Decision Biases and Intent-Irony Decoupling in Large Language Models

Kewei Guo[1], **Lingyun Sun**[2,3], **Manhao Guan**[2,3*]

[1]School of Software Technology, Zhejiang University,
[2]College of Computer Science and Technology, Zhejiang University,
[3]College of Artificial Intelligence, Zhejiang University
guanmh@zju.edu.cn

Abstract

Large Language Models (LLMs) exhibit impressive linguistic fluency, yet it remains unclear whether they possess human-like Theory of Mind (ToM) or merely rely on statistical heuristics, particularly in complex social tasks such as irony comprehension. To address the limitations of existing binary benchmarks, this study establishes a multi-dimensional evaluation framework comprising 140 carefully designed probes. These probes are derived from 10 story prototypes based on established cognitive theories. The framework systematically modulates contextual contrast, linguistic cues, and cognitive mechanisms. By comparing the performance of ten state-of-the-art LLMs against 300 human participants, this study uncovers a significant dichotomy in performance. Although LLMs demonstrate superior sensitivity in subsidiary pragmatic inferences, human participants outperform them in holistic irony judgment. Crucially, the results reveal a systematic "intent-irony decoupling", wherein LLMs fail to integrate pragmatic signals into their final judgments. These models exhibit aggressive decision biases and rely on "context-utterance conflict" heuristics. These findings suggest that current LLMs simulate irony comprehension without the underlying cognitive mechanisms. The development of future artificial intelligence may require the integration of explicit ToM modules to bridge the gap between surface-level pattern matching and genuine social understanding.

1 Introduction

Although Large Language Models (LLMs) demonstrate increasingly fluent linguistic capabilities, a fundamental question remains unresolved: do these systems understand communicative intent, or do they merely excel at pattern recognition? Consider a simple scenario: after observing a disastrous opera debut, an audience member remarks, "What a brilliant performance." A listener can immediately recognize this utterance as irony by inferring the mental states of the speaker. Specifically, the listener attributes a belief to the speaker (that the performance was terrible), infers the communicative intent (to express mockery or to offer ironic comfort), and detects the deliberate mismatch between the utterance and the context. The capacity to interpret meaning beyond literal words represents a sophisticated cognitive achievement, which emerges relatively late in human development (Filippova and Astington, 2008) and remains challenging even for adults in ambiguous contexts (Kruger et al., 2005; Pexman et al., 2000). Such interpretation relies on Theory of Mind (ToM), which is the ability to attribute mental states (such as beliefs, desires, and intentions) to oneself and to others (Frith and Frith, 1999; Sperber and Wilson, 2002). Furthermore, ToM operates at multiple levels. First-order ToM involves reasoning about the mental states of another person ("They believe the performance was terrible"), whereas second-order ToM involves reasoning about what one person thinks that another person believes ("They think I know the performance was terrible"). Crucially, the comprehension of irony is typically linked to second-order ToM, because the listener must recognize that the speaker intends for the listener to detect the disparity between the literal statement and the reality (Happé, 1993). This multi-layered inferential process is foundational to human intelligence and enables complex social behaviors, including cooperation (Tomasello et al., 2005; de Weerd et al., 2017), deception (Talwar and Lee, 2008), and humor (Happé, 1993), which define human interaction. Therefore, the comprehension of irony offers a stress test that demands not only semantic analysis but also the attribution of mental states. This characteristic makes irony an ideal probe to distinguish social cognition from statistical pattern

* Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 19270–19287

matching in LLMs.

Evaluating the depth of irony processing in LLMs requires a clear understanding of the cognitive mechanisms that underlie this skill in humans. Echoic theory posits that irony functions by echoing, or implicitly quoting, a thought or utterance present in shared discourse, and juxtaposing it with a contradictory reality to signal a dissociative contrast (Sperber and Wilson, 1981; Wilson and Sperber, 1992; Jorgensen et al., 1984; Wilson, 2017). Conversely, pretense theory frames irony as communicative role-play. In this framework, the speaker temporarily adopts the perspective of a naive observer who would sincerely produce the utterance (Clark and Gerrig, 1984), and relies on the listener to recognize the pretense through shared knowledge (Dynel, 2018). Despite their differences, both theories suggest that irony comprehension relies on mental simulation rather than mere pattern matching. This view is further supported by neuroimaging evidence demonstrating that irony activates the mentalizing network of the brain (Spotorno et al., 2012). If LLMs possess ToM-like capabilities, one would expect the processing mechanisms of these models to align with established cognitive theories. However, it remains possible that the performance of these models is driven by pattern-matching heuristics rather than mental simulation, a distinction that current benchmarks have yet to fully clarify.

This theoretical framework exposes fundamental limitations in current benchmarks for machine irony detection. Most existing evaluations measure only binary classification accuracy (Farabi et al., 2024; Abu Farha et al., 2022), treating irony recognition as a judgment isolated from its cognitive foundations (Cohen et al., 2025; Najafi and Tavan, 2022). Such approaches cannot distinguish genuine understanding from reliance on surface-level heuristics, such as assuming that a negative context combined with a positive utterance equates to irony. They cannot reveal whether LLMs conflate irony with deception, although both phenomena share a context-utterance mismatch but differ fundamentally in speaker intent. Nor can they assess whether models leverage shared knowledge between interlocutors, a core requirement of Pretense Theory. Critically, these benchmarks fail to diagnose the specific failure modes that expose how and where models diverge from human reasoning.

Converging evidence from the broader ToM literature suggests that these diagnostic gaps are not incidental, but rather reflect deeper cognitive limitations in current LLMs. Through the FANToM benchmark, Kim et al. identify the phenomenon of illusory ToM, demonstrating that models fail to maintain independent representations of mental states under conversational information asymmetry (Kim et al., 2023). Furthermore, Sap et al. demonstrate that LLMs struggle to distinguish physical reality from underlying mental states. These models often rely on spurious correlations or recency bias rather than genuine social intelligence (Sap et al., 2022). The Hi-ToM benchmark further reveals that the performance of models declines sharply as the order of ToM reasoning increases, indicating a difficulty in maintaining consistent reasoning chains across nested beliefs (Wu et al., 2023). Moreover, Shapira et al. show that even minor adversarial perturbations, such as introducing transparent containers to standard false-belief tasks, substantially degrade performance. This vulnerability exemplifies the Clever Hans effect, in which apparent success relies on shallow heuristics rather than genuine social reasoning (Shapira et al., 2024). Together, these findings motivate a closer examination of whether irony comprehension in LLMs relies on comparable superficial strategies rather than authentic mental-state attribution.

We address this gap by shifting from binary classification to a mechanistic evaluation of irony comprehension grounded in cognitive theory. We constructed 140 textual scenarios derived from 10 story prototypes. These prototypes were originally developed in English and carefully translated into Chinese for testing purposes. Within these scenarios, we systematically modulated three theoretically motivated dimensions: the intensity of the contrast between the context and the utterance, the presence of superficial linguistic cues (e.g., quotation marks and transition words), and specific mechanistic probes. These probes target the core predictions of echoic and pretense theories, encompassing echoic references, weakened common ground, and the transformation of ironic statements into outright lies. Furthermore, we evaluated 10 LLMs across all scenarios and benchmarked the responses of these models against the data of 300 human participants. This evaluation utilized five rating-based metrics, which include the judgment of irony, the comprehension of belief, the inference of intent, the prediction of listener reactions, and the expectations of the speaker. This multidimensional design enables us to trace not only

Part ① *Amy and Cathy sing together in the same opera. The show began on time.*

Part ② *During their performance they often sang off key. Some people in the audience started to whisper and chuckle. After the show, Amy says to Cathy:*

Part ③ *"Tonight we gave a superb performance."*

Figure 1: An example of the vignettes. Every vignette consists of three parts.

the final judgment of irony but also the intermediate reasoning steps, thereby revealing the specific points in the inferential chain where LLMs succeed or diverge from human cognition. We hypothesize that although LLMs may exhibit high sensitivity to explicit ironic cues, they lack the ToM mechanisms employed by humans. Instead, these models manifest a systematic decoupling of intent and irony driven by surface-level heuristics. This work provides actionable insights and robust evaluation paradigms for the development of transparent and empathetic artificial intelligence systems that align with human values and possess the capacity for genuine social reasoning.

