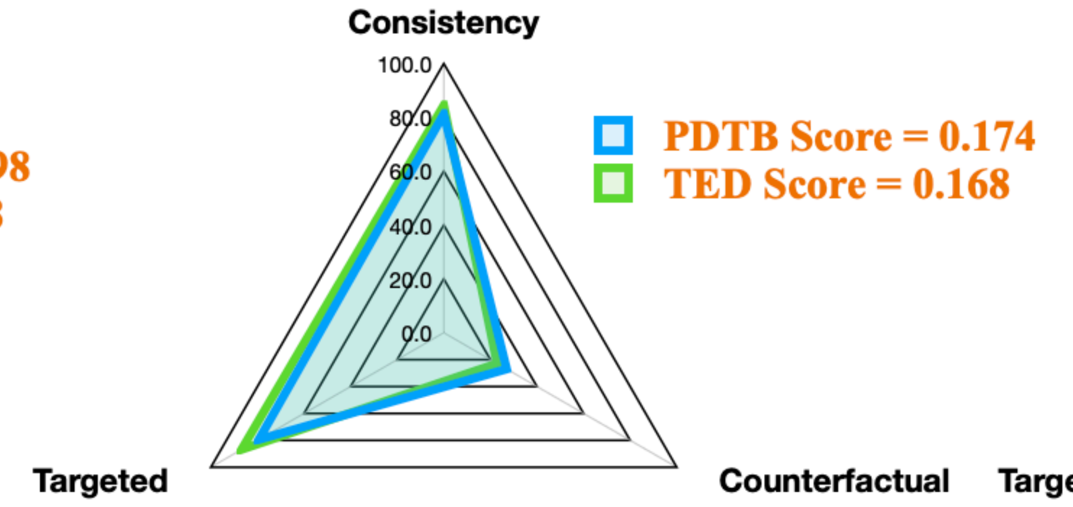
LLaMA2-7B



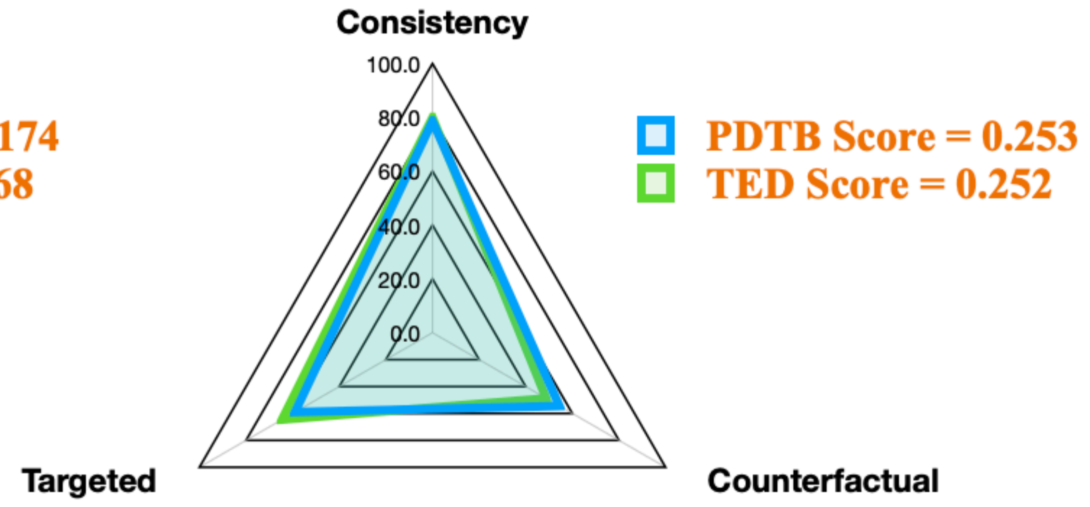
LLaMA2-13B



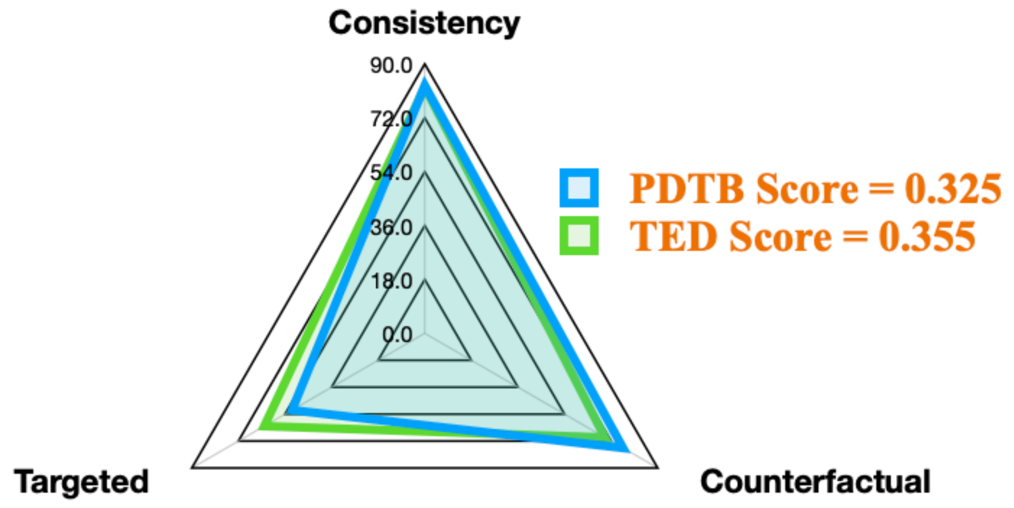
LLaMA2-7B-Chat



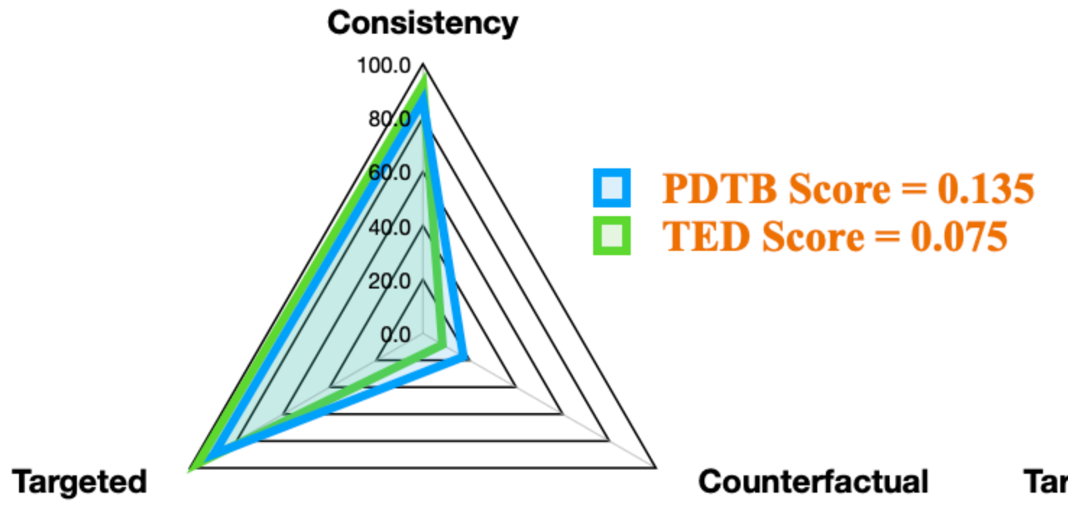
LLaMA2-13B-Chat



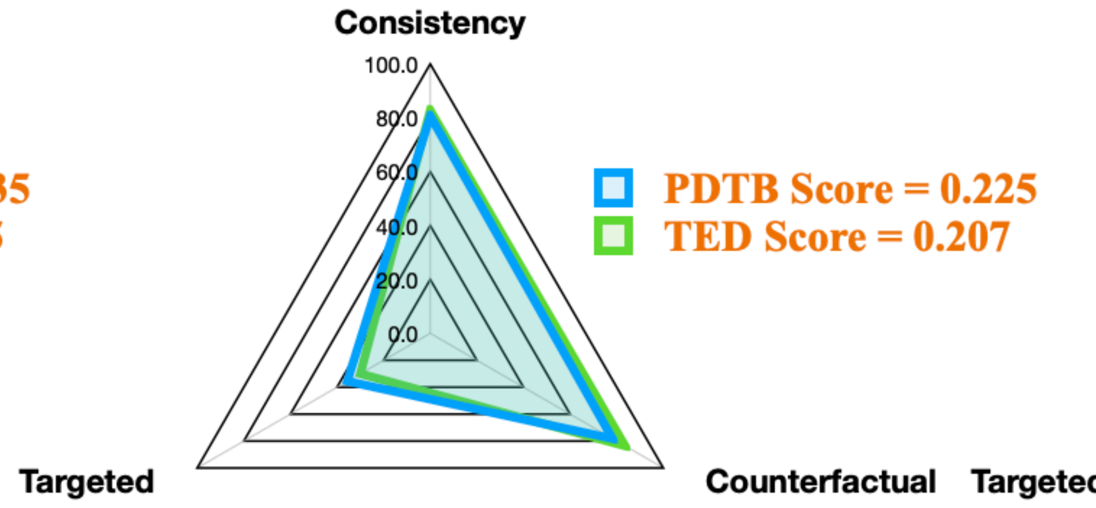
Vicuna-13B



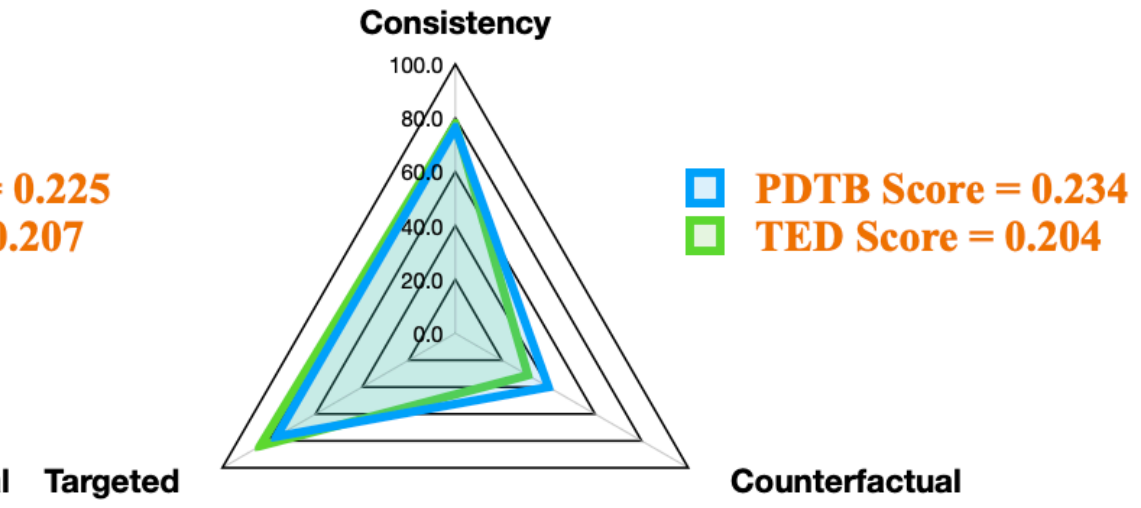
Wizard-13B



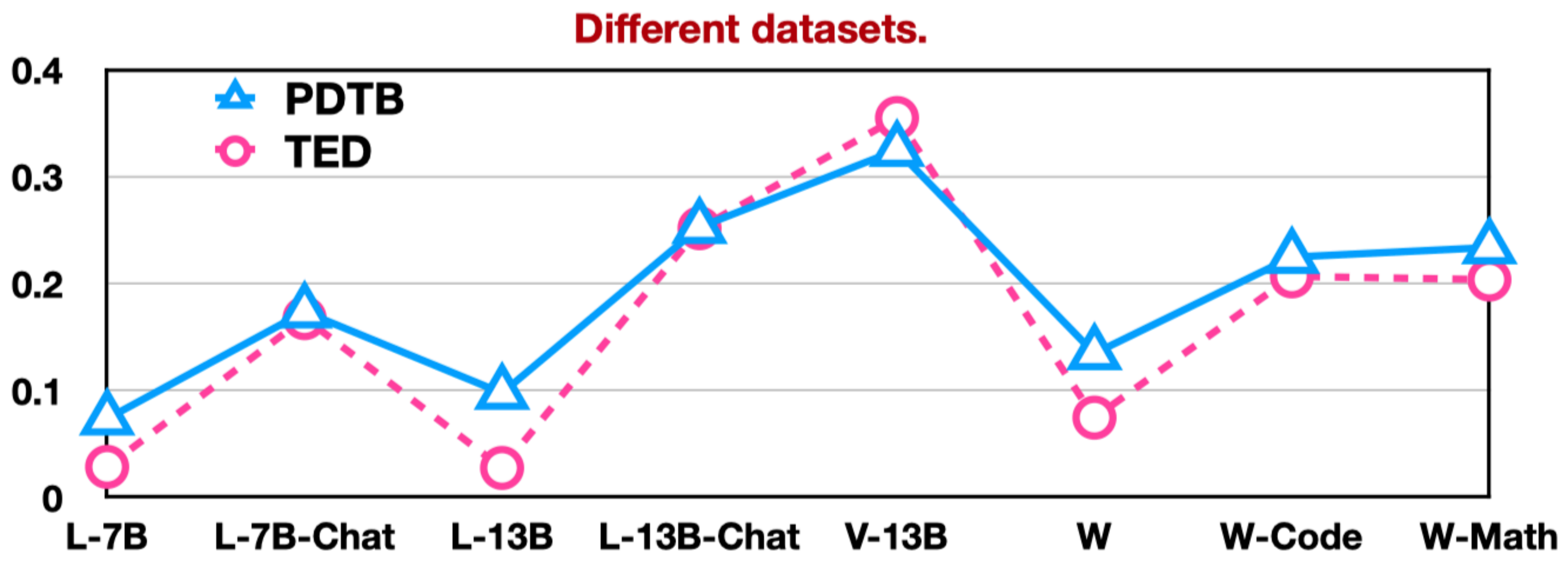
Wizard-13B-Code

