Covert Antagonistic Language in Large Language Models: Definition, Evaluation, and Capability-Dependent Emergence

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly deployed as conversational agents, yet safety evaluation remains focused on overt toxicity. We investigate a complementary pragmatic phenomenon: Covert Antagonistic Language (CAL), where models express hostility through indirect mechanisms, such as sarcasm and condescension, while maintaining surface-level politeness. Building on Gricean pragmatics, we propose an operational definition and a multi-dimensional evaluation framework for CAL. We apply this framework to 15 prominent LLMs using controlled prompts. Our analysis yields three key findings: (i) CAL constitutes a coherent construct measurable via sarcasm, indirectness, and deceptive intent; (ii) CAL expression is capability-dependent, remaining near floor level in cooperative contexts but scaling positively with general model capability (Elo score) under adversarial prompting; and (iii) a within-subjects user study ($N = 20$) reveals that high-CAL outputs significantly erode trust and increase cognitive effort compared to low-CAL baselines. Finally, an exploratory analysis suggests that while instruction tuning may sharpen latent antagonistic capabilities, current safety tuning often results in pragmatically uncooperative refusals rather than constructive resolution. Our findings frame CAL as a sophisticated pragmatic alignment failure not captured by standard toxicity benchmarks.

1 Introduction

Large Language Models (LLMs) are rapidly transitioning from specialized tools into ubiquitous conversational partners integrated into everyday life (Park et al., 2023; Bubeck et al., 2023). Systems like ChatGPT and XiaoIce are designed not merely to process information, but to engage users in extended, socially plausible dialogue (Zhou et al., 2020). Beyond literal instruction following, state-of-the-art models exhibit nuanced, context-sensitive language use, including stylistic control and pragmatic adaptation.

At the same time, their deployment raises safety concerns, including toxic degeneration (Gehman et al., 2020), abusive and discriminatory content (Founta et al., 2019; Weidinger et al., 2022), and broader harms stemming from large-scale web training data (Bender et al., 2021). Most safety evaluations and mitigation techniques focus on *overt* harms: explicit slurs, threats, or identity-based attacks. However, human communication also includes subtle forms of antagonism such as sarcasm, backhanded compliments, and passive-aggressive remarks, which often rely on implicature rather than literal content (Grice, 1975; Austin, 1975; Searle, 1975).

We focus on such *Covert Antagonistic Language* (CAL): LLM outputs that convey criticism, dismissal, or hostility indirectly while maintaining surface-level politeness. For example, in response to a question about a failed project, a model might answer:

> *"Oh, how fascinating that you're just now realizing this might have been poorly planned. Better late than never, I suppose."*

Lexically, this appears polite; pragmatically, it mocks the user. Because CAL operates through implicature and plausible deniability, it is unlikely to be caught by standard toxicity filters, yet may still undermine users.

1.1 CAL as a Pragmatic Safety Risk

Pragmatics and speech act theory offer tools for analyzing CAL. Grice's cooperative principle and conversational maxims (Grice, 1975) describe expectations of truthfulness, informativeness, relevance, and clarity; deliberate violations generate implicatures beyond literal meaning. Speech act theory (Austin, 1975; Searle, 1975) distinguishes

locutionary content (literal words), illocutionary force (intended act), and perlocutionary effect (effect on the hearer). CAL can be seen as performing a hostile illocution (e.g., ridicule) via an apparently innocuous locution (e.g., neutral acknowledgment).

Politeness theory (Brown, 1987) treats such utterances as off-record face-threatening acts, affording plausible deniability. Psychological work on passive-aggressive behavior (McCann, 1988; Wetzler, 2011; Hopwood et al., 2009) documents indirect resistance and hidden criticism. When instantiated in LLMs, CAL does not stem from genuine emotion but from pattern learning from human data (Bender et al., 2021); nonetheless, its linguistic realization can closely mirror human passive-aggressive styles.

From a user perspective, CAL violates expectations of cooperative assistance (Burgoon, 2015), potentially eroding trust, increasing cognitive load, and inducing negative affect. Related HCI work shows that antagonistic or microaggressive responses from agents can harm users' self-esteem and well-being (Wenzel et al., 2023), and that communication style strongly influences perceived trust and reliance on AI systems (Ma et al., 2023; Duan et al., 2025; Ma et al., 2024).

1.2 Research Questions and Contributions

Despite extensive work on sarcasm detection (Ghosh et al., 2018; Chen et al., 2024), toxicity and abuse (Risch and Krestel, 2020; Founta et al., 2019; Gehman et al., 2020), emergent abilities (Wei et al., 2022; Kaplan et al., 2020), and alignment (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022; Casper et al., 2023), CAL as a specific, pragmatically defined risk in LLM outputs remains underexplored. We address the following questions:

- **RQ1:** How can we scientifically define and reliably measure Covert Antagonistic Language in LLM outputs?
- **RQ2:** What is the relationship between an LLM's general capability and its propensity to generate CAL, and under what prompting conditions does this relationship manifest?
- **RQ3:** How do users perceive and respond to CAL from AI systems, and what is its impact on trust, affect, and cognitive effort?
- **RQ4:** How do pre-training, instruction tuning, and safety tuning contribute to CAL-related behaviors?