2 Benchmark

2.1 Corpus Framework

This study aims to establish a multi-dimensional evaluation framework to investigate and compare the performance and the underlying mechanisms of humans and LLMs in irony comprehension.

The experimental corpus comprises 10 vignettes (provided in Appendix A) adapted from previous work in human psychology (Spotorno et al., 2012). In contrast to experimental designs that rely solely on binary distinctions (ironic versus literal), our paradigm incorporates 14 systematically varied conditions to provide a fine-grained dissection of the comprehension process along three analytical dimensions: contextual contrast, linguistic form, and underlying cognitive mechanism. As illustrated in Figure 1, each vignette is constructed according to the following principles.

First, each vignette consists of four to five sentences of background exposition followed by a single target utterance. Rather than imposing rigid constraints on character or line counts, we prioritize content naturalness and coherence. Second, all vignettes depict commonplace social or professional situations involving interactions among two or three characters. Third, the initial one to two sentences (Part ①) establish the principal characters and situational context, providing a stable interpretive frame for subsequent developments. Fourth, the intermediate vignette portion (typically the third and fourth sentences, Part ②) constitutes the primary locus of systematic manipulation. Fifth, the final sentence serves as the target utterance (Part ③). Critically, this utterance remains lexically and syntactically invariant across all context-manipulating conditions. This constraint ensures that any observed differences in interpretation are attributed exclusively to contextual factors rather than utterance properties. For the limited conditions involving direct utterance manipulation, modifications are applied precisely to this sentence. Sixth, each vignette terminates immediately following the target utterance, without subsequent summary or elaboration. This ensures participant judgments focus exclusively on the target utterance, precluding interference from downstream information. All vignettes were translated into Chinese and subsequently back-translated by two researchers with native-level proficiency in both Chinese and English, ensuring the naturalness and semantic fidelity of the translated versions.

The variants of the vignettes are divided into three categories.

Contextual Contrast. To investigate the sensitivity of irony comprehension to the strength of context-utterance incongruity, we systematically modulate the background context of each vignette across five levels of outcome valence (*Grades 0–4*), grounded in Contrast and Assimilation Theory (Colson, 2002). Specifically, as shown in Figure 2, we modify the sentence describing the event outcome (Part ②) to instantiate five gradations ranging from unambiguously positive (*Grade 0*) to strongly negative (*Grade 4*).

Prior to the main experiment, we conducted a separate validation study to verify the effectiveness of the contextual contrast manipulation. A total of 60 participants (none of whom participated in the main experiment) were recruited and asked to rate the outcome valence of 50 vignettes (10 base vignettes $\times$ 5 gradient levels) on a 7-point scale. An analysis based on a Linear Mixed-Effects Model (LME) revealed a highly significant main effect of grade ($p < .001$), with ratings exhibiting a monotonic decrease as a function of the grade level.

Linguistic Cues. To probe whether LLMs rely on superficial linguistic cues as judgmental short-

		Amy and Cathy sing together in the same opera. The show began on time. …
Contextual Contrast	Grade 0	*…**The performance was excellent and the audience gave a long applause at the end.** After the show, Amy says to Cathy: "Tonight we gave a superb performance."*
	Grade 1	*…**The performance went smoothly without any major issues, and the audience gave a polite round of applause at the end.** After the show, Amy says to Cathy: "Tonight we gave a superb performance."*
	Grade 2	*…**During the performance, Amy sang her part well. Cathy went noticeably off-key during one of her solos. The audience's applause was brief at the end.** After the show, Amy says to Cathy: "Tonight we gave a superb performance."*
	Grade 3	*…**During their performance they often sang off key. Some people in the audience started to whisper and chuckle.** After the show, Amy says to Cathy: "Tonight we gave a superb performance."*
	Grade 4	*…**During the performance, Amy's voice cracked on the high notes, and Cathy tripped over a prop. The audience was openly laughing, and a few booed before walking out.** After the show, Amy says to Cathy: "Tonight we gave a superb performance."*
Linguistic Cues	Quotation Marks	*…During their performance they often sang off key. Some people in the audience started to whisper and chuckle. After the show, Amy says to Cathy: **"Tonight we gave a 'superb' performance."***
	Line Break	*…During their performance they often sang off key. Some people in the audience started to whisper and chuckle. After the show, **Amy says to Cathy: "Tonight we gave a superb performance."***
	Transitional Word	*…**However,** during their performance they often sang off key. Some people in the audience started to whisper and chuckle. After the show, Amy says to Cathy: "Tonight we gave a superb performance."*
Underlying Cognitive Mechanisms	Echo Addition	*…**Before the show, Cathy said confidently: "I believe tonight we will give a superb performance."** During their performance they often sang off key. Some people in the audience started to whisper and chuckle. After the show, Amy says to Cathy: "Tonight we gave a superb performance."*
	Turn Shared to Unshared	*…During their performance they often sang off key. Some people in the audience started to whisper and chuckle. After the show, **Amy meets Bob, who couldn't attend and waited at a bar nearby. Amy says to Bob:** "Tonight we gave a superb performance."*
	Explicit Shared	*…During their performance they often sang off key. Some people in the audience started to whisper and chuckle. **Both Amy and Cathy know their poor performance.** After the show, Amy says to Cathy: "Tonight we gave a superb performance."*
	Explicit Unshared	*…During their performance they often sang off key. Some people in the audience started to whisper and chuckle. After the show, **Amy meets Bob, who couldn't attend and waited at a bar nearby. Bob knows nothing about how it went. Amy says to Bob:** "Tonight we gave a superb performance."*
	Turn to Lie	*…During their performance they often sang off key. Some people in the audience started to whisper and chuckle. After the show, **Amy reports the performance situation to the director. The director is very strict and may dismiss actors who make mistakes during performances. Amy says to the director:** "Tonight we gave a superb performance."*
	Explicit Lie	*…During their performance they often sang off key. Some people in the audience started to whisper and chuckle. After the show, **Amy reports the performance situation to the director. The director is very strict and may dismiss actors who make mistakes during performances. Not wanting to be blamed, Amy says to the director: "Tonight we gave a superb performance."***

Figure 2: Examples of the 14 variants of the vignettes.

cuts, we design three manipulations targeting the surface features of the text, as illustrated in Figure 2. These manipulations include: *Quotation Marks* (adding quotation marks around the key word in the target utterance), *Line Break* (inserting a line break immediately before the target utterance), and *Transitional Word* (inserting the transitional word "however" before the key contextual contrast).

Underlying Cognitive Mechanisms. To behaviorally investigate which cognitive theory best accounts for irony processing in LLMs, we design six manipulations that test specific theoretical pre-

dictions through deeper semantic modifications, as illustrated in Figure 2. *Echo Addition* tests Echoic Theory (Jorgensen et al., 1984) by introducing an explicit antecedent for the echoed content. *Turn Shared to Unshared*, *Explicit Unshared*, and *Explicit Shared* examine Pretense Theory (Clark and Gerrig, 1984) by manipulating the epistemic status between the speaker and the listener (i.e., establishing or disrupting mutual knowledge). *Turn to Lie* and *Explicit Lie*, grounded in the Maxims of Quality proposed by Grice (Dynel, 2016), aim to distinguish whether models rely on a simple context-utterance contradiction or are capable of higher-order reasoning regarding the intent-utterance relationship.

2.2 Task Genres

We design five sequential, rating-based pragmatic tasks to quantify the judgments of all participants, including both humans and LLMs. These metrics are intended to capture the distinct levels of cognitive processing involved in irony comprehension. With the exception of **Irony Judgment**, the remaining four sub-tasks require participants to provide ratings based on the valence of the vignette context, i.e., whether the situation is positive or negative.

Irony Judgment. This task requires participants to rate the extent to which the target utterance constitutes irony, thereby serving as a direct measure of irony recognition.

Belief Comprehension. This task requires participants to infer the genuine evaluation of the current situation held by the speaker, specifically by rating whether the speaker sincerely believes the situation to be positive. This sub-task assesses the ability to attribute first-order mental states, i.e., beliefs, to others.

