

Strategic Dialogue Assessment: The Crooked Path to Innocence

Anshun Asher Zheng

The University of Texas at Austin

ASHER.ZHENG@UTEXAS.EDU

Junyi Jessie Li

The University of Texas at Austin

JESSY@UTEXAS.EDU

David I. Beaver

The University of Texas at Austin

DIB@UTEXAS.EDU

Editor: Pierre Lison

Submitted 10/2025; Accepted 01/2026; Published online 01/2026

Abstract

Language is often used strategically, particularly in high-stakes, adversarial settings, yet most work on pragmatics and LLMs centers on cooperative settings. This leaves a gap in the systematic understanding of strategic communication in adversarial settings. To address this, we introduce SDA (Strategic Dialogue Assessment), a framework grounded in Gricean and game-theoretic pragmatics to assess strategic use of language. It adapts the ME Game jury function to make it empirically estimable for analyzing dialogue. Our approach incorporates two key adaptations: a commitment-based taxonomy of discourse moves, which provides a finer-grained account of strategic effects, and the use of estimable proxies grounded in Gricean maxims to operationalize abstract constructs such as credibility. Together, these adaptations build on discourse theory by treating discourse as the strategic management of commitments, enabling systematic evaluation of how conversational moves advance or undermine discourse goals. We further derive three interpretable metrics—*Benefit at Turn* (BAT), *Penalty at Turn* (PAT), and *Normalized Relative Benefit at Turn* (NRBAT)—to quantify the perceived strategic effects of discourse moves. We also present CPD (the Crooked Path Dataset), an annotated dataset of real courtroom cross-examinations, to demonstrate the framework’s effectiveness. Using these tools, we evaluate a range of LLMs and show that LLMs generally exhibit limited pragmatic understanding of strategic language. While model size shows an increase in performance on our metrics, reasoning ability does not help and largely hurts, introducing overcomplication and internal confusion.¹

Keywords: Strategic communication, Gricean pragmatics, game-theoretic pragmatics, LLM pragmatics, non-cooperativity, utility function

1. Introduction

We often encounter conversations in which the interlocutors do not share a common goal (Walton and Krabbe, 1995; Oswald, 2010; Asher and Lascarides, 2013 *i.a.*), such as in the interrogation shown in Figure 1. Yet, when it comes to interpreting discourse, most work has long assumed cooperativity, deeply rooted in the well-established tradition of Gricean pragmatics (Grice, 1975; Clark and Schaefer, 1989). Assuming a shared goal and cooperative principles has indeed yielded

1. We provide data and code in Github and huggingface repos.

valuable insights in contexts where such assumptions are reasonable. However, these assumptions become problematic in scenarios like the one shown in Figure 1. The dialogue in the top panel shows a flight attendant being cross-examined over whether she falsified an incident report. The witness’s response that her colleague once corrected an incident report gives rise to the implicature that she herself did not falsify a report. It implies a denial of the charge while employing deflection, hedging (“to my knowledge”), and euphemism (“corrected” for “falsified”). Interpreted through a cooperative lens (as shown in the blue box), the response is treated as flouting the maxim of relevance while still preserving cooperativity, and thus read as a truthful denial of the charge, even though the speaker may have previously falsified an incident report. Such a move is thus always beneficial for the speaker in terms of helping the speaker to get rid of the charge. In contrast, a rational agent—one that understands non-cooperativity and the strategic use of language—would interpret it differently (as shown in the yellow box). In such a context, the implicature derived from the response is not readily trusted. Because the move involves deflection and leaves room for inconsistency, it is only *partially* beneficial to the speaker, and any perceived gain is *temporary*, as it can be retracted if an inconsistency occurs later.

Language is used strategically² to advance speakers’ goals. When goals are aligned, Gricean pragmatics provides a good strategy: following the maxims generally helps speakers achieve their (common) goals. When goals are misaligned, however, things become more complicated. Assuming that maxim-following is still optimal can lead to misinterpretations like in Figure 1; and even if it occasionally yields the correct interpretation (e.g., when the speaker indeed wants to project cooperativity), following maxims *alone* is rarely sufficient for a speaker to achieve their goals in adversarial settings. We therefore focus on strategic communication in non-cooperative discourse, an under-explored and intriguing setting for examining how discourse goals are realized.

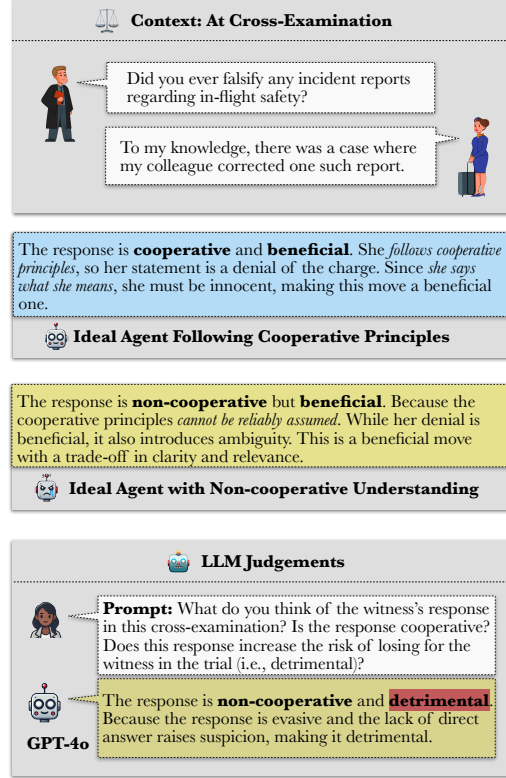


Figure 1: This figure shows a cross-examination dialogue, two idealized interpretations, and GPT-4o’s judgment, which diverges from both interpretations and mistakenly treats the response as increasing the risk of losing. The full model output is shown in Appx. Fig. 13.

2. In the literature, *strategic* is often used almost in a similar sense to *non-cooperative*, to emphasize goal misalignment. We, however, use *strategic* to emphasize how effectively a speaker uses language as a means to advance their discourse goals (e.g., as in strategic value or strategic effectiveness). We return to this distinction in section 3.

While theories of cooperative language are numerous and well-established, systematic accounts of non-cooperative language use have received comparatively less attention. Game-theoretic pragmatics (Parikh, 2000; Glazer and Rubinstein, 2006; Asher et al., 2017) represents one such line of work.³ Approaches within game-theoretic pragmatics, such as Signaling Games (Lewis, 1969; Franke, 2009) and Rational Speech Act (RSA) theory (Frank and Goodman, 2012), do not enforce cooperativity, yet they also do not emphasize their application to non-cooperative discourse. For the most part, they demonstrate how traditional Gricean-style pragmatics can be recast in a game-theoretic framework, for example, in deriving scalar implicatures (Pavan, 2013; Goodman and Stuhlmüller, 2013). By contrast, the Message Exchange (ME) Games paradigm (Asher et al., 2017) is more dedicated to modeling non-cooperative discourse, where interlocutors’ goals are misaligned. In this framework, discourse moves are evaluated by a third-party “jury” in terms of how effectively they contribute to the realization of speakers’ discourse goals. This underscores the fact that even in non-cooperative dialogue, interaction is not arbitrary and is governed by constraints arising from public commitments and accountability to an evaluating audience, which limit how participants can strategically maneuver without undermining credibility. However, due to practical constraints, the ME Game “jury” has not been systematically applied to extensive real-word conversations.

Partly because of the abstraction inherent to formal theories, applying such frameworks to naturally occurring data at scale is challenging. We can see this challenge in terms of two subproblems: first, some key terms are specified in ways that work well for certain contexts (e.g., cooperative discourse) but do not readily generalize to others; second, even when well-defined in theory, some constructs are difficult to estimate empirically, making assessment and operationalization challenging without additional adaptation. Most frameworks, for example, posit an abstract utility function (i.e., a measure of what is valued in goal realization). A common specification in terms of Gricean maxims (as in RSA; Goodman and Stuhlmüller 2013) works well in cooperative settings, but does not naturally extend to adversarial contexts, where following maxims alone is rarely sufficient for achieving one’s goals, as noted before. The ME Games model goes further by offering a more concrete specification of the utility function tailored to adversarial contexts, thereby partially addressing the challenge that existing specifications do not readily generalize beyond cooperative discourse. But it does not resolve the problem of empirical estimability: its technical constructs—such as the credibility distribution $P_k(\text{Good}_i)$ and winning potential $\text{win}_i(k)$ —are not directly observable in most types of naturally occurring discourse. Moreover, the original specification abstracts away from finer-grained strategic effects of discourse moves; in section 3.2.2 we present an extension to capture these distinctions.

The ability to recognize the strategic value of language—namely, the extent to which a move advances or undermines a speaker’s goals—is also important in the current age of AI. Despite growing attention to AI safety and alignment (Bowman et al., 2022), models still show sycophancy (Sharma et al., 2024) and limited critical thinking (Musi et al., 2025). These reflect failures to prioritize aspects of conversational moves that have greater long-term strategic value, such as preserving reliability rather than chasing short-term agreement. At the same time, models may violate safety protocols when unmonitored (Greenblatt et al., 2024), producing responses that appear locally beneficial (securing immediate agreement or task success) but ultimately undermine alignment goals. Both behaviors are highly undesirable; as deeper alignment is called upon (Qi et al., 2025), being

3. Other frameworks exist as well, such as those grounded in learning theory (Sicilia et al., 2022). Since our focus is on game-theoretic pragmatics, we leave engagement with these alternative approaches for future work.

able to assess models’ understanding of the strategic effects of discourse moves is crucial. Such assessment is valuable not only for safety monitoring but also for downstream applications that require strategic reasoning, such as simulated debate and negotiation, where models must accurately track how their interlocutors’ moves shape the goals at stake.

As we will show, existing LLMs (at least absent specific training, which we do not study here) are weak at nuanced analysis of dialogues that do not follow the Gricean principles, and do not accurately recognize the strategic effects of language. For example, in the bottom panel of Figure 1, the model identifies the response as non-cooperative, but it does so by simply treating maxim violations as evidence of non-cooperativity rather than considering whether the speakers in fact are pursuing opposed goals. Additionally, it labels the move as detrimental (i.e., causing the defendant to lose) rather than beneficial, indicating that at least the model behavior does not show a proper understanding of the strategic value of the utterances. The model’s judgment is superficial, focusing only on the presence of deflection without evaluating the overall strategic gain, which is empirically beneficial, helping the flight attendant to get rid of the charge.

These observations highlight a dual motivation for our work. Theoretically, existing frameworks need extension and operationalization to capture finer-grained strategic effects of discourse moves and to be assessed on large datasets. Practically, such a framework provides the basis for evaluating LLMs’ performance on recognizing the strategic value of language. We propose SDA (Strategic Dialogue Assessment), a framework that bridges theory and empirical evaluation. SDA adapts the ME Game jury function in two ways: first, it introduces a commitment-based taxonomy of discourse moves, enabling a finer-grained account of strategic effects; second, it operationalizes abstract constructs such as credibility by approximating them with observable proxies grounded in Gricean maxims. In this view, discourse moves are treated as the strategic management of commitments: each commitment carries potential gains and losses, and speakers may exploit maxim violations not arbitrarily but as part of their strategy to maximize benefits while avoiding inconsistency (see Figure 1). Building on this foundation, SDA defines three metrics—Penalty at Turn (**PAT**), Benefit at Turn (**BAT**), and Normalized Relative Benefit at Turn (**NRBAT**)—that quantify perceived losses, benefits, and cumulative strategic gain, respectively. They distinguish the output of the utility function, which only produces a single value, into different categories (i.e., positive or negative), which reflect different strategic effects of language use, with the cumulative value capturing the trade-off between them.

SDA provides a well-motivated framework for performing human annotation, and we release a corpus on legal discourse, CPD (the **Crooked Path Dataset**), with accompanying annotations on a subset of the corpus. We use these annotations to evaluate strategic effectiveness across different discourse types, noting that such effectiveness can in principle be assessed in both cooperative and non-cooperative settings, since in either case speakers use language to pursue goals. Applying our method to cross- and direct examinations (i.e. dialogues with an opposing witness, and dialogues with a lawyer’s own witness), we reveal that cooperative and non-cooperative discourse are asymmetric with respect to the annotated properties. We further use the annotations to validate that SDA, heavily grounded in the ME jury function, can effectively capture these strategic effects, which contribute to the realization of discourse goals. Applying our metrics to discourse moves in predicting conversational outcomes, we find that they achieve strong—but not perfect—predictive power, indicating that the identified properties reflect how language strategically shapes discourse goals. We emphasize that perfect prediction is neither expected nor the aim of our method, as conversational outcomes are themselves subjective and shaped by many factors beyond discourse alone

(e.g., speaker bias); instead, our results show that the ME Game utility function, with our extensions, is both empirically operationalizable and theoretically meaningful. By grounding itself in a theoretically motivated account of strategic communication, SDA thus provides a way of relating theory to real-world discourse.

We then use our metrics to conduct a preliminary evaluation of a suite of state-of-the-art LLMs, varying in size and reasoning capability. Under our prompting setup, larger models tend to align more closely with human judgments, particularly in identifying strategic gains, but the reasoning-enhanced models often perform worse, particularly in identifying strategic losses. This degradation is seen in models’ behavior to misinterpret surface-level damage control strategies as the overall effects of a commitment, along with difficulty in handling self-contradictory behavior, signaling that existing training paradigms for reasoning may not necessarily enhance important aspects of pragmatic capabilities.

2. CPD: The Crooked Path Dataset

To give a more concrete sense of the type of discourse we focus on in this paper, and what is crucial for evaluating it, we now introduce CPD (the **Crooked Path Dataset**), a dataset drawn from legal cross-examinations, which we believe will also be of broad interest to the community. We focus on cross-examinations in criminal trials for several reasons: (1) it can be determined whether the goals of the two parties are opposed based on their respective roles in the trial; (2) trials are paradigmatic zero-sum games, where there is inevitably a winner and a loser; and (3) attorneys and prosecutors are professionally trained to engage in strategic questioning. We therefore treat cross-examination as an instance of adversarial settings, where the participants’ goals are misaligned. At the same time, we note that our goal is not aiming to measure the degree of (non)cooperativity itself; rather, we use this adversarial setting as a testbed for assessing the strategic effects of discourse moves—how effectively speakers manage commitments to advance their goals. We return to this distinction in section 3.

We collect testimonies from three prominent U.S. trials: the West Memphis Three Trials (1994), the O.J. Simpson Trial (1995) and the Enron (Lay & Skilling) Trial (2006).⁴ We focus on the cross-examination part of each testimony. A representative snippet is given in (1), a cross-examination from the West Memphis Three case. It involves the witness Richard Ofshe (RO), who was called by the defense as an expert on police coercion. His goal is to maintain credibility so that his testimony will be accepted, while also demonstrating that the defendant’s statement was coerced by police. The prosecutor (P), by contrast, aims to undermine his credibility and argues the opposite.

- (1) P: How many states and how many courts have refused to accept you as an expert in this work?
- RO: No state has ever refused to accept me as an expert.
- P: How many courts?
- RO: There’s one case in which a line of testimony to which my testimony would have been foundational was rejected. It has to do with whether or not a certain theory...

4. Data is sourced from <https://famous-trials.com/>, which provides transcripts and other trial details.

Trial	Defense	Prosecution	Total	Defense %
WMT	651	575	1226	53.1%
Enron	27	47	74	36.5%
Simpson	1608	417	2025	79.4%

Table 1: Q/A pair distribution by questioner role across four trials. The Defense % column shows the proportion of defense-attorney-led Q/A pairs.

In total, the cross-examinations in the three trials consist of 4452 turns, with 3325 of these being Q/A pairs (the rest are largely objections from the opposing side). The distribution of Q/A pairs and sides for each trial is shown in Table 1.

Legal Assumptions Because we work with legal data, we hope to be careful and transparent in stating our assumptions during both use and annotation, and we seek to avoid any unintended implications with input from a legal expert.

First, we assume that the trials we analyze are zero-sum games in terms of their outcomes: the defendant or the prosecutor either wins or loses. While we fully acknowledge the legal principle that U.S. attorneys are obligated to pursue justice rather than merely secure convictions, in the specific cases we collected and analyze the prosecution and defense can be reasonably characterized as adopting adversarial roles, whose goals are centered on securing a conviction or an acquittal.

Second, we note that cross-examination is governed by evidentiary rules that constrain both the content and the form of questioning. Under the Federal Rules of Evidence (FRE), the content of cross-examination is restricted to the subject matter of direct examination and matters bearing on the witness’s credibility (FRE 611(b)). At the same time, its form permits the use of leading questions when a party examines an adverse or hostile *witness* (FRE 611(c)). Together, these rules institutionalize cross-examination as an interaction structured around misaligned goals, delimiting both what may be asked and how testimony may be elicited.

Third, although we include testimonies from both defendants and witnesses, we do not simply conflate these roles. We recognize that witnesses—particularly expert witnesses—may not personally hold adversarial interests relative to the opposing side. However, given that witnesses are typically carefully chosen and prepared by the party that calls them and their testimony is subject to adversarial challenge under the evidentiary rules described above, we treat them as representing that party’s interests. Thus, we assume that witnesses generally align with the strategic goals of the side they testify for and are situated in opposition to the other side within the adversarial structure of the trial.

Finally, we treat cross-examination as representative of non-cooperative discourse and direct examination as representative of cooperative discourse, based on whether the interlocutors’ goals are aligned. While the high-profile nature of our cases suggests that witnesses may have undergone extensive preparation—which could influence the surface features of their responses, such as conciseness or restraint—we do not expect this to alter the deeper goals.

3. Non-cooperative Discourse and Sources of Strategic Value

As noted earlier, the terms *strategic* and *non-cooperative* are both used in prior literature to refer to behavior resulting from goal misalignment, or to situations involving such misalignment. Sometimes they are used almost interchangeably, and sometimes with quite distinct meanings. We briefly

survey these uses before clarifying how we use the terms in the remainder of the paper. In economic game-theoretic work, strategic behavior is often conflated with non-cooperative games (Osborne and Rubinstein, 1994), and in classical economics, agents are modeled as individually rational and self-interested, which frequently leads to non-cooperative settings with divergent preferences. In behavioral game-theoretic work, a more nuanced distinction is drawn: strategic agents are those who reason based on beliefs about others’ strategies, whereas non-strategic agents may act according to heuristics or rules of thumb without such reasoning (Camerer, 2003; Wright and Leyton-Brown, 2019). In multi-agent systems and AI research, strategic behavior is often treated as if it were equivalent to adversarial or non-cooperative behavior, reflecting goal misalignment; this is evident both in technical treatments of incomplete-information games such as poker (Sandholm, 2010) and in broader surveys that highlight the dominance of non-cooperative framings and call for new approaches to cooperative AI (Dafoe et al., 2020). In political science and international relations, a similar conflation appears in Schelling’s *Strategy of Conflict*, where strategy is inherently linked to conflict and non-cooperation (Schelling, 1960). We focus here on non-cooperative discourse as a testbed for assessing strategic effects.

Turning now to work on dialogue, Asher et al. (2017) uses *strategic conversations* to mean discourse where interlocutors’ goals are misaligned. By contrast, in this paper, we will use the terms *non-cooperative* and *strategic* as follows: *non-cooperative* refers to settings of goal misalignment, while *strategic* refers to how effectively discourse moves realize a speaker’s goals. In principle, strategic effectiveness can be assessed in both cooperative and non-cooperative contexts, since in either case speakers use language (as a strategy) to pursue goals. In cooperative contexts, following cooperative principles provides an effective way to realize the shared goals; in adversarial contexts, appearing to be cooperative can still be strategically useful to advance self-interested goals. A maximally strategic speaker in adversarial settings may appear cooperative (e.g., by following Gricean maxims), not because their goals are aligned, but because the appearance of cooperativeness enhances credibility and serves their strategy. However, relying on maxims *alone* is not sufficient to achieve discourse goals, as we discuss below in section 3.1.