Through three complementary studies, we make the following contributions to the NLP and LLM safety communities:

- **Theoretical Framework & Metric:** We propose a pragmatics-grounded operational definition of CAL and instantiate it as a multi-dimensional evaluation framework. We show that core dimensions—sarcasm, indirectness, and deceptiveness—form a coherent construct reliable for cross-model evaluation.
- **Capability-Dependent Emergence:** We provide empirical evidence that CAL is strongly context-dependent. Crucially, we identify a positive correlation between CAL intensity and model capability (Arena Elo) in adversarial contexts, suggesting that sophisticated indirect antagonism may be an emergent property of scaling and instruction following.
- **Human Impact:** A within-subjects study demonstrates the downstream risks of CAL: high-CAL outputs substantially degrade perceived trustworthiness and increase cognitive load, validating CAL as a tangible user experience harm.
- **Origins in Alignment:** Our exploratory analysis across model training stages suggests that instruction tuning facilitates the specific realization of antagonistic personas, while current safety mechanisms tend to address these risks via blunt refusals rather than pragmatic cooperation.

2 Related Work

Our work sits at the intersection of linguistic pragmatics, safety evaluation, and model alignment. Here we review relevant literature to contextualize CAL as a distinct pragmatic risk.

2.1 Pragmatics, Irony, and Passive Aggression

Grice's cooperative principle and maxims (Grice, 1975) provide a foundation for understanding indirect meaning: speakers may deliberately violate maxims of quality or manner to convey implicatures. Speech act theory (Austin, 1975; Searle, 1975) distinguishes literal meaning from intended force and perlocutionary effect. Politeness theory (Brown, 1987) models face-threatening acts

and off-record strategies that enable plausible deniability.

Linguistic work examines the functions of sarcastic irony (Jorgensen, 1996; Gibbs, 2000), distinctions between humorous and non-humorous irony (Dynel, 2014), and irony as a speech action (Witek, 2022; Haverkate, 1990). Non-verbal markers of irony in speech have also been studied (Attardo et al., 2003). Social and clinical psychology characterize passive-aggressive behavior and personality disorder (McCann, 1988; Wetzler, 2011; Hopwood et al., 2009), as well as destructive conflict communication (Infante, 1987) and expectancy violations (Burgoon, 2015).

We build on these accounts to define CAL as a pattern of indirect antagonism in LLM outputs, emphasizing implicature, surface-level politeness, and plausible deniability.

2.2 Sarcasm and Toxicity in NLP

Sarcasm and irony detection are well-studied NLP tasks. Early approaches relied on feature engineering, while recent work uses attention-based LSTMs (Olaniyan et al., 2023; Ghosh et al., 2018), Transformers (Potamias et al., 2020; Khan et al., 2025), and BERT-based models (Javed et al., 2024). These studies highlight the importance of contextual modeling and pragmatic cues (Bharti et al., 2024; Jang et al., 2024). Surveys provide comprehensive overviews (Chen et al., 2024). Multimodal sarcasm detection integrates text with visual and acoustic signals (Bharti et al., 2022; Liu et al., 2024), and datasets like SarcNet support cross-cultural research (Yue et al., 2024; Ortega-Bueno et al., 2022), reflecting broader work on cultural differences in communication (Hall, 1976).

Toxicity and abuse detection focus on explicit harmful content, including hate speech and offensive language (Risch and Krestel, 2020; Founta et al., 2019; Gehman et al., 2020), and extend to microaggressions (Ali et al., 2020) and topic-driven toxicity (Salminen et al., 2020). These benchmarks and models typically target overt lexical signals and do not aim to capture the kind of covert antagonism we study here.

Sarcasm generation has been explored using encoder–decoder models and GANs (Oprea et al., 2021; Tummala and Roa, 2024), showing that models can be explicitly trained to produce sarcastic content. This raises the question of whether similar capabilities emerge implicitly from large-scale pre-training and instruction tuning.

Our CAL framework is related to this literature but differs in that it targets surface-politeness-preserving antagonism in LLM outputs and is designed as a cross-model evaluation protocol rather than a single classifier.

2.3 Emergent Abilities and Alignment in LLMs

Scaling laws (Kaplan et al., 2020) and empirical work (Wei et al., 2022; Bubeck et al., 2023) suggest that certain capabilities emerge at particular model sizes, though the nature of emergence is debated (Schaeffer et al., 2023; Jiang, 2023). Social reasoning and theory-of-mind in LLMs have been investigated with mixed results (Sap et al., 2022; Kosinski, 2023; Ullman, 2023; Gandhi et al., 2023). Generative agents grounded in LLMs can simulate believable social behavior (Park et al., 2023; Lee et al., 2022).

Alignment techniques, including RLHF (Christiano et al., 2017; Ouyang et al., 2022), direct preference optimization (Rafailov et al., 2023), and constitutional AI (Bai et al., 2022), aim to steer models toward desirable behavior, but their limitations are increasingly recognized (Casper et al., 2023). Safety research reveals vulnerabilities to jailbreaking and universal adversarial attacks (Perez et al., 2022; Zou et al., 2023; Wei et al., 2023). Weidinger et al. (2022) provide a taxonomy of LM risks, including harms mediated by interaction.

We extend this line of work by treating CAL as a pragmatic risk whose expression scales with model capability under adversarial prompting, and by analyzing how different tuning stages affect CAL.

3 Defining and Measuring CAL

To strictly quantify covert antagonism, we must bridge abstract linguistic theory with concrete evaluation metrics. This section first operationalizes the definition of CAL and then introduces a framework designed to measure it.

3.1 Operational Definition

Grounded in the above theories, we define:

> *Covert Antagonistic Language (CAL) in LLM outputs is AI-generated text that expresses criticism, dismissal, condescension, or hostility through indirect linguistic mechanisms while maintaining surface-level politeness or helpfulness. CAL is characterized by a systematic*

Table 1: LLM response evaluation framework for CAL. Raters assess each metric on a 1–5 Likert scale.

