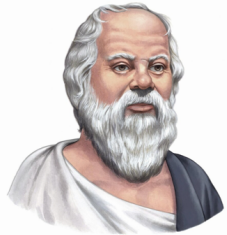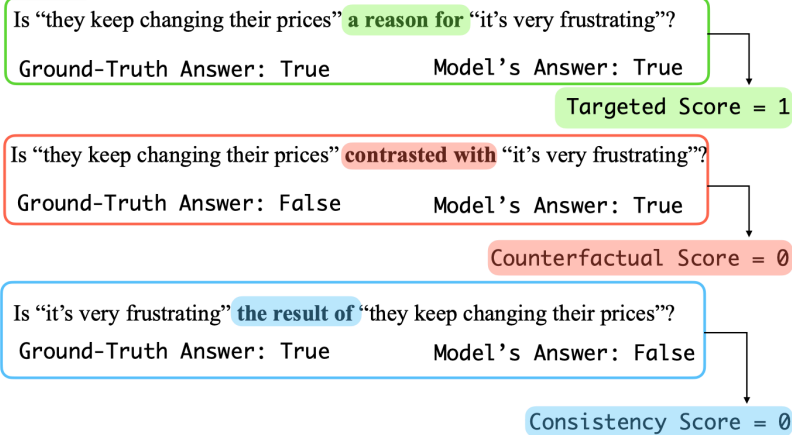# Discursive Socratic Questioning:
Evaluating the Faithfulness of Language Models' Understanding of Discourse Relations

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

Socrates 🇬🇷
470 – 399 BC

Want to know how well do LLMs understand discourse relations? We propose an end-to-end automatic scoring framework for discourse relations, leveraging LLMs in a Socratic style.
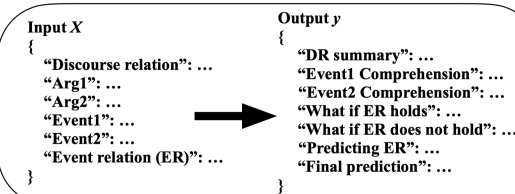

Is "they keep changing their prices" **a reason for** "it's very frustrating"?

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

Discourse relation: 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.

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": ...

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

What to ask: In-context learning to annotate salient signals.

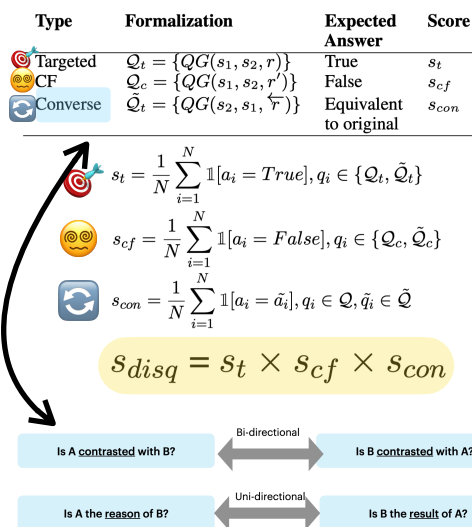**Algorithm 1** DiSQ interrogates a language model.

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:    $a_i = \text{LM}(q = q_i, c = d)$ ▷ 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:    $\tilde{a}_i = \text{LM}(q = \tilde{q}_i, c = d, (q_i, a_i) \in \mathcal{H})$ ▷ 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:** $\mathcal{H}$