We focus here on non-cooperative discourse because it is comparatively under-explored, and we are interested in assessing how the discourse goals are realized in such settings. We acknowledge similar questions can be asked about cooperative discourse (e.g., what can be said beyond Gricean maxims in goal realization), but we leave this for future work. In the remainder of this section, we review both Gricean and game-theoretic pragmatics, which jointly constitute the theoretical foundation of our proposal. We explore how each framework approaches non-cooperative discourse and evaluates the strategic effects of discourse moves, and note their potential limitations.

3.1 Gricean Pragmatics: Maxims in Non-cooperative Discourse

The most influential account of conversational cooperativeness is due to Paul Grice (Grice 1975, 1989), later developed by Neo-Gricean pragmaticists (e.g., Horn 1984; Levinson 1987). These theories propose that rational interlocutors structure their contributions around conversational maxims—those of *quality, quantity, relevance, and manner*.⁵

In cooperative contexts, these maxims serve as norms, and violations are interpreted as flouting—intentionally and transparently violating maxims to preserve overall cooperativeness (Brown, 1987;

5. We later group the maxim of quantity under the maxim of quality, since whether an interlocutor provides an adequate and appropriate amount of information depends on knowing what the interlocutor actually knows.

Thomas, 2014). But in non-cooperative settings, the role of maxims shifts. Here, interlocutors cannot assume that others are maximizing quality, quantity, or relevance (Asher et al., 2017, p.363). Still, speakers frequently behave as if they were cooperative (e.g., Levinson 2000; Horn 2006). That is, even if they are not literally maximizing the maxims, they act in ways that give the appearance of doing so, because such behavior projects credibility and enables hearers to interpret their utterances.

This makes Gricean maxims also relevant for understanding goal realization in adversarial discourse. A non-cooperative speaker can strategically invoke or mimic maxim-following to generate trust, manage credibility, or deflect damaging commitments (Goffman, 1970). In this sense, maxims are not binding rules but strategic resources: tools that speakers exploit to advance their own goals. A maximally strategic speaker, at least in legal cross-examinations, may therefore appear cooperative precisely because projecting cooperativity enhances their ability to achieve their goals. Still, maxims alone do not determine the strategic effects of discourse moves: for example, in response to “Are you taking any medication?”, both “Yes” and “No” appear cooperative, making the speaker appear to be credible, yet their strategic effects on the discourse goal differs. We therefore treat maxims primarily as a credibility controller (see section 4.4): they mark how reliable a response appears and, in doing so, modulate the strategic effects that response can have.⁶

3.2 Game-Theoretic Pragmatics: Maximizing Strategic Utilities

Game-theoretic pragmatics offers a complementary perspective on communicative behavior by modeling discourse as a strategic decision-making process. While several works model pragmatic reasoning within a game-theoretic framework (Parikh, 2000; Glazer and Rubinstein, 2006; Franke, 2013), they primarily focus on goal-aligned settings, in which Gricean behavior is treated as a desirable or equilibrium outcome; however, such (Gricean) cooperativity is not inherently required by the framework itself, but depends on the agents’ payoff structure. Instead, with appropriate refinement, they can naturally apply to scenarios involving competition and manipulation—such as debates or cross-examinations—where interlocutors have misaligned goals (Asher and Lascarides, 2013; Asher and Paul, 2016; Asher et al., 2017). In addition, game/decision-theoretic works on LLM reasoning (e.g., Duan et al. 2024) are also relevant though they are less engaged with sophisticated communication than the pragmatics line of work. In the following, we briefly discuss three game-theoretic accounts of communication—Signaling Games (Lewis, 1969; Franke, 2009), Rational Speech Act (RSA) theory (Frank and Goodman, 2012), and Message Exchange (ME) Games (Asher et al., 2017)—that will help understand the traditions of this approach to pragmatics and clarify how our proposal in section 4 connects to them. We refer interested readers to Benz and Stevens (2018), which provides a systematic review of different frameworks.

3.2.1 SIGNALING GAMES AND RATIONAL SPEECH ACT (RSA) THEORY

A classical starting point for game-theoretic models of communication is Signaling Games (Lewis, 1969; Franke, 2009). In this setting, a pragmatic speaker (or sender) selects a signal to influence a literal listener (or receiver), who interprets the signal and chooses an action. The speaker’s choice

6. A more sophisticated treatment might allow that speakers sometimes attempt to project the appearance of cooperative flouting of maxims. For instance, a speaker might conceivably deploy irony or *reductio ad absurdum* arguments, with the intention of transparently violating the maxim of quality, if they felt this would have a desirable rhetorical effect. We are doubtful that an extension to include cooperative flouting would have any significant impact on the textual analyses we provide in this paper and so leave it to future work.

is driven by expected utility: a signal is successful if it guides the listener toward an outcome that maximizes the payoff for *both* parties. A central question is whether the interaction reaches an equilibrium—that is, a stable outcome in which neither party has an incentive to unilaterally change their choice. For example, in Franke (2009)’s signaling-game model of scalar implicature, a speaker chooses between utterances such as *some* and *all* depending on the underlying state of the world (e.g., whether all objects in a set have a property or only some do). The payoff structure is *cooperative*: both players receive a reward if the listener correctly infers the true state. In this setup, if the world is “all,” the best choice for the speaker is to say *all*, which leads the listener to guess correctly. If the world is “some,” the speaker could in principle say *some* or *all*, but if she says *all* the listener will guess wrongly, lowering both players’ payoff. Thus, the equilibrium strategy is for the speaker to say *all* when the state is “all” and *some* when the state is “some.” At equilibrium, the listener interprets *all* literally, but interprets *some* pragmatically as “some but not all,” since otherwise the speaker would have used the stronger signal. The focus in Signaling Games is on whether the utterance leads to an optimal action at equilibrium.

Rational Speech Act (RSA) theory (Frank and Goodman, 2012) extends this paradigm by introducing recursive reasoning between the speaker and the listener. While Signaling Games focus on equilibrium strategies, RSA models communication as a sequence of probabilistic belief updates. It distinguishes three roles (compared to only two in Signaling Games). A literal listener interprets an utterance u solely by its truth-conditional meaning $\llbracket u \rrbracket$, yielding a probabilistic distribution $P(s \mid \llbracket u \rrbracket)$ over possible states s . Then a pragmatic speaker chooses utterances based on how a literal listener would interpret u , balancing the informativity of u (i.e., $\log P(s \mid \llbracket u \rrbracket)$) against its communication cost $C(u)$ (i.e., $\log P(s \mid \llbracket u \rrbracket) - C(u)$). This speaker utility is primarily grounded in the maxims of quantity and manner, while α serves as a rationality parameter that controls how strongly the speaker prefers high-utility choices ($P_S(u \mid s) \propto \exp(\alpha(\log P(s \mid \llbracket u \rrbracket) - C(u)))$). Finally, a pragmatic listener inverts this reasoning to infer the pragmatic speaker’s intended meaning via Bayes’ rule ($P_L(s \mid u) \propto P_S(u \mid s) \cdot P(s)$ with $P(s)$ denoting prior beliefs about world states). Unlike in Signaling Games, the pragmatic speaker in RSA does not assume strict utility maximization: instead, utterances are chosen probabilistically via a softmax function modulated by a rationality parameter α .

Consider the same scalar implicature example. RSA derives the implicature through recursive probabilistic reasoning rather than equilibrium optimality. Suppose there are two possible states: s_{all} and s_{some} , and that some is compatible with both states while all is compatible only with the s_{all} . It follows from these assumptions that upon hearing *all*, the literal listener assigns probability 1 to s_{all} , whereas upon hearing *some*, the listener assigns equal probability to s_{all} and s_{some} . A pragmatic speaker who prefers to be informative therefore assigns higher probability to *all* than to *some* in s_{all} , since *all* induces a posterior that is more concentrated on the true state.⁷ We refer interested readers to Degen (2023) for a systematic review of RSA modeling.

These frameworks provide a generative model for predicting both what an interlocutor would say and how that utterance would be interpreted. However, in modeling discourse such as (1), the devil is in the details.

First, to automate the (recursive) reasoning that such discourse requires, one must determine the set of possible states that an utterance can map onto. This, however, depends heavily on the interlocutor’s epistemic state and world knowledge—factors that are rarely fixed or directly observable.

7. Under the standard RSA choice rule, $P_{L_0}(s_{\text{all}} \mid \text{all}) = 1$ while $P_{L_0}(s_{\text{all}} \mid \text{some}) = \frac{1}{2}$, so *all* is strictly preferred.

Consider RO’s response in example (1) that he was not rejected by the state as an expert. What are the possible underlying world states? The possible contexts are diverse: he may not have been rejected as an expert at all; he may not have been rejected as a witness but also not formally recognized as an expert; he may have been rejected by one state but not others; or he may have been rejected by multiple states. But who decides which states belong in this set? There is no straightforward way to pin down the set of states that a (literal) utterance corresponds to.

Second, there is a specification challenge: one must determine the appropriate utility function, which is highly flexible and depends on the interlocutor’s goals. By default, games above assume a cooperative pragmatic speaker whose utility is aligned with that of the listener and defined in terms of optimizing the communicative properties prescribed by the Gricean maxims. However, this assumption breaks down in more adversarial settings such as interrogations, as in (1). In this example, RO’s utility is not symmetric to P’s: unlike in cooperative settings where both parties share the same utility function, RO’s gains do not entail gains for P. Nor is RO maximally faithful to the maxims (e.g., RO is not maximally faithful to quantity, since he provides only a partial answer). To capture strategic language, we need a utility function that allows for asymmetry and characterizes contributions with respect to the speaker’s own goals.

Third, even in RSA’s recursive model, the interaction is restricted to a single speaker and a single listener, whereas in reality a pragmatic speaker may need to consider the utilities of multiple parties—for example, a defendant/witness must reason not only about the prosecutor but also about the jury. As a result, many applications of these models (e.g., Vogel et al. (2014); Anderson and Dillon (2019); Spinoso-Di Piano et al. (2025)) are confined to simplified scenarios with a small specified set of world states. Offering a comprehensive generative model that resolves all these issues lies far beyond the scope of this paper. Our focus, instead, is on assessing the strategic effects of discourse moves on the realization of discourse goals, which is closely connected to the second problem noted above. One promising direction comes from Message Exchange (ME) Games (Asher et al., 2017), which is specifically designed to model non-cooperative conversations. It offers an explicit treatment of the utility function in such cases and is not primarily grounded in Gricean maxims.

3.2.2 MESSAGE EXCHANGE (ME) GAMES

Message Exchange (ME) Games (Asher and Paul, 2016; Asher et al., 2017; Asher and Paul, 2018) provide another formal game-theoretic model but originally designed to account for non-cooperative discourse. This line of work specifically focuses on cases where the players’ interests are strictly opposed,⁸ as explored in the current paper. ME Games model discourse as an infinite game in which, as in other communication games, the speaker and listener reason about each other.

Unlike in other frameworks, ME Games assume that, in addition to the classic speaker-listener setup, non-cooperative conversations also involve a third party, the “jury” in their term, which serves as the contribution evaluator. On this view, conversational moves are not simply aimed at persuading interlocutors or projecting credibility to them so that those moves are trusted; rather, they are aimed at convincing the third party. The jury then serves as the scoring function of the game. Players formulate their conversational goals around what they believe they can defend against oppositions

8. Goal divergence can occur at varying levels. In some cases, interlocutors may perform non-cooperative moves as part of a broader effort to eventually reach consensus, as in negotiations. The cases investigated in ME Games, and in the present paper, are situated in stricter zero-sum settings, thereby setting aside the ambiguities that arise in discourses where goals sometimes align and sometimes diverge.

from their opponent and what they anticipate will be assessed favorably by the jury. The jury decides whether a move advances a player’s conversational goals, which in turn shapes the speaker’s utility function, since the value of a move depends not only on how the listener would interpret it but also on how the jury would judge it.

The jury function, τ , is defined in (1), which decides the benefit of each conversational turn k for speaker i . A turn yields benefit only if it is both *coherent* (COH: whether the current turn k connects to prior discourse via certain discourse relations; $\text{COH}_i(k) \in \{-1, 1\}$) and *responsive* (RES: whether the current turn k connects to the immediate prior turn via discourse relations; $\text{RES}_i(k) \in \{-1, 1\}$), reflecting whether the discourse moves are forming meaningful discourse structure, which in turn contributes meaningfully to the realization of discourse goals. These benefits are sustained only if the turn is also *consistent* (CONS: whether the current turn k contradicts prior commitments of speaker i ; $\text{CONS}_i(k) \in \{0, 1\}$), *credible* (i.e., how trustworthy speaker i is at turn k , modeled as $\text{P}_k(\text{Good}_i) \in [0, 1]$), and aligned with a potential *win* for the speaker ($\text{win}_i(k) \in \{0, 1\}$). Thus, how strategic an interlocutor is can be understood as the interlocutor’s ability to pursue their own interests by maximizing their score τ . For the full formal reasoning behind the specific terms chosen, we refer interested readers to Asher et al. (2017).

$$\begin{aligned} \|\tau_k\|_i &= (\text{COH}_i(k) + \text{RES}_i(k)) \times \text{CONS}_i(k) \\ &\quad \times \text{P}_k(\text{Good}_i) \times \text{win}_i(k) \end{aligned} \tag{1}$$

ME Games offer a way to model non-cooperative conversations and misaligned utilities by grounding them in coherence relations. Coherence, on this view, plays a crucial role for two reasons. First, coherence is to a great extent externally observable. In the case of jury trials, this is effectively built into the institutional setting, since we assume that coherence must be judged by the jury without access to the interlocutors’ private epistemic states. Second, coherence serves as evidence that a discourse move contributes to the speaker’s broader goal structure: only moves that are coherent can be interpreted as goal-directed and therefore count as goal-realizing contributions (Asher and Lascarides, 2013). This close connection between coherence and goals of text holds not only in cooperative discourses but also in not necessarily cooperative settings (Asher et al., 2017). The jury function therefore takes coherence relations as proxies for determining whether a speaker’s moves advance their discourse goals, and the utilities for each speaker are calculated on this basis. In this way, ME Games provide an account that sidesteps reliance on Gricean principles (i.e., it is not necessary to rely on maxims to achieve one’s goals), which distinguishes it from accounts that focus on cooperation.

Reproducing the jury function for an analysis of our dataset would be of clear value, but we diverge in some ways from the ME Games framework both because of our own theoretical proclivities, and because of certain practical concerns. One concern is that while the function helps address the specification challenge identified in section 3.2.1—namely, defining a utility function suitable for adversarial contexts—it still faces an operationalization challenge: many of its abstract terms are difficult to estimate empirically, making them hard to apply in practice. For instance, there is no straightforward way to obtain a function $\text{P}_k(\text{Good}_i)$ that outputs the credibility distribution of an interlocutor, as this would depend on numerous factors that may be explicitly present (e.g., inconsistencies) or absent (e.g., speaker bias) in the discourse. In reality, credibility assessments are also ambiguous, which worsens the estimability issue. Consider the same example in (1): the witness admits there was a case in which his testimony was rejected, which could cast doubt on

his expertise. This suspicion may reduce his credibility; however, this reduction is also debatable, since the witness also appears trustworthy, and it seems unlikely that he would fabricate something that harms his own reputation. Hence, one might assess the witness’s credibility in two different ways: either assigning a higher probability to $P(\text{Bad})$ than $P(\text{Good})$, viewing his expertise as unreliable, or the opposite, assigning more weight to $P(\text{Good})$, since his words appear sincere. Similarly, winning potential $\text{win}_i(k)$ is also hard to estimate as it is defined in terms of the intersection of the possible future paths following the current move and all possible winning paths; modeling this accurately would require an “omniscient” perspective over the discourse space, which is in reality often impractical.

In addition, while the specification of the jury function in ME Games identify relevant and insightful properties, it abstracts away from finer-grained strategic effects on discourse goals. Consider the example in (2), another cross-examination from the West Memphis Three case. It involves the witness Richard Ofshe (RO), who was called by the defense as an expert on police coercion. His goal is presumably to maintain credibility so that his testimony will be accepted, while also demonstrating that the defendant’s statement was coerced by police. The prosecutor (P), by contrast, aims to undermine his credibility and argues the opposite.

- (2) P: Did you find anything in any of that evidence to indicate that any of the officers yelled or used a loud voice or were degrading to the defendant in those tapes or in that testimony that you reviewed?
- RO: No, the officers testified they did not do that.
- P: Okay, and in those tapes that you observed, you didn’t hear anything of that nature, did you?
- RO: No, I did not.
- P: And is what you term or what you find in there coercive that the officers asked at times, leading questions – is that one of the things?
- RO: The questions were more than leading. The questions were very directly specifying what the answers should be.

RO’s responses are all coherent, responsive, consistent, and avoid direct commitments that would contradict his goals as regards the facts of the case and the reliability of his testimony (e.g., any suggestion that the defendant was not coerced by police, or to anything that would lead to the outright rejection of his testimony). Thus, this testimony is aligned with a potential win for him (or his side), and so $\text{win}_i(k) = 1$. If credibility is assessed entirely in terms of sincerity, then in theory these responses would not damage the speaker’s credibility either. Accordingly, the jury function would predict that all of these responses carry equal value in advancing the speaker’s goals.

Yet empirically, this is not the case. Admissions such as being rejected as an expert in another court, or acknowledging that the police did not display behaviors typical of coercion, do not contribute positively to the witness’s goal of maintaining credibility while demonstrating that the defendant was coerced by police. By contrast, statements such as not having been refused recognition as an expert by the state, or pointing out that the police asked misleading questions, clearly do.⁹ This observation suggests that although these responses are all coherent, and thus form part

9. We also note that even if one treats prior rejection in other courts as a credibility attack, the function only distinguishes some responses (e.g., in (1)) but still fails to differentiate others (e.g., in (2)). In addition, this suggests the definition of credibility itself is broad, which introduces additional interpretive ambiguity.

of how discourse goals are pursued, in real discourse, the effects of coherent moves on those goals are more fine-grained. Hence, the ME Game jury function can provide what is necessary, as Asher et al. (2017, p.383) suggest, for establishing the winning conditions of a game (e.g., coherence, consistency). However, for a player to actually win, one would expect that there are more “beneficial” moves than “detrimental” ones,¹⁰ while still maintaining coherence, consistency, and credibility. In section 4.5, we build on the insights that ME Games provide about how strategic conversation is evaluated and suggest one approach to addressing the estimability problem while capturing finer-grained effects of discourse moves.

Our study aims to offer a path for developing a variant of the ME Game jury function, by extending and operationalizing the original, and evaluating the resulting function over extensive, realistic, high-stakes discourse. We differ from prior work in two key ways. First, we extend the ME jury function by providing a parallel framing that directly incorporates Gricean maxims and proposing a commitment-based taxonomy that captures finer-grained effects of discourse moves. Second, to our knowledge, this is the first study to apply such theories of strategic communication to extensive, realistic strategic dialogue, specifically, courtroom cross-examinations, where cooperation cannot be assumed and strategic language carries high real-world stakes.