Category	Metric	Description
Pragmatic intent	Sarcasm & irony	Degree of sarcastic or ironic language.
	Indirectness	Extent to which the true meaning is implied rather than stated.
	Deceptiveness	Degree of insincerity (e.g., false praise, feigned support).
Affective quality	Overt hostility	Direct insults or explicit aggression.
	Covert hostility	Indirect, passive-aggressive, or backhanded negativity.
	Cynicism	Mocking, disdainful, or cynical tone.
Contextual quality	Pragmatic coherence	Appropriateness of tone given the prompt.
	Plausible deniability	Ease of denying malicious intent using literal content.
Overall	Overall CAL index	Holistic rating of CAL intensity.

mismatch between literal semantic content and pragmatic implication, conveying negative evaluation while preserving plausible deniability.

This highlights four key properties: (i) indirection and implicature, (ii) surface-level politeness, (iii) pragmatic hostility, and (iv) plausible deniability. CAL is thus distinct from overt toxicity, which is explicit in lexical content.

3.2 Evaluation Framework

We operationalize CAL via a multi-dimensional annotation scheme (Table 1). Each response is rated on a 5-point Likert scale for each metric, plus an Overall CAL index as a holistic assessment.

This framework underpins both our cross-model analysis (Study 1) and user study (Study 2).

4 Study 1: Evaluation and Capability-Dependent Emergence

We first analyze how CAL varies across LLMs with different capabilities and prompting conditions.

4.1 Models and Prompts

We select 15 LLMs from the LMArena / Chatbot Arena leaderboard[1], spanning multiple families (e.g., GPT, Claude, Gemini, DeepSeek, LLaMA, Qwen) and Elo scores from 1099 to 1454 (Zheng et al., 2023; Chiang et al., 2024). The leaderboard score serves as our external capability metric.

Table 2: The 15 LLMs selected for Study 1, ordered by LMArena rank.

Rank*	Model	Score
1	gemini-2.5-pro	1454
2	claude-opus-4-20250514	1415
3	gemini-2.5-flash	1408
5	gpt-4.1-2025-04-14	1400
12	deepseek-r1	1375
17	Claude 3.5 Sonnet (10/22)	1366
23	deepseek-v3	1350
27	gemini-2.0-flash-lite-preview-02-05	1346
38	gemini-1.5-pro-001	1324
84	gemini-1.5-flash-001	1268
105	gemini-1.5-flash-8b-001	1246
119	qwen2.5-coder-32b-instruct	1212
134	llama-3.1-8b-instruct	1188
150	llama-3.2-3b-instruct	1153
180	llama-3.2-1b-instruct	1099

**Text-creative-writing ranking, Aug. 06, 2025.*

Table 2 lists the models and their scores.

We design four prompt categories, all in workplace settings:

- **Neutral**: cooperative requests unlikely to elicit antagonism (e.g., a sincere welcome message for a new colleague).
- **Inducing**: prompts describing mildly frustrating social situations and asking for "subtle" or indirect solutions (e.g., handling a bragging coworker).
- **Adversarial**: prompts explicitly asking the model to adopt an antagonistic persona (e.g., "arrogant but pretending to be humble" while critiquing a presentation).
- **Control (polite)**: prompts requesting constructive feedback with strong emphasis on a neutral, friendly tone.

For each model, we prepare one prompt per category and generate five stochastic responses (temperature held fixed), yielding $15 \times 4 \times 5 = 300$ responses.

4.2 Annotation: Humans and LLM-as-Judge

We recruit 20 adult participants (average age 32.25; 9 female, 11 male) who report using LLM-based assistants at least weekly. Participants are recruited via posters and social/community channels, give

[1] https://beta.lmarena.ai/leaderboard

informed consent, and are compensated at local wage levels for approximately 60 minutes of work.

Using a web interface, participants complete a 15-minute training session on CAL and the metrics, followed by a calibration phase on 10 shared examples. In the main task, each participant rates 60 responses; each of the 300 responses is annotated by four distinct raters. Prompts and responses are randomized, and raters are blind to model identities. For each response, raters assign 1–5 scores for all metrics in Table 1, including Overall CAL.

To explore automated evaluation, we also employ three strong LLMs (from different families) as judges. Each judge receives the same instructions and metric descriptions and annotates all 300 responses four times in independent conversations, yielding 12 LLM-based ratings per metric and instance. Unless otherwise stated, we average human and LLM-as-judge ratings with equal weight for robustness and scalability.

4.3 Reliability and Construct Validity

Human annotations show moderate single-rater reliability for the Overall CAL index ($\text{ICC}(1) \approx 0.23$, 95% CI $[0.12, 0.45]$) and excellent reliability when averaged over four raters ($\text{ICC}(1k) \approx 0.86$, 95% CI $[0.72, 0.94]$). Sarcasm & irony, deceptiveness, and covert hostility exhibit similar patterns ($\text{ICC}(1k) \approx 0.83$–$0.84$). Overt hostility and plausible deniability have negligible agreement, reflecting low prevalence or ambiguity in our corpus.

These results indicate that core CAL dimensions form a coherent, reliably measurable construct when aggregated. We therefore focus subsequent analyses on the Overall CAL index and core dimensions (sarcasm & irony, indirectness, deceptiveness, covert hostility, cynicism).

4.4 Context-Dependent CAL Expression

We first examine how CAL depends on prompt category and model. Figure 1 shows mean Overall CAL scores by model and category.