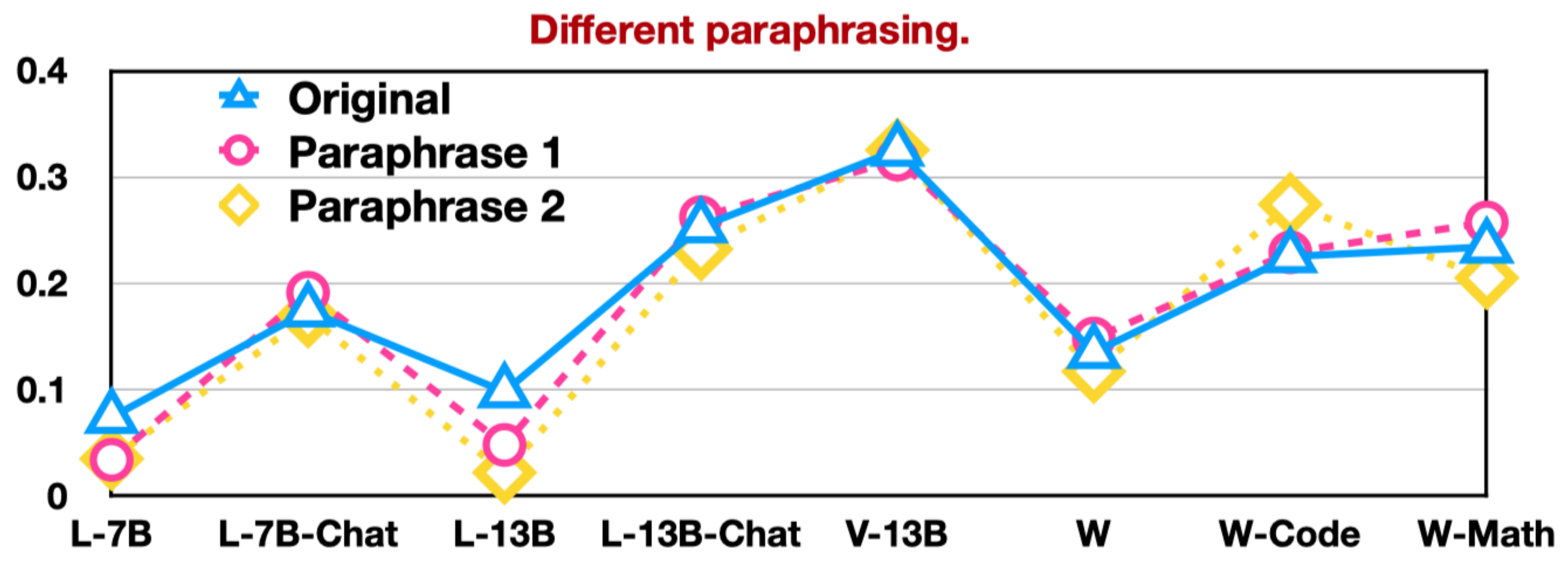


Wizard-13B-Math



Consistency of DiSQ Scores





DiSQ Scores under different datasets.

DiSQ Scores under different paraphrasing.

Impact of Discourse Relations on DiSQ Scores

Minority classes are still challenging for LLMs.

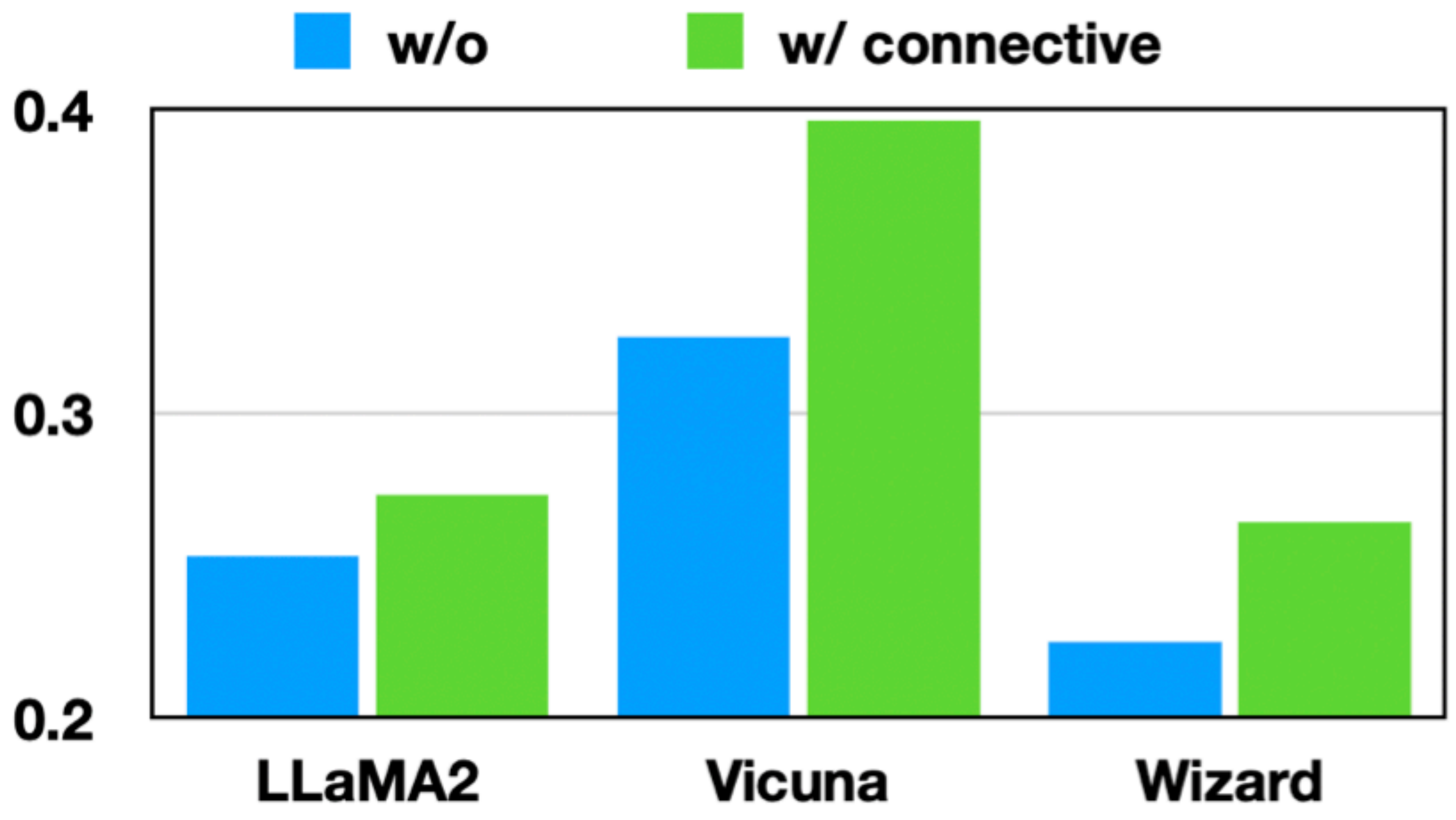
	Models	Overall	Comp.Conce	Comp.Contra	Cont.Reason	Cont.Result	Exp.Conj	Exp.Equiv	Exp.Inst	Exp.Detail	Exp.Subst	Temp.Async	Temp.Sync
	1. Random Basline	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
PDTB	**2A. LLaMA2-7B**	0.074	0.029	0.083	0.094	0.095	0.076	0.056	0.087	0.067	0.156	0.035	0.048
	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	0.040
	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	**0.092**
	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	0.028
	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**	0.069
	7A. Wizard	0.135	**0.221**	0.256	0.067	0.107	0.170	0.072	0.167	0.128	0.108	0.097	0.082
	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**	**0.109**
	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	**0.135**
	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	0.000
	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**	0.000

