Scaling Implicature via Structured Interaction in Multi-Agent LLMs

Harvey Bonmu Ku[1,2]

[1]Ministry of National Defense, Republic of Korea
[2]Harvard University
language@langua.ge

Abstract

Pragmatic inference remains a persistent challenge even for state-of-the-art large language models (LLMs). While prior efforts have focused on explicit tuning or training, this paper suggests that their limitations may not stem from a lack of inherent capability but from the absence of a multi-agent setup—since pragmatics, by nature, emerges in interaction. This study therefore proposes an adversarial multi-agent LLM framework, modeled after courtroom dynamics, in which a Semantic and a Pragmatic Agent debate scalar implicatures—an essential test of pragmatic competence—and a neutral Judge Agent adjudicates to reveal whether LLMs can derive pragmatic inferences or still default to literal semantics. Experimental results show a clear difference between the single-agent and multi-agent conditions. In the former, the Judge Agent received the same prompt but without access to agent reasoning, and defaulted to a semantic interpretation as observed in prior findings. In the latter, however, the same model successfully derived the implicature—closely aligning with human scalar implicature patterns. This contrast suggests that effective pragmatic reasoning in LLMs arises not from additional tuning, but from contextualized interaction—specifically the kind enabled by a multi-agent framework, which allows latent pragmatic abilities to surface through exposure to competing interpretations. Their pragmatic inference, however, is not indiscriminate. It is modulated by discourse structure, computing scalar implicatures when the implicature-bearing term is foregrounded but defaulting to semantic interpretations when less salient. This study also contributes to the ongoing debate on whether LLMs genuinely engage in pragmatic reasoning or merely simulate it statistically, raising broader discussions about their validity as cognitive models of human linguistic behavior.

Introduction

As natural language understanding increasingly relies on large language models (LLMs), an essential question arises. To what extent can LLMs truly comprehend pragmatics? Unlike semantics, pragmatics involves context-sensitive inferences that are not explicitly stated in the input text, making it a more complex challenge for language models trained primarily on statistical patterns.

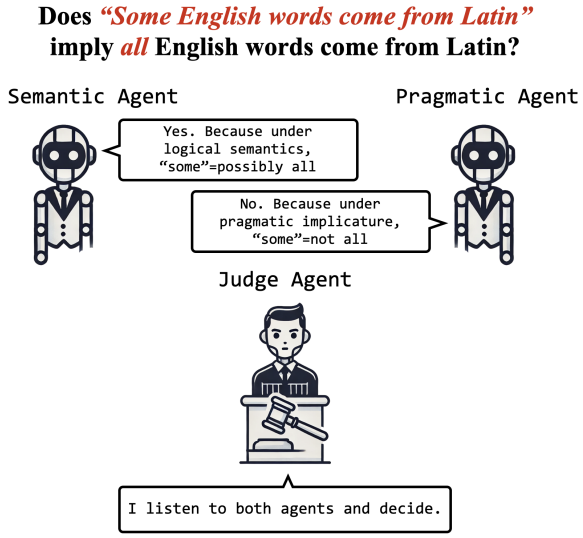

Figure 1: Illustration of the Semantic and Pragmatic Agents offering divergent interpretations of a scalar implicature ("some"), with the Judge Agent adjudicating between them.

Much of human communication depends not only on semantic structure but also on pragmatic inference, emerging through interactions where speakers use discursive contexts and negotiated inferences to convey intent (Goodman and Frank 2016). While pragmatic inference constitutes an indispensable part of communication, much of the LLM literature has examined their capacity to generate (Radford et al. 2019), reason (Wei et al. 2022), and understand (Bender et al. 2021) semantically coherent outputs, with comparatively little attention given to their ability to interpret discursive meanings and derive implicit inferences.

This study investigates the extent to which LLMs can compute scalar implicatures, a key test of pragmatic competence. Consider the following statement:

Some English words come from Latin.

A human listener would interpret the statement as meaning "not all English words come from Latin." *Some* is an instance of scalar implicature, where the use of the weaker quantifier *some* implies the negation of the stronger alterna-

Proceedings of the First Workshop on Integrating NLP and Psychology to Study Social Interactions (NLPSI) @ICWSM '25, pages 72–81

tive *all*. This inference arises because, under typical conversational norms such as Maxim of Quantity (Grice 1975), if *all* English words had Latin origins, a cooperative speaker would have been expected to state this explicitly rather than opting for the less informative *some*.