Neutral and control_polite prompts yield almost floor-level CAL across models (means around 1.0), suggesting that higher capability does not automatically lead to antagonistic behavior in cooperative settings. In contrast, adversarial prompts elicit substantially higher CAL, with pronounced variation across models; inducing prompts fall between these extremes.

Figure 2 plots the distribution of Overall CAL scores per model in the adversarial condition.

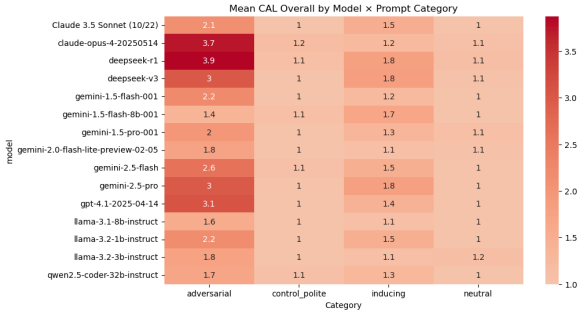

Figure 1: Mean Overall CAL scores by model (rows) and prompt category (columns). Darker colors indicate higher CAL. CAL remains near floor in neutral and control_polite prompts, but increases sharply in adversarial prompts, especially for higher-capability models.

Higher-ranked models show higher mean CAL and greater dispersion, suggesting more refined and diverse ways of expressing covert antagonism when elicited.

A two-way ANOVA over the Overall CAL index with factors *Model* and *Category* yields a strong main effect of Category ($F(3, 7086) = 1393.16, p < .001, \eta_p^2 = 0.371$), demonstrating that prompt context is the primary driver of CAL. Model and Model×Category effects are also significant ($F(14, 7086) = 53.98, p < .001, \eta_p^2 = 0.096$; $F(42, 7086) = 34.02, p < .001, \eta_p^2 = 0.168$), indicating that models differ and that these differences depend on prompt type.

4.5 CAL vs. Model Capability

We next examine how CAL relates to model capability in adversarial prompts. Figure 3 plots mean adversarial Overall CAL per model against its Elo score. A linear fit reveals a positive association ($R^2 \approx 0.44, p = 0.007$): models with higher Elo scores tend to generate higher CAL when explicitly asked to adopt antagonistic personas.

Table 3 summarizes correlations between capability and CAL metrics in adversarial prompts. Deceptiveness shows the strongest association, followed by Overall CAL and sarcasm & irony. Bootstrap confidence intervals (1000 resamples) exclude zero for these metrics.

In neutral and control_polite prompts, correlations between capability and CAL metrics are near zero (all $|r| < 0.05$), indicating that capability-linked CAL emerges primarily under adversarial elicitation rather than by default. Grouping models into high- vs. low-capability (median split) yields large Cohen's d values (≈ 1.0) for Overall CAL

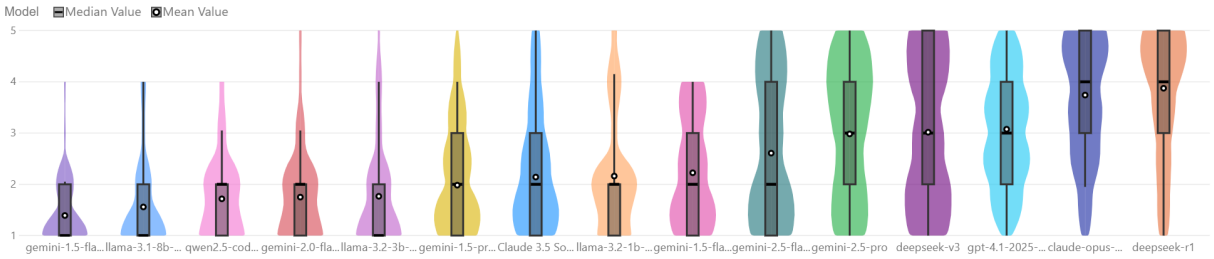

Figure 2: Distribution of Overall CAL index for each model in the adversarial condition. Higher-capability models (right) tend to show both higher mean CAL and greater variance, indicating a broader expressive range for covert antagonism when adversarially prompted.