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*.

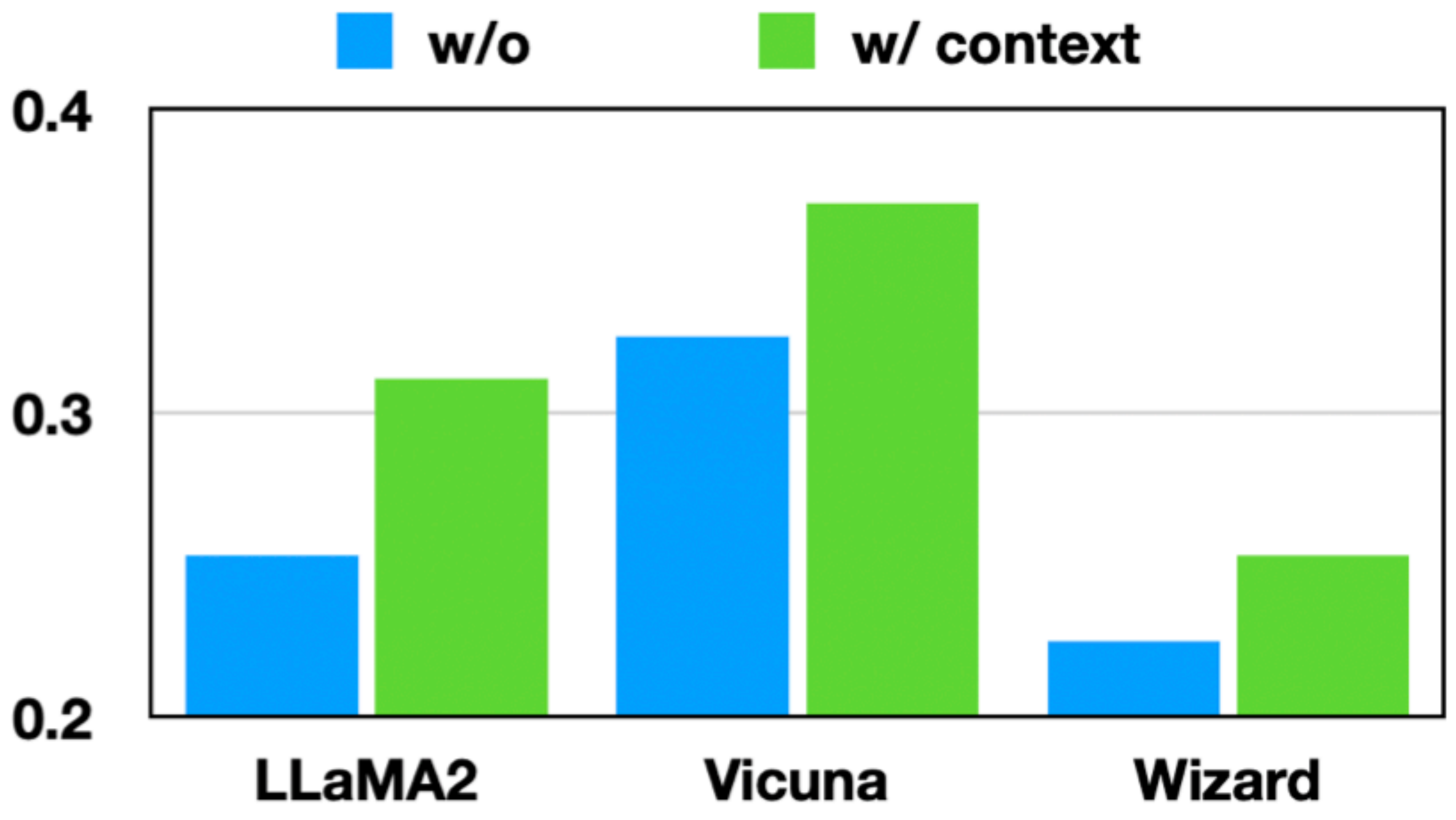
"so"



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*.