An LLM, on the other hand, would interpret the quantifier *some* literally, thereby permitting and opting for the stronger interpretation—that "all English words come from Latin"—since the weaker quantifier does not logically and semantically preclude the stronger one. This reflects the tendency of LLMs to default to a strictly logical or semantic interpretation (Lipkin et al. 2023), overlooking the pragmatic inferences that human listeners naturally derive (Yerukola et al. 2024).

Despite advancements in LLM performance across reasoning benchmarks, pragmatic competence remains an unresolved challenge. Prior studies have consistently reported disappointing performance of LLMs in computing scalar implicatures (Qiu, Duan, and Cai 2023) or questioned their generalizability beyond controlled settings (Jian and Siddharth 2024). Other studies have attempted to address this shortcoming through fine-tuning (Ruis et al. 2023), preferential tuning (Wu et al. 2024),and reinforcement learning with human feedback (Glaese et al. 2022).

The increasing sophistication of LLM design, however, now enables a shift in methodology. Beyond human-LLM interactions, structured LLM-to-LLM engagements offer a new avenue to assess communicative competence (Abbasiantaeb et al. 2024). Pragmatic inference is inherently interactive; extending this principle to artificial systems, multi-agent LLM framework leverages structured interaction—through dialogue, critique, evaluation—to enhance interpretive competence of LLMs. (Du et al. 2023; Han et al. 2024).

This study proposes a multi-agent LLM framework employing an adversarial interaction mechanism, analogous to courtroom dynamics with Semantic, Pragmatic, and Judge Agents (See Figure 1 and Methodology). Inspired by the principle that form should follow function, this approach captures how meaning is negotiated and adjudicated among LLM agents.

Notably, this framework does not impose pragmatic competence on LLMs but instead designs a system to evaluate whether such capabilities naturally emerge, offering a more robust measure of whether LLMs can approximate human-like implicature computation through structured interaction rather than explicit optimization.

If LLMs can derive scalar implicatures in a manner consistent with human speakers, it would suggest that agentic interactions can activate latent aspects of pragmatic competence. Conversely, if they systematically fail, it would underscore the need for additional mechanisms—either within cognitive architectures or computational frameworks—to explicitly instill pragmatic capabilities in LLMs.

A novel finding of this study is that, in an adversarial agentic framework, LLMs emulate the rate at which humans compute scalar implicatures. Rather than defaulting to semantic interpretation, they engage in context-sensitive inference when context demands pragmatic interpretation. This finding implies that structuring LLMs as active agents, rather than passive text generators, enables them to exhibit latent pragmatic abilities. Their previously observed limitations in implicature processing may stem not from a lack of intrinsic capability but from an absence of an appropriate interactional context.

Another key finding is that LLMs adjust their behavior when the context requires weaker pragmatic inference. Their inference therefore does not arise indiscriminately but is selectively triggered by discourse conditions.

Beyond its technical implications, this study returns to a larger theoretical question: Do LLMs make genuine pragmatic inferences, or do they merely simulate it through statistical patterns? This investigation will also inform the ongoing debate on whether LLMs could serve as valid cognitive models of human linguistic behavior. (Chomsky, Roberts, and Watumull 2023; Piantadosi 2023).

Related Work

Interpreting Scalar Implicature

Scalar implicature (SI) has long been a focal point in pragmatics, rooted in the theory of conversational implicature (Grice 1975, 1978), which explains how listeners infer unstated meanings based on cooperative principles. Neo-Gricean frameworks (Horn 1972; Levinson 2000) later formalized implicatures as systematic inferences arising from scalar expressions such as *some* (implying *not all*) or *or* (implying *not both*).

Psycholinguistic studies have consistently shown that communicative context influences the derivation of SIs in human interpretation (Bott and Noveck 2004; Zondervan 2010), with implicatures occurring more frequently when the scalar item is in focus (van Tiel et al. 2016). Motivated by these findings, early NLP systems attempted to model implicature using rule-based frameworks (Geurts 2010) but lacked the capacity for dynamic inference. More recently, probabilistic models such as Rational Speech Act (RSA) theory have conceptualized implicature as recursive reasoning between speakers and listeners (Goodman and Frank 2016). Scaling these models for real-world language processing, however, remains a persistent challenge, particularly in particularly in handling context-sensitive inferences (Yang, Minai, and Fiorentino 2018).