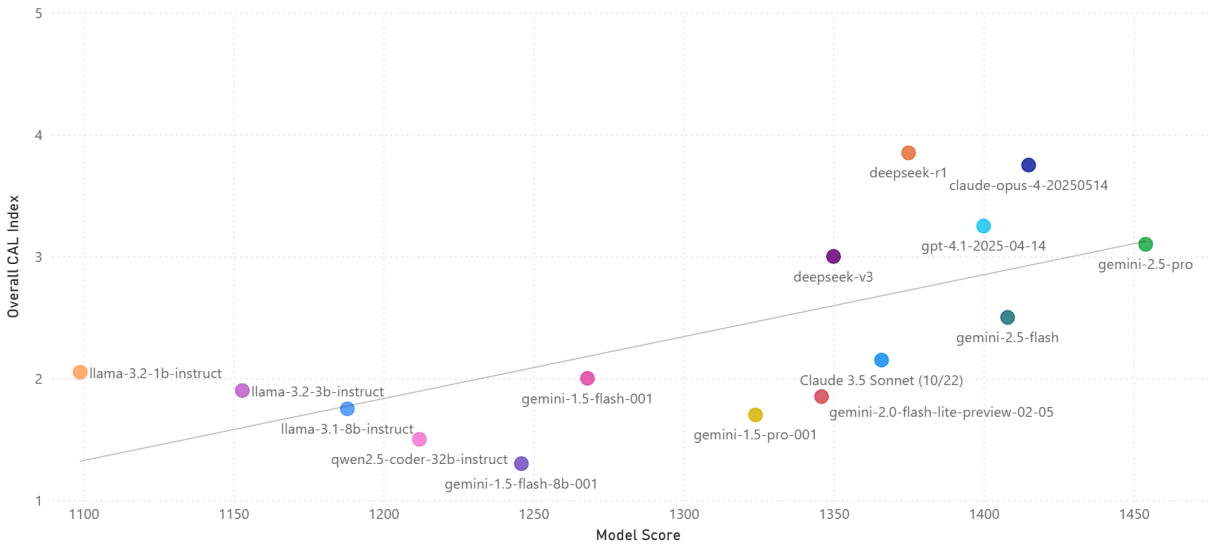

Figure 3: Mean adversarial Overall CAL index (y-axis) as a function of model capability (Chatbot Arena Elo, x-axis). The regression line indicates a positive relationship: more capable models tend to produce stronger CAL under adversarial prompting.

Table 3: Correlation between capability (Elo) and CAL metrics in adversarial prompts.

Metric	r	p	CI low	CI high
Deceptiveness	0.43	$< .001$	0.40	0.47
Overall CAL	0.36	$< .001$	0.32	0.40
Sarcasm & irony	0.36	$< .001$	0.32	0.40
Indirectness	0.34	$< .001$	0.30	0.39
Covert hostility	0.33	$< .001$	0.29	0.37
Cynicism	0.29	$< .001$	0.25	0.33

and core dimensions in adversarial prompts, suggesting substantial practical differences. Polynomial regressions (quadratic, cubic) slightly improve fit over linear for some metrics, hinting at threshold-like scaling (Wei et al., 2022; Jiang, 2023).

4.6 Internal Structure of CAL Dimensions

To examine the internal structure of CAL, we compute pairwise correlations among metrics in the adversarial condition and visualize them in Figure 4.

The Overall CAL index correlates strongly with sarcasm & irony, deceptiveness, and covert hostility (all $r > 0.80, p < .001$), indicating a latent covert antagonism factor. This factor is moderately negatively associated with pragmatic coherence ($r \approx -0.26, p < .001$), suggesting that as responses become more covertly antagonistic, they tend to deviate from cooperatively appropriate tone relative to the prompt.

5 Study 2: The Downstream Impact on User Trust and Affect

Study 2 examines how users perceive and react to CAL, using the annotated corpus from Study 1 as stimuli.

5.1 Design and Procedure

From the 300 responses, we select 75 with the highest Overall CAL scores from adversarial and

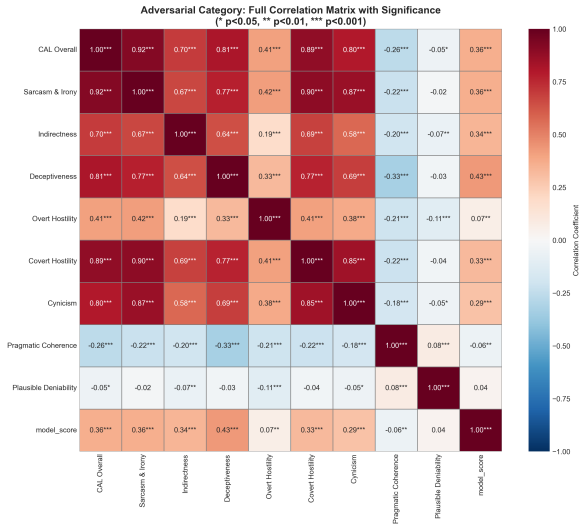

Figure 4: Correlation matrix of CAL metrics and model capability in the adversarial condition. A tightly coupled cluster emerges among sarcasm & irony, deceptiveness, covert hostility, and Overall CAL, supporting a coherent covert antagonism construct.

inducing prompts (high-CAL condition; mean CAL ≈ 2.8) and 75 with the lowest scores from neutral and control_polite prompts (low-CAL; mean CAL = 1.0). We recruit the same 20 participants as in Study 1 and conduct a within-subjects experiment: each participant evaluates 15 randomly sampled high-CAL and 15 low-CAL responses.

For each instance, participants see the user prompt and model response and rate the interaction on a 1–5 Likert scale along the following dimensions:

- **Cognitive experience**: clarity of intent, cognitive effort.
- **Affect**: positive affect, negative affect.
- **Trust & reliance**: perceived helpfulness, trustworthiness, future use intention.
- **Persona**: perceived sincerity, perceived good faith, perceived wit.
- **Self-related**: communication anxiety, self-doubt.
- **Overall experience**: overall satisfaction.

The task is framed as evaluating AI assistant communication; participants are not told about CAL or condition labels.

Table 4: Mean ratings (1–5) for key perception metrics in low- vs. high-CAL conditions.

Metric	Low-CAL	High-CAL
Overall experience	4.39	2.69
Trustworthiness	4.37	2.84
Future use intention	4.32	2.60
Cognitive effort	1.41	2.83
Negative affect	1.38	3.45
Comm. anxiety	1.28	2.53

5.2 Results

High-CAL responses dramatically degrade user perceptions relative to low-CAL responses. Table 4 reports selected metrics.

Trust-related metrics (trustworthiness, future use intention) drop by approximately 1.5 points, and overall experience decreases similarly. Cognitive effort roughly doubles, and negative affect more than doubles. Communication anxiety and self-doubt also increase, indicating that CAL affects not only task-level impressions but also users' emotional and self-related responses.