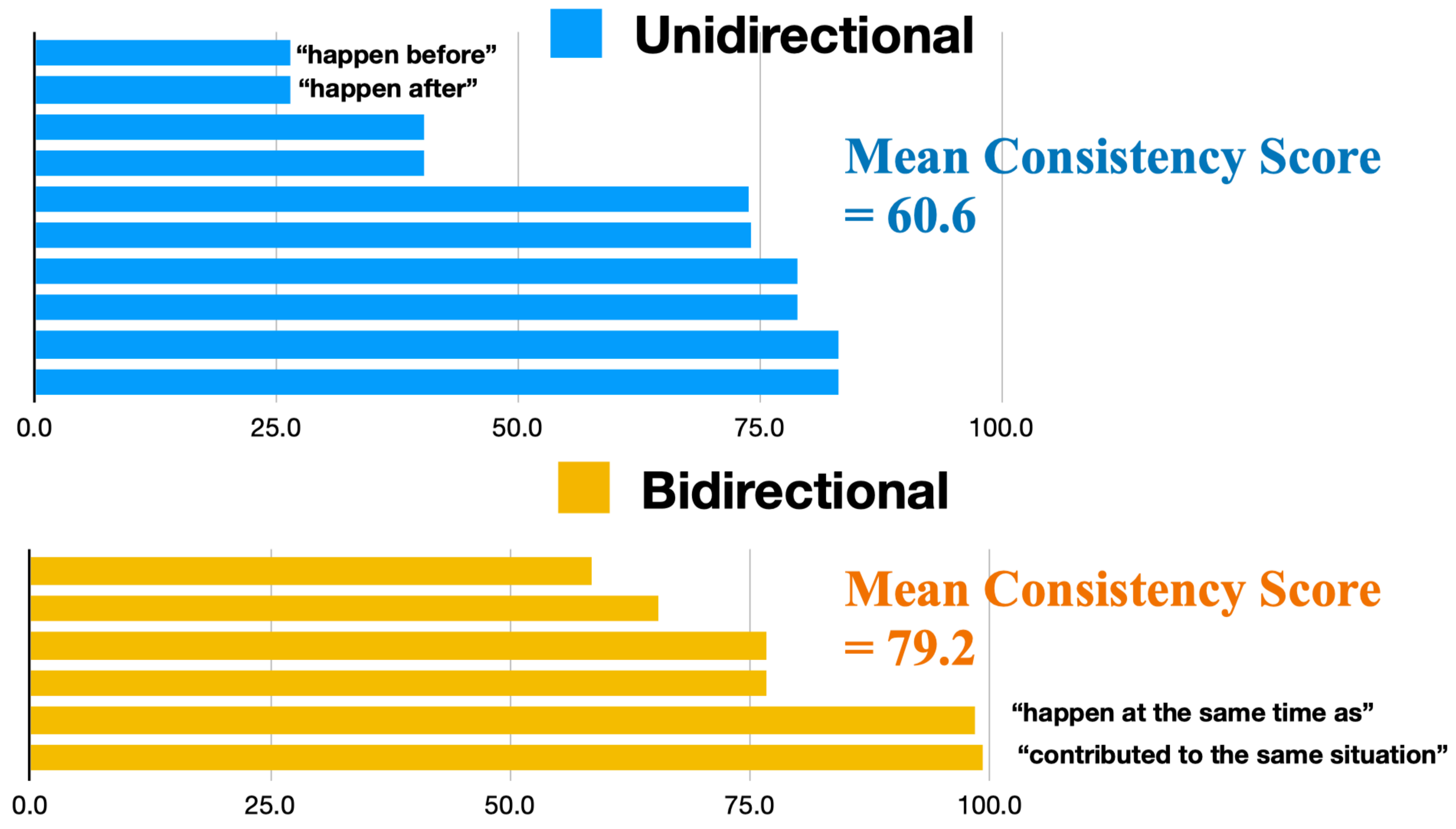
Subsequent context



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

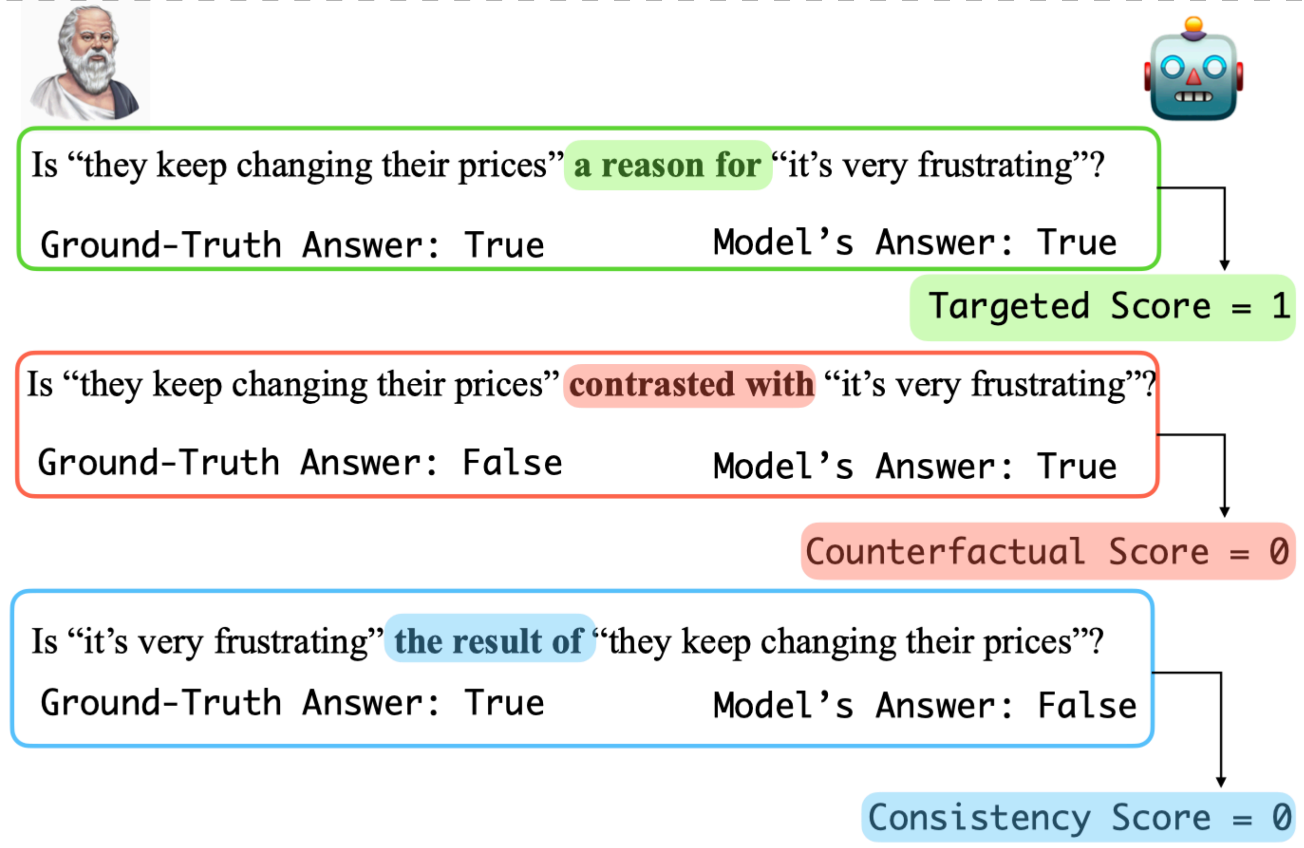


LLaMA model might overfit to verbatim keywords.

See You at Poster Session 6
10:30 – 12:00 Wednesday

💿💻👉

