Strategic Dialogue Assessment: The Crooked Path to Innocence

Anshun Asher Zheng, Junyi Jessy Li, David, I. Beaver

The University of Texas at Austin

Scenario

Did you ever falsify any incident reports regarding in-flight safety?

To my knowledge, there was a case where my colleague corrected one such report.

The response is **non-cooperative** but **beneficial**. Because the cooperative principles *cannot be reliably assumed*. While her denial is beneficial, it also introduces ambiguity. This is a beneficial move with a trade-off in clarity and relevance.

GPT-4o

The response is **non-cooperative** and **detrimental**. Because the response is evasive and the lack of direct answer raises suspicion, making it detrimental.

Models understand (non-)cooperativity via a Gricean lens, but ***does not*** recognize the ***strategic value*** the language carries

SDA: STRATEGIC DIALOGUE ASSESSMENT

WHAT IS "STRATEGIC"?

Strategic refers to how ***effective*** an agent realizes their (conversational) ***goal***

BUT…HOW UTTERANCES SYSTEMATICALLY CONTRIBUTE TO GOALS?

LET'S TAKE A LOOK AT A REAL COURTROOM DIALOGUE

Central issue: Whether police interrogation is ***coercive***
Prosecutor: It's ***not*** coercive but routine
Expert: It's ***coercive*** so the defendant's statements are not reliable

[...]In what you term or what you find in there coercive [...]?

[...] The questions were very directly specifying what the answers should be.

➡ ***coercive***

Did you find anything in the statement [...] to indicate that the officers gave him the information about which boy was castrated?

[...] In their statements? Perhaps there is no such record.

➡ ***Lack evidence of coercive interrogation***

Ok, you also talked to Mr. Smith for three hours?

No. I talked with him for the length of time it took to produce the transcript here.

➡ ***Not relevant to the central issue***

[...] what coercive tactics do you allege that the police made in this case – or did?

In order to answer your question, first I need to break the interrogation down [...] so that I can cut out parts of it and focus on a particular part.

➡ ***Not relevant to the QUD***

Observation

Commitments
- Interpretable
  - What content they convey
  - How they are conveyed
- Non-interpretable

w.r.t. QUDs

Beneficial
Detrimental
Neutral

Grice maxims
Consistency
etc.

None

Operationalization

$$\text{Utility}(i) = \text{Commitment}(i) \times \text{Manner\_of\_C}(i)$$

The utility of an utterance is ***what it conveys*** modified by ***how it is conveyed***

$$\text{BaT}_i = \begin{cases} f_c(C_i), \text{if } C_i \in \{\text{BENEFICIAL}, \text{NEUTRAL}\} \\ f_c(C_i) \times (\text{Rel}_i + \text{Man}_i + \text{Qual}_i), \\ \qquad \text{if } C_i = \text{DETRIMENTAL} \\ 0, \qquad \text{otherwise} \end{cases}$$

$$\text{PaT}_i = \begin{cases} |f_c(C_i)| + \text{Const}_i \times \sum_{j=1}^{i} \text{BaT}_j, \\ \qquad \text{if } C_i \in \{\text{DETRIMENTAL}, \text{NONE}\} \\ |f_c(C_i)| \times (\text{Rel}_i + \text{Man}_i + \text{Qual}_i) \\ \quad + \text{Const}_i \times \sum_{j=1}^{i} \text{BaT}_j, \text{ otherwise} \end{cases}$$

CPD: THE CROOKED PATH DATASET

We collected testimonials in cross-examinations of three prominent U.S. trials*: the West Memphis Three Trials (1994), the O.J. Simpson Trial (1995) and the Enron (Lay & Skilling) Trial (2006)*, which amount to 3325 Q/A pairs.

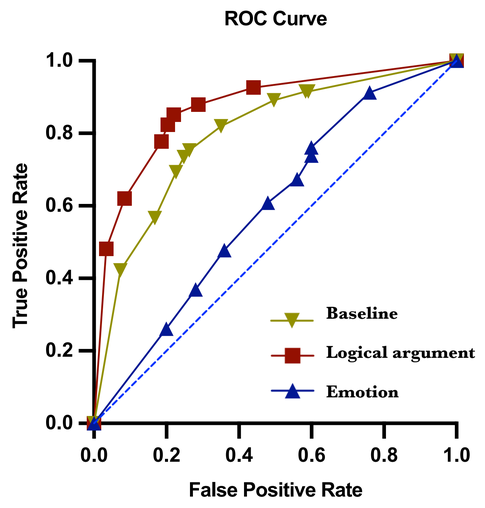
DOES OUR FORMULATION CAPTURE GOAL REALIZATION?