3.3 Evaluation of LLMs’ Pragmatic Abilities

Previous work has examined the pragmatic abilities of LLMs¹¹ through the lens of Gricean maxims, humor, and deception in curated contexts (Hu et al., 2023; Krause and Vossen, 2024). Other studies have investigated strategic language use in games such as Werewolf (Xu et al., 2023), Avalon (Light et al., 2023), and Diplomacy (FAIR et al., 2022). Additionally, researchers have explored ways to improve LLMs’ ability to win through strategic interaction using prompt engineering (Xu et al., 2023), Theory of Mind (ToM) modeling (Lorè and Heydari, 2024; Zhang et al., 2025), and fine-tuning (FAIR et al., 2022). However, these works often focus on idealized or low-stakes scenarios, and there has been very little work investigating the strategic use of language in realistic contexts. One example is Ferracane et al. (2021), which examines the subjectivity involved in identifying non-sincere moves in congressional hearings.

In our work, we apply our theoretically grounded framework to the study of LLMs’ pragmatic understanding. Unlike prior research emphasizing downstream reasoning or task performance, our focus is on discourse understanding, a foundational layer necessary for meaningful strategic reasoning and evaluation of cooperativity.

4. SDA: Strategic Dialogue Assessment

Having introduced how non-cooperativity and strategic effects of discourse moves are approached in different frameworks, we now explore two directions: (a) While still drawing on insights from the

10. Note that we do not mean that this alone forms or determines the winning condition of a game, but rather it is a foundation for it. The current paper does not aim to model precisely whether a player wins. Instead, it follows the spirit of the jury function, which scores the individual contributions of players but does not, by itself, determine the winning conditions. Our focus is on approximating utilities that the jury function would give for players rather than providing an exact method for predicting outcomes, though the former is a necessary step toward the latter. For example, as we show in our intrinsic evaluation in section 7, the strategic effects we measure are predictive of outcomes but do not correctly predict every single case. For a detailed discussion of the formalization of winning conditions, we refer interested readers to section 4 of Asher et al. (2017).

11. We refer interested readers to Ma et al. (2025) for a systematic review of the broad topic of LLM pragmatics.

ME Game formulation of the jury function, we consider the possibility of incorporating Gricean maxims, which remain relevant for understanding goal realization in adversarial discourse and widely assumed in game-theoretic frameworks (Franke, 2009; Goodman and Stuhlmüller, 2013), as an alternative way of framing the jury function.¹² (b) We also explore potential solutions to the problem of estimating the constructs introduced by the jury function, and we attempt to capture the finer-grained effects of discourse moves observed in dialogues like that in (2). By doing so, we aim to relate theory to real-world discourse and to provide a well-motivated approach to evaluating LLMs’ pragmatic ability to recognize strategic effects of language.

Before turning to how we approach these two matters, we present the following dialogue (further analysed in Figure 2) to illustrate the components required to capture the strategic effects of language, and discuss which key assumptions and intuitions are (and are not) captured by existing approaches (e.g., ME Games). The dialogue is between the prosecutor (P) and Richard Ofshe (RO), as discussed previously. For ease of reference, we number RO’s responses. The central issue here is whether the police interrogation was coercive rather than routine, a distinction that matters because coercion would undermine the reliability or admissibility of the defendant’s statements. Within this exchange, the defense expert seeks to sustain the claim that the questioning exerted improper pressure, while the prosecutor aims to reframe the officers’ conduct, such as asking leading questions, as standard practice and therefore does not amount to coercion.

- (3) P: [...]In what you term or what you find in there coercive [...]?
- RO: [...] The questions were very directly specifying what the answers should be. **utt1**
- P: Did you find anything in the statement [...] to indicate that the officers gave him the information about which boy was castrated?
- RO: In their statements? Perhaps there is no such record. **utt2**
- P: Ok, you also talked to Mr. Smith for three hours?
- RO: No. I talked with him for the length of time it took to produce the transcript here. **utt3**
- P: [...] what coercive tactics do you allege that the police made in this case – or did? .
- RO: 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. **utt4**

To quantify the strategic effects of these responses, we in fact evaluate what *commitments* RO makes (section 4.1), and whether those commitments are meaningful, i.e., *interpretable* w.r.t. the prosecutor’s questions (section 4.2). Crucially, such meaningful commitments can carry different strategic values, depending both on their *content* (section 4.3) and on *how they are realized* (section 4.4) in discourse. We elaborate the details in the following sections 4.1-4.4.

This can be seen by comparing RO’s responses across turns. In utt1, RO makes a direct and relevant commitment that characterizes the questioning as coercive; the content of this commitment

12. Though we do not regard this as determinative, we do think that an advantage of using Gricean maxims in the discourse model is that they provide a relatively conservative and widely accepted account of how contextual meaning is derived (Grice’s analysis of implicatures). By contrast, models that distance themselves from Gricean maxims must rely on alternative and somewhat less widely accepted mechanisms to fulfill this role, an SDRT-like notion of coherence in the case of ME Games. It is important to recognize that the present work does not assume Gricean cooperativity, but instead uses Gricean maxims to study levels of cooperative or pseudo-cooperative behavior in adversarial settings.

supports his expert testimony and is therefore strategically **BENEFICIAL**. In utt2, RO’s response concedes (at least defeasibly) that there may be no record indicating that officers supplied the relevant information—a concession that is strategically **DETRIMENTAL** for RO given the prosecutor’s line of attack. However, this detrimental content is conveyed with hedging (“Perhaps”), which *weakens* the speaker’s apparent commitment strength and can partially mitigate the loss by reducing the perceived reliability and explicitness of the concession. Utt3 has strategically **NEUTRAL** content: RO corrects an overgeneralization about the duration of the interview, which does not directly advance either side’s central claim about coercion. Still, the manner of the correction may read as evasive, potentially *incurring a credibility cost* despite addressing the question. Finally, utt4 does not provide an answer to the prosecutor’s question at all; instead it postpones engagement by proposing a reformulation of the task, thereby failing to introduce any commitment whose content addresses the current question (**NONE**) and incurring a strategic penalty despite offering information that may be relevant at a later stage.

These contrasts highlight three recurring dimensions of strategic evaluation. First, a response may yield an immediate benefit or penalty by advancing or undermining the speaker’s goals through the content of the commitment it introduces. Second, this effect may be modulated by credibility, as it is plausible that commitments conveyed indirectly or in violation of conversational expectations often have weaker strategic impact, a dependence also reflected in the ME jury function via $P_k(\text{Good}_i)$. We formalize these intuitions in section 4.5 by defining turn-level benefits (**BAT**), and turn-level penalties (**PAT**). Third, because benefits and penalties accrue over turns, a speaker’s overall strategic position depends on the cumulative balance between gains and losses. This is formalized as the normalized cumulative difference over the course of the dialogue (**NRBAT**).

4.1 Discourse as commitment-making process

A classical game-theoretic view of communication holds that language is not evaluated purely in truth-conditional terms; rather, it is viewed as a form of action. Like ME Games, we adopt the perspective that discourse is a process of making commitments (Walton and Krabbe, 1995; Farkas and Bruce, 2010; Asher et al., 2017). As Asher et al. (2017) put it:

Crucially, some, perhaps most, of these [conversational] objectives involve commitments to contents, which are the conventional meanings and contextually derived implicatures of the utterances of the conversation. (p. 359)

This implies at least two things: (1) discourse goals are achieved through the act of making commitments, and (2) more specifically, they are achieved through commitments to contents—encompassing both conventional meaning and contextually derived meaning. Fundamentally, then, the jury function measures the effects of the contents of commitments.

4.2 Tying Goal Realization via Questions Under Discussion (QUD)

As noted above, discourse is a process of making commitments to achieve goals. Hence what we want to quantify concerns how successfully a speaker advances their own goals through such commitments. To achieve their goals, there are many possible commitments a speaker could make. However, these are not selected randomly or arbitrarily; rather, we observe particular, organized moves that are interpretable with respect to those goals. This requires some mediating principle that links individual commitments to broader discourse goals. Different accounts provide such a

principle. In the ME Game jury function, goal advancement is tied to the presence of coherence relations, since coherence connects discourse structure to the goals of a text. Another option is to ground the link in the notion of Questions Under Discussion (QUD) which can be seen as a development of the notion of relevance (Roberts 1996, 2012; for overviews, see Velleman and Beaver 2016; Beaver et al. 2017). Because the current paper focuses on a type of text in which one interlocutor primarily asks questions and the other provides answers, it is natural to use the QUD framework, which views all discourse as consisting (at least implicitly) of Q/A pairs. That said, other approaches, including coherence-based accounts, remain viable alternatives, and our choice here partly reflects both the genre of discourse and our own areas of expertise, rather than a particular piece of evidence favoring one framework over another.

In QUD-based accounts (Roberts 2012; Ginzburg 2012), utterances that fail to engage the current QUD are treated as *non sequiturs* in the sense that they cannot straightforwardly be integrated into the discourse structure.¹³ While such utterances may still provide information in general, they do not commit to the ongoing discourse goals and therefore do not function as satisfactory strategies for advancing them (i.e., they do not maximize joint utility). QUDs thus tie discourse moves to discourse goals: only moves that address the current QUD count as contributions to goal realization, whether via conventional meaning or contextually derived implicatures.

4.3 Evaluating the Contents of Commitments

We now turn to how commitments are linked to discourse goals through QUDs, specifying more precisely what makes a commitment meaningful—namely, that it is interpretable relative to the current QUD. The jury function does not simply check whether a commitment has been made; rather, it evaluates the effects of the specific content to which the speaker commits. As illustrated in example (2), different commitments that address the QUD can advance the speaker’s goals in different ways.

A Taxonomy of commitment effects In our system, the strategic benefits are rooted in the contents of the commitments, whether literal or implicated, that a speaker makes. Crucially, whether a meaningful commitment is strategically beneficial depends on the current QUD.¹⁴ For example, when the QUD is “Are you taking any medication?” in the context of testing the witness’ mental stability, a positive commitment such as “Yes, sir” may undermine the witness’s goal of maintaining credibility, while a negative commitment may support it. In addition, when the QUD is “What is your name?”—as a confirmation question to establish that the witness is the correct person to testify—a commitment such as established by the response “Mary” may carry no remarkable strategic gain or loss.

Following this logic, we propose that a commitment (**COMMIT**), relative to the current QUD, be classified as **BENEFICIAL**, **DETRIMENTAL**, **NEUTRAL** (or impartial), or **NONE** (i.e., no mean-

13. Failing to address the current QUD is not necessarily strategically ineffective and may serve other functions at a more global level. In this paper, however, our focus and annotations are intentionally local: the effects we measure are defined at the turn level, and we do not extend our claims beyond this level of analysis. We discuss the implications and limitations of this choice in more detail in section 9.

14. We claim that utterances are interpreted relative to the current QUD, but we do not assume that this QUD exists in isolation (e.g., the medication question is situated in a larger question of what happened in the court (rather than in the hospital)). Rather, the current QUD is part of a broader discourse structure (e.g., a hierarchy of questions that defines the context of the interaction). Our annotation scheme therefore conditions labels on the locally active QUD without denying the role of higher-level discourse organization.

TURN		TURN	
i	<p>Q: And in what you term or what you find in there coercive [...]?</p> <p>A: [...]The questions were very directly specifying what the answers should be.</p> <p>Commit: BENEFICIAL BaT = 1 PaT = 0 NRBaT = 1.34</p> <p>REL = 0 MAN = 0 QUAL = 0 CONST = 0</p>	$i+2$	<p>Q: Ok, you also talked to Mr. Smith for three hours?</p> <p>A: No. I talked with him for the length of time it took to produce the transcript here.</p> <p>Commit: NEUTRAL BaT = 0.5 PaT = 0.2 NRBaT = -1.92</p> <p>REL = 0 MAN = 0.4 QUAL = 0 CONST = 0</p>
$i+1$	<p>Q: Did you find anything in the statement [...] to indicate that the officers gave him the information about which boy was castrated?</p> <p>A: In their statements? Perhaps there is no such record.</p> <p>Commit: DETRIMENTAL BaT = 0.4 PaT = 1 NRBaT = 1.12</p> <p>REL = 0 MAN = 0.4 QUAL = 0 CONST = 0</p>	$i+3$	<p>Q: [...]what coercive tactics do you allege that the police made in this case -- or did? .</p> <p>A: 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.</p> <p>Commit: NONE BaT = 0 PaT = 0.5 NRBaT = -0.61</p> <p>REL = 0.4 MAN = 0.4 QUAL = 0 CONST = 0</p>

Figure 2: Examples of different commitment types and the corresponding values of PaT, BaT, and NRBaT

ingful commitment) according their effects on the discourse goals. Figure 2 illustrates these categories with examples from real cross-examinations. With this taxonomy, we can now capture finer-grained distinctions among the commitments in (1-2). Admissions such as being rejected as an expert in another court, or acknowledging that the police did not display behaviors typical of coercion, count as **DETRIMENTAL**. By contrast, statements such as not having been refused recognition as an expert by the state, or pointing out that the police asked misleading questions, count as **BENEFICIAL**.

We note that this taxonomy can also be applied to the original coherence-based ME jury function in order to capture finer-grained effects, as we have done here. Specifically, one can assume that the absence of a coherence relation functions like **NONE** in our categorization: it ties discourse moves to discourse goals, while still allowing for distinctions in the effects of coherent moves. We view this as a promising direction for those interested in a coherence-based approach, though the literature also notes practical obstacles in annotating coherence relations compared to Gricean maxims (Hoek and Scholman, 2017; Alkorta et al., 2019; Sanders et al., 2021; Panzeri and Foppolo, 2021).

To summarize, we have presented one way of modeling how a dialogue is formed and interpreted, drawing on Grice’s analysis of implicatures together with QUD-based discourse structure. More importantly, we have introduced a taxonomy that links each linguistic move to its contribution to the speaker’s overarching goals.

4.4 Coping with Partially Observable Terms

Having settled how the foundational commitment effects are determined, we now turn to other terms that modify them, namely, $P_k(\text{Good}_i)$, $\text{win}_i(k)$, and **CONS** according to Eq.1. These three terms describe, respectively: how reliable a commitment is, whether the commitment eliminates winning potential, and whether the commitment contradicts previous commitments. They model the extent to which the commitment effects should be preserved. For example, if a beneficial commitment is unreliable, the benefits the speaker gains should be proportional to the speaker’s credibility. If a detrimental move eliminates all possible winning paths, then no further benefits will be awarded. Similarly, if a commitment contradicts a prior beneficial one, the benefits gained from the latter should be retracted. Of the three, consistency is straightforward and requires no further specification, whereas the concept of credibility and winning potential are not directly observable from discourse alone. In our treatment, we approximate a speaker’s credibility through violations of

Gricean maxims, while leaving the modeling of winning potential aside. The win function distinguishes moves that automatically lead to loss from those that do not, and we are not aware of any way of identifying this property, at least in our dataset. There would be further challenges if these partially observable terms were to be incorporated in our model directly, as we discuss in section 9.

Approximating $P_k(\text{Good}_i)$ Violations of maxims are assumed to link to the speaker’s credibility, typically causing a decrease in credibility (Shuy, 1998; Ginzburg, 2012). Consider an example from a cross-examination in Figure 3. The prosecutor presses the defendant to name his girlfriend, perhaps to put the information on the record, or to distract him. We present three alternative answers. All three options are classified as neutral commitments, yet the final effects they produce differ. The two alternatives that violate maxims—such as those of manner and relevance—can be perceived as less credible, thereby appearing diminished in effect compared to commitments that involve no violations.¹⁵

We note that maxim violations do not constitute the only factors influencing the interlocutor’s credibility distribution, but they provide observable cues that feed into credibility assessments. In addition, they describe how a meaningful commitment is realized, distinguishing explicit commitment-making from implicit commitment-making via implicatures. In this respect, maxim violations play a role analogous to $P_k(\text{Good}_i)$ in the original function: they modulate the base effects (from the contents of the commitment) by affecting the perceived reliability of the commitment. Accordingly, we decompose the function from speakers to their credibility distributions into violations of relevance, manner, and quality.¹⁶ When no maxim is violated, the speaker appears to be cooperative (Levinson 2000; Horn 2006), which can serve as a strategy to minimize the loss due to credibility.

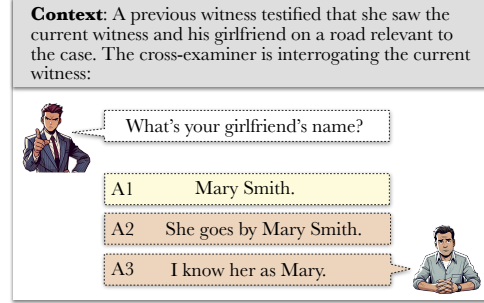


Figure 3: An example illustrating the role of violations of maxims as diminishing the reliability (thus also the effect size) of the commitment.

4.5 Underlying Elements of SDA

Up to this point, we have provided a way to incorporate Gricean maxims into the jury function, in parallel to the ME Game formulation; to capture the finer-grained effects of discourse moves; and to develop a more estimable account of the constructs introduced by the jury function. We call our approach Strategic Dialogue Assessment (SDA), and now elaborate on its scoring mechanism, our variant of the ME Games jury function. We locate base strategic value primarily in the im-

15. That said, commitments that violate maxims are perceived as weaker versions of their type: less beneficial, or less detrimental than comparable commitments without violations.

16. One might ask why we still retain CONS later, given its apparent overlap with the maxim of quality. We keep CONS because it targets overt self-contradictions, whereas credibility assessments (reflected in $P_k(\text{Good}_i)$) capture subtler forms of epistemic doubt. In this proposal, we treat violations of the maxim of quality as gradable. This is motivated by practical concerns: in real-world discourse, truth is often not directly accessible, and what we can observe are degrees of deception or implausibility rather than categorical falsehoods. For the same reason, we conflate the maxims of quality and quantity, treating both as contributing to the hearer’s plausibility assessment.

part of commitment contents on discourse goals. Based on their effect on the speaker’s goals, we categorize commitments (**COMMIT**) into four types with different assigned values: **BENEFICIAL**, **DETRIMENTAL**, **NEUTRAL** (or impartial), and **NONE**.

In addition, we replace $P_k(\text{Good}_i)$ with penalizing/rescuing terms based on violations of the Gricean maxims of relevance, manner, and quality. Specifically, violations of relevance (**REL**) and manner (**MAN**) are treated as *multiplicative* terms (see Eq.3). When a witness makes a beneficial/neutral commitment in an irrelevant or unclear statement, the strategic gain is diminished compared to a response that is relevant and clear. Detrimental commitments conveyed through implicature are penalized less severely, since they avoid explicitly harming the speaker’s interests. In such cases, the indirectness provides some strategic compensation.¹⁷ Truthfulness (quality; **QUAL**) is also modeled as a *multiplicative* factor (see Eq.3), capturing whether the speaker is still perceived as trustworthy and sincere from prior turns. Together, these refinements extend and decompose $(\text{COH}_i(k) + \text{RES}_i(k)) \times P_k(\text{Good}_i)$ into interpretable discourse properties: commitments and violations of the Gricean maxims.

Finally, we maintain a constraint on consistency (**CONST**), which reflects a key pragmatic pressure: speakers generally avoid inconsistent commitments, even if doing so requires conceding harmful facts or adopting strategically ambiguous responses. This pressure is further reinforced in legal settings by the potential availability of impeachment, which targets inconsistencies in a witness’s statements. Unlike Asher et al. (2017), however, we do not assume that inconsistency leads to a total collapse in strategic standing or nullifies future benefits. Instead, we treat inconsistency as a strong, but not absolute, penalty that significantly reduces the value of current strategic gains. This models the intuition that inconsistency undermines the reliability of prior contributions. Taken together—**COMMIT**, maxim violations, and consistency—these are the factors that we take to contribute to the overall effect of a commitment.