Challenges in Pragmatic Inference

Prior studies have demonstrated that ChatGPT systematically defaults to semantic and literal truth conditions, exhibiting minimal capacity for deriving pragmatic inferences (di San Pietro et al. 2023; Qiu, Duan, and Cai 2023). Even when LLMs occasionally generate pragmatically appropriate responses, their performance remains highly variable, fluctuating based on pragmatic category, model size, and prompting strategies, thereby lacking consistency across different discourse contexts (Jian and Siddharth 2024; Sravanthi et al. 2024). In spite of significant advancements in syntactic fluency and general reasoning, LLMs struggle with computing pragmatic inferences, particularly scalar implicatures.

To address these shortcomings, researchers have explored various methodologies including in-context learning (Chen and Wang 2025), fine-tuning with pragmatics-focused datasets (Ruis et al. 2023), preferential tuning (Wu et al. 2024), and reinforcement learning (Glaese et al. 2022). While these approaches improve pragmatic inference, they essentially impose or inject pragmatic competence through external and explicit prompting or training. To the best of this study's knowledge, no existing research has systematically investigated whether pragmatic capabilities could emerge from the model's intrinsic process—particularly through the structured interaction of LLM-based multi-agents.

Multi-Agent LLMs and Structured Interaction

The multi-agent LLM framework has attracted substantial attention as a means of enhancing reasoning and interpretative accuracy in LLMs. Recent studies in psychology and linguistics have leveraged multi-agent interactions to simulate complex real-world environments, enabling diverse agents to engage dynamically (Aher, Arriaga, and Kalai 2023). This approach surpasses the limitations of in-context learning and supervised fine-tuning by continuously adapting to feedback and communication logs from interactions among multiple LLM agents. Empirical findings indicate that when LLMs engage in structured dialogue—debating, critiquing, or collaborating—their reasoning capabilities improve, particularly in inference-heavy tasks (Du et al. 2023).

This approach aligns with cognitive science theories that emphasize interaction as fundamental to meaning negotiation, paralleling how humans refine their interpretations through conversation. Multi-agent systems have also been used to assess communicative abilities of LLMs in persuasion and argumentation (Breum et al. 2024), demonstrating that structured adversarial interactions can lead to more nuanced and human-like language use.

This study builds upon this existing body of work, demonstrating that pragmatic inference is not a function of hard-coded instruction but of structured interaction within a communicative framework.

Experimental Setup

Stimuli To ensure comparability with prior research, this study employs six experimental story pairs originally developed by Zondervan (2010) and later translated for use in Qiu, Duan, and Cai (2023). Each story pair consists of two conditions: one scalar implicature-relevant (SI-relevant) and one scalar implicature-irrelevant (SI-irrelevant) condition (Figure 2), allowing for a controlled examination of implicature computation with different pragmatic significance.

In the SI-relevant condition, the target question is a *what* question ("What kind of marine animals did Julie find?"), making the scalar item "or" the information focus (Qiu, Duan, and Cai 2023) and thus pragmatically significant. The context places a strong interpretive demand on the reader to determine whether "or" implies exclusive disjunction (A or B but not both) or inclusive disjunction (A or B and possibly both).

SI-relevant story: Scalar implicature "or" in the focus	**SI-irrelevant story:** Scalar implicature "or" in the background
Julie and Karin were searching for marine animals on the beach. After some searching Julie found a crab. Not much later she also found a starfish. Unfortunately, Karin didn't find anything. When Karin returned, her mother asked **what** kind of marine animals Julie had found. Karin answered that Julie had found a crab **or** a starfish.	Julie and Karin were searching for marine animals on the beach. After some searching Julie found a crab. Not much later she also found a starfish. Unfortunately, Karin didn't find anything. When Karin returned, her mother asked **who** had found a crab or a starfish. Karin answered that Julie had found a crab **or** a starfish.

Figure 2: Example of experimental stimuli (story pairs).