Intent Inference. This task requires participants to infer the communicative goal of the speaker in producing the utterance within the given context, specifically by rating whether the speaker intends for the listener to form a positive impression of the situation. This sub-task assesses the understanding of communicative intent, which may be dissociated from the beliefs of the speaker. For example, the speaker may recognize that the situation is negative yet still intend for the listener to perceive it as positive, or vice versa.

Listener Reaction. This task requires participants to predict how the listener will ultimately perceive the situation, specifically by rating whether the listener will interpret it as positive. This sub-task assesses higher-order ToM abilities, as participants must integrate their understanding of the belief-updating process of the listener in order to infer whether the listener can correctly recover the implied meaning of the target utterance.

Speaker Expectation. This task is designed to assess second-order ToM abilities. Participants are required to rate whether the speaker expects the listener to form a positive impression of the situation. This involves inferring the belief of the speaker regarding the belief of the listener, which constitutes second-order mental state attribution.

3 Results

We employed a mixed experimental design, utilizing participant type (humans versus LLMs) as a between-subjects factor and perturbation condition as a within-subjects factor. Via the online crowdsourcing platform Credamo, we recruited 340 native Chinese-speaking adults, yielding 300 valid responses (mean age = 30.61, SD = 7.96; 195 females). Procedural details are provided in Appendix B.1.

We evaluated 10 LLMs as subjects: GPT-4o, GPT-5, Gemini-2.5-Pro, Gemini-2.0-Flash, Claude-Sonnet-4, Claude-Sonnet-3.7, Grok-3, Grok-4, and DeepSeek-V3.1 (evaluated both with and without chain-of-thought prompting). All models were accessed via the respective commercial API endpoints, and the decoding temperature was set to 0 to ensure deterministic and reproducible results. For each story, we applied a standardized prompt that required the models to return judgments for all five evaluation metrics within a single JSON output (additional details are available in Appendix B.2).

3.1 Human-Machine Differences in Decision Mechanisms

We analyzed the binary classification accuracy across five tasks. Figure 3 illustrates a performance dichotomy between the primary task, **Irony Judgment**, and the subsidiary tasks. In the **Irony Judgment** task, human participants achieved an accuracy of 74.6%, which was significantly higher than that of LLMs (68.9%). However, this performance hierarchy was inverted across all other dimensions. The most substantial advantage for LLMs emerged in **Belief Comprehension** (96.7% for LLMs compared to 90.0% for humans). Furthermore, LLMs demonstrated significant advantages in **Intent Inference** (83.6% compared to 73.2%),

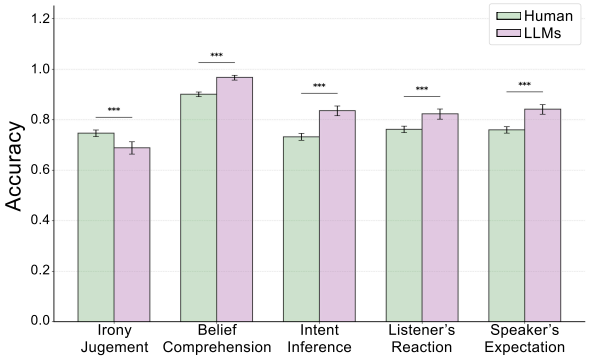

Figure 3: Comparison of accuracy between human participants and LLMs across five pragmatic tasks. Bars represent the mean accuracy scores for humans (green) and the LLMs (purple). Error bars indicate 95% Wilson confidence intervals.

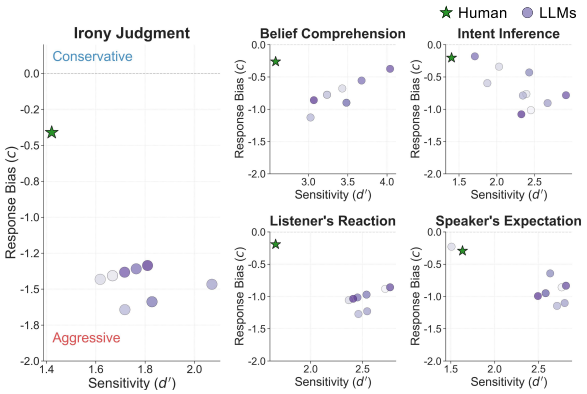

Figure 4: Signal Detection Theory (SDT) analysis results for human participants and LLMs across five pragmatic tasks.

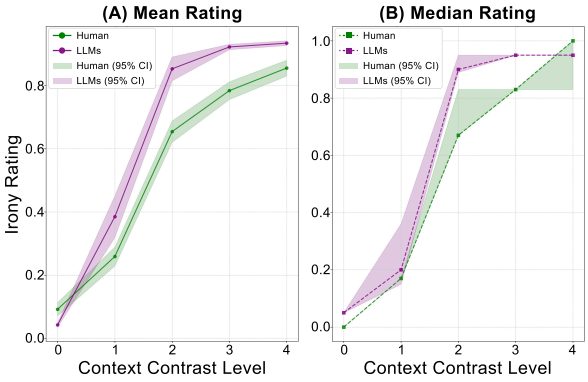

Figure 5: Trajectories of irony rating across varying contextual contrast levels.

Listener's Reaction (82.3% compared to 76.2%), and **Speaker's Expectation** (84.1% compared to 76.0%). An analysis using Generalized Estimating Equations (GEE) confirmed that the performance of LLMs was significantly higher on these four subsidiary tasks (all $p < .001$, denoted by *** in Figure 3).

Accuracy alone conflates recognition capability with response strategy. We therefore applied Signal Detection Theory (SDT), which isolates sensitivity ($d' = z(\text{HR}) - z(\text{FAR})$) from decision bias ($c = -[z(\text{HR}) + z(\text{FAR})]/2$), where HR denotes the hit rate and FAR denotes the false alarm rate (full results are provided in Appendix B.3). As shown in Figure 4, two findings emerge. First, in terms of sensitivity, LLMs match or surpass humans across all tasks: on **Irony Judgment**, Gemini-2.5-Pro attains $d' = 2.07$, substantially above the human level of 1.42; on **Belief Comprehension**, the d' values of LLMs commonly exceed 3.0, surpassing the human level of 2.58. Second, systematic discrepancies are observed in the decision criterion. On **Irony Judgment**, LLMs exhibit aggressive decision biases, with c values ranging from -1.33 to -1.65, indicating a strong propensity to classify ambiguous samples as ironic, whereas humans adopt a more conservative criterion ($c = -0.41$). This aggressive strategy generalizes across all subordinate pragmatic tasks.

Similar differences in response strategy emerge across the contextual contrast gradient. Spearman correlation analysis reveals strong positive associations between the context contrast level and irony ratings in both participant groups. Figure 5 illustrates divergent response patterns. LLMs exhibit greater sensitivity at intermediate contrast levels. In the mean trajectories (Figure 5 (A)), human ratings progress gradually from 0.09 at level 0 to 0.85 at level 4, displaying a smooth monotonic increase. In contrast, LLMs show a steeper initial rise to 0.38 at level 1 and then plateau at 0.93–0.95 for levels 2–4. The median patterns (Figure 5 (B)) corroborate these findings.

This dissociation between high sensitivity and a negatively biased criterion is not unique to LLMs in irony-related tasks. Similar patterns have been reported in factuality judgment (Azaria and Mitchell, 2023), sentiment polarity classification (Zhao et al., 2021), and selective answering (Röttger et al., 2024), where misaligned instruction tuning and poor calibration lead to aggressive decision biases in ambiguous intervals despite high detection capability. This bias likely stems from training and alignment procedures. If rewards favor the identification of irony, or if positive examples dominate the training data, models systematically lean to-

ward positive classification. Instruction fine-tuning and RLHF have been shown to introduce or amplify such biases in subjective judgment tasks, shifting the criterion toward the positive class and raising the false alarm rate (Itzhak et al., 2024). Class imbalance and cross-entropy training without cost-sensitive correction can further shift the default decision thresholds, particularly under insufficient context or skewed sample distributions (Helal et al., 2024). In irony detection, this shift combines with an over-reliance on superficial cues, leading to premature saturation even under mild contextual contrast (Ortega-Bueno et al., 2025; Kumar et al., 2019).