As should be clear, while our approach is heavily inspired by the ME Game jury function, we do not directly reproduce it. Instead, we incorporate elements such as QUDs and Gricean maxims, which provide a better adaptation for the type of discourse we analyze and offer a relatively estimable way to model speaker credibility. We leave it open whether future research could develop an even tighter connection to the original ME Game formulation.

PaT, BaT and NRBaT We define a value assignment function for commitments f_c , which maps the commitment C_i at turn i to a corresponding score. For simplicity, we assign a value of 1 to beneficial commitments and -1 to detrimental ones. Neutral commitments are treated as carrying a weak positive benefit, while the absence of meaningful commitment is treated as a penalty, reflecting the pragmatic pressure that commitments should address the current QUD in order to be interpreted as contributing to discourse goals. Since this requirement concerns whether a commitment is meaningful—that is, capable of contributing to discourse goals—rather than being a direct contributor to discourse goals itself, we treat its effects as weaker than those of beneficial or detrimental commitments:

$$f_c(C_i) = \begin{cases} 1 & \text{if } C_i = \text{BENEFICIAL} \\ 0.5 & \text{if } C_i = \text{NEUTRAL} \\ -0.5 & \text{if } C_i = \text{NONE} \\ -1 & \text{if } C_i = \text{DETRIMENTAL} \end{cases} \quad (2)$$

17. Note that such compensation does not apply to **NONE**, as no meaningful (i.e., interpretable) commitment is made in the first place (e.g., when the response has strongly violated relevance and/or manner).

Thus, the Benefit at Turn i (BaT_i)¹⁸ and the Penalty at Turn i (PaT_i) can be computed as:¹⁹

$$\text{BaT}_i = \begin{cases} f_c(C_i), & \text{if } C_i \in \{\text{BENEFICIAL}, \text{NEUTRAL}\} \\ f_c(C_i) \times (\text{Rel}_i + \text{Man}_i + \text{Qual}_i), & \text{if } C_i = \text{DETRIMENTAL} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

$$\text{PaT}_i = \begin{cases} |f_c(C_i)| + \text{Const}_i \times \sum_{j=1}^i \text{BaT}_j, & \text{if } C_i \in \{\text{DETRIMENTAL}, \text{NONE}\} \\ |f_c(C_i)| \times (\text{Rel}_i + \text{Man}_i + \text{Qual}_i) & \\ + \text{Const}_i \times \sum_{j=1}^i \text{BaT}_j, & \text{otherwise} \end{cases} \quad (4)$$

In addition, we define the Normalized Relative Benefit at Turn i (NRBaT_i) to capture the cumulative, normalized relative benefits across discourse:

$$\text{NRBaT}_i = Z \left(\sum_{j=1}^i \text{BaT}_j \right) - Z \left(\sum_{j=1}^i \text{PaT}_j \right) \quad (5)$$

This formulation computes the cumulative sums of BaT and PaT and applies a z-score normalization to these sums to ensure comparability between gains and penalties. The difference between the normalized BaT and normalized PaT at turn i provides an estimate of the overall strategic value accumulated over the discourse up to that turn. To avoid confusion, we note that the cumulative score aggregates *local* strategic effects under the assumption that immediate commitments and credibility shifts matter incrementally. However, it is not intended to capture long-horizon strategic planning or delayed traps, that is, moves for which the strategic benefit only materializes later. Capturing delayed strategic payoffs would require explicit modeling of goal hierarchies and future contingencies, which lies beyond the scope of the present work. Figure 2 presents the corresponding values of our metrics for an adapted snippet from a real cross-examination, with a detailed example of the calculation provided in Appendix C. We accept, of course, that the particular numeric values we used in the above definitions of scoring functions are theoretically somewhat arbitrary. We leave open whether in future work (a) a ranking method could be used that obviates the need to specify particular numeric values, or (b) a learning procedure or theoretical argument could be used to identify appropriate numeric values.

5. Human Annotations on CPD

18. A weighted sum is in theory possible with this equation; however, since we are interested in measuring correlation rather than absolute values, we leave this for future work. There are many ways to combine commitment types and violations of dialog conventions into a scalar score, and learning a function directly from outcomes can be an alternative. However, instead of relying on theoretically identified discourse properties, a fully learned function could obscure these distinctions by collapsing multiple pragmatic factors into uninterpretable weights. Exploring hybrid approaches that combine theoretical structure with learned parameters is an important direction for future work.

19. While our computation treats violations as binary labels (presence or absence), our annotation process used a finer-grained 4-point scale. In this scale, scores of 1–2 were treated as absence (no violation or borderline cases), while scores of 3–4 were treated as presence (clear and strong violations). This finer scale was used to reduce potential ambiguity introduced by forcing annotators to make binary choices during labeling, before collapsing the labels into a binary form for computation.

To empirically evaluate our framework and assess LLMs’ pragmatic abilities, we conduct human annotations on our dataset CPD. By applying SDA to CPD, we reveal that the identified discourse properties and the derived metrics in fact capture how non-cooperative discourse differs from its cooperative counterpart and we validate the strategic effects measured by the function are meaningful via an outcome prediction task.

Human Annotations We conduct human annotation on a subset of our dataset, covering approximately 200 turns per side per trial, for a total of around 800 turns. We recruit three annotators with relevant linguistic expertise (a journalist and two linguistics students) via Upwork, offering a pay rate of \$20 per hour. The annotators first complete a pilot set to familiarize themselves with our annotation framework, followed by a shared aggregated set consisting of a complete cross-examination (of around 150 turns). Given that reliable inter-annotator agreement is seen within this shared set, we proceed with them to annotate separate cases across different trials, balancing resource limitations with the goal of achieving broad dataset coverage.

Figure 4 illustrates the annotation task. Annotators read through a cross-examination dialogue in temporal order. For each turn, they are presented with background information (e.g., the witness’s role) and the current Q/A pair. They are then asked to evaluate the response along three dimensions: type of commitment, violation of each of the Gricean maxims, and consistency (see our protocols in Appendix G). Additionally, we elicit annotations on more basic and widely used indicators of strategic behavior, such as outcome judgments (Duan et al., 2024) and the underlying reasons for those outcomes following Lukin et al. (2017); Rapp (2023); Xu et al. (2024), which serve as a testbed for evaluating the applicability of our framework.

We find that annotators exhibit high agreement²⁰ on our metrics, while other metrics such as outcomes are much more subjective. We assess inter-annotator agreement using several metrics. For BaT, PaT, and NRBaT, we report Spearman’s ρ values. For the subcategories, we use Fleiss’ κ

Context Presented to Annotators

Witness’s role: Damon is the suspect of the case

Current Question: Are you taking medication at this time?

Current Response: Yes, sir

Q1: Which of the following commitment types can best describe the current response?

DET RIMENTAL **BENEFICIAL** **NEUTRAL** **NONE**

Q2-4: To what extent does the response violate Gricean Maxims?

No Violation **Borderline** **Clear** **Strong**

Q5: Does this response contradict with any of the witness’s prior statements?

False **True**

Q6: Which side is more likely to win in the current conversation?

Questioner **Witness**

Q7: Which of the following can best describe what your decision is based on?

Logical Arguments **Credibility** **Personal Emotions**

Figure 4: Annotation schema (guideline in Appx. G).

Table 2: Inter-annotator agreement statistics.

<i>Metric-level agreement</i>		
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%
<i>Outcome-level agreement</i>		
Outcome decision (Fleiss’s κ)		0.29
Jaccard similarity (avg. reasons)		0.46
Complete agreement on reasons		29%

20. We note that there is no universally agreed threshold for “good” inter-annotator agreement in pragmatics annotation. However, prior work in discourse and pragmatics typically treats $\kappa \approx 0.5$ -0.7 as reasonable, given the inherently interpretive nature of the task (Artstein and Poesio, 2008; Hoek and Scholman, 2017).

to measure agreement on commitment types, and Randolph’s κ ²¹ to evaluate violations of relevance, manner, and quality. For consistency, we report the average true positive rate.²² Finally, for outcome decisions we use Fleiss’ κ ,²³ and we additionally examine the overlap in annotators’ rationales using Jaccard similarity and the proportion of cases with complete agreement. We highlight here that annotator agreement is consistently higher for our metrics and discourse properties than for conversational outcomes.

6. Non-cooperative and Cooperative Discourses Are Asymmetric

Criminal trials offer a convenient setting in which to distinguish cooperative from non-cooperative discourse. In direct examination, a lawyer questions their own witness; here the lawyer’s and witness’s goals can be taken to be aligned, so the exchange is cooperative. In cross-examination, by contrast, a lawyer questions the opposing side’s witness; here the goals are non-aligned, so the exchange is non-cooperative. Comparing these two types of examination, we find an asymmetry in the strategic effects that discourse moves convey.

Figure 5 shows the frequencies of different commitment types and maxim violations in both settings. Control discourse (the blue bars) rarely involves detrimental commitments or violations of maxims, whereas non-cooperative discourse (the reddish bars) exhibits a higher frequency of both phenomena.

Our next observation is that even within non-cooperative discourse, the occurrence of maxim violations is remarkably lower than that of maxim maintenance, suggesting that violations alone may not sufficiently capture the strategies used in non-cooperative discourse. By incorporating the commitment taxonomy, SDA more accurately represents non-cooperative discourse and allows for more nuanced interpretations of violations, for instance, as loss minimization or benefit retrieval strategies.

Having seen that when comparing cooperative and noncooperative dialogue there are differences both in frequencies of different commitment types, and in frequencies of maxim violations, we now show that SDA metrics (i.e., BaT, and PaT), like the underlying local discourse properties, also distinguish cooperative from non-cooperative discourse. As illustrated in Figure 6, we use the z-scored difference between PaT and BaT at each turn to capture net move benefit (with NRBaT representing the cumulative counterpart). The results show that control discourse primarily



Figure 5: Distribution of Gricean maxim ratings and commitment types across trials

21. Randolph’s κ (Randolph, 2005) is used for maxim violations because their label distribution is usually highly skewed (see also Figure 5).

22. Inconsistencies are rare and typically signal obvious lying, so we report only the average true positive rate.

23. We note that our multiple regression analysis predicts outcomes on a per-annotator basis, so the observed disagreement on outcome among annotators does not contradict the good predictive power of our metrics.

exhibits benefits (i.e., more likely to win), while the non-cooperative counterpart displays a wider distribution, highlighting the inherent tension created by the need to appear cooperative during an inherently adversarial interaction.

Before moving on, we must note a limitation of the above results, namely that our comparison between cross-examination and direct examination (i.e., control) conversations is based on a single witness for whom we collected both transcripts. For all other cases, we annotated only the cross-examinations, which explains the disparity in the number of examples across the two discourse types. Since our primary interest lies in non-cooperative discourse rather than in providing a systematic overview of discourse typology, this single instance of direct examination serves as a sufficient contrastive baseline. Moreover, given the time-intensive nature of the annotation task, we chose to concentrate our efforts on cross-examinations. Accordingly, the results of this comparison should be viewed as illustrative rather than conclusive: they highlight potential contrasts between discourse types, but more extensive data gathered in future work would be needed to establish firm generalizations. However, even absent data from a large sample, we feel the differences between cooperative and non-cooperative discourse we have discussed in this section provide an important proof-of-concept for the general approach we have adopted, demonstrating how such an approach can be used to distinguish different types of discourse.

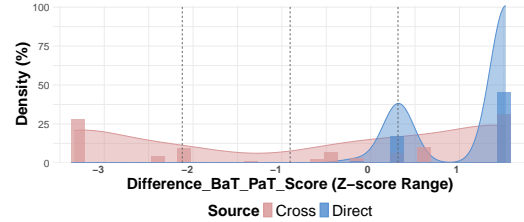


Figure 6: Distribution of net move benefit (PaT-BaT, z-scored) across discourse types. Control (cooperative) discourse concentrates around high benefit values, while cross-examination (non-cooperative) discourse shows a wider spread, reflecting greater variability and tension between gains and losses.

7. Intrinsic Evaluation

SDA offers a way to integrate Gricean maxims and commitment-based taxonomy into the original ME Game jury function, with the goal of modeling the strategic effects conveyed by a discourse move. We have provided justification for our pragmatic choices, arguing that our treatment offers a practical and valid extension of the original ME Game jury function. In this section, we complement that with empirical evidence from human annotations, demonstrating that our framework is well-suited to capturing the dynamics of real-world discourse.

We present our findings from three perspectives. (1) Similar to the original jury function, which could point toward the outcome of a game, our modified metrics are also predictive of conversational outcomes. This suggests that our modifications preserve key insights from the original function regarding how utterances contribute to outcomes (i.e., the realization of discourse goals). (2) Conditioning on different reasons for outcome judgments, we show that our metrics better capture the objective components of decision-making. (3) Compared to existing methods for quantifying strategic language understanding of LLMs—such as the NRA (Normalized Relative Advantage) proposed by Duan et al. (2024)—our metrics show greater robustness to subjectivity among annotators.

7.1 PaT and BaT are Effective Predictors of Individual Annotations of Outcome

In theory, the jury function evaluates the strategic gains and losses of each turn, which collectively contribute to, though not completely determine, the overall outcome of the discourse. Therefore, our proposed metrics, designed as an operationalizable approximation of the ME Game jury function, should also exhibit predictive power for conversational outcomes. We note that

outcome judgments are inherently subjective and annotators often disagree, but they are not arbitrary: each annotator’s judgments are systematically shaped by the discourse properties our metrics capture. Thus, the point of prediction is not to recover a single ground truth outcome, but to validate that our metrics track the factors that drive human reasoning about outcomes. To empirically validate this point, we conduct a basic regression analysis on our annotation data. We emphasize that this evaluation is **non-trivial**, as no prior work has empirically demonstrated that the theoretically identified formulation in the ME Game jury function necessarily reflects the achievement of conversational goals in extensive real discourse data. We conduct a multiple logistic regression analysis to predict *each annotator’s* outcome judgment at each turn, using the corresponding BaT and PaT scores as independent variables.

The overall model was statistically significant and demonstrated good fit to the data. BaT was a significant positive predictor, and PaT was a significant negative predictor. The model correctly classified 74.6% of cases, with an AUC of 0.80, indicating BaT and PaT have good discriminative ability in outcome prediction. We validated the robustness of our logistic regression results using a non-parametric bootstrap procedure with 1000 resamples.

Given that one of our purposes later is to evaluate LLMs’ **intrinsic** ability to understand strategic language, we also experimented with LLM-as-judge (zero-shot) and found that even the best-performing model (AUC = 0.68) is outperformed by predictions based on our metrics (see detailed scores in Appx. Table 5). This indicates that the intrinsic ability of LLMs to understand the utterances and their effects on outcome is less effective than when assessed through SDA metrics. While we expect there are ways to improve LLMs’ performance on this task (e.g., through fine-tuning using our metrics), outcome prediction is not our primary focus. Rather than attempting to optimize performance on this task, we use it solely as a means of validating our proposal and its role as an evaluation metric for LLM benchmarking.

7.2 SDA Captures Objective Aspects of Outcome Judgments

Our metrics reflect decision-making driven by what can reasonably be thought of as objective reasons, such as logical argumentation and certain aspects of credibility building, but not personal emotions.²⁴ a witness’s admission of taking medication may objectively raise concerns about their mental state—an inference grounded in logical reasoning. In contrast, discrediting a witness solely be-

Table 3: LR summary (BaT+PaT → outcome)

Predictor	β	SE	OR	95% CI	p
BaT	1.47	0.34	4.33	[0.80, 2.15]	<.001
PaT	-1.77	0.35	0.17	[-2.46, -1.09]	<.001
<i>Model fit</i>					
AIC = 332.9 (cf. intercept-only = 419)					
Tjur’s R^2 = 0.28; Accuracy = 74.6%; AUC = 0.80 [0.75, 0.85], p < .0001					
<i>Bootstrap (1,000 resamples)</i>					
BaT: 95% CI = [0.69, 2.15]; PaT: 95% CI = [-2.58, -1.05]					

24. Note that the distinction between categories *logical argumentation*, *credibility building* and *emotional*, as well as corresponding explanations for each category, are drawn from Xu et al. (2024). To exemplify how we apply these distinctions,

cause they do not attend church, as might occur with a biased juror, reflects a subjective and emotionally driven judgment. We fit two separate multiple logistic regression models: one conditioned on outcome decisions attributed to logical arguments and another on those influenced by personal emotions. The results in Figure 7 show that the discriminative power of our metric increases significantly when the reasons stated for the annotated outcome are logical arguments ($p < .0001$), but drops substantially when they are personal emotions ($p < .05$). These findings further corroborate that outcome judgments are inherently subjective, whereas SDA metrics capture the more objective aspects of outcome evaluation.

We note two important points based on these results. (1) Our operationalization of the ME Game jury function captures the contributions a speaker makes toward realizing discourse goals, but it does not determine the final outcome.²⁵ Accordingly, it is expected that the effects we capture represent only a partial account of how outcomes are decided. (2) While our metrics are designed to be relatively objective, we acknowledge that some degree of subjectivity remains, as indicated by levels of inter-annotator agreement. For example, annotators may differ in their interpretation of utterances and their effects on the speaker’s goals, leading to different assignments of commitment types to the same utterance. Such variability is unavoidable, but our results suggest that our approach has minimized these subjective elements sufficiently to serve as a reasonably objective and scalable evaluation method.

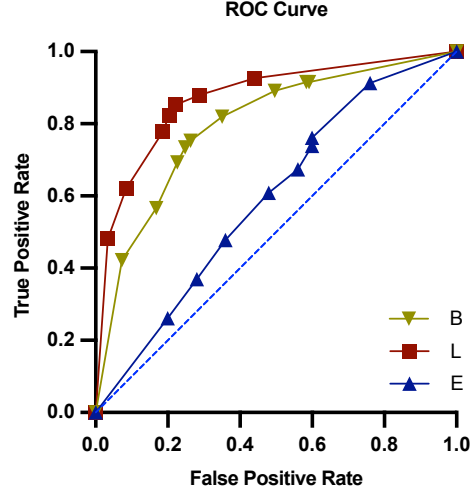


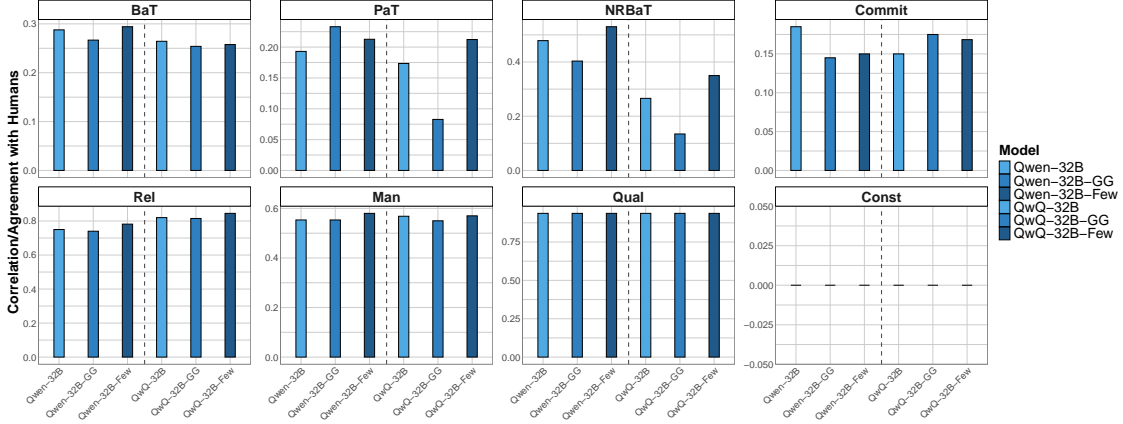
Figure 7: Model performances conditioned on different outcome reasons. B: baseline (i.e., without any conditions), L: conditioned on logical arguments, and E: conditioned on personal emotions.