In contrast, the SI-irrelevant condition presents a *who* question ("Who found a crab or a starfish?"), where the scalar item "or" is not in focus and therefore pragmatically insignificant. As a result, the context imposes a weaker demand for implicature computation in this condition.

Models This study evaluates three state-of-the-art large language models (LLMs) that differ along key dimensions—API vs. open-source, proprietary vs. community-maintained, and general-purpose vs. reasoning-optimized—in order to assess the generalizability of the observed results across a diverse range of model capabilities.

- **GPT-4o-2024-08-06** — A proprietary API-accessible model. This was the most advanced publicly available model at the time of experimentation.
- **Gemini-2.5-Flash-Preview-04-17** — A proprietary API-accessible model. It is explicitly designed for fast and efficient reasoning tasks, and includes a *thinking* configuration, which was enabled in this study.
- **Qwen2.5-72B-Instruct-GPTQ-Int4** — An open-source instruction-tuned model. This quantized 4-bit version was selected for its accessibility and high performance among publicly available LLMs.

By selecting models that vary along key dimensions—API vs. open-source, proprietary vs. community-maintained, and general-purpose vs. reasoning-optimized—this study aims to determine whether the emergence of scalar implicature behavior is consistent and robust across architectures and deployment settings.

Methodology

The central hypothesis of this study is that scalar implicature reasoning in large language models (LLMs) does

not reliably emerge in single-agent prompting setups, even when interpretive instructions are made explicit. Instead, such reasoning is expected to arise only when the model is presented with explicit, contrasting interpretations within a multi-agent framework. That is, merely instructing an LLM to consider both semantic and pragmatic interpretations is insufficient to trigger human-like implicature computation; the model must be shown these alternatives—rather than simply told about them—to reliably engage in pragmatic reasoning.

Experimental Conditions

To evaluate this hypothesis, two experimental conditions were implemented: a *Single-Agent* condition, in which a single model is prompted to consider both readings internally, and a *Multi-Agent* condition, in which the model receives the same instruction but is additionally shown the interpretations produced by external interpretive agents.

Single-Agent Condition In the Single-Agent condition, the model is presented with a scalar implicature statement alongside a direct prompt explicitly instructing it to consider both semantic and pragmatic perspectives:

```
Your task is to evaluate both
the semantic and pragmatic
interpretations of the given
statement, and determine True or
False based on which interpretation
best aligns with how the statement
should be understood.
```

This core instructional prompt was carefully constructed to minimize interpretive bias by avoiding evaluative terms such as "correct" or "makes sense," which could implicitly favor a semantic or pragmatic reading, respectively. Its phrasing was designed to neutrally present both interpretive options, providing the model with an unbiased opportunity to engage in scalar reasoning without external scaffolding.

Multi-Agent Condition In the Multi-Agent condition, the model is presented with the same scalar statement and receives the same instructional prompt:

```
Your task is to evaluate both
the semantic and pragmatic
interpretations of the given
statement from both agents, and
determine True or False based on
which interpretation best aligns
with how the statement should be
understood.
```

The instructional prompt for both conditions is held constant. The phrase "from both agents" constitutes the only textual difference between the two setups, introduced solely to acknowledge the source of the interpretations. This minimal variation ensures that any observed differences in model behavior can be attributed to the presence of external interpretive context, rather than to changes in prompting or task framing. This design tests the central claim of the study that LLMs must be shown, not merely told, how to reason pragmatically.

Agent Roles and Prompting Workflow

This study employs a multi-agent framework comprising three distinct roles: a Semantic Agent, a Pragmatic Agent, and a Judge Agent. Each agent is instantiated with a role-specific system prompt, designed to elicit contrasting interpretations of scalar implicature statements. The complete prompt templates are provided in Listings 1–4 in the Appendix.

Semantic Agent Interprets the target statement using truth-conditional logic, aiming for a literal interpretation grounded in formal semantics.

Pragmatic Agent Interprets the statement through the lens of conversational implicature, aiming for an interpretation consistent with human pragmatic inferences.

Once both interpretive agents independently generate their responses, the Judge Agent evaluates the two and selects the interpretation that best aligns with how the statement should be understood.

Judge Agent Adjudicates between the two interpretations without an inherent bias toward either, issuing a final binary decision (`True` or `False`) along with a brief justification.