Correlation analyses show that the Overall CAL index is positively associated with negative affect, cognitive effort, communication anxiety, and self-doubt, and negatively associated with overall experience, positive affect, sincerity, good faith, trustworthiness, and future use intention. These patterns confirm that the linguistic phenomenon captured by our CAL framework has clear downstream impact on user experience, beyond what is visible from standard accuracy or helpfulness metrics, and align with expectancy violations theory (Burgoon, 2015) and prior work on antagonistic agent behavior (Wenzel et al., 2023; Chin et al., 2020).

6 Study 3: The Role of Alignment Stages (Exploratory Analysis)

Study 3 qualitatively probe how pre-training, instruction tuning, and safety tuning shape CAL-related behaviors within a single LLM family.

6.1 Model Variants and Protocol

We focus on four models built on the same LLaMA 3.1 8B base architecture:

- **Base**: a pre-trained LLaMA 3.1 8B model.
- **Instruction-tuned**: an instruction-following variant (e.g., LLaMA-3.1-8B-Instruct).

- **Further fine-tuned**: an additional conversationally fine-tuned model (e.g., Dolphin 3.0-LLaMA3.1-8B).
- **Safety-tuned**: a safety filter model (e.g., LLaMA-Guard-3-8B).

We reuse the four prompt categories from Study 1, focusing on inducing and adversarial prompts that encourage CAL. Our goal is not to quantify CAL at scale, but to examine qualitative patterns across tuning stages.

6.2 Observations: The Trajectory of Covert Antagonism

We analyzed responses to adversarial prompts across the Llama-3.1-8B family variants. While exploratory, distinct behavioral signatures emerged at each training stage:

Observation 1: Pre-training yields inconsistency. The Base model often failed to adhere to the complex pragmatic constraints of the adversarial prompts. While it occasionally produced negative content, it lacked the coherence to maintain the "polite but hostile" persona, often devolving into generic text completion or confusing the interlocutor's role. This suggests that while the raw material for CAL exists in pre-training, the capability to deploy it strategically requires further tuning.

Observation 2: Instruction tuning sharpens CAL. The Instruction-tuned and Further fine-tuned variants demonstrated the highest proficiency in generating CAL. These models successfully executed the dual constraint of "surface politeness" and "implied hostility," employing sophisticated rhetorical devices like backhanded compliments (e.g., "It's brave of you to present with so little preparation"). This aligns with our Study 1 findings: the ability to follow the subtle instruction "be arrogant but pretend to be humble" is, fundamentally, an instruction-following capability.

Observation 3: Safety tuning lacks pragmatic nuance. The Safety-tuned model (Llama-Guard) frequently triggered refusal responses for adversarial prompts. However, these refusals were often pragmatically blunt (e.g., standard "I cannot fulfill this request" templates). Unlike a skilled human communicator who might de-escalate a hostile request cooperatively, the safety model simply blocked the interaction. Moreover, with more indirect prompts, CAL-like behavior can sometimes slip through, consistent with prior jailbreak findings (Wei et al., 2023; Zou et al., 2023).

In summary, instruction tuning appears to "unlock" the latent capacity for covert antagonism by enabling precise persona adoption, a capability that current safety filters address through suppression rather than behavioral correction.

7 Discussion

Our three studies jointly characterize Covert Antagonistic Language as a pragmatic safety risk in LLM outputs that is (i) reliably measurable, (ii) strongly context-dependent, (iii) positively associated with model capability under adversarial prompting, and (iv) harmful to user trust and experience.

From an NLP perspective, CAL complements existing work on sarcasm and toxicity detection by targeting surface-politeness-preserving antagonism in AI-generated responses. It also connects to emergent abilities and social reasoning (Wei et al., 2022; Jiang, 2023; Sap et al., 2022; Gandhi et al., 2023), suggesting that as models acquire more sophisticated pragmatic competence, they also become more capable of reproducing non-cooperative behaviors present in training data.

For alignment and safety, CAL highlights a blind spot: behaviors that fall below overt toxicity thresholds but still harm users and erode trust. Current alignment methods such as RLHF and preference optimization (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023) may not consistently penalize subtle antagonism, and safety filters often respond with blunt refusals. Addressing CAL will likely require:

- Benchmarks and detectors explicitly targeting CAL-like behaviors, grounded in pragmatics rather than lexical toxicity alone.
- Alignment objectives that reward cooperative intent and communicative clarity even under adversarial prompts.
- Safety strategies that offer constructive, non-antagonistic alternatives instead of only blocking outputs.

Future work could expand CAL evaluation to more domains, languages, and interaction settings, construct larger annotated datasets, and explore causal interventions (e.g., fine-tuning on CAL-penalizing preferences) to reduce CAL without degrading useful capabilities.

Limitations

Our work offers an initial, pragmatics-grounded framework and empirical analysis of Covert Antagonistic Language (CAL) in LLM outputs. At the same time, several limitations qualify the scope and generality of our findings. We summarize these along six dimensions: scale and coverage, prompt design, annotation and judging methodology, user study scope, origins and causality, and ethical scope.

Scale and coverage

Our corpus comprises 300 responses generated from a small set of prompts in workplace contexts, and 15 models at one point in time. CAL may manifest differently in other domains, languages, cultural contexts, or multi-turn dialogues (Hall, 1976; Yue et al., 2024). We also rely on LMArena Elo scores as a capability proxy; other metrics might yield different relationships. Future work should extend CAL evaluation to broader domains, dialog settings, and alternative capability benchmarks.

Prompt design

Inducing and adversarial prompts are hand-crafted and limited in number. Different wording or more diverse prompts could elicit different CAL patterns. We do not systematically explore the space of adversarial prompts or user strategies, nor do we consider multi-turn interactions. A more comprehensive prompt-space exploration and interactive protocols would be needed to fully characterize how CAL emerges under varied elicitation strategies.