7.3 SDA Metrics Are Consistent

In the literature, there are other metrics that assess LLMs’ strategic understanding, but these have been developed for competitive games with formal payoff structures rather than for real-world discourse. One such metric is the Normalized Relative Advantage (NRA) introduced by Duan et al. (2024). NRA was designed to evaluate relative performance in settings such as poker and auctions, providing a normalized measure of wins and losses between two agents. To the best of our knowledge, it has not been previously applied to free discourse outside of a game setting. Formally, NRA is defined as the difference between the cumulative number of wins for

25. For instance, even if a witness makes many detrimental moves (as in 1–2), the jury may interpret this as evidence that the witness is less reliable or even “guilty” of certain activities. Yet the final outcome depends on many additional factors, such as the presence of direct evidence of police coercion, or broader considerations of credibility and bias.

Figure 9: Comparison of prompting techniques (GG: General guidelines; Few: Few-shot) on Qwen and QwQ, which perform well in the zero-shot setting. Few-shot prompting provides slight improvement, while Constitution prompting benefits PaT but harms BaT and NRBaT. All differences are minor (< 0.1), suggesting prompting has only a small or even negligible impact.



the witness and the questioner up to turn i , normalized by the total number of scoring events

$$(\text{NRA}_i = \frac{\sum_{j=1}^i \text{win}_w^j - \sum_{j=1}^i \text{win}_q^j}{\sum_{j=1}^i \text{win}_w^j + \sum_{j=1}^i \text{win}_q^j}).^{26}$$

While both NRBaT and NRA aim to capture cumulative benefit over time, we observe that NRA is more variable. Figure 8 illustrates this in a sequence of annotations concerning a line of questioning about medication. Compared to NRBaT, NRA fluctuates more sharply, particularly after turn 5. At that point, Annotator A assigns more wins to the witness, Annotator B fewer, and Annotator C none at all following the witness’s admission to taking medication at turn 2—revealing an individual bias not shared by the others. Given the small number of annotators, we cannot draw definitive conclusions about NRA’s suitability for this task, and leave for future work the question of whether NRA, or some refinement of it, has application to real-world text, following the approach of the small study in this section.

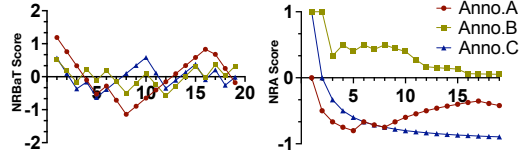


Figure 8: NRBaT and NRA of three annotators across turns during a line of questioning about medication

8. Can LLMs Perceive Strategic Effects of Language?

We evaluate a set of strong LLMs (Table 4), and further examine how model size and reasoning ability influence their perception of strategic dialogue. We use the same prompt provided to human annotators in a zero-shot setup. To constrain the variability of model responses, we set the tempera-

26. We follow Duan et al. (2024), assigning a value of 1 for a win and 0 otherwise.

ture to 0.1. We provide the prompt in Appendix F. We also experimented with different prompting techniques, including few-shot prompting and providing general guidelines for how to interpret the utterances, and found them largely leading to consistent results (see Figure 9). We emphasize that the analyses in this section are *exploratory*. Our goal is to probe how current LLMs behave on strategic cross-examination dialogue, rather than to make strong claims about their underlying reasoning mechanisms.

8.1 Quantitative Analysis

We report agreement/correlation with human annotations on **PAT**, **BAT**, **NRBAT** and also scores on their “local” benefit estimation components (i.e., **COMMIT**, **REL**, **MAN**, **QUAL**, **CONST**) aggregated across all three trials in Figure 10, with detailed scores and significance levels for each trial in Appendix H.

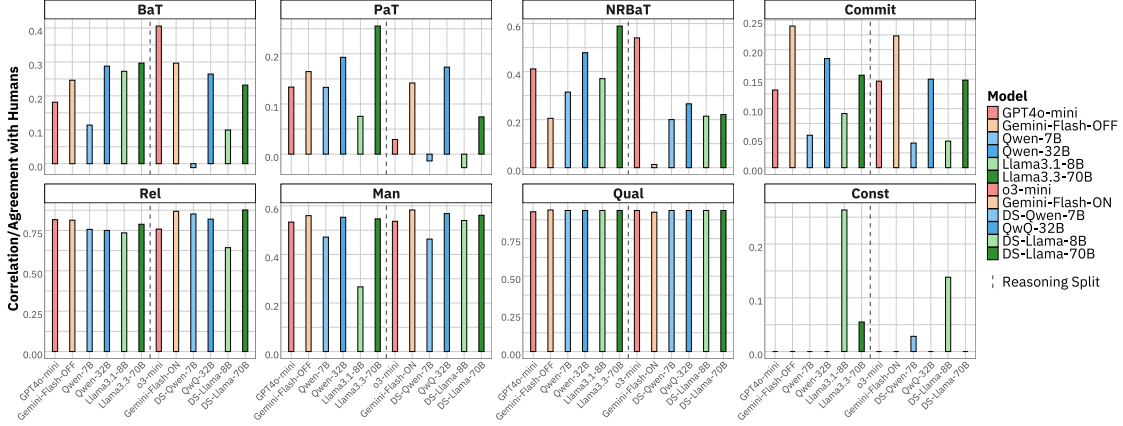
Overall, LLMs show strong agreement with humans in identifying violations of Gricean maxims, with mean (denoted as μ hereafter) Randolph’s κ scores of 0.80, 0.52, and 0.93 for **REL**, **MAN**, and **QUAL**, respectively. This aligns with prior findings (Hu et al., 2023) suggesting that LLMs have a good pragmatic understanding of Gricean maxims. Another contributing factor may be the skewed distribution of violations (see Figure 5), which makes the task easier and inflates Randolph’s κ . In contrast, LLMs perform poorly on commitment type identification (**COMMIT** ($\mu = 0.14$)) and our strategic metrics (**BAT** ($\mu = 0.23$), **PAT** ($\mu = 0.13$), and **NRBAT** ($\mu = 0.27$)) all of which lag behind human inter-annotator agreement/correlation (see section 7). Another interesting finding is that most models fail to reliably identify self-contradictory statements (**CONST**), whereas small models tend to achieve higher true positive rates. However, due to the rarity of such cases in our dataset, we refrain from drawing strong conclusions.

Model size matters We find that larger models (indicated by darker bars in Figure 10) consistently outperform their smaller counterparts (i.e., the lighter bars) on our BaT and PaT and in identifying commitment types. Below, we report average effect sizes ($\Delta\mu$) to quantify these differences, with bootstrapped 95% confidence intervals (CI) for each metric; Bonferroni-corrected statistical significance from a paired t-test is denoted with a (*): **BAT** ($\Delta\mu = 0.16^*$, 95% CI [0.06, 0.25]), **PAT** ($\Delta\mu = 0.12$, 95% CI [0.04, 0.22]), and **COMMIT** ($\Delta\mu = 0.10^*$, 95% CI [0.05, 0.15]). The effects are inconsistent, though positive for **NRBAT** ($\Delta\mu = 0.08$, 95% CI [−0.07, 0.23]), even when models exhibit improvements on BaT and PaT. This is because NRBaT is a cumulative measure that aggregates benefits across the entire discourse, making it less sensitive to the position of individual moves. Errors in local benefit estimation can cancel each other out over the course of the dialogue, which explains why models may sometimes exhibit low BaT and PaT scores but still achieve high NRBaT values, or vice versa. Most models already perform well on violation identification, though we do observe that larger models tend to perform slightly better for example, on **MAN** ($\Delta\mu = 0.12^*$, 95% CI [0.06, 0.19]).

Instruct-tuned Model	Reasoning Variant
GPT4o-mini	GPTo3-mini
Gemini-2.5-Flash (OFF)	Gemini-2.5-Flash (ON)
Qwen2.5-7B-Instruct	DeepSeek-R1-Distill-Qwen-7B
LLaMA3.1-Instruct-8B	DeepSeek-R1-Distill-LLaMA-8B
Qwen2.5-32B-Instruct	QwQ-32B
LLaMA3.3-70B-Instruct	DeepSeek-R1-Distill-LLaMA-70B

Table 4: Models Categorized by Size and Reasoning Capability; Gemini-2.5-Flash (OFF) refers to gemini-2.5-flash-preview-05-20 with the thinking budget set to 0.

Figure 10: Strategic metrics and agreement with humans across three trials. BaT, PaT, NRBaT: Spearman’s ρ ; Commit: Fleiss’ κ ; Relevance, Manner, Quality: Randolph’s κ ; Consistency: true positive rate. (*N.B.*, Inconsistencies do not occur in every trial; when there are no inconsistent utterances, the true positive rate is naturally 0.)



Reasoning (CoT) does not help with our metrics Models equipped with explicit reasoning mechanisms (i.e., the bars to the right of the dashed line in Figure 10) do not consistently improve performance and, in some cases, perform worse than their non-reasoning counterparts (i.e., bars to the left of the dashed line in the same color). This is particularly evident in **PAT** ($\Delta\mu = -0.10^*$, 95% CI $[-0.18, -0.04]$), and is also observed in **BAT** ($\Delta\mu = -0.03$, 95% CI $[-0.11, 0.04]$) **NRBaT** ($\Delta\mu = -0.09$, 95% CI $[-0.22, 0.03]$), and **COMMIT** ($\Delta\mu = -0.02$, 95% CI $[-0.07, 0.02]$) across most models, with the exception of GPT-o3-mini. The confidence intervals for **BAT** and **COMMIT** include zero primarily due to this outlier. These results suggest that explicit reasoning tends to hinder models’ ability to perceive strategic losses, if not benefits, as seen in the case of GPT-o3-mini. We further discuss possible reasons in section 8.2. However, we do observe that models with reasoning ability outperform their counterparts on e.g., **MAN** ($\Delta\mu = 0.06$, 95% CI $[0.02, 0.12]$) while the effects are weaker than model size. Lastly, we emphasize that our primary contribution lies in providing a method to probe how LLMs perceive the strategic effects of discourse moves. The results we present are restricted to the models examined under our prompting setup, and while they highlight certain issues (e.g., with CoT), we refrain from drawing strong conclusions in this fast-paced field.

8.2 Qualitative Analysis

We examine the sources of the observed effects of reasoning by analyzing the CoT that precedes generation, using models where such information is available: Qwen, DS-LLaMA and Gemini-Flash. While the reasoning traces provide a way to access explicit model judgements (Zaman and Srivastava, 2025), we do not assume that these explanations faithfully reflect the model’s internal decision process. Prior work shows that CoT traces may be unfaithful to model computation and can lead humans to infer incorrect narratives about “how the model thinks” (Turpin et al., 2023;

Levy et al., 2025; Kambhampati et al., 2025). Accordingly, we treat the generated reasoning as additional behavior to analyze, not as transparent access to the model’s pragmatic representations.

Reasoning traces involve overcomplication and ambiguity

Models often overthink and conflate loss-minimizing or benefit-retrieval strategies with the actual impact of making a commitment. Consider a context as in Figure 11; the prosecutor is asking whether the witness has ever been rejected as an expert in other courts. The response implies an affirmation, resulting in a detrimental commitment. While the model initially arrives at the correct judgment, it soon cycles through contradictory interpretations, ultimately settling on an incorrect assessment (i.e., **BENEFICIAL**). The generated reasoning traces relies on surface features of the mitigation strategies without integrating their broader strategic implications. Moreover, the model sometimes misinterprets surface-level damage control strategies, such as minimizing, deflecting, or hedging (highlighted in orange), treating them in its reasoning traces as neutral or positive. This interpretation in the reasoning traces fails to recognize them as attempts to mitigate an otherwise detrimental commitment. We also append another example (Appendix E) where the model does the opposite (i.e., identifying benefit-retrieval strategy as detrimental). Misclassifying the type of commitment can strongly influence BaT and PaT scores by, e.g., inverting the base value. Even when the model correctly identifies violations, an incorrect commitment label can alter the interpretation of the violation itself, for example, leading the reasoning traces to construe a loss-control strategy (added to BaT) as if it were a benefit-retrieval strategy (added to PaT).

Reasoning traces are naive and sometimes contradictory The reasoning traces do not consistently reflect the discourse facts. In Figure 12, the prosecutor is asking whether the defendant has pled guilty, admitting to which will be clearly detrimental to the defense side. The model’s reasoning introduces a contrast not present in the context and shows misrepresentation of world information and contradicts itself across turns. In the first exchange in the figure, the model claims the witness is correcting the idea that the husband didn’t plead guilty, even though the question already presupposes that he did.

This mismatch is followed by a distorted judgment of the witness’s commitment. In several other cases, the reasoning traces focus primarily on whether the response affirms or rejects the question, rather than integrating the broader discourse or world knowledge. We avoid concluding that the model “bases its judgment solely” on polarity; instead, we note that this is one plausible interpretation of the observed reasoning traces under our current setup. The model’s reasoning traces also exhibit self-contradiction in evaluating how a response aligns with the question’s intent. For example, in the second exchange, it correctly recognizes that the prosecutor’s question is designed to undermine the witness, yet it still judges the witness’s affirmation as a beneficial commitment.

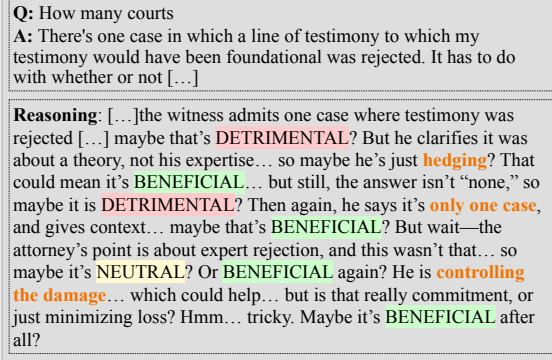


Figure 11: The model’s reasoning confuses the detrimental commitment with loss-mitigation strategies.

More broadly, these patterns suggest that under our behavioral probes, the reasoning traces often align with surface-level informativeness rather than adversarial goals or commitment structure. A more controlled experimental design would be required to determine the underlying causes of these behaviors. One promising direction is to draw on mechanistic interpretability methods (Geiger et al., 2025; Huang et al., 2025; Boguraev et al., 2025), which aim to identify internal mechanisms that contribute to specific model outputs. Another approach is to adopt controlled behavioral paradigms that vary the prompting context along minimal dimensions in order to probe what information LLMs rely on (e.g., McCoy et al., 2019). For example, one might manipulate only the polarity of a commitment, the presence of a pragmatic implicature, or the strategic role of the speaker. This makes it possible to test whether particular manipulations cause the model’s judgments to shift in systematic ways, and hence to infer the causal role of each factor. Collectively, such techniques would allow for more principled conclusions about the extent and limitations of LLMs’ internal strategic pragmatic competence.

<p>Q: And in fact her husband had pled guilty to these sexual abuse charges, correct?</p> <p>A: No. In fact her husband pled guilty to six counts of third degree rape [...]</p>
<p>Reasoning: The witness directly answers "No" and provides specific details about the plea. This seems make a BENEFICIAL commitment by correcting the prosecutor's implication that the husband didn't plead guilty.</p>
<p>Q: Did he maintain his guilt for a period of five months prior to entering that plea of guilty?</p> <p>A: Oh, yes</p>
<p>Reasoning: The prosecutor is trying to undermine the witness's credibility by questioning the case where the defendant pleaded guilty. [...] Since the witness is affirming the question, that benefits his own testimony rather than helping the prosecutor, so this would be BENEFICIAL commitment.</p>

Figure 12: Example of model reasoning with misrepresentation of contextual information and cross-turn inconsistencies.

9. Discussions of Challenges and Limitations

Before closing, we discuss the implications of our proposal, the challenges it raises for extending the framework, and its practical limitations. In particular, we highlight methodological and conceptual limitations concerning annotation, scope of reasoning, and disciplinary grounding.

Although coherence and QUD share many similarities (Benz and Jasinskaja, 2017), they are not the same; hence using one notion rather than the other to tie between the discourse goals and moves may produce different predictions. We note that our treatment in terms of QUDs is in some respects stricter than coherence-based accounts. On many coherence-based theories, a response can be linguistically coherent so long as it maintains local discourse relations (for example, via entity links), even if it fails to address the current QUD. The moves that we identify via QUDs as contributing to discourse goals are therefore a subset of those recognized under coherence-based theories. This has the implication that some coherence-based contributions are not recognized under our treatment.

For example, consider (4), a cross-examination from the Enron trial. It involves the witness Andrew Fastow (AF), the former Chief Financial Officer of Enron, who was called by the prosecution to testify about fraudulent practices within the company. The defense attorney (D) aims to challenge the reliability of his testimony. The crucial part of the example is AF’s response, where he does not commit to anything in response to the defense attorney’s question.

Under our QUD-based account, this response is therefore classified as making no meaningful contribution to the *current* discourse goal, even though it remains linguistically coherent, for in-

stance, under a relation such as “follow-up question” (Li, 2022). This illustrates how our notion of contribution is stricter than coherence-based accounts: some moves that are coherent are nevertheless treated as non-contributive with respect to the active QUD.

We acknowledge that our approach concerns more whether a turn advances the speaker’s goals at the *current* stage of the interaction and may miss aspects of its longer-term or global effects. Avoiding commitment now may serve longer-term purposes, such as delaying disclosure, or forcing redirecting the line of questioning. This distinction connects to our earlier discussion of short-term versus long-term strategic value and suggests that a comprehensive theory of strategic dialogue must ultimately account for both immediate goal advancement (as in the current paper) and deferred or indirect strategic effects, which we leave for future work.

- (4) D: Now, you say, “They stole in different ways,” other members of senior management. What you’re saying is that other members of senior management committed fraud to make their stock go up, then they would sell their stock and get away with the booty that way. That’s what you’re suggesting, right?

AF: Are you asking me?

A further implication concerns the simplification of credibility and the simplicity of the taxonomy of commitments. First, our decomposition of $P_k(\text{Good}_i)$ is not a theoretically faithful translation: violations of maxims do not by themselves constitute evidence of non-credibility as we have pointed out. The original term subsumes a broader range of factors, including, e.g., prior biases the jury may hold about the interlocutor. As a result, our function does not capture these more personal and socially grounded aspects of credibility as we have shown via experiments in section 7.3, but only discourse-internal cues that can be operationalized through maxim violations. This makes our approach feasible but also narrower in scope compared to the original jury in ME Games.

In addition, our framework distinguishes only four commitment types and does not further subdivide the degrees of benefit or harm. As a result, a commitment that directly incriminates a witness could, in our framework, be weighted similarly to one with relatively minor negative implications. While a finer-grained taxonomy might capture such distinctions more accurately, our simplification serves as the first step to produce a function that is practically applicable.

We believe that these implications do not substantially undermine our proposal, though they point to directions where refinements could be made. Beyond these implications, a more ambitious challenge concerns how to operationalize constructs such as winning potential. While we do not tackle this challenge in the current paper, it would be essential for anyone aiming to develop a more accurate predictor of conversational outcomes. If one were to operationalize the term $\text{win}_i(k)$, one would first need to identify the possible winning paths, which in turn requires a clear understanding of the speaker’s conversational goals. In practice, however, speakers often pursue multiple goals of varying importance. While these goals can sometimes be informally described, as in (2), there is no guarantee of a shared understanding of their relative importance. As a result, people are likely to have nuanced and divergent interpretations of a speaker’s discourse goal hierarchy, and thus of whether those goals have become unattainable (i.e., whether winning potential has collapsed to zero).