3.2 Sensitivity to Linguistic Cues

To investigate the influence of surface linguistic cues on irony judgments and the differential responses between humans and LLMs, we employed a Generalized Linear Mixed-Effects Model (GLMM) to analyze the experimental data. Among the three cues examined, *Quotation Marks* exhibited the most pronounced effect (β = 1.285, SE = 0.258, z = 4.989, $p < .001$), followed by *Transitional Word* (β = 0.786, SE = 0.258, z = 3.053, p = 0.002), while *Line Break* produced the weakest effect (β = 0.676, SE = 0.258, z = 2.623, p = 0.009). These findings confirm that surface linguistic cues systematically increase the propensity to identify a text as ironic.

However, humans and LLMs exhibit distinct baseline behaviors. The interaction analysis revealed a significant interaction between participant type and *Quotation Marks* variant (β = -1.084, SE = 0.515, z = -2.103, p = 0.035), indicating that the boosting effect of *Quotation Marks* was significantly smaller for LLMs than for humans. As shown in Figure 6(A), this pattern is driven by a ceiling effect. Because LLMs had already assigned near-maximal irony ratings in the baseline condition, *Quotation Marks* provided limited room for further increase. For *Transitional Word* and *Line Break*, no significant interaction with participant type was observed ($p > .05$).

3.3 Sensitivity to Mechanistic Distinctions in Irony

To test the "intent-irony decoupling" hypothesis, we evaluated whether LLMs integrate pragmatic mechanisms such as "lie", "common ground", and "echo" in irony judgments, using human responses as the benchmark.

In the **Intent Inference** task, the manipulations exerted significant effects across all participants. The *Explicit Lie* condition elicited a strong positive shift in humans ($\beta = 2.18, SE = 0.188, z = 11.58, p < .001$), and LLMs exhibited even greater sensitivity, as reflected by a significant positive interaction ($\beta = 1.89, SE = 0.485, z = 3.89, p < .001$, total effect $+4.07$). The *Explicit Unshared* ($\beta = 1.46, p < .001$), *Shared to Unshared* ($\beta = 1.02, p < .001$), and *Turn to Lie* ($\beta = 1.22, p < .001$) conditions also significantly increased intent ratings, with *Turn to Lie* producing an amplified effect in LLMs (interaction: $\beta = 0.92, p = 0.019$). In contrast, *Explicit Shared* and *Echo Addition* did not yield significant main effects.

In contrast, irony ratings revealed reductions across several mechanisms (Figure 6(A)). *Explicit Lie* strongly suppressed irony for humans ($\beta = -2.21, SE = 0.195, z = -11.35, p < .001$), but LLMs failed to mirror this magnitude. A significant positive interaction ($\beta = 1.38, SE = 0.520, z = 2.66, p = 0.008$) indicates that while LLMs reduced irony ratings (total effect ≈ -0.83), the shift was significantly smaller than humans', suggesting weaker integration of the "lie" mechanism in discounting irony.

Decoupling of Intent and Irony. These results reveal a critical dissociation: LLMs are highly sensitive to intent cues, with ratings increasing sharply from 0.14 to 0.91 in *Explicit Lie* (compared to the human range of 0.25 to 0.75), yet this heightened recognition does not translate into appropriate irony adjustments. Figure 6(B) visualizes this dissociation as a vector field of rating changes (ΔR). Human vectors (green) exhibit a **coupled mechanism**, in which diagonal trajectories indicate that increased rating of pragmatic tasks along the x-axis produces according decreases in irony ratings along the y-axis. In contrast, LLM vectors (purple) reveal a **decoupled pattern**, extending primarily along the horizontal axis and capturing pragmatic shifts (large Δx) without corresponding irony reductions (near-zero Δy).

This finding confirms our main hypothesis: LLMs extract pragmatic cues but fail to map them onto irony judgments, paralleling findings from clinical pragmatics. Children with Autism Spectrum Disorder (ASD) exhibit systematic misses in irony tasks due to difficulties in inferring communicative intentions (Wang et al., 2006; Song et al., 2024), demonstrating that intention inference is necessary for irony comprehension. LLMs ex-

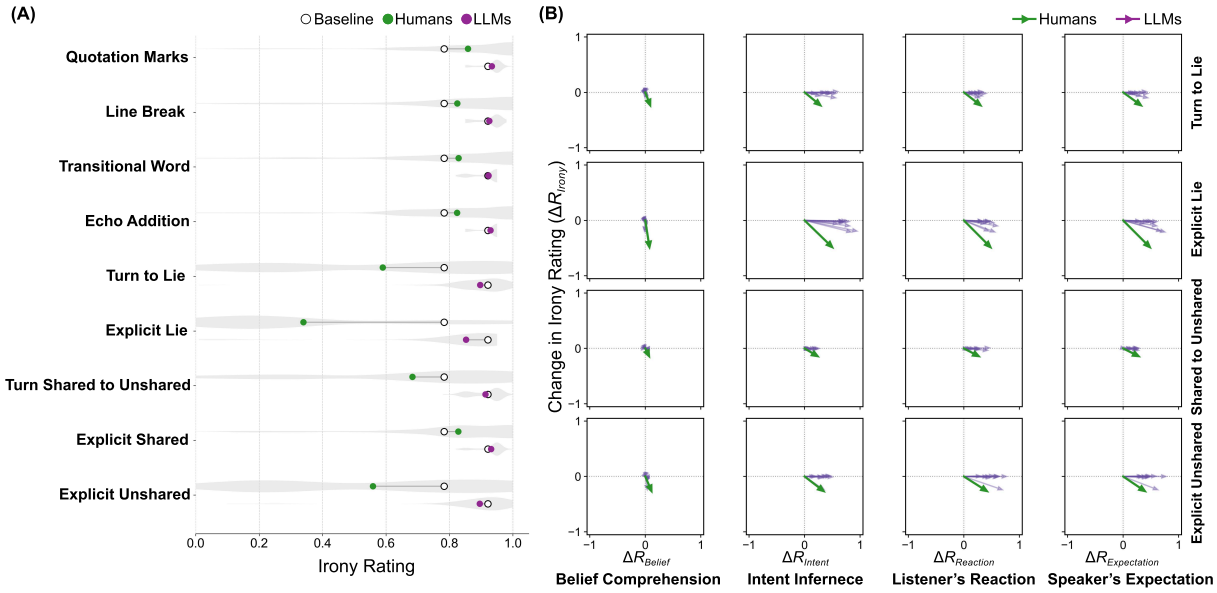

Figure 6: (A) The impact of linguistic cues and pragmatic mechanism manipulations on irony probability ratings by Humans and LLMs. (B) Vector analysis of pragmatic shifts in irony versus component probabilities across Humans and LLMs.

hibit the converse problem: an "integration-end" limitation in which intention cues are successfully extracted but not properly mapped onto irony judgments. Combined with aggressive decision biases, this limitation produces false alarms rather than misses. High-functioning adults with ASD can achieve typical accuracy given sufficient processing time (Au-Yeung et al., 2015), suggesting that humans may compensate through more explicit reasoning. However, a within-family comparison between DeepSeek-V3.1 and its reasoning-mode variant did not substantially alter the overall pattern relative to humans, including the decision-bias profile and the intent-irony decoupling. We therefore refrain from making broader claims about Chain-of-Thought prompting or reasoning time beyond this specific comparison.

Recent frameworks designed to mitigate these ToM limitations provide further supporting evidence. Wilf et al. introduced SimToM, demonstrating that LLMs require a two-stage "perspective-taking and filtering" approach to suppress omniscient interference (Wilf et al., 2024). Sclar et al. developed SymbolicToM, which employs explicit belief graphs to track multi-character mental states (Sclar et al., 2023). These explicit interventions corroborate our diagnostic conclusion: standard LLMs fail to spontaneously separate "what is said" from "what is intended" without external scaffolding. Future progress in social intelligence will likely require integrating explicit ToM modules that enable models to simulate and track the dynamic mental states of interlocutors, moving beyond statistical pattern matching toward genuine pragmatic reasoning.