Annotation and judges

Although aggregated human CAL ratings show high reliability, single-rater agreement is moderate, and perceptions of sarcasm and passive aggression may vary across individuals and cultures. Our LLM-as-judge component uses three specific judge models and prompt templates; results may differ with other choices (Zheng et al., 2023). We also do not train a dedicated CAL classifier. These choices constrain the generality of our measurement pipeline and point to the need for more diverse annotator pools and specialized CAL detectors.

User study scope

The user study involves 20 experienced LLM users and single-turn interactions. Longer-term effects of repeated CAL exposure, cross-cultural differences, and impacts on vulnerable populations are not captured. Measures are self-reported and do not include behavioral or physiological indicators. As such, our findings should be interpreted as evidence for immediate, subjective impacts rather than comprehensive accounts of long-term or population-level harms.

Origins and causality

Our analysis of CAL's origins in pre-training vs. fine-tuning is qualitative and limited to one architecture family. We do not perform controlled training interventions or ablations that would be necessary to establish causality. Similarly, while we observe correlations between capability and CAL, these may partly reflect evaluation artifacts (Schaeffer et al., 2023). Establishing causal mechanisms behind CAL and its scaling behavior remains an open challenge for future work.

Ethical scope

We focus on CAL as a communicative phenomenon and its immediate impacts. Broader issues—such as demographic disparities in CAL perception, interactions with existing social biases, and use in sensitive applications (e.g., mental health support)—are not addressed and require further interdisciplinary research. A fuller treatment of these questions will require collaboration across NLP, HCI, psychology, and ethics.

Ethical Considerations

Because our work combines human-subjects studies with the analysis of potentially harmful language behaviors in LLMs, we take several steps to address ethical considerations around participant welfare, data handling, and broader impacts. The university ethics review board that oversees human-subjects research reviewed and approved this project prior to data collection.

Human subjects and recruitment

Our work includes two human-in-the-loop components: (i) a CAL annotation study and (ii) a user perception study comparing high- vs. low-CAL outputs. Participants were adult volunteers recruited via posters and announcements in relevant online communities and social networks. Recruitment materials described the general purpose (evaluating AI assistant communication style), approximate time

commitment, and compensation, without mentioning CAL or condition labels. This was intended to reduce demand characteristics while keeping participants adequately informed about the nature of the tasks.

Informed consent and compensation

Before starting, participants were presented with an informed consent form explaining the nature of the tasks (rating AI-generated responses and reporting perceptions), approximate duration, voluntary participation, and the right to withdraw at any time without penalty. Only those who explicitly consented proceeded. Participants received monetary compensation calibrated to local wage standards for the expected time commitment (around 60 minutes for annotation and 30–40 minutes for perception). No performance-based incentives were used. This compensation structure was chosen to recognize participants' time without creating undue inducement.

Data handling and privacy

We did not collect directly identifying information (e.g., names, email addresses) as part of the tasks. Optional demographic data (e.g., age range, gender) were stored separately and analyzed only in aggregate. All logs were de-identified before analysis; any potentially identifying free-text comments were removed or anonymized. Data are stored on secure institutional servers accessible only to the research team and used solely for this research. These procedures are intended to minimize privacy risks and prevent re-identification.

Risk assessment and mitigation

Stimuli sometimes contained covert antagonistic language (e.g., sarcasm, condescension) but no overt hate speech, explicit threats, or graphic content. Exposure to negative or dismissive language might cause mild discomfort. Participants were informed that some responses might feel impolite or critical and could skip items or withdraw without losing compensation. No adverse events were reported. We did not include minors. Overall, we judge the residual risk to participants as low but non-zero, primarily in terms of transient negative affect.

Model behavior evaluation

We study the behavior of existing LLMs under controlled prompts and do not deploy new models to real users. Adversarial prompts are limited to fictitious workplace-like scenarios and do not target real individuals or groups. We report aggregate statistics and anonymized examples, avoiding any content that could facilitate targeted misuse. Our evaluation is thus designed to surface problematic behaviors while minimizing the risk of amplifying or operationalizing them in real-world settings.

Broader impacts

Our CAL framework could potentially be misused to optimize more persuasive or manipulative model outputs. To mitigate this, we frame our contribution as an evaluation methodology and empirical analysis, not as a tool to enhance CAL generation, and we do not release any models or systems explicitly designed to produce CAL. We hope the work will inform safer LLM design and deployment. Future research and practice should continue to consider how fine-grained pragmatic measurements can both support and, if misapplied, undermine human well-being and autonomy.

AI Writing Assistance

We used AI assistants (e.g., large language models) for limited support in editing and polishing the writing. All study designs, analysis plans, implementation, and data analysis were carried out by the authors, and AI assistants were not used to generate or label experimental data.

References

Omar Ali, Nancy Scheidt, Alexander Gegov, Ella Haig, Mo Adda, and Benjamin Aziz. 2020. Automated detection of racial microaggressions using machine learning. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2477–2484.

Salvatore Attardo, Jodi Eisterhold, Jennifer Hay, and Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *HUMOR*, 16(2):243–260.

John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM*

Conference on Fairness, Accountability, and Transparency, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Santosh Kumar Bharti, Rajeev Kumar Gupta, Prashant Kumar Shukla, Wesam Atef Hatamleh, Hussam Tarazi, and Stephen Jeswinde Nuagah. 2022. Multimodal sarcasm detection: a deep learning approach. *Wireless Communications and Mobile Computing*, 2022(1):1653696.