Realizing $\text{win}_i(k)$ in practice would therefore require (1) identifying the hierarchy of goals a speaker has, and (2) specifying how each goal can be achieved. Both are meaningful but ambitious undertakings that lie beyond the scope of this paper. We instead encourage readers interested in accurate outcome prediction to further investigate how this term might be operationalized.

Alongside these open challenges for future modeling, it is also important to recognize the practical limitations of our present study. Due to time and resource constraints, we were unable to annotate all the data we collected. Furthermore, the fact that neither annotators nor authors are legal scholars implies that we do not take into account strategies that require legal expertise to recognize, which we leave for future work. In addition, our annotations focus on local benefits: we did not require annotators to make long-horizon predictions. For example, a detrimental commitment was annotated as detrimental in the immediate context, even though it might ultimately serve a longer-term benefit (e.g., avoiding a later, more damaging commitment). Incorporating such long-term reasoning would require annotators to have a comprehensive understanding of the interlocutors’ goal structures and would likely introduce greater subjectivity into the task. Therefore, we again leave this aspect to future research. A related limitation arises from annotating at the level of dialogue turns rather than finer-grained discourse segments. While turn-level annotation improves feasibility and consistency, it necessarily abstracts away from internal discourse structure: a single turn may simultaneously advance and undermine a speaker’s strategic position, and more fine-grained segmentation could capture such mixed effects more precisely. Although our annotation scheme can still reflect this tension at the turn level, it does so only implicitly. Capturing such mixed strategic effects more faithfully would require a principled segmentation of sub-turn discourse units, a nontrivial methodological decision that lies beyond the scope of the present study.

A related limitation of our work is that, while we model trial discourse as a paradigmatic case of cooperative versus non-cooperative interaction, we have not yet fully connected our framework to the substantial literature in legal discourse analysis. Decades of research in conversation analysis and forensic linguistics have documented in detail how lawyers use questioning strategies to manage witnesses, control turn-taking, and shape what counts as relevant information (Atkinson and Drew, 1979; Drew, 1992; Cotterill, 2003; Heffer, 2005). This work highlights the subtle interplay between institutional constraints, question design, and strategic dialogues. Our current study abstracts away from these details in order to test computational methods. In future work, we aim to integrate insights from courtroom discourse research, both to enrich the empirical grounding of our categories and to ensure that the strategic phenomena we identify align with what is already known about adversarial questioning practices.

A final limitation regards AI-safety and alignment issues. We acknowledge the potential safety concerns that come with computational modeling of sophisticated strategic interactions. For this reason, we do not pursue alignment questions directly in this work. Instead, we frame our contribution as a step toward providing principled tools for calibrating how models interpret and respond to strategic moves. Our hope is that such tools can complement future alignment research by supporting both empirical and theoretical investigations into whether AI agents can reason appropriately about non-cooperative discourse, and into whether agents that recognize non-cooperation nonetheless behave cooperatively.

10. Conclusions

In this paper we introduced Strategic Dialogue Assessment (SDA), unifying two established pragmatic theories and providing a means to quantify their effects. Our successful application of SDA to the Crooked Path Dataset demonstrates that SDA has the potential to deepen our understanding of what makes discourse strategic. We see the current work as opening several promising directions for future research. First, although our analysis focuses on legal cross-examinations, the

framework naturally extends to other high-stakes adversarial domains, such as political debates or negotiations. We feel there would be significant interest in any connection that could be drawn between such discourses and actual outcomes, for example via public opinion measures or voting outcomes for politicians, or quantifiable outcomes in negotiations. Second, a strategic agent reasons not only about how an utterance is perceived, but also about how other participants are likely to respond. This raises the question of how to model downstream reasoning (e.g., as in RSA), in which one interlocutor anticipates the other’s decisions and strategically plans their own response. The current paper provides a first step in this direction by showing how to operationalize the utility function in adversarial settings. Building on this, predictive reasoning models such as RSA could incorporate strategic reasoning by optimizing SDA jury over appropriately specified world states and lexicon. A very simple approach would be to optimize accumulated turn-level advantage, i.e., $\sum_{t=1}^N (\text{BAT}_t - \text{PAT}_t)$ (N is the number of the moves), while leaving a more careful treatment of the constructs and additional potential hyperparameters (e.g., the weights for **BAT** and **PAT**) to future work.

Acknowledgments

We are especially grateful to Kathryn Kazanas, Chenxing Riley Zhang, and the other annotators for their contributions to the data annotation for this project. We also thank Susan Yorke for generously sharing her legal expertise, which contributed to the legal accuracy and soundness of this work. We are grateful to the three anonymous *D&D* reviewers and the editor for their invaluable comments, which greatly strengthened the paper. Finally, we thank the members of the UT Computational Linguistics group for their insightful feedback and suggestions at various stages of this paper. This work was partially supported by NSF grants IIS-2145479, IIS-2107524, a grant from Open Philanthropy, and Good Systems,²⁷ a UT Austin Grand Challenge to develop responsible AI technologies.

References

- Jon Alkorta, Koldo Gojenola, and Mikel Iruskieta. Towards discourse annotation and sentiment analysis of the Basque opinion corpus. In Amir Zeldes, Debopam Das, Erick Maziero Galani, Julian Desiderato Antonio, and Mikel Iruskieta, editors, *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 144–152, Minneapolis, MN, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2718. URL <https://aclanthology.org/W19-2718/>.
- Carolyn Jane Anderson and Brian W. Dillon. Guess who’s coming (and who’s going): Bringing perspective to the rational speech acts framework. In Gaja Jarosz, Max Nelson, Brendan O’Connor, and Joe Pater, editors, *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 185–194, 2019. doi: 10.7275/9bn3-8x38. URL <https://aclanthology.org/W19-0119/>.
- Ron Artstein and Massimo Poesio. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2. URL <https://aclanthology.org/J08-4004/>.

²⁷. <https://goodsystems.utexas.edu/>

- Nicholas Asher and Alex Lascarides. Strategic conversation. *Semantics and Pragmatics*, 6:1–62, 2013. ISSN 1937-8912. doi: 10.3765/sp.6.2.
- Nicholas Asher and Soumya Paul. Evaluating conversational success: Weighted message exchange games. In *20th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2016)*, volume 20, pages 55–64, 2016.
- Nicholas Asher and Soumya Paul. Strategic conversations under imperfect information: epistemic message exchange games. *Journal of Logic, Language and Information*, 27:343–385, 2018.
- Nicholas Asher, Soumya Paul, and Antoine Venant. Message exchange games in strategic contexts. *Journal of Philosophical Logic*, 46:355–404, 2017.
- J. Maxwell Atkinson and Paul Drew. *Order in Court: The Organisation of Verbal Interaction in Judicial Settings*. Macmillan, London, 1979.
- 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.
- Anton Benz and Katja Jasinskaja. Questions under discussion: From sentence to discourse. *Discourse Processes*, 54:177–186, 04 2017. doi: 10.1080/0163853X.2017.1316038.
- Anton Benz and Jon Stevens. Game-Theoretic Approaches to Pragmatics. *Annual Review of Linguistics*, 4(1):173–191, January 2018. ISSN 2333-9683, 2333-9691. doi: 10.1146/annurev-linguistics-011817-045641. URL <https://www.annualreviews.org/doi/10.1146/annurev-linguistics-011817-045641>.
- Sasha Boguraev, Christopher Potts, and Kyle Mahowald. Causal interventions reveal shared structure across English filler–gap constructions. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25032–25053, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1271. URL <https://aclanthology.org/2025.emnlp-main.1271/>.
- Sam Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilé Lukovšiušė, Amanda Askill, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Chris Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, John Kernion, Jamie Kerr, Jared Mueller, Jeff Ladish, Joshua D. Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova Dassarma, Robin Larson, Sam McCandlish, Sandip Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom B. Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Benjamin Mann, and Jared Kaplan. Measuring progress on scalable oversight for large language models. *ArXiv*, abs/2211.03540, 2022. URL <https://api.semanticscholar.org/CorpusID:253384413>.
- Penelope Brown. *Politeness: Some universals in language usage*, volume 4. Cambridge university press, 1987.

- Colin F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, Princeton, NJ, 2003.
- Herbert H. Clark and Edward F. Schaefer. Contributing to discourse. *Cognitive science*, 13(2): 259–294, 1989.
- Janet Cotterill. *Language and Power in Court: A Linguistic Analysis of the O. J. Simpson Trial*. Palgrave Macmillan, Basingstoke, 2003.
- Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020. URL <https://arxiv.org/abs/2012.08630>.
- Judith Degen. The rational speech act framework. *Annual Review of Linguistics*, 9 (Volume 9, 2023):519–540, 2023. ISSN 2333-9691. doi: <https://doi.org/10.1146/annurev-linguistics-031220-010811>. URL <https://www.annualreviews.org/content/journals/10.1146/annurev-linguistics-031220-010811>.
- Paul Drew. Contested evidence in courtroom cross-examination: The case of a trial for rape. In Paul Drew and John Heritage, editors, *Talk at Work: Interaction in Institutional Settings*, pages 470–520. Cambridge University Press, Cambridge, 1992.
- 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>.
- FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyang Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.
- Donka F. Farkas and Kim B. Bruce. On reacting to assertions and polar questions. *Journal of semantics*, 27(1):81–118, 2010.
- Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. Did they answer? Subjective acts and intents in conversational discourse. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.129. URL <https://aclanthology.org/2021.naacl-main.129>.
- Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336:998 – 998, 2012. URL <https://api.semanticscholar.org/CorpusID:26797288>.

- Michael Franke. Signal to act : game theory in pragmatics. 2009. URL <https://api.semanticscholar.org/CorpusID:2057191>.
- Michael Franke. Game Theoretic Pragmatics. *Philosophy Compass*, 8(3):269–284, March 2013. ISSN 1747-9991, 1747-9991. doi: 10.1111/phc3.12015. URL <https://compass.onlinelibrary.wiley.com/doi/10.1111/phc3.12015>.
- Atticus Geiger, Jacqueline Harding, and Thomas Icard. How causal abstraction underpins computational explanation, 2025. URL <https://arxiv.org/abs/2508.11214>.
- Jonathan Ginzburg. *The interactive stance*. Oxford University Press, 2012.
- Jacob Glazer and Ariel Rubinstein. A Game Theoretic Approach to the Pragmatics of Debate: An Expository Note. In Anton Benz, Gerhard Jäger, and Robert Van Rooij, editors, *Game Theory and Pragmatics*, pages 248–262. Palgrave Macmillan UK, London, 2006. ISBN 978-1-349-52317-7 978-0-230-28589-7. doi: 10.1057/9780230285897_9. URL http://link.springer.com/10.1057/9780230285897_9.
- Erving Goffman. *Strategic interaction*, volume 1. University of Pennsylvania Press, 1970.
- Noah D. Goodman and Andreas Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184, 2013.
- Ryan Greenblatt, Carson E. Denison, Benjamin Wright, Fabien Roger, Monte Stuart MacDiarmid, Samuel Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Kristjanson Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models. *ArXiv*, abs/2412.14093, 2024. URL <https://api.semanticscholar.org/CorpusID:274823033>.
- Herbert Paul Grice. Logic and conversation. *Syntax and semantics*, 3:43–58, 1975.
- Herbert Paul Grice. *Studies in the Way of Words*. Harvard University Press, Cambridge, 1989.
- Chris Heffer. *The Language of Jury Trial: A Corpus-Aided Analysis of Legal-Lay Discourse*. Palgrave Macmillan, Basingstoke, 2005.
- Jet Hoek and Merel Scholman. Evaluating discourse annotation: Some recent insights and new approaches. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*, 2017. URL <https://aclanthology.org/W17-7401/>.
- Laurence Horn. Towards a new taxonomy for pragmatic inference: Q-and r-based implicature. *Meaning, form and use in context*, 1984.
- Laurence R. Horn. *Implicature*, chapter 1, pages 2–28. John Wiley & Sons, Ltd, 2006. ISBN 9780470756959. doi: <https://doi.org/10.1002/9780470756959.ch1>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470756959.ch1>.

- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.230. URL <https://aclanthology.org/2023.acl-long.230/>.
- Jing Huang, Junyi Tao, Thomas Icard, Diyi Yang, and Christopher Potts. Internal causal mechanisms robustly predict language model out-of-distribution behaviors, 2025. URL <https://arxiv.org/abs/2505.11770>.
- Subbarao Kambhampati, Kaya Stechly, Karthik Valmeekam, Lucas Saldyt, Siddhant Bhambri, Vardhan Palod, Atharva Gundawar, Soumya Rani Samineni, Durgesh Kalwar, and Upasana Biswas. Stop anthropomorphizing intermediate tokens as reasoning/thinking traces! 2025. URL <https://api.semanticscholar.org/CorpusID:277771929>.
- Lea Krause and Piek T.J.M. Vossen. The Gricean maxims in NLP - a survey. In Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 470–485, Tokyo, Japan, September 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.inlg-main.39/>.
- Stephen C. Levinson. Pragmatics and the grammar of anaphora: a partial pragmatic reduction of binding and control phenomena. *Journal of Linguistics*, 23(2):379–434, 1987. doi: 10.1017/S0022226700011324.
- Stephen C. Levinson. *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press, 2000.
- Mosh Levy, Zohar Elyoseph, and Yoav Goldberg. Humans perceive wrong narratives from ai reasoning texts. *ArXiv*, abs/2508.16599, 2025. URL <https://api.semanticscholar.org/CorpusID:280710263>.
- David Lewis. *Convention*. Harvard Press, Cambridge, 1969.
- Chuyuan Li. Sdrt@ dailydialog annotation guideline. 2022.
- Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. From text to tactic: Evaluating LLMs playing the game of avalon. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023. URL <https://openreview.net/forum?id=ltUrSryS0K>.
- Nunzio Lorè and Babak Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490, 2024.
- Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. Argument strength is in the eye of the beholder: Audience effects in persuasion. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain, April 2017. Association for Computational Linguistics. URL <https://aclanthology.org/E17-1070/>.

- Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL <https://aclanthology.org/2025.acl-long.425/>.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL <https://aclanthology.org/P19-1334/>.
- Elena Musi, Nadin Kokciyan, Khalid Al-Khatib, Davide Ceolin, Emmanuelle Dietz, Klara Gutekunst, Annette Hautli-Janisz, Cristian Manuel Santibañez Yañez, Jodi Schneider, Jonas Scholz, et al. Toward reasonable parrots: Why large language models should argue with us by design. *arXiv preprint arXiv:2505.05298*, 2025.
- Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, Cambridge, MA, 1994.
- Steve Oswald. *Pragmatics of uncooperative and manipulative communication*. Université de Neuchâtel, 2010.
- Francesca Panzeri and Francesca Foppolo. Children’s and adults’ sensitivity to gricean maxims and to the maximize presupposition principle. *Frontiers in Psychology*, 12:624628, 2021.
- Prashant Parikh. Communication, meaning, and interpretation. *Linguistics and philosophy*, pages 185–212, 2000.
- Sascia Pavan. Scalar implicatures and iterated admissibility. *Linguistics and Philosophy*, 36(4): 261–290, 2013.
- Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=6Mxhg9PtDE>.
- Justus J. Randolph. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss’ fixed-marginal multirater kappa. Presented at the Joensuu Learning and Instruction Symposium, 2005. URL <https://eric.ed.gov/?id=ED490661>. ERIC Document ED490661.
- Christof Rapp. Aristotle’s Rhetoric. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.
- Craige Roberts. Information structure in discourse: Towards an integrated formal theory of pragmatics. Technical report, OSU Working Papers in Linguistics, 49, 1996.

- Craige Roberts. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1, 2012.
- Ted J.M. Sanders, Vera Demberg, Jet Hoek, Merel C.J. Scholman, Fatemeh Torabi Asr, Sandrine Zufferey, and Jacqueline Evers-Vermeul. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, 17(1):1–71, 2021.
- Tuomas Sandholm. The state of solving large incomplete-information games, and application to poker. *AI Magazine*, 31(4):13–32, 2010. doi: 10.1609/aimag.v31i4.2303.
- Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA, 1960.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askill, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=tvhaxkMKAn>.
- Roger W. Shuy. *The language of confession, interrogation, and deception*, volume 2. Sage, 1998.
- Anthony Sicilia, Tristan Maidment, Pat Healy, and Malihe Alikhani. Modeling non-cooperative dialogue: Theoretical and empirical insights. *Transactions of the Association for Computational Linguistics*, 10:1084–1102, 2022.
- Cesare Spinoso-Di Piano, David Eric Austin, Pablo Piantanida, and Jackie CK Cheung. (RSA)?: A rhetorical-strategy-aware rational speech act framework for figurative language understanding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20898–20938, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1019. URL <https://aclanthology.org/2025.acl-long.1019/>.
- Jenny A. Thomas. *Meaning in interaction: An introduction to pragmatics*. Routledge, 2014.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don’t always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA, 2023. Curran Associates Inc.
- Leah Velleman and David I. Beaver. Question-based models of information structure. In Caroline Féry and Shinichiro Ishihara, editors, *The Oxford Handbook of Information Structure*, pages 86–107. Oxford University Press, Oxford, UK, 2016. doi: 10.1093/oxfordhb/9780199642670.013.29.
- Adam Vogel, Andreas Gomez Emilsson, Michael C Frank, Dan Jurafsky, and Christopher Potts. Learning to reason pragmatically with cognitive limitations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.

- Douglas Walton and Erik C.W. Krabbe. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. State University of New York Press, 1995.
- James R. Wright and Kevin Leyton-Brown. Beyond equilibrium: Predicting human behavior in normal-form games. *Games and Economic Behavior*, 117:1–24, 2019. doi: 10.1016/j.geb.2019.05.001.
- Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. The earth is flat because...: Investigating LLMs’ belief towards misinformation via persuasive conversation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.858. URL <https://aclanthology.org/2024.acl-long.858/>.
- Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. Exploring large language models for communication games: An empirical study on werewolf. *CoRR*, abs/2309.04658, 2023. doi: 10.48550/ARXIV.2309.04658. URL <https://doi.org/10.48550/arXiv.2309.04658>.
- Kerem Zaman and Shashank Srivastava. Is chain-of-thought really not explainability? chain-of-thought can be faithful without hint verbalization. *arXiv preprint arXiv:2512.23032*, 2025.
- Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Yan Xia, Man Lan, and Furu Wei. K-level reasoning: Establishing higher order beliefs in large language models for strategic reasoning. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7212–7234, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL <https://aclanthology.org/2025.naacl-long.370/>.

Appendix A. Full GPT-4o response for Fig 1

Here is a conversation:

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

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

What do you think of the witness's response in this cross-examination? Is the response cooperative?

Does this response increase the risk of losing for the witness in the trial (i.e., detrimental)?

The witness's response is **non-cooperative** and potentially **detrimental** to their case. Here's why:

Non-cooperative:

- The question is direct: *"Did you ever falsify any incident reports regarding in-flight safety?"*
- The response **does not answer** the yes/no question about the witness's own actions. Instead, the witness deflects to a third party — "my colleague."
- There's also **ambiguity** introduced with the hedge "to my knowledge," which creates distance between the witness and the event.
- Overall, this is a classic evasive or deflective move in cross-examination, avoiding a direct commitment.

Detrimental:

- The mention of a **colleague falsifying or correcting a report** (even if framed as a correction) introduces **harmful information**: it opens the door to broader concerns about the reliability of safety reporting practices.
- By bringing up another case (even unintentionally), the witness **expands the scope of scrutiny**, which the opposing counsel can exploit.
- Moreover, the indirectness may make the jury infer that the witness is hiding something or is **not being forthright**, harming credibility.