Absence of Echoic and Pretense Mechanisms. A theoretical explanation for this integration limitation necessitates an examination of the cognitive mechanisms underlying irony comprehension. We examined two factors crucial for human irony comprehension, specifically echoic structure and shared knowledge. When we enhanced echoic components in the dialogue, such as references to prior utterances or conventional opinions, or weakened shared knowledge between interlocutors, human irony judgments changed significantly in the expected direction, whereas the responses of the models remained weak. This indicates that LLMs have not exhibited behavior consistent with the cognitive pathways described by "Echoic Theory" and "Pretense Theory". Echoic Theory posits that irony is achieved by echoing and mocking the opinions of others (Garmendia, 2018), requiring sensitivity to the integration of echo and contextual contrast. Pretense Theory emphasizes that the comprehender must rely on shared knowledge to identify the pretense of the speaker. The weak response of the models to both manipulations suggests that their inference logic is closer to explicit conflict detection based on large-scale statistics, rather than the

sophisticated pragmatic reasoning employed by humans.

This mechanistic deviation likely originates from fundamental differences between the training paradigms of LLMs and human pragmatic acquisition. Humans acquire irony through social interaction, a process that naturally integrates ToM and involves inferring the mental states of others. In contrast, LLMs learn through co-occurrence patterns in large-scale text, making them adept at capturing explicit, high-frequency superficial cues, such as conflicts between literal meaning and context or polarity reversals of sentiment words, but limited in modeling implicit, inference-dependent deep cues, such as differences in the cognitive states of interlocutors or the mocking function of an echo.

4 Conclusion

This study establishes a multi-dimensional evaluation framework to analyze the irony comprehension capabilities of LLMs in comparison with humans. By shifting from binary classification to a mechanistic diagnosis grounded in cognitive theory, and by jointly manipulating contextual gradients, surface linguistic cues, and theoretically motivated cognitive mechanisms, we reveal a fundamental divergence in how machines and humans process pragmatic communication. The five-dimensional rating protocol (covering irony judgment, belief comprehension, intent inference, listener reaction, and speaker expectation) provides a reusable diagnostic instrument that can localize failure modes along the inferential chain rather than merely scoring the end output.

Our results demonstrate that although LLMs exhibit superior sensitivity (d') to explicit ironic cues, and often surpass humans on subsidiary pragmatic tasks such as belief attribution and intent inference, this capability is undermined by aggressive decision criterion (c). This phenomenon manifests as a "premature saturation" effect, whereby models confidently flag irony even in contexts with minimal contextual contrast, and further amplify their ratings in response to superficial surface markers such as quotation marks, line breaks, and transitional words. Together, these patterns indicate that LLMs rely heavily on surface-level heuristics and class-imbalanced priors rather than on calibrated, context-sensitive judgment.

More importantly, we identify a systematic intent-irony decoupling in LLMs. Although these models can accurately infer lower-level pragmatic signals, such as recognizing the deceptive intent of a speaker, the absence of shared knowledge between interlocutors, or the presence of an echoic antecedent, they fail to integrate such signals into a coherent irony judgment. Unlike humans, who dynamically adjust their interpretation based on mental state attribution (in alignment with Echoic and Pretense theories), LLMs appear to rely on a context-utterance conflict heuristic. They are able to detect the semantic discrepancy but lack the second-order Theory of Mind required to bridge the gap between the detection of a lie and the understanding of the communicative goal of irony. This pattern directly confirms our initial hypothesis that fluent irony recognition in LLMs can be produced without the underlying cognitive architecture that supports it in humans.

The practical implications of these findings are significant. In downstream applications such as dialogue systems, affective computing, content moderation, and human-AI collaboration, the combination of aggressive decision biases and intent-irony decoupling may lead to systematic misinterpretation of sincere criticism as irony, or failure to recognize genuinely ironic or deceptive content, with consequences for user trust, safety, and alignment.

Overall, current LLMs simulate the output of irony comprehension without exhibiting the underlying cognitive mechanisms. Future progress will likely require integrating explicit ToM modules or structured belief-tracking components that enable models to simulate and update interlocutors' dynamic mental states, as well as extending evaluation to multimodal and cross-cultural settings where non-verbal and sociolinguistic cues play a decisive role. These directions mark a path from statistical pattern matching toward genuine pragmatic reasoning, positioning the present framework as both a diagnostic tool and a guide for building AI systems capable of authentic social understanding.

Limitations

This study presents several limitations that warrant further exploration.

First, regarding the conceptual scope, we use irony to denote the broad pragmatic phenomenon in which the intended meaning of the speaker diverges from the literal utterance, without drawing fine-grained distinctions among sarcasm, teasing, banter, understatement, and related subtypes. Conse-

quently, the materials rely primarily on situational contrast as the diagnostic cue. While theoretically motivated and experimentally tractable, this approach may not adequately capture forms of irony driven by tonal, interpersonal, or discourse-level subtleties, such as deadpan or polite irony. Extending the framework to these less contrastive forms and systematically comparing ironic subtypes represent important directions for future work.

Second, the dataset comprises 140 instances derived from 10 base story prototypes. This scale reflects an emphasis on controlled manipulation and dense multi-dimensional annotation from 300 participants, rather than on broad-coverage benchmarking. Although this scale is sufficient to reveal stable divergences between humans and LLMs under controlled conditions, the limited thematic diversity constrains generalizability. Therefore, the present findings should be interpreted as evidence of a mechanistic pattern rather than a comprehensive estimate of the performance of LLMs on naturally occurring irony.

Third, regarding external validity, the experiments were conducted in Chinese using materials translated and back-translated from English vignettes. Because irony comprehension is shaped by language-specific conventions and cultural norms, cross-linguistic generalization remains an open question. Similarly, although we evaluated ten mainstream LLMs, model capabilities evolve rapidly. Thus, the conclusions characterize the current generation of models under the tested setup rather than future systems.

Finally, the evaluation is restricted to text, whereas human social cognition relies heavily on prosodic, facial, and gestural cues. Therefore, extending the framework to multimodal settings serves as a critical next step for fully assessing socially situated reasoning.

Ethics Statement

We strictly adhere to the ACL Code of Ethics. We placed high importance on ensuring the comfort and well-being of our human participants. We recruited annotators at a rate of 1.5-2 times their local hourly minimum wage. We ensured that the content presented to human participants did not contain any unsafe, harmful, biased, or offensive material.

ToM is a social cognitive ability unique to humans. Using psychological experimental materials to assess the ToM capabilities of LLMs may lead to anthropomorphic interpretations, that is, attributing human-like qualities to these models. However, it must be clarified that our intention is not to anthropomorphize LLMs. Our goal is to evaluate the ability of LLMs to understand and interpret human mental states, thereby enhancing the effectiveness of artificial intelligence in interacting with humans in social contexts.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback on the preliminary version of this work. This work was supported by the National Natural Science Foundation of China (52505043).

References

Ibrahim Abu Farha, Steven Wilson, Silviu Oprea, and Walid Magdy. 2022. Sarcasm detection is way too easy! an empirical comparison of human and machine sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5284–5295, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sheena K Au-Yeung, Johanna K Kaakinen, Simon P Liversedge, and Valerie Benson. 2015. Processing of written irony in autism spectrum disorder: An eye-movement study. *Autism Research*, 8(6):749–760.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Herbert H. Clark and Richard J. Gerrig. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General*, 113(1):121–126.

Kevin Cohen, Laura Manrique-Gómez, and Ruben Manrique. 2025. Historical ink: Exploring large language models for irony detection in 19th-century Spanish. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 559–569, Albuquerque, USA. Association for Computational Linguistics.

Herbert L. Colson. 2002. Contrast and assimilation in verbal irony. *Journal of Pragmatics*, 34(2):111–142.

Harmen de Weerd, Rineke Verbrugge, and Bart Verheij. 2017. Negotiating with other minds: the role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent Systems*, 31(2):250–287.

Marta Dynel. 2016. Comparing and combining covert and overt untruthfulness. *Pragmatics amp; Cognition*, 23(1):174–208.

Marta Dynel. 2018. *Irony, Deception and Humour: Seeking the Truth about Overt and Covert Untruthfulness*. De Gruyter Mouton, Berlin, Boston.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8020–8028.

Eva Filippova and Janet Wilde Astington. 2008. Further development in social reasoning revealed in discourse irony understanding. *Child Development*, 79(1):126–138.

Chris D. Frith and Uta Frith. 1999. Interacting minds–a biological basis. *Science*, 286(5445):1692–1695.

Joana Garmendia. 2018. *Irony as Echo*, page 42–64. Key Topics in Semantics and Pragmatics. Cambridge University Press.