Conclusion



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}?	

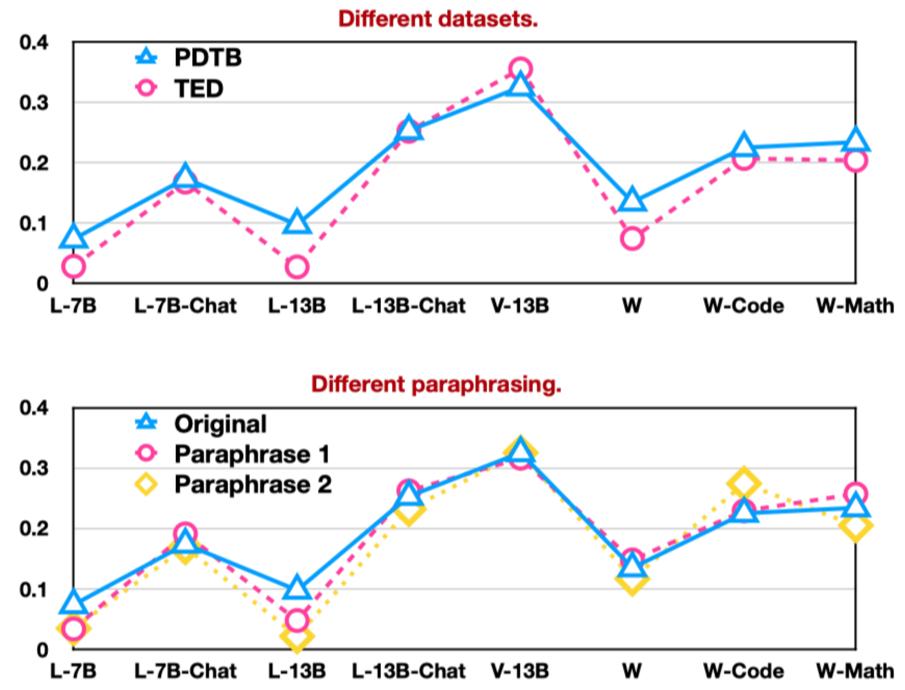
DiSQ is a new formalization using QA to evaluate models’ faithfulness in understanding discourse.

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.