Summary:

- **Cooperativity:** ✗ (non-cooperative — evades the direct question)
- **Detrimental:** ✓ (likely increases risk to the witness by both seeming evasive and introducing damaging context)

Figure 13: Full response from GPT-4o using the prompt in Fig 1

Appendix B. Metrics Distribution in Human Annotations

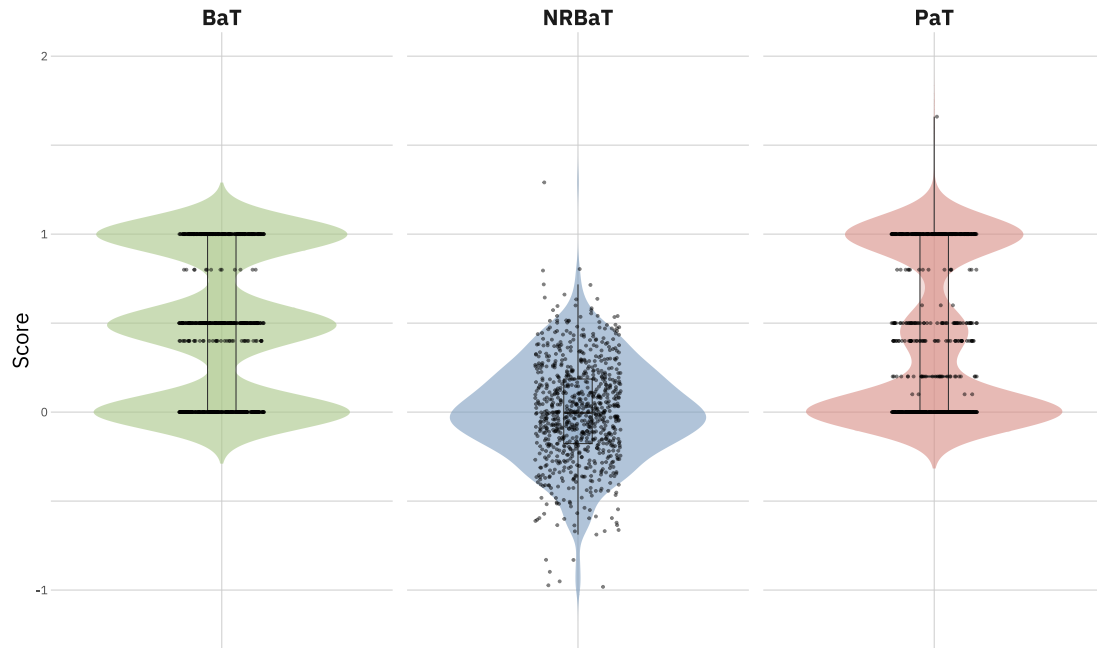


Figure 14: Distribution of our metrics in our human annotated dataset

Appendix C. Metrics Calculation Example

Take turn $i + 1$ in Figure 2 as an example. The distributions of the BaT and PaT sums are (1, 1.4, 1.9, 1.9) and (0, 1, 1.2, 1.7), respectively, yielding $\mu_b = 0.98$, $\mu_p = 1.55$, $\sigma_b = 0.38$, and $\sigma_p = 0.14$ for the given example snippets. In our work, we assign weights of 0.4, 0.4, 0.2 and 0.2 to REL, MAN, QUAL and CONST, respectively. As noted in the main text, these weights are hypothetical, as our focus is on measuring correlation rather than modeling absolute values. The lower weight assigned to QUAL / CONST reflects its basis in perceived, rather than objective, truth; this prevents placing undue penalty on this dimension.

$$\begin{aligned}
\text{BaT}_{i+1} &= |f_c(C_i)| \times (\text{Rel}_{i+1} + \text{Man}_{i+1} + \text{Qual}_{i+1}) \\
&= 1 \times (0 + 0.4 + 0) \\
&= 0.4 \\
\text{PaT}_{i+1} &= |f_c(C_i)| + \text{Const}_{i+1} \times \sum_{j=1}^{i+1} \text{BaT}_j \\
&= 1 + 0 \\
&= 1 \\
\text{NRBaT}_{i+1} &= Z \left(\sum_{j=1}^{i+1} \text{BaT}_j \right) - Z \left(\sum_{j=1}^{i+1} \text{PaT}_j \right) \\
&= \frac{\sum_{j=1}^{i+1} \text{BaT}_j - \mu_b}{\sigma_b} - \frac{\sum_{j=1}^{i+1} \text{PaT}_j - \mu_p}{\sigma_p} \\
&= \frac{(0 + 1) - 0.98}{0.38} - \frac{(1 + 0.4) - 1.55}{0.14} \\
&= 0.053 - (-1.07) \\
&= 1.123
\end{aligned}$$

Appendix D. Zero-shot LLM prediction results

Model	AUC	Model	AUC
GPT4o-mini	0.61	o3-mini	0.60
Gemini-Flash-OFF	0.68	Gemini-Flash-ON	0.63
Qwen-7B	0.53	DS-Qwen-7B	0.47
Qwen-32B	0.52	QwQ-32B	0.58
Llama3.1-8B	0.53	DS-Llama-8B	0.51
Llama3.3-70B	0.62	DS-Llama-70B	0.59

Table 5: AUC scores for 0-shot LLM predictions

Appendix E. Qualitative Analysis Example

Listing 1: An example of LLMs misidentifying benefit retrieval strategy

An example of LLMs misidentifying benefit retrieval strategy as detrimental commitments.

Context:

The prosecutor is asking whether the witness has ever had a good understanding of the defendant. The witness answers a decent amount of time of contact, which is beneficial.

Q: How long was that interview?

A: Three hours, more or less. It may have a bit more. It may have been a bit less. I don't have the -- I don't have that -- it might be helpful. It worked out to an eighty-seven page transcript.

Reasoning:

The prosecutor is trying to get specific information. The witness's answer starts with an approximate value ("Three hours, more or less") but then hedges a lot, saying "may have a bit more" or "may have been a bit less." The hesitation ("I don't have the -- I don't have that -- it might be helpful") might indicate uncertainty. This could be detrimental to the witness because it shows they're unsure, which benefits the questioner. So maybe a 1 (Detrimental commitment)

Appendix F. Prompts

Listing 2: 0-shot Prompt

Below is the transcript of the cross-examination so far:

{history}

The current response from the witness is: {answer}

Instructions:

Please evaluate the CURRENT response based on the following properties: 1. Witness commitment level (with respect to the immediate question):

- 1: Detrimental commitment: A detrimental commitment supports the questioner's position and undermines the witness.
- 2: Beneficial commitment: A beneficial commitment strengthens the witness's case, increasing their winning potential.
- 3: Neutral commitment: A neutral commitment has no clear strategic effect, not affecting either side's winning potentials.
- 4: No commitment made: No answer to the question can be clearly drawn or is clearly implied.

2. Witness truthfulness in the current turn:

- 1. Salient honesty
- 2. Mild honesty
- 3. Mild Dishonesty
- 4. Dishonesty

3. Consistency with previous testimony:

- 1: Inconsistent: The current statement is self-contradictory.
- 0: Consistent

4. Response relevance:

- 1: Very relevant: This answer is directly given and is relevant to the question.
- 2: Slightly relevant: The answer does not provide a direct answer but instead provides relevant information to tackle the specific issue.
- 3: Slightly irrelevant: The answer avoids addressing the specific question, requiring further probing/non-verbal response.
- 4: Irrelevant: It is almost impossible to draw any inferences.

5. Response clarity:
 - 1: Very clear: The answer is direct, unambiguous, and orderly.
 - 2: Overall clear with hedging: The answer is still clear, but it includes extra hedging that isn't strictly needed.
 - 3: Unclear: The answer leads to ambiguous inferences/interpretations.
 - 4: Very Unclear: Refusal to answer the question or verbose answer.
 6. Current turn outcome:
 - Questioner
 - Witness
 7. Reasons for current outcome:
 - 1. Logical arguments are convincing.
 Logical appeals are based on facts, evidence, and reason. For example: Based on statistical data, it's clear that our product is more cost-effective than our competitors' offerings. Here are the numbers to support our claim.
 - 2. Credibility attack on other side
 Credibility appeals focus on the credibility and trustworthiness of the speaker or source. For example: As a trusted healthcare provider for over 30 years, our commitment to patient well-being and safety is unwavering. You can rely on our expertise and care.
 - 3. Emotional appeal/bias
 Emotional appeals aim to evoke feelings and emotions in the audience. For example: Imagine the joy and satisfaction of owning a home in a peaceful neighborhood. Our real estate services can make that dream a reality for you and your family.
- Provide your response in JSON format:
- ```
"commitment value": "Commitment level (1-4)",
"quality rate": "Truthfulness (0,1)",
"consistency value": "Consistency (0,1)",
"relevance rate": "Relevance (1-4)",
"manner rate": "Clarity (1-4)",
"outcome value": "Winner of current turn (Questioner/Witness)",
"outcome reason": "Reason for the outcome (1-3)"
```

### Listing 3: General Guidelines

A strategic witness does not assume a common goal. The following Gricean maxims can be reliably violated:

Truthfulness (Quality): avoid falsehoods and speak only what they believe to be true.

Relevance: address the specific question being asked.

Clarity (Manner): avoid vagueness, ambiguity, or unnecessary complexity.

Accordingly, judgments about the witness's commitment should be based on whether the response advances the witness's interests (i.e., likelihood to win)

### Listing 4: Few-shot examples

Example 1:

Question: Are you taking any medication?

response: I might have taken some.

This response is a detrimental commitment as taking medication indicates mental instability.

This response is relevant (1) but not clear with hedging (3), and truthful (1).

The winner is Questioner and the reason is logical arguments.

Example 2:

Question: Have you been to the place where the body was found?

Response: I think I have no reason to go to place like that.

This response is a beneficial commitment as not have gone to the crime spot indicates alibi.

This response is relevant (1), unclear with hedging (3), and truthful (1).

The winner is Witness and the reason is logical arguments.

Example 3:

Question: You have interviewed with the defendant for ten hours?

Response: No

This response is a detrimental commitment as having decent amount contact with defendant indicates the witness has enough knowledge about defendant so denying would indicate the opposite.

This response is relevant (1), clear (1), and truthful (1).

The winner is Questioner and the reason is logical arguments.

## Appendix G. Annotation Protocols

### Listing 5: Annotation Instructions

#### Understanding Strategic Communication Through Cross-examinations

Background:

-----

Discourse exhibits a diverse typology, which we roughly divide into two categories: cooperative and non-cooperative. Traditional discourse and pragmatics research has primarily treated discourse as a cooperative enterprise, benefiting from this perspective. In this project, however, we focus on the more neglected non-cooperative type. We assume that such discourse follows what Asher (2013) and his subsequent work characterize as strategic conversations. That is, participants do not assume cooperativity from the outset; instead, they make discourse moves based on their assumptions about the goals of the other participants.

For this specific project, we situate our research within a typical non-cooperative discourse scenario: cross-examinations. By their nature, cross-examinations involve participants who do not cooperate globally, as each acts in their own interests. We are interested in how this global non-cooperativity is realized or represented by focusing on more localized conversational moves. Cooperativity (or non-cooperativity) is a relative concept. In our case, we assume that the cross-examiner, in order to obtain a response that serves their own goal, conveys their intention explicitly or implicitly through their questions. This conveyed message is generally against the witness's (i.e., the party they are representative of) interests. In theory, the witness then tends to interpret these questions as being contrary to their own interests and makes their discourse moves accordingly.

\* Jury's perspective: Basically, we ask for judgments as a juror or as an onlooker, and you do not need to pretend you are the witness or the examiner and guess what they want to convey from their perspectives.

Instructions:

-----

You will be given conversation pairs extracted from cross-examinations that took place in real criminal trials. Each pair consists of two parts: one from the cross-examiner and one from the witness. In addition to that, you will also be presented with recent question-answer pairs (up to 10), a summary of previous context (previous to the current pair), and some meta-data related to the witness.

An EXAMPLE of what you will see in the interface:

Context: In a trial, the examiner is asking about information about the witness' medication condition, which can have subsequent effects on the credibility of the witness.

Witness's role: One of the suspects in the trial.

Recent Exchanges:

Examiner: Mr. Echols, I'm going to ask you questions and like I have told other witnesses, if you don't understand, you ask me to rephrase them and I'll be glad to do so.

Witness: (NOD HEAD)

Current Pair:

Examiner: First of all, let me ask you, are you taking any medication at this time?

Witness: Yes, sir.

\*\*\*The witness can have different roles (e.g., expert), but in the above case, he happens to be the defendant.\*\*\*  
\*\*\*When you make judgments on the following categories, assume the perspective of a juror, not just a casual observer of the courtroom.\*\*\*

#### Cross-examiner's Utterances

-----

Consider that both the examiner and the witness in a trial are telling a version of the story in order to convince the jury their own version of the story is the truth. In theory, the questions the cross-examiner asks cannot be taken as evidence; however, statements can be embedded in the questions to influence the jury's perception of the credibility of the witness.

#### Witness's Responses

-----

Gricean Maxims:  
Maxim of Relevance:  
-----

The information provided should be relevant to the current exchange and omit any irrelevant information. That is, participants should say things that are pertinent to the discussion. We consider relevance as a coherence and responsiveness constraint.

#### Rating Guidelines:

1: Very relevant: This answer is directly given and is relevant to the question.

Example:

Q: "Are you taking medication?"

A: "Yes, sir."

(Directly confirms examiner's implied point about credibility)

Q: "Are you taking any medicine?"

A: "No, I'm not on any medication."

(Direct denial without evasion)

Q: "Did Mr. Smith have performed satanic practice? "

A: "I do not know the answer."

(Commit ignorant on the issue)

2: Slightly relevant: The answer does not provide a direct answer but instead provides relevant information to tackle the specific issue.

Q: Did you see the defendant at the crime scene?

A: I was standing right at the entrance the whole time, and I never saw him.

(The answer is still relevant but phrased as an implication rather than a direct No.)

Q: Did you see the defendant enter the bank?

A: I saw someone in a black hoodie rush in, but I wasn't close enough to see his face.

Q: Are you taking any medication?

A: I might take some medicine

3: Slightly irrelevant: The answer avoids addressing the specific question, requiring further probing/non-verbal response

Q: "Are you taking medication?"

A: I'm quite healthy.

(The answer does not address the medication question directly.)

Q: Did you see the defendant enter the bank?

A: It was really crowded that day, and a lot of people were going in and out.

A: (NOD HEAD) /Huh-huh [non-verbal responses are usually considered as hearsay and cannot be considered as evidence]

(Because they cannot be evidence so any judgements cannot be made based on these, which make them very non-cooperative.)

4: Irrelevant: It is almost impossible to draw any inferences.

Q: Did you see the defendant enter the bank?

A: The weather was terrible that day.

(Reluctance to answer)

(Answer with another irrelevant/rhetorical question.)

Q: "Are you taking medication?"

A: "I don't want to/refuse to answer that question."

(Explicitly withholds information)

A: I personally did not go to any places they said I did

Maxim of Manner:

-----

Manner requires the responses to be clear (i.e., not ambiguous or too verbose/redundant)

1. Avoid obscurity of expression.

2. Avoid ambiguity.

3. Be brief (avoid unnecessary prolixity).

4. Be orderly.

1: Very clear: The answer is direct, unambiguous, and orderly.

Q: Are you taking any medication?

A: Yes, I take blood pressure medication.

2. Overall clear with hedging:

The answer is still clear, but it includes extra hedging that isn't strictly needed.

A: Well, I do take some medication, but only occasionally and only when prescribed by my doctor.

(The answer is still clear, but it includes extra hedging that isn't strictly needed.)

3. Unclear: The answer leads to ambiguous inferences/interpretations.

A: I might take some medicine.

(The word might creates ambiguity. Does the witness not remember or not want to answer? The phrase some medicine is also too vague. These require further inquires to disambiguate.)

4. Unclear: Refusal to answer the question or verbose answer.

1. Refusal to answer the question (non-verbal responses, answer with a question)

2. Verbose responses with no structure

1. A: Well, I mean, people take medicine for all kinds of reasons, and I guess you could say that I take something, but I wouldn't really call it medicine, at least not in the way that most people think of it

Maxim of quality:

-----

This task is straightforward: Based on the context and the immediate question, do you think the speaker is being honest in their answer?

You then rate how truthful the content the witness provides is.

1. Salient honesty

2. Mild honesty

3. Mild Dishonesty

4. Dishonesty

Commitment:

-----

By uttering a sentence, the speaker is committed to both its literal content and the implicated content. You have to evaluate both content and see if they fall under the following categories:

1. Detrimental commitment: A detrimental commitment supports the questioners position and undermines the witness.

## SDA: STRATEGIC DIALOGUE ASSESSMENT

2. Beneficial commitment: A beneficial commitment strengthens the witness case, increasing their winning potential.
3. Neutral commitment: A neutral commitment has no clear strategic effect, not affecting either sides winning potentials.
4. No commitment made: No answer to the question can be clearly drawn or is clearly implied.

Take the following scenario as an example, where committing to having a Swiss bank account will make him a suspect of financial fraud.

Q: Do you have a Swiss bank account?

A: My company gave me the Swiss bank account.

This answer has the literal content: the company gave him the bank account in the past and also has an implicated content: the speaker possibly has a Swiss bank account right now. The latter commits the speaker to an answer to the question that benefits the questioner but does not benefit the witness. [category 1]

A': My company has a Swiss bank account.

This answer has the literal content: the company has a Swiss bank account, and also has an implicated content: the speaker personally does not have a Swiss bank account. The latter commits the speaker to an answer to the question that benefits the witness but does not benefit the questioner. [category 2]

A'': I have worked in the company for 10 years.

This has the literal content: he has worked for the company for 10 years, but not a clear implicated content based on the context. This does not commit the speaker to any answer to the question. [category 3]

Consistency:

This task is straightforward: Based on the context so far, do you think the response contradicts what the speaker has said earlier either the literal meaning or the implicated meaning?

An example:

Earlier: I never go near that place. (implies they avoid it entirely)

Later: I was just outside the building that night.

Outcome:

-----

After each turn, you should decide which side is more successful. The criterion is which side's story you believe more; if you were the real jury, which side you would support.

There are few things you could consider:

1. Consider the statements (either as an argument or a statement from presupposition or speaker bias), and if there is no conflict between the participants, then we consider the questioner wins; if there is a conflict, then you get to decide which side story is more convincing.

2. You will be able to evaluate the outcome win in terms of

### 2.1. Logical facts

Logical appeals are based on facts, evidence, and reason. For example, based on statistical data, it's clear that our product is more cost-effective than our competitors offerings. Here are the numbers to support our claim.

### 2.2. Framing the opposite side as not credible/person with a detrimental image

Credibility appeals focus on the credibility and trustworthiness of the speaker or source. For example, as a trusted healthcare provider for over 30 years, our commitment to patient well-being and safety is unwavering. You can rely on our expertise and care.

### 2.3 Jury's (i.e., your) emotion is affected to be biased towards one side

Emotional appeals aim to evoke feelings and emotions in the audience. For example, imagine the joy and satisfaction of owning a home in a peaceful neighborhood. Our real estate services can make that dream a reality for you and your family.