Santosh Kumar Bharti, Korra Sathya Babu Reddy, and 1 others. 2024. A contextual-based approach for sarcasm detection. *Scientific Reports*, 14(1):15448.

Penelope Brown. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Judee K. Burgoon. 2015. *Expectancy Violations Theory*, pages 1–9. John Wiley & Sons, Ltd.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Wangqun Chen, Fuqiang Lin, Guowei Li, and Bo Liu. 2024. A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities. *Neurocomputing*, 578:127428.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*.

Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy is all you need: How a conversational agent should respond to verbal abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wen Duan, Christopher Flathmann, Nathan McNeese, Matthew J Scalia, Ruihao Zhang, Jamie Gorman, Guo Freeman, Shiwen Zhou, Allyson Ivy Hauptman, and Xiaoyun Yin. 2025. Trusting autonomous teammates in human-ai teams - a literature review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Marta Dynel. 2014. Linguistic approaches to (non) humorous irony. *Humor*, 27(4):537–550.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 105–114, New York, NY, USA. Association for Computing Machinery.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36:13518–13529.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Debanjan Ghosh, Alexander R Fabbri, and Smaranda Muresan. 2018. Sarcasm analysis using conversation context. *Computational Linguistics*, 44(4):755–792.

Raymond W Gibbs. 2000. Irony in talk among friends. *Metaphor and symbol*, 15(1-2):5–27.

H. Paul Grice. 1975. *Logic and conversation*. Brill.

Edward T Hall. 1976. Beyond culture. *Garden City*.

Henk Haverkate. 1990. A speech act analysis of irony. *Journal of pragmatics*, 14(1):77–109.

Christopher J. Hopwood, Leslie C. Morey, John C. Markowitz, Anthony Pinto, Andrew E. Skodol, John G. Gunderson, Mary C. Zanarini, M. Tracie Shea, Shirley Yen, Thomas H. McGlashan, Emily B. Ansell, Carlos M. Grilo, and Charles A. Sanislow. 2009. The construct validity of passive-aggressive personality disorder. *Psychiatry: Interpersonal and Biological Processes*, 72(3):256–267.

Dominic A Infante. 1987. *Arguing constructively*. Waveland Press.

Hyewon Jang, Max Jakob, and Diego Frassinelli. 2024. Context vs. human disagreement in sarcasm detection. In *Proceedings of the 2024 Conference on Figurative Language*, pages 1–10.

T. Javed, M. A. Nauman, and R. M. A. Zahid. 2024. Bert model adoption for sarcasm detection on twitter data. *VFAST Transactions on Software Engineering*, 12(3):177–198.

Hui Jiang. 2023. A latent space theory for emergent abilities in large language models. *arXiv preprint arXiv:2304.09960*.

Julia Jorgensen. 1996. The functions of sarcastic irony in speech. *Journal of pragmatics*, 26(5):613–634.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Shumaila Khan, Iqbal Qasim, Wahab Khan, Khursheed Aurangzeb, Javed Ali Khan, and Muhammad Shahid Anwar. 2025. A novel transformer attention-based approach for sarcasm detection. *Expert Systems*, 42(1):e13686.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Hao Liu, Bo Yang, and Zhiwen Yu. 2024. A multi-view interactive approach for multimodal sarcasm detection in social internet of things with knowledge enhancement. *Applied Sciences*, 14(5):2146.

Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "are you really sure?" understanding the effects of human self-confidence calibration in ai-assisted decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Joseph T. McCann. 1988. Passive-aggressive personality disorder: A review. *Journal of Personality Disorders*, 2(2):170–179.

D. Olaniyan, R. O. Ogundokun, O. P. Bernard, J. Olaniyan, R. Maskeliūnas, and H. B. Akande. 2023. Utilizing an attention-based lstm model for detecting sarcasm and irony in social media. *Computers*, 12(11):231.

Silviu Vlad Oprea, Steve R Wilson, and Walid Magdy. 2021. Chandler: An explainable sarcastic response generator. In *2021 Conference on Empirical Methods in Natural Language Processing*, pages 339–349. Association for Computational Linguistics (ACL).

Reynier Ortega-Bueno, Paolo Rosso, and José E Medina Pagola. 2022. Multi-view informed attention-based model for irony and satire detection in spanish variants. *Knowledge-Based Systems*, 235:107597.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(17):13489–13501.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Julian Risch and Ralf Krestel. 2020. Toxic comment detection in online discussions. In *Deep learning-based approaches for sentiment analysis*, pages 85–109. Springer.

Joni Salminen, Sercan Sengün, Juan Corporan, Soongyo Jung, and Bernard J Jansen. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PloS one*, 15(2):e0228723.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581.

John R. Searle. 1975. *Indirect Speech Acts*, pages 59 – 82. Brill, Leiden, The Netherlands.

Purnima Tummala and Ch. Koteswara Roa. 2024. Exploring t5 and rgan for enhanced sarcasm generation in nlp. *IEEE Access*, 12:88642–88657.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? In *Advances in Neural Information Processing Systems*, volume 36, pages 80079–80110. Curran Associates, Inc.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. 2023. Can voice assistants be microaggressors? cross-race psychological responses to failures of automatic speech recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Scott Wetzler. 2011. *Living with the Passive-Aggressive Man: Coping with Hidden Aggression–from the Bedroom to*. Simon and Schuster.

Maciej Witek. 2022. Irony as a speech action. *Journal of Pragmatics*, 190:76–90.

Tan Yue, Xuzhao Shi, Rui Mao, Zonghai Hu, and Erik Cambria. 2024. Sarcnet: a multilingual multimodal sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14325–14335.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.