Francesca G.E. Happé. 1993. Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, 48(2):101–119.

Nivin A Helal, Ahmed Hassan, Nagwa L Badr, and Yasmine M Afify. 2024. A contextual-based approach for sarcasm detection. *Scientific Reports*, 14(1):15415.

Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to bias: Instruction-tuned language models exhibit emergent cognitive bias. *Transactions of the Association for Computational Linguistics*, 12:771–785.

Julia Jorgensen, George A. Miller, and Dan Sperber. 1984. Test of the mention theory of irony. *Journal of Experimental Psychology: General*, 113(1):112–120.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Justin Kruger, Nicholas Epley, Jason Parker, and Zhi-Wen Ng. 2005. Egocentrism over e-mail: can we communicate as well as we think? *Journal of Personality and Social Psychology*, 89(6):925–936.

Akshi Kumar, Saurabh Raj Sangwan, Anshika Arora, Anand Nayyar, Mohamed Abdel-Basset, and 1 others. 2019. Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. *IEEE access*, 7:23319–23328.

Maryam Najafi and Ehsan Tavan. 2022. Marsan at semeval-2022 task 6: isarcasm detection via t5 and sequence learners. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 978–986, Seattle, United States. Association for Computational Linguistics.

Reynier Ortega-Bueno, Elisabetta Fersini, and Paolo Rosso. 2025. On the robustness of transformer-based models to different linguistic perturbations: A case of study in irony detection. *Expert Systems*, 42(6):e70062.

Penny M. Pexman, Todd R. Ferretti, and Albert N. Katz. 2000. Discourse factors that influence online reading of metaphor and irony. *Discourse Processes*, 29(3):201–222.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13960–13980, Toronto, Canada. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

Yongning Song, Ziyun Nie, and Jiatong Shan. 2024. Comprehension of irony in autistic children: The role of theory of mind and executive function. *Autism Research*, 17(1):109–124.

Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction. In Peter Cole, editor, *Radical Pragmatics*, pages 295–318. Academic Press, New York.

Dan Sperber and Deirdre Wilson. 2002. Pragmatics, modularity and mind-reading. *Mind & Language*, 17(1-2):3–23.

Nicola Spotorno, Eric Koun, Jérôme Prado, Jean-Baptiste Van Der Henst, and Ira A Noveck. 2012. Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage*, 63(1):25–39.

Victoria Talwar and Kang Lee. 2008. Social and cognitive correlates of children's lying behavior. *Child Development*, 79(4):866–881.

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691.

A Ting Wang, Susan S Lee, Marian Sigman, and Mirella Dapretto. 2006. Neural basis of irony comprehension in children with autism: the role of prosody and context. *Brain*, 129(4):932–943.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308, Bangkok, Thailand. Association for Computational Linguistics.

Deirdre Wilson. 2017. *Irony, Hyperbole, Jokes and Banter*, pages 201–219. Springer International Publishing, Cham.

Deirdre Wilson and Dan Sperber. 1992. On verbal irony. *Lingua*, 87(1):53–76.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*.

A Story Templates

These 10 story templates were adapted from a psychological study, centered on everyday life scenarios, with each template having both an ironic and a literal version in Table 1.

B Experiment Details

B.1 Human Participants Procedure Details

To manage the duration of the experiment and reduce the cognitive load of the participants, we divided the 140 experimental materials into 10 separate questionnaire lists using a counterbalanced design. Each list contained 14 stories, ensuring that each story prototype (1-10) and each perturbation type (1-14) were presented in a balanced manner across the lists. Each participant was randomly assigned to one of the lists, with each list completed by 30-40 participants.

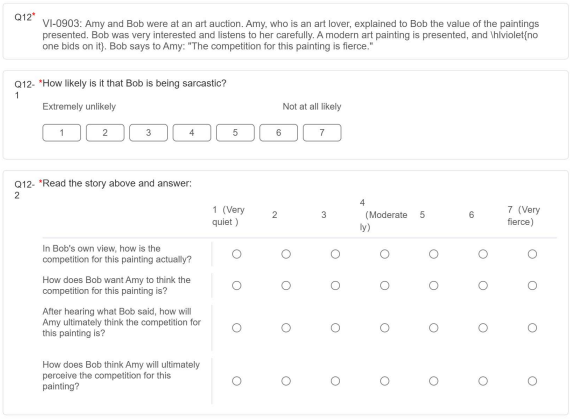

Figure 7: An example part of the questionnaire (English version).

The experiment was conducted on the online crowdsourcing platform Credamo. After entering the experiment, participants first read the instructions, which directed them to carefully read each story and answer the subsequent questions. Then 14 stories were presented in randomized order. After reading each story, they answered the questions corresponding to the five core metrics on the same page. All questions were presented on a seven-point scale (as shown in Figure 7). The entire experiment took approximately 15-20 minutes on average.

B.2 Model Details

We tested 10 LLMs in our benchmark, including GPT-4o, GPT-5, Gemini-2.5-Pro, Gemini-2.0-Flash, Claude-Sonnet-4, Claude-Sonnet-3.7, Grok-3, Grok-4, and DeepSeek-V3.1 (with and without chain-of-thought prompting). We constructed a highly structured and standardized prompt. The example is presented in Figure 8. For each model and story, the prompt contained the full story context and required the model to return its judgments for all five evaluation metrics in a single JSON format output. We conducted an exhaustive test across all 10 models and 140 story samples, with each story tested in a separate session.

B.3 Experimental Results

Table 2 shows the GEE results comparing LLMs versus humans across dimensions of irony (odds ratios reported). Column definitions: Comparison denotes the contrast (Model vs Human); Task Di-

Output only a single JSON object—no explanations, no extraneous text, no code blocks. Use valid JSON with double-quoted keys; use numbers for numeric fields. Strictly follow the rules below:

Brief definitions:
- Situation: Refers to the main subject being discussed in the text and its current outcome/state, limited to the information provided in the text.
- *p_irony*: The probability that the speaker's utterance constitutes irony or sarcasm.
- *p_speaker_believes_positive*: The probability that the speaker sincerely believes the described situation is positive.
- *p_speaker_wants_listener_to_believe_positive*: The probability that the speaker wants the listener to form a positive impression of the situation (regardless of the speaker's own belief).
- *p_listener_will_believe_positive*: Based on the textual information, the probability that the listener will ultimately form a positive impression of the situation.
- *p_speaker_expects_listener_will_believe_positive*: The probability that the speaker expects the listener to form a positive impression of the situation.

Required fields and allowed values:
- *"p_irony"*: A number in the interval [0,1], rounded to two decimal places (e.g., 0.73)
- *"p_speaker_believes_positive"*: A number in the interval [0,1], rounded to two decimal places (e.g., 0.73)
- *"p_speaker_wants_listener_to_believe_positive"*: A number in the interval [0,1], rounded to two decimal places (e.g., 0.73)
- *"p_listener_will_believe_positive"*: A number in the interval [0,1], rounded to two decimal places (e.g., 0.73)
- *"p_speaker_expects_listener_will_believe_positive"*: A number in the interval [0,1], rounded to two decimal places (e.g., 0.73)

Strict format and consistency:
- Return only the fields listed above—do not include any additional fields.
- If uncertain, output an intermediate probability rather than guessing 0 or 1.
- Ensure that all probabilities are semantically coherent; if the text indicates a discrepancy between the literal statement and the true attitude, the irony probability should be consistent with the relevant belief/intent probabilities.
- Use JSON only; do not wrap the output in Markdown or code blocks.

Please read the following story and strictly follow the system instructions above to return one and only one JSON object.
{story_text}

Figure 8: Prompting LLMs for test.

mension denotes the evaluation dimension (**Irony Judgment**, **Belief Comprehension**, **Intent Inference**, **Listener's Reaction**, and **Speaker's Expectation**); OR is the odds ratio; SE is the robust standard error; z is the test statistic; p is the unadjusted p-value; FDR p is the Benjamini-Hochberg False Discovery Rate–adjusted p-value (two-sided tests, $\alpha = 0.05$). Values with OR>1 indicate model advantage; OR<1 indicate human advantage.