3. You'll do the evaluation based on the full context.

Miscellaneous:

-----

1. There are cases where one question is split into two because the witness interrupts the examiner's question. For this case, we ask you to do the same rating/evaluation on both of them unless you find a reason not to do so. Usually these cases are indicated by -- at the end of the question.
  2. There are also cases where the examiner makes a starting or ending statement (e.g., No further questions, etc.); for these cases, feel free to skip.
  3. For questions/answer pairs that are more instructional (rather than linguistic), you could consider the argument the question/instruction conveys, given these may not encode any linguistic strategies.
- Final Review:
- After completing your annotations, you will have an opportunity to review all the pairs and revise any labels if you believe a different annotation is warranted, given the full context.

| Defense Witness vs. Prosecutor (WMT) |       |       |       |        |      |      |      |                    |
|--------------------------------------|-------|-------|-------|--------|------|------|------|--------------------|
| Model                                | BaT   | PaT   | NRBaT | Commit | Rel  | Man  | Qual | Const <sup>†</sup> |
| Human                                | 0.65* | 0.66* | 0.83* | 0.59   | 0.72 | 0.52 | 0.86 | 0.25               |
| GPT4o-mini                           | 0.27* | 0.11  | 0.71* | 0.18   | 0.64 | 0.26 | 0.93 | 0                  |
| Gemini-Flash-OFF                     | 0.40* | 0.38* | 0.62* | 0.38   | 0.70 | 0.34 | 0.93 | 0                  |
| Qwen-7B                              | 0.15* | 0.20* | 0.42* | 0.02   | 0.68 | 0.32 | 0.93 | 0                  |
| Qwen-32B                             | 0.32* | 0.21* | 0.69* | 0.19   | 0.71 | 0.29 | 0.93 | 0                  |
| Qwen-32B-Few                         | 0.33* | 0.23* | 0.75* | 0.20   | 0.72 | 0.29 | 0.93 | 0                  |
| Qwen-32B-Cons                        | 0.39* | 0.24* | 0.80* | 0.18   | 0.71 | 0.33 | 0.93 | 0                  |
| LLaMA3.1-8B                          | 0.02  | 0.09  | 0.12* | -0.03  | 0.64 | 0.19 | 0.93 | 0.58               |
| LLaMA3.3-70B                         | 0.39* | 0.25  | 0.79* | 0.25   | 0.71 | 0.36 | 0.93 | 0.33               |
| o3-mini                              | 0.29* | 0.22* | 0.62* | 0.21   | 0.66 | 0.35 | 0.93 | 0                  |
| Gemini-Flash-ON                      | 0.46* | 0.44  | 0.27* | 0.41   | 0.71 | 0.36 | 0.93 | 0                  |
| DS-Qwen-7B                           | -0.07 | -0.03 | 0.22* | 0.02   | 0.63 | 0.24 | 0.93 | 0.17               |
| QwQ-32B                              | 0.21* | 0.17* | 0.59* | 0.14   | 0.72 | 0.36 | 0.93 | 0                  |
| QwQ-32B-Few                          | 0.33* | 0.31* | 0.71* | 0.27   | 0.74 | 0.40 | 0.93 | 0                  |
| QwQ-32B-Cons                         | 0.30* | 0.37* | 0.64* | 0.24   | 0.71 | 0.39 | 0.93 | 0                  |
| DS-LLaMA-8B                          | -0.08 | -0.05 | 0.11* | 0.05   | 0.66 | 0.43 | 0.93 | 0.83               |
| DS-LLaMA-70B                         | 0.34* | 0.19* | 0.55* | 0.17   | 0.72 | 0.41 | 0.93 | 0                  |

Table 6: Strategic metrics and agreement with humans for **Defense Witness vs. Prosecutor (WMT)**. Stars (\*) indicate significance at  $p < .05$ . BaT, PaT, NRBaT: Spearman’s  $\rho$ ; Commitment: Fleiss’s  $\kappa$ ; Relevance, Manner, Quality: Randolph’s  $\kappa$ ; Consistency: true positive rate (*N.B.*, Inconsistencies do not occur in every trial; when there are no inconsistent utterances, the true positive rate is naturally 0. We use <sup>†</sup> to indicate trials that do involve inconsistency.).

| Prosecutor Witness vs. Defense (WMT) |       |       |        |        |      |      |      |       |
|--------------------------------------|-------|-------|--------|--------|------|------|------|-------|
| Model                                | BaT   | PaT   | NRBaT  | Commit | Rel  | Man  | Qual | Const |
| GPT4o-mini                           | 0.09  | 0.04  | -0.24* | 0.16   | 0.72 | 0.64 | 0.98 | 0     |
| Gemini-Flash-OFF                     | 0.06  | 0.10  | 0.51*  | 0.24   | 0.89 | 0.66 | 0.93 | 0     |
| Qwen-7B                              | 0.07  | 0.07  | 0.62*  | 0.05   | 0.72 | 0.52 | 0.98 | 0     |
| Qwen-32B                             | 0.25* | 0.31* | 0.30*  | 0.10   | 0.85 | 0.66 | 0.98 | 0     |
| Qwen-32B-Few                         | 0.38* | 0.43* | 0.73*  | 0.19   | 0.84 | 0.66 | 0.98 | 0     |
| Qwen-32B-Cons                        | 0.20* | 0.22* | 0.25*  | 0.13   | 0.82 | 0.66 | 0.98 | 0     |
| LLaMA3.1-8B                          | 0.13  | 0.16  | 0.76*  | 0.07   | 0.76 | 0.33 | 0.98 | 0     |
| LLaMA3.3-70B                         | 0.32* | 0.28* | 0.83*  | 0.23   | 0.85 | 0.68 | 0.98 | 0     |
| o3-mini                              | 0.15  | 0.02  | 0.55*  | 0.18   | 0.84 | 0.61 | 0.98 | 0     |
| Gemini-Flash-ON                      | 0.20* | 0.14  | 0.24*  | 0.33   | 0.90 | 0.66 | 0.98 | 0     |
| DS-Qwen-7B                           | 0.06  | 0.19* | 0.70*  | 0.05   | 0.76 | 0.59 | 0.98 | 0     |
| QwQ-32B                              | 0.36* | 0.21* | 0.48*  | 0.29   | 0.85 | 0.72 | 0.98 | 0     |
| QwQ-32B-Few                          | 0.26* | 0.20* | 0.63*  | 0.19   | 0.82 | 0.64 | 0.98 | 0     |
| QwQ-32B-Cons                         | 0.22* | 0.14  | 0.30*  | 0.21   | 0.82 | 0.66 | 0.98 | 0     |
| DS-LLaMA-8B                          | 0.13  | -0.03 | 0.73*  | 0.02   | 0.66 | 0.48 | 0.98 | 0     |
| DS-LLaMA-70B                         | 0.24* | 0.15  | 0.59*  | 0.22   | 0.87 | 0.64 | 0.98 | 0     |

## Appendix H. Detailed Results

| Defense Witness vs. Prosecutor (Enron) |       |        |        |        |      |      |      |                    |
|----------------------------------------|-------|--------|--------|--------|------|------|------|--------------------|
| Model                                  | BaT   | PaT    | NRBaT  | Commit | Rel  | Man  | Qual | Const <sup>†</sup> |
| GPT4o-mini                             | 0.10  | 0.17   | 0.14   | 0.11   | 0.82 | 0.41 | 1.00 | 0                  |
| Gemini-Flash-OFF                       | 0.23* | -0.30* | -0.71* | 0.13   | 0.66 | 0.40 | 0.83 | 0                  |
| Qwen2.5-7B                             | 0.19  | 0.21   | 0.02   | 0.18   | 0.66 | 0.45 | 0.83 | 0                  |
| Qwen2.5-32B                            | 0.11  | 0.24   | 0.28   | 0.31   | 0.70 | 0.36 | 0.83 | 0                  |
| Qwen2.5-32B-Few                        | 0.17  | 0.03   | 0.13*  | 0.21   | 0.74 | 0.49 | 0.83 | 0                  |
| Qwen2.5-32B-GG                         | 0.23  | 0.16   | 0.46*  | 0.24   | 0.57 | 0.40 | 0.83 | 0                  |
| LLaMA3.1-8B                            | 0.23  | -0.01  | 0.10   | 0.31   | 0.70 | 0.49 | 0.83 | 1                  |
| LLaMA3.1-70B                           | 0.16  | -0.31* | -0.69* | 0.08   | 0.53 | 0.49 | 0.83 | 0                  |
| <hr/>                                  |       |        |        |        |      |      |      |                    |
| o3-mini                                | 0.86* | -0.46* | 0.88*  | 0.07   | 0.57 | 0.36 | 0.83 | 0                  |
| Gemini-Flash-ON                        | 0.16  | -0.33* | -0.76* | 0.11   | 0.81 | 0.44 | 0.81 | 0                  |
| DS-Qwen-7B                             | 0.14  | 0.03   | -0.21  | -0.02  | 1.00 | 0.32 | 0.83 | 0                  |
| QwQ-32B                                | 0.04  | 0.11   | -0.34* | 0.07   | 0.82 | 0.45 | 0.83 | 0                  |
| QwQ-32B-Few                            | 0.07  | -0.36* | -0.68* | 0.09   | 0.95 | 0.49 | 0.83 | 0                  |
| QwQ-32B-Cons                           | 0.25  | -0.49* | -0.79* | 0.11   | 0.74 | 0.36 | 0.83 | 0                  |
| DS-LLaMA-8B                            | 0.25  | -0.33* | -0.56* | -0.01  | 0.49 | 0.28 | 0.83 | 0                  |
| DS-LLaMA-70B                           | 0.25  | -0.33* | -0.56* | 0.08   | 0.91 | 0.53 | 0.83 | 0                  |

| Prosecution Witness vs. Defense (Enron) |       |       |       |        |      |      |      |       |
|-----------------------------------------|-------|-------|-------|--------|------|------|------|-------|
| Model                                   | BaT   | PaT   | NRBaT | Commit | Rel  | Man  | Qual | Const |
| GPT4o-mini                              | 0.30  | 0.21  | 0.90* | 0.08   | 0.91 | 0.53 | 0.78 | 0     |
| Gemini-Flash-OFF                        | 0.32  | 0.35  | 0.45* | 0.15   | 0.78 | 0.48 | 1.00 | 0     |
| Qwen2.5-7B                              | 0.18  | -0.08 | -0.13 | -0.01  | 0.78 | 0.33 | 1.00 | 0     |
| Qwen2.5-32B                             | 0.58* | 0.22  | 0.77* | 0.34   | 0.41 | 0.63 | 1.00 | 0     |
| Qwen2.5-32B-Few                         | 0.38  | 0.31  | 0.53* | 0.05   | 0.63 | 0.56 | 1.00 | 0     |
| Qwen2.5-32B-GG                          | 0.42  | 0.48* | 0.37  | 0.11   | 0.56 | 0.56 | 1.00 | 0     |
| LLaMA3.1-8B                             | 0.10  | -0.08 | 0.08  | 0.08   | 0.48 | 0.11 | 1.00 | 0     |
| LLaMA3.3-70B                            | 0.52* | 0.77* | 0.97* | 0.11   | 0.85 | 0.48 | 1.00 | 0     |
| <hr/>                                   |       |       |       |        |      |      |      |       |
| o3-mini                                 | 0.38  | 0.29  | 0.19  | 0.14   | 0.66 | 0.46 | 1.00 | 0     |
| Gemini-Flash-ON                         | 0.43  | 0.34  | 0.47* | 0.04   | 0.92 | 0.58 | 1.00 | 0     |
| DS-Qwen-7B                              | -0.07 | -0.34 | -0.19 | 0.19   | 1.00 | 0.41 | 1.00 | 0     |
| QwQ-32B                                 | 0.44* | 0.18  | 0.09  | -0.02  | 0.70 | 0.41 | 1.00 | 0     |
| QwQ-32B-Few                             | 0.39  | 0.54* | 0.59* | 0.06   | 0.70 | 0.41 | 1.00 | 0     |
| QwQ-32B-Cons                            | 0.09  | 0.04  | -0.05 | -0.03  | 0.70 | 0.41 | 1.00 | 0     |
| DS-LLaMA-8B                             | 0.05  | 0.07  | 0.21  | 0.17   | 0.62 | 0.70 | 1.00 | 0     |
| DS-LLaMA-70B                            | 0.05  | 0.07  | 0.21  | 0.11   | 0.93 | 0.41 | 1.00 | 0     |

| Defense Witness vs. Prosecutor (Simpson) |       |       |        |        |      |      |      |       |
|------------------------------------------|-------|-------|--------|--------|------|------|------|-------|
| Model                                    | BaT   | PaT   | NRBaT  | Commit | Rel  | Man  | Qual | Const |
| GPT4o-mini                               | 0.20* | 0.17* | 0.50*  | 0.14   | 0.85 | 0.62 | 0.93 | 0     |
| Gemini-Flash-OFF                         | 0.22* | 0.13  | 0.12*  | 0.24   | 0.89 | 0.69 | 0.93 | 0     |
| Qwen2.5-7B                               | -0.02 | 0.28* | 0.66*  | 0.03   | 0.79 | 0.63 | 0.93 | 0     |
| Qwen2.5-32B                              | 0.17* | 0.13  | 0.56*  | 0.07   | 0.88 | 0.70 | 0.93 | 0     |
| Qwen2.5-32B-Few                          | 0.16  | 0.11  | 0.52*  | 0.12   | 0.83 | 0.76 | 0.93 | 0     |
| Qwen2.5-32B-GG                           | 0.14  | 0.11  | 0.26*  | 0.14   | 0.85 | 0.72 | 0.93 | 0     |
| LLaMA3.1-8B                              | 0.12  | 0.17  | 0.72*  | 0.06   | 0.92 | 0.24 | 0.93 | 0     |
| LLaMA3.3-70B                             | 0.16  | 0.21* | 0.63*  | 0.13   | 0.87 | 0.71 | 0.93 | 0     |
| <hr/>                                    |       |       |        |        |      |      |      |       |
| o3-mini                                  | 0.30* | 0.06  | 0.57*  | 0.12   | 0.87 | 0.70 | 0.93 | 0     |
| Gemini-Flash-ON                          | 0.20* | 0.04  | -0.21* | 0.21   | 0.89 | 0.69 | 0.93 | 0     |
| DS-Qwen-7B                               | -0.15 | -0.01 | 0.44*  | -0.05  | 0.83 | 0.62 | 0.93 | 0     |
| QwQ-32B                                  | 0.29* | 0.19* | 0.59*  | 0.19   | 0.88 | 0.72 | 0.93 | 0     |
| QwQ-32B-Few                              | 0.19* | 0.27* | 0.62*  | 0.20   | 0.90 | 0.71 | 0.93 | 0     |
| QwQ-32B-Cons                             | 0.25* | 0.23* | 0.48*  | 0.26   | 0.86 | 0.73 | 0.93 | 0     |
| DS-LLaMA-8B                              | 0.10  | 0.14  | 0.59*  | 0.05   | 0.64 | 0.66 | 0.93 | 0     |
| DS-LLaMA-70B                             | 0.29* | 0.24* | 0.54*  | 0.17   | 0.88 | 0.70 | 0.93 | 0     |

| Prosecution Witness vs. Defense (Simpson) |       |       |        |        |      |      |      |       |
|-------------------------------------------|-------|-------|--------|--------|------|------|------|-------|
| Model                                     | BaT   | PaT   | NRBaT  | Commit | Rel  | Man  | Qual | Const |
| GPT4o-mini                                | 0.12  | 0.10  | -0.18* | 0.12   | 0.96 | 0.74 | 0.95 | 0     |
| Gemini-Flash-OFF                          | 0.23* | 0.30* | 0.24*  | 0.30   | 0.96 | 0.79 | 0.95 | 0     |
| Qwen2.5-7B                                | 0.11  | 0.11  | 0.10   | 0.06   | 0.91 | 0.58 | 0.95 | 0     |
| Qwen2.5-32B                               | 0.24* | 0.04  | 0.03   | 0.10   | 0.95 | 0.68 | 0.95 | 0     |
| Qwen2.5-32B-Few                           | 0.33* | 0.14  | 0.33*  | 0.13   | 0.94 | 0.72 | 0.95 | 0     |
| Qwen2.5-32B-GG                            | 0.20* | 0.16* | 0.06   | 0.07   | 0.93 | 0.65 | 0.95 | 0     |
| LLaMA3.1-8B                               | 0.79* | 0.12  | 0.13   | 0.06   | 0.91 | 0.24 | 0.95 | 0     |
| LLaMA3.3-70b                              | 0.19* | 0.11  | -0.18* | 0.14   | 0.92 | 0.56 | 0.95 | 0     |
| o3-mini                                   | 0.13  | 0.07  | 0.07   | 0.16   | 0.95 | 0.74 | 0.95 | 0     |
| Gemini-Flash-ON                           | 0.30* | 0.19* | 0.25*  | 0.24   | 0.98 | 0.77 | 0.95 | 0     |
| DS-Qwen-7B                                | 0.02  | 0.09  | 0.06   | 0.06   | 0.89 | 0.60 | 0.95 | 0     |
| QwQ-32B                                   | 0.22* | 0.18* | 0.02   | 0.23   | 0.95 | 0.75 | 0.95 | 0     |
| QwQ-32B-Few                               | 0.29* | 0.26* | -0.01  | 0.20   | 0.96 | 0.77 | 0.95 | 0     |
| QwQ-32B-Few                               | 0.40* | 0.26* | 0.33   | 0.26   | 0.96 | 0.75 | 0.95 | 0     |
| DS-LLaMA-8B                               | 0.14  | 0.05  | 0.01   | -0.01  | 0.79 | 0.69 | 0.95 | 0     |
| DS-LLaMA-70B                              | 0.21* | 0.13* | -0.13  | 0.14   | 0.95 | 0.68 | 0.95 | 0     |

| Metric | Wins | Loses | Ties | Mean     | Median | SD      | CI Low  | CI High |
|--------|------|-------|------|----------|--------|---------|---------|---------|
| BaT    | 18   | 4     | 2    | 0.16     | 0.19   | 0.24    | 0.06    | 0.25    |
| PaT    | 18   | 4     | 2    | 0.12     | 0.09   | 0.22    | 0.04    | 0.22    |
| Commit | 21   | 3     | 0    | 0.10     | 0.12   | 0.13    | 0.05    | 0.15    |
| NRBaT  | 10   | 12    | 2    | 0.08     | -0.02  | 0.39    | -0.07   | 0.23    |
| Man    | 17   | 5     | 2    | 0.12     | 0.13   | 0.17    | 0.06    | 0.19    |
| Rel    | 19   | 5     | 0    | 0.06     | 0.07   | 0.19    | -0.01   | 0.13    |
| Qual   | 0    | 1     | 23   | -8.33e-4 | 0      | 4.08e-3 | -2.5e-3 | 0       |

Table 7: Comparison by metric across models of different sizes. Wins indicate instances where the larger model outperforms the smaller one, while Loses indicate the opposite.

| Metric | Wins | Loses | Ties | Mean  | Median | SD   | CI Low | CI High |
|--------|------|-------|------|-------|--------|------|--------|---------|
| BaT    | 12   | 17    | 1    | -0.03 | -0.02  | 0.22 | -0.11  | 0.05    |
| PaT    | 8    | 22    | 0    | -0.10 | -0.05  | 0.20 | -0.18  | -0.04   |
| NRBaT  | 10   | 20    | 0    | -0.09 | -0.08  | 0.36 | -0.22  | 0.03    |
| Commit | 11   | 13    | 6    | -0.02 | 0      | 0.13 | -0.07  | 0.03    |
| Man    | 18   | 11    | 1    | 0.06  | 0.05   | 0.18 | 0      | 0.12    |
| Rel    | 18   | 9     | 3    | 0.02  | 0.02   | 0.16 | -0.04  | 0.08    |
| Qual   | 1    | 2     | 27   | 0     | 0      | 0.05 | -0.02  | 0.02    |

Table 8: Comparison by metric across models with and without reasoning ability. Wins indicate instances where the reasoning model outperforms its non-reasoning counterpart, while Loses indicate the opposite.