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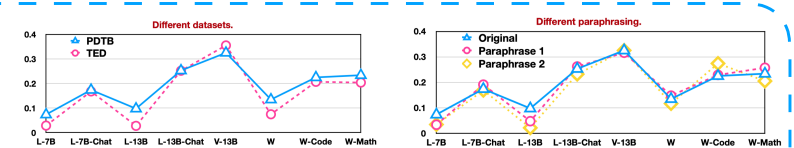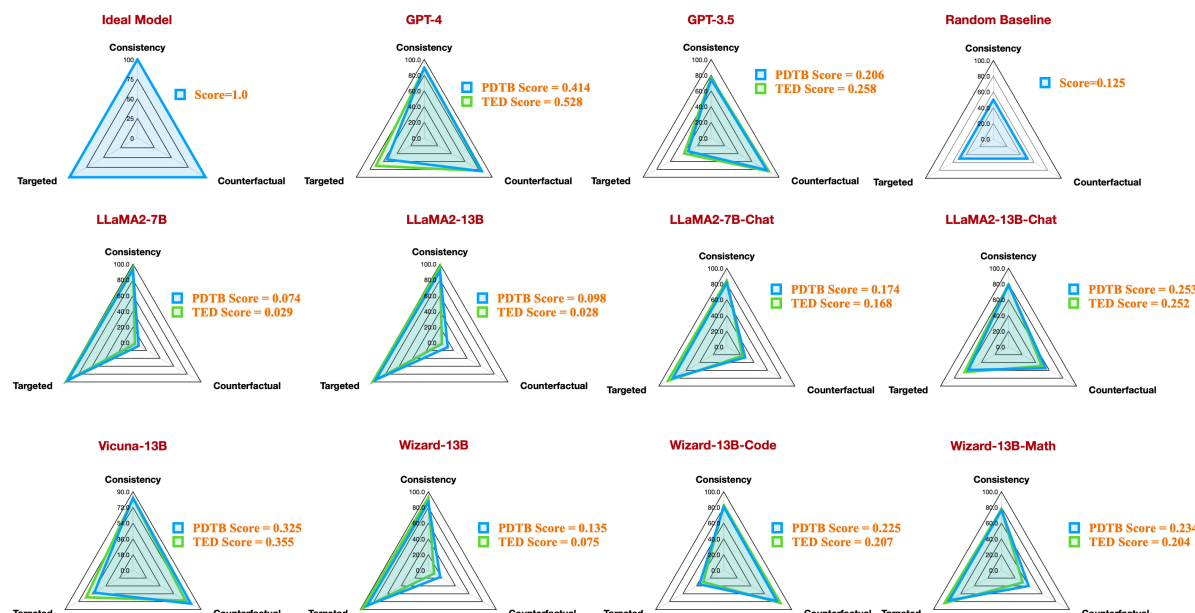(Given you answered A is the result of B.) Is B the reason of A?

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}

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

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

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

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

Is A contrasted with B? ↔ Bi-directional ↔ Is B contrasted with A?
Is A the reason of B? ↔ Uni-directional ↔ Is B the result of A?

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

How to ask: DiSQ is composed of (1) targeted questions, (2) counterfactual questions and (3) converse questions. DiSQ Scores is a **multiplication of the three aspects.**

Question statistics for PDTB dataset.

Evaluation setup: PDTB and TED-MDB dataset. Zero-shot QA based on DiSQ.

Evaluated models: (1) Closed-source models: GPT-4 / GPT-3.5. (2) Open-source models: LLaMA family; Vicuna model; Wizard family.

Different datasets. (PDTB / TED)

Different paraphrasing. (Original / Paraphrase 1 / Paraphrase 2)

RQ2: DiSQ Scores are consistent in terms of dataset and paraphrase variations.

Models	Overall	Comp.Conce	Comp.Contra	Cont.Reason	Expa.Conj	Expa.Equiv	Expa.Instant	Expa.Detail	Expa.Subst	Temp.Async	Temp.Sync
1. Random Baseline	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	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	0.092
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**	0.002	0.513	**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
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	0.257	0.262	**0.232**	**0.135**
10A. GPT-3.5	0.206	0.151	0.278	0.082	0.161	0.246	0.257	0.262	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

(PDTB)

RQ3: Minority classes are challenging.

(w/o / w/ connective)   (w/o / w/ context)   LLaMA2, Vicuna, Wizard

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

Unidirectional — Mean Consistency Score = 69.6
Bidirectional — Mean Consistency Score = 79.2

RQ4: Linguistic Features: Benefits from **discourse connectives**, **discourse context**, and **historical QAs**.

Ideal Model — Score=1.0
Consistency / Targeted / Counterfactual

GPT-4 — PDTB Score = 0.414 / TED Score = 0.528

GPT-3.5 — PDTB Score = 0.206 / TED Score = 0.258

Random Baseline — Score=0.125

LLaMA2-7B — PDTB Score = 0.074 / TED Score = 0.029

LLaMA2-13B — PDTB Score = 0.098 / TED Score = 0.028

LLaMA2-7B-Chat — PDTB Score = 0.174 / TED Score = 0.168

LLaMA2-13B-Chat — PDTB Score = 0.253 / TED Score = 0.252

Vicuna-13B — PDTB Score = 0.325 / TED Score = 0.355

Wizard-13B — PDTB Score = 0.135 / TED Score = 0.075

Wizard-13B-Code — PDTB Score = 0.225 / TED Score = 0.207

Wizard-13B-Math — PDTB Score = 0.234 / TED Score = 0.204

Overall performance (RQ1): (1) Gap between Close- and Open source models; (2) Benefits from further fine-tuning; (3) Consistency between two datasets.

NUS | Computing — National University of Singapore

WING — Web Information Retrieval / Natural Language Processing Group

Sichuan University 1896

A*

ACL 2024 — Bangkok, Thailand