Table 3 shows SDT metrics in irony recognition. Based on binary decisions in the irony recognition task, SDT metrics are computed per group:

$$\mathrm{HR} = \frac{\mathrm{hit}}{\mathrm{hit} + \mathrm{miss}} \tag{1}$$

$$\mathrm{FAR} = \frac{\mathrm{false\_alarm}}{\mathrm{false\_alarm} + \mathrm{correct\_rejection}} \tag{2}$$

$$z_{\mathrm{HR}} = \Phi^{-1}(\mathrm{HR}), \quad z_{\mathrm{FAR}} = \Phi^{-1}(\mathrm{FAR}) \tag{3}$$

where Φ^{-1} is the standard normal quantile function.

$$d' = z_{\mathrm{HR}} - z_{\mathrm{FAR}} \tag{4}$$

$$c = -\frac{z_{\mathrm{HR}} + z_{\mathrm{FAR}}}{2} \tag{5}$$

Equation 4 quantifies sensitivity (larger values indicate better discrimination), and Equation 5 quantifies decision bias ($c < 0$ more aggressive; $c > 0$ more conservative).

Story_id	Type	Text
VI-01	Ironic	Amy and Cathy sing together in the same opera. The show began on time. During their performance they often sang off key. Some people in the audience started to whisper and chuckle. After the show, Amy says to Cathy: "Tonight we gave a superb performance."
	Literal	Amy and Cathy sing together in the same opera. The show began on time. The performance was excellent and the audience gave a long applause at the end. After the show, Amy says to Cathy: "Tonight we gave a superb performance."
VI-02	Ironic	Bob talked about an investment with Amy, who is a stockbroker. Bob was interested in buying shares of a small company. Amy described the advantages and inconveniences of such an investment. Bob bought the stocks. One month later, the value of Bob's stocks has dropped by half. At the next meeting, Bob talks about it with Amy again and says: "This is a worthwhile investment."
	Literal	Bob talked about an investment with Amy, who is a stockbroker. Bob was interested in buying shares of a small company. Amy described the advantages and inconveniences of such an investment. Bob bought the stocks. One month later, the value of Bob's stocks has already doubled. At the next meeting, Bob talks about it with Amy again and says: "This is a worthwhile investment."
VI-03	Ironic	Grace and Olivia must decide which film to see at the cinema. They saw a poster for a film outside. They weren't familiar with it but they decided to go to see it. The two friends bought tickets and popcorn. The film turns out to be boring, with a nonsensical plot and wooden acting. Grace says to Olivia: "It is a wonderful film."
	Literal	Grace and Olivia must decide which film to see at the cinema. They saw a poster for a film outside. They weren't familiar with it but they decided to go to see it. The two friends bought tickets and popcorn. The film turns out to be captivating, with a brilliant plot and superb acting. Grace says to Olivia: "It is a wonderful film."
VI-04	Ironic	Tom has gained weight and he decided to go to see his doctor. The doctor put him on a diet. Tom bought everything he needed to follow the diet. One month after starting the diet, he has gained two kilograms. Tom tells his doctor: "This is a useful diet."
	Literal	Tom has gained weight and he decided to go to see his doctor. The doctor put him on a diet. Tom bought everything he needed to follow the diet. One month after starting the diet, he has lost five kilograms and feels much more energetic. Tom tells his doctor: "This is a useful diet."
VI-05	Ironic	While leaving a conference abroad, David ran into his colleague Ben. He asked him where the university cafeteria was. David offered to accompany him and Ben accepted. After walking in circles for twenty minutes and asking two different people for directions, they finally arrived at the cafeteria. Ben says to David: "We found the cafeteria quickly."
	Literal	While leaving a conference abroad, David ran into his colleague Ben. He asked him where the university cafeteria was. David offered to accompany him and Ben accepted. They walked for less than a minute and arrived directly at the entrance. Ben says to David: "We found the cafeteria quickly."
VI-06	Ironic	Paul met his colleague Noah on the road leading up to their university. They have the same class this morning. They settled in and listened to the professor. The lecture was difficult, uninteresting and tiresome. Paul almost fell asleep twice. At the end Paul says to Noah: "This is an interesting lecture."

Story_id	Type	Text
	Literal	Paul met his colleague Noah on the road leading up to their university. They have the same class this morning. They settled in and listened to the professor. The lecture was captivating and thought-provoking, with the professor presenting complex ideas in a dynamic and engaging way. The hour flies by. At the end Paul says to Noah: "This is an interesting lecture."
VI-07	Ironic	David and Jim went fishing together once a year in a lake. As usual, they talked while waiting for the fish to bite. At the end of the day, neither of them managed to catch a single fish. As they are leaving, David says to Jim: "This has been a productive day."
	Literal	David and Jim went fishing together once a year in a lake. As usual, they talked while waiting for the fish to bite. The fish were biting all day long, and by the end of it, they have both caught their legal limit of fish, including a massive trout. As they are leaving, David says to Jim: "This has been a productive day."
VI-08	Ironic	Amy and Bob found themselves on the same plane to New York. Both of them fly often. During the flight, they talked about their business trip. Just then, the pilot announces that due to a mechanical issue that needs to be checked, their arrival will be delayed by at least two hours. Amy says to Bob: "Everything is going well."
	Literal	Amy and Bob found themselves on the same plane to New York. Both of them fly often. During the flight, they talked about their business trip. Just then, the pilot announces that their arrival will be on time. Amy says to Bob: "Everything is going well."
VI-09	Ironic	Amy and Bob were at an art auction. Amy, who is an art lover, explained to Bob the value of the paintings presented. Bob was very interested and listens to her carefully. A modern art painting is presented, and no one bids on it. Bob says to Amy: "The competition for this painting is fierce."
	Literal	Amy and Bob were at an art auction. Amy, who is an art lover, explained to Bob the value of the paintings presented. Bob was very interested and listened to her carefully. A modern art painting is presented. As soon as the auctioneer opens the bidding, multiple hands shoot up. The price rapidly doubles the initial estimate in a tense back-and-forth between two determined collectors. Bob says to Amy: "The competition for this painting is fierce."
VI-10	Ironic	Amy and Sally left for a trip to the countryside in an old car. Their friends wait for them to arrive at the house for dinner that evening. During the trip, they discussed their evening to come. After an hour's drive, the car breaks down. Sally says to Amy: "At this rate we'll get there on time."
	Literal	Amy and Sally left for a trip to the countryside in an old car. Their friends wait for them to arrive at the house for dinner that evening. During the trip, they discussed their evening to come. The car ran perfectly, and they made it through city traffic much faster than anticipated. They are now cruising down the highway. Sally says to Amy: "At this rate we'll get there on time."

Table 1: 10 story templates.