Metric-level agreement

BaT (Spearman's ρ)	**0.65**
PaT (Spearman's ρ)	**0.66**
NRBaT (Spearman's ρ)	**0.83**
COMMIT (Fleiss' κ)	**0.59**
REL (Randolph's κ)	**0.72**
MAN (Randolph's κ)	**0.52**
QUAL (Randolph's κ)	**0.86**
CONST (avg. TPR)	**25%**

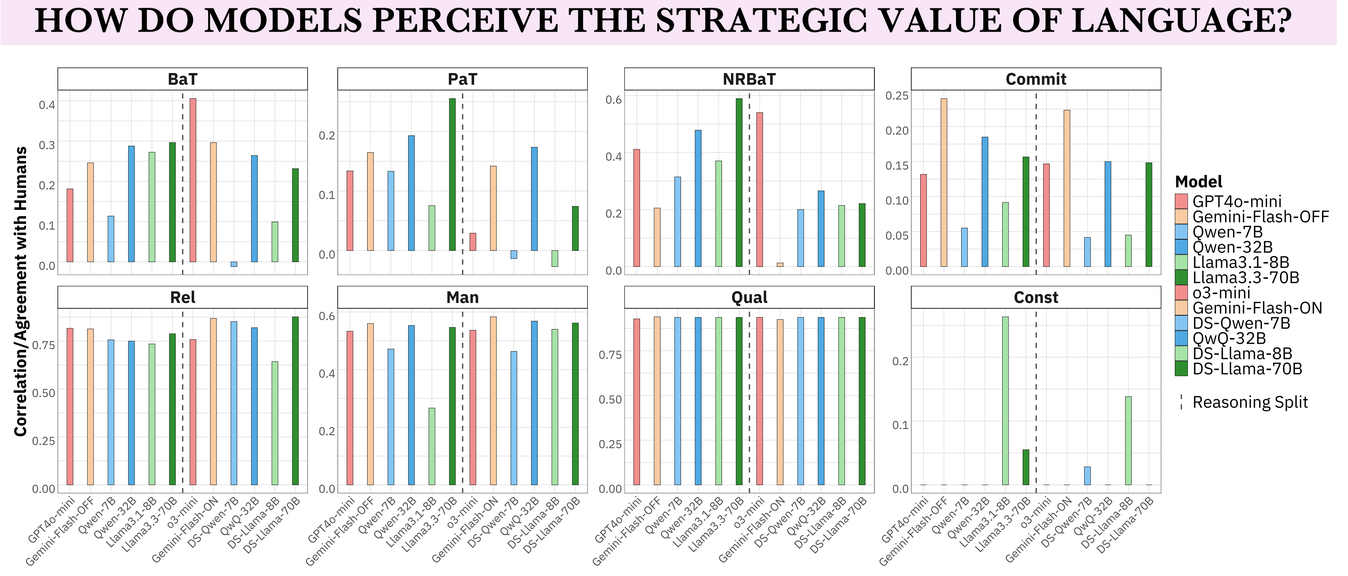
Outcome-level agreement

Outcome decision (Fleiss's κ)	**0.29**
Jaccard similarity (avg. reasons)	**0.46**
Complete agreement on reasons	**29%**



(1) Annotators show good agreement over our metrics on our dataset
(2) Our metrics are able to predict each annotator's outcome of the conversation
(3) Sensitive to more objective component of decision making.

HOW DO MODELS PERCEIVE THE STRATEGIC VALUE OF LANGUAGE?



(1) Model size shows an increase in performance on our metrics

(2) Reasoning ability does not help but largely hurts

REFERENCES

[1] Nicholas Asher, Soumya Paul, and Antoine Venant. Message exchange games in strategic contexts. Journal of Philosophical Logic, 46:355–404, 2017.
[2] David I. Beaver, Craige Roberts, Mandy Simons, and Judith Tonhauser. Questions under discussion: Where information structure meets projective content. Annual Review of Linguistics, 3(1):265– 284, 2017.
[3] Herbert Paul Grice. Logic and conversation. Syntax and semantics, 3:43–58, 1975.
[4] Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. GTBench: Uncovering the strategic reasoning capabilities of LLMs via game-theoretic evaluations. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL https://openreview.net/forum?id=ypggxVWIv2.

Check our D&D paper @