Supplementary Slides

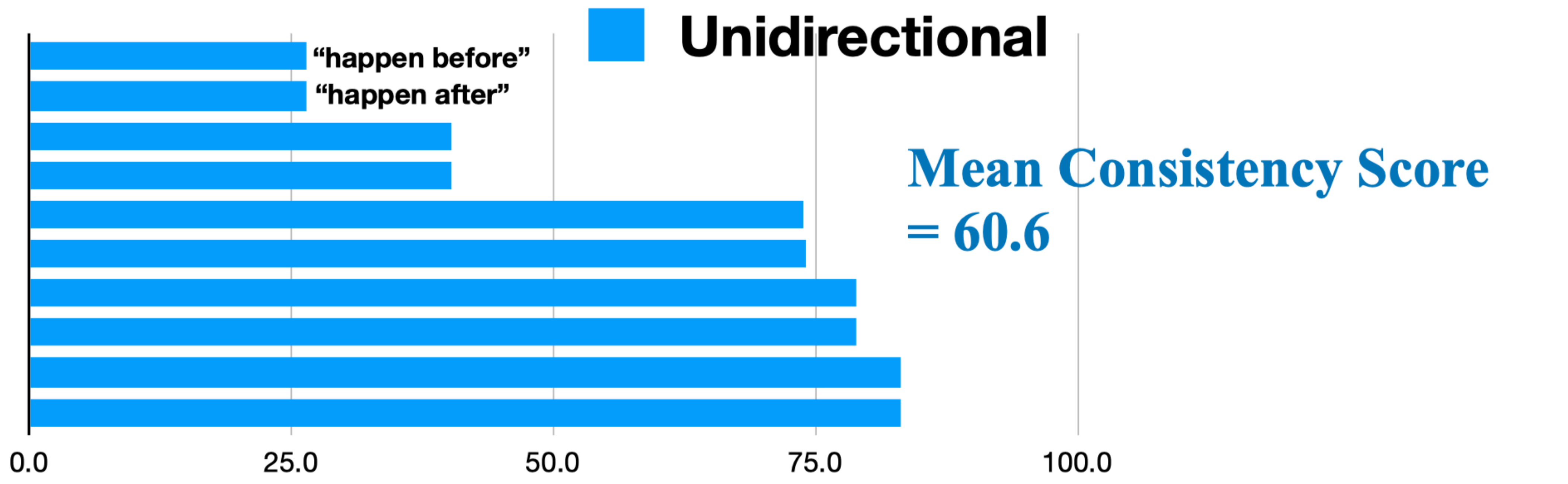
Evaluation

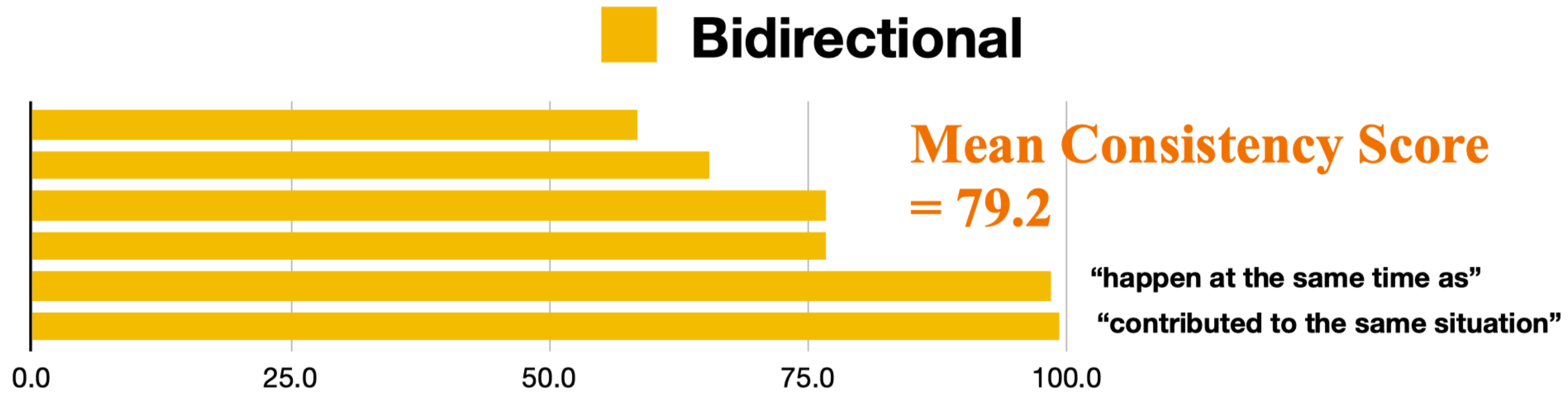
RQ4: Impact of Linguistic Features

	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.

Evaluation

RQ3: Impact of Discourse Relations on DiSQ Scores

	Exp.Inst.Arg2-as-instance	Exp.Detail.Arg1-as-detail	Exp.Detail.Arg2-as-detail	Exp.Subst.Arg2-as-subst	Cont.Result	Cont.Reason	Comp.Conc.Arg1-as-denier	Comp.Conc.Arg2-as-denier	Temp.Async.Prec	Temp.Async.Succ
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.