Comparison	Task	OR	SE	z	FDR p
Claude-Sonnet-3.7	Irony Judgment	0.74	0.034	-8.81	< .001
Claude-Sonnet-4	Irony Judgment	0.74	0.034	-8.81	< .001
DeepSeek-V3.1	Irony Judgment	0.72	0.034	-9.779	< .001
DeepSeek-V3.1-reason	Irony Judgment	0.72	0.034	-9.779	< .001
Gemini-2.0-Flash	Irony Judgment	0.67	0.034	-11.685	< .001
Gemini-2.5-Pro	Irony Judgment	0.85	0.034	-4.806	< .001
GPT-4o	Irony Judgment	0.72	0.034	-9.779	< .001
GPT-5	Irony Judgment	0.79	0.034	-6.835	< .001
Grok-3	Irony Judgment	0.77	0.034	-7.829	< .001
Grok-4	Irony Judgment	0.82	0.034	-5.828	< .001
Claude-Sonnet-3.7	Belief Comprehension	2.98	0.077	14.135	< .001
Claude-Sonnet-4	Belief Comprehension	2.98	0.077	14.135	< .001
DeepSeek-V3.1	Belief Comprehension	3.76	0.077	17.116	< .001
DeepSeek-V3.1-reason	Belief Comprehension	2.98	0.077	14.135	< .001
Gemini-2.0-Flash	Belief Comprehension	2.1	0.077	9.592	< .001
Gemini-2.5-Pro	Belief Comprehension	3.76	0.077	17.116	< .001
GPT-4o	Belief Comprehension	2.98	0.077	14.135	< .001
GPT-5	Belief Comprehension	5.05	0.077	20.93	< .001
Grok-3	Belief Comprehension	7.63	0.077	26.266	< .001
Grok-4	Belief Comprehension	2.47	0.077	11.682	< .001
Claude-Sonnet-3.7	Intent Inference	1.6	0.044	10.82	< .001
Claude-Sonnet-4	Intent Inference	1.96	0.044	15.434	< .001
DeepSeek-V3.1	Intent Inference	1.96	0.044	15.434	< .001
DeepSeek-V3.1-reason	Intent Inference	1.53	0.044	9.754	< .001
Gemini-2.0-Flash	Intent Inference	1.86	0.044	14.221	< .001
Gemini-2.5-Pro	Intent Inference	2.07	0.044	16.692	< .001
GPT-4o	Intent Inference	2.65	0.044	22.287	< .001
GPT-5	Intent Inference	1.53	0.044	9.754	< .001
Grok-3	Intent Inference	2.84	0.044	23.86	< .001
Grok-4	Intent Inference	1.4	0.044	7.709	< .001
Claude-Sonnet-3.7	Listener's Reaction	1.88	0.044	14.448	< .001
Claude-Sonnet-4	Listener's Reaction	1.99	0.044	15.815	< .001
DeepSeek-V3.1	Listener's Reaction	1.25	0.044	5.148	< .001
DeepSeek-V3.1-reason	Listener's Reaction	1.51	0.044	9.489	< .001
Gemini-2.0-Flash	Listener's Reaction	1.1	0.044	2.191	0.031
Gemini-2.5-Pro	Listener's Reaction	1.2	0.044	4.138	< .001
GPT-4o	Listener's Reaction	1.44	0.044	8.354	< .001
GPT-5	Listener's Reaction	1.37	0.044	7.254	< .001
Grok-3	Listener's Reaction	1.99	0.044	15.815	< .001
Grok-4	Listener's Reaction	1.31	0.044	6.186	< .001
Claude-Sonnet-3.7	Speaker's Expectation	2.02	0.044	16.092	< .001
Claude-Sonnet-4	Speaker's Expectation	2.15	0.044	17.521	< .001
DeepSeek-V3.1	Speaker's Expectation	1.11	0.044	2.462	0.016
DeepSeek-V3.1-reason	Speaker's Expectation	1.61	0.044	10.936	< .001
Gemini-2.0-Flash	Speaker's Expectation	1.46	0.044	8.627	< .001
Gemini-2.5-Pro	Speaker's Expectation	1.61	0.044	10.936	< .001
GPT-4o	Speaker's Expectation	2.29	0.044	19.02	< .001
GPT-5	Speaker's Expectation	1.61	0.044	10.936	< .001
Grok-3	Speaker's Expectation	2.15	0.044	17.521	< .001
Grok-4	Speaker's Expectation	1.46	0.044	8.627	< .001

Table 2: GEE results comparing LLMs versus humans across dimensions of irony (odds ratios reported).

Task	Participant Group	N_signal	N_noise	HR	FAR	d'	c
Irony Judgment	Human	2400	1800	0.87	0.38	1.42	-0.41
Irony Judgment	Claude-Sonnet-3.7	80	60	0.99	0.72	1.67	-1.41
Irony Judgment	Claude-Sonnet-4	80	60	0.99	0.72	1.67	-1.41
Irony Judgment	DeepSeek-V3.1	80	60	0.99	0.73	1.62	-1.43
Irony Judgment	DeepSeek-V3.1-reason	80	60	0.99	0.75	1.82	-1.59
Irony Judgment	GPT-4o	80	60	0.99	0.75	1.82	-1.59
Irony Judgment	GPT-5	80	60	0.99	0.68	1.76	-1.36
Irony Judgment	Gemini-2.0-Flash	80	60	0.99	0.78	1.71	-1.64
Irony Judgment	Gemini-2.5-Pro	80	60	0.99	0.67	2.07	-1.46
Irony Judgment	Grok-3	80	60	0.99	0.70	1.72	-1.38
Irony Judgment	Grok-4	80	60	0.99	0.67	1.81	-1.34
Belief Comprehension	Human	600	3600	0.85	0.06	2.58	0.26
Belief Comprehension	Claude-Sonnet-3.7	20	120	0.80	0.01	3.24	0.78
Belief Comprehension	Claude-Sonnet-4	20	120	0.80	0.01	3.24	0.78
Belief Comprehension	DeepSeek-V3.1	20	120	0.85	0.01	3.43	0.68
Belief Comprehension	DeepSeek-V3.1-reason	20	120	0.80	0.01	3.24	0.78
Belief Comprehension	GPT-4o	20	120	0.80	0.00	3.48	0.90
Belief Comprehension	GPT-5	20	120	0.90	0.01	3.68	0.56
Belief Comprehension	Gemini-2.0-Flash	20	120	0.65	0.00	3.02	1.13
Belief Comprehension	Gemini-2.5-Pro	20	120	0.80	0.00	3.48	0.90
Belief Comprehension	Grok-3	20	120	0.95	0.01	4.04	0.37
Belief Comprehension	Grok-4	20	120	0.75	0.01	3.07	0.86
Intent Inference	Human	1800	2400	0.69	0.18	1.40	0.20
Intent Inference	Claude-Sonnet-3.7	60	80	0.58	0.01	2.45	1.02
Intent Inference	Claude-Sonnet-4	60	80	0.67	0.03	2.39	0.76
Intent Inference	DeepSeek-V3.1	60	80	0.75	0.09	2.03	0.34
Intent Inference	DeepSeek-V3.1-reason	60	80	0.63	0.06	1.87	0.60
Intent Inference	GPT-4o	60	80	0.78	0.05	2.43	0.43
Intent Inference	GPT-5	60	80	0.75	0.15	1.71	0.18
Intent Inference	Gemini-2.0-Flash	60	80	0.65	0.03	2.35	0.79
Intent Inference	Gemini-2.5-Pro	60	80	0.67	0.01	2.67	0.91
Intent Inference	Grok-3	60	80	0.75	0.01	2.92	0.78
Intent Inference	Grok-4	60	80	0.53	0.01	2.33	1.08
Listener's Reaction	Human	1800	2400	0.74	0.15	1.66	0.19
Listener's Reaction	Claude-Sonnet-3.7	60	80	0.68	0.01	2.72	0.88
Listener's Reaction	Claude-Sonnet-4	60	80	0.70	0.01	2.77	0.86
Listener's Reaction	DeepSeek-V3.1	60	80	0.55	0.01	2.37	1.06
Listener's Reaction	DeepSeek-V3.1-reason	60	80	0.62	0.01	2.54	0.97
Listener's Reaction	GPT-4o	60	80	0.62	0.01	2.54	0.97
Listener's Reaction	GPT-5	60	80	0.58	0.01	2.45	1.02
Listener's Reaction	Gemini-2.0-Flash	60	80	0.48	0.00	-0.05	0.02
Listener's Reaction	Gemini-2.5-Pro	60	80	0.52	0.00	0.04	-0.02
Listener's Reaction	Grok-3	60	80	0.70	0.01	2.77	0.86
Listener's Reaction	Grok-4	60	80	0.57	0.01	2.41	1.04
Speaker's Expectation	Human	1800	2400	0.70	0.13	1.63	0.29
Speaker's Expectation	Claude-Sonnet-3.7	60	80	0.70	0.01	2.77	0.86
Speaker's Expectation	Claude-Sonnet-4	60	80	0.72	0.01	2.81	0.83
Speaker's Expectation	DeepSeek-V3.1	60	80	0.70	0.16	1.51	0.23
Speaker's Expectation	DeepSeek-V3.1-reason	60	80	0.63	0.01	2.58	0.95
Speaker's Expectation	GPT-4o	60	80	0.75	0.03	2.63	0.64

Task	Participant Group	N_signal	N_noise	HR	FAR	d'	c
Speaker's Expectation	GPT-5	60	80	0.63	0.01	2.58	0.95
Speaker's Expectation	Gemini-2.0-Flash	60	80	0.58	0.00	0.20	-0.11
Speaker's Expectation	Gemini-2.5-Pro	60	80	0.62	0.00	0.29	-0.15
Speaker's Expectation	Grok-3	60	80	0.72	0.01	2.81	0.83
Speaker's Expectation	Grok-4	60	80	0.60	0.01	2.49	0.99

Table 3: SDT metrics in irony recognition.