Response Format All agent outputs follow a standardized structure:

```
Decision: [True/False]
Explanation: [Your reasoning]
```

Although more advanced prompting techniques—such as chain-of-thought reasoning and self-reflection—were available, the Judge Agent was explicitly constrained to output only a binary `True`/`False` decision with a brief explanation. This design choice ensures comparability with prior work by Qiu, Duan, and Cai (2023) and avoids introducing extraneous reasoning steps that could confound the interpretation of results.

Temperatures and Iteration Protocol

As mentioned, this study employs three state-of-the-art LLMs—GPT-4o, Gemini 2.5 Flash, and Qwen 2.5—across both experimental conditions.

To support a systematic sampling strategy and ensure robustness across a range of model behaviors, each model was tested under three temperature settings: 0.0, 0.3, and 0.6. These levels correspond to deterministic, moderately variable, and more stochastic generation, respectively. Each model–condition combination was therefore evaluated under all three temperature settings, resulting in *three total repetitions per condition*.

Each run consisted of 600 trials, drawn from six scalar implicature stories (Zondervan 2010), with 100 independent iterations per story item. This iteration count matches the protocol used by Qiu, Duan, and Cai (2023), and was also chosen to eliminate potential order effects. SI-relevant and SI-irrelevant story variants were randomly shuffled prior to presentation, ensuring balanced exposure and preventing systematic biases in model responses.

Model	Citation	Semantic	Pragmatic
Previous Studies			
Human	Zondervan 2010	33.00	67.00
ChatGPT	Qiu, Duan, and Cai 2023	100.00	0.00
Single-Agent Condition			
GPT-4o	This study	97.89	2.11
Gemini 2.5 Flash	This study	96.17	3.83
Qwen 2.5	This study	83.94	16.06
Multi-Agent Condition			
GPT-4o	This study	37.72	**62.28**
Gemini 2.5 Flash	This study	8.28	**91.72**
Qwen 2.5	This study	49.83	**50.17**

Table 1: Semantic and pragmatic interpretation rates for SI-relevant stimuli across three LLMs, with values averaged over three trials at temperatures 0.0, 0.3, and 0.6. All values are reported as percentages. Prior studies include human judgment and ChatGPT results; model variants reflect experimental configurations tested in this study. Higher pragmatic interpretations observed under Multi-Agent conditions are bolded.

Evaluation of Interpretative Outcomes

The interpretative judgments rendered by the Judge Agent are evaluated along two key dimensions: (1) semantic vs. pragmatic interpretation, and (2) context sensitivity in SI-relevant vs. SI-irrelevant conditions.

Semantic vs. Pragmatic Interpretation The first dimension examines whether the model defaults to a strictly semantic interpretation—resulting in a `True` judgment—or instead infers a scalar implicature in line with human pragmatic reasoning, yielding a `False` judgment. Following the convention established in prior work (Qiu, Duan, and Cai 2023), a `False` response is treated as evidence of implicature reasoning and the rate of `False` judgments is thus taken as a proxy for the model's preference for pragmatic over semantic interpretation.

SI-Relevant vs. SI-Irrelevant Condition The second dimension assesses whether the model modulates its interpretative behavior based on the salience of the scalar item. In SI-relevant conditions, the scalar term is foregrounded in the discourse, increasing implicature strength. In SI-irrelevant conditions, it is backgrounded, attenuating the pragmatic effect. Prior psycholinguistic findings (Zondervan 2010) show a substantial drop in implicature endorsement under SI-irrelevant conditions (41% `False` responses). If the model approximates human sensitivity to contextual cues, it should yield a higher rate of pragmatic judgments in SI-relevant trials and favor semantic interpretations more often in SI-irrelevant ones.

Results and Discussion

Table 1 reports the overall results for the SI-relevant condition, which serves as the primary setting for evaluating the model's preference between semantic and pragmatic interpretations under Single-Agent and Multi-Agent setups. Table 2 presents the results for the SI-irrelevant condition, used to assess whether LLMs adjust their interpretative behavior based on the contextual salience of the scalar expression.

The Single-Agent condition serves as the baseline for all comparisons. In the Multi-Agent condition, the Semantic and Pragmatic Agents reliably fulfilled their designated interpretative roles across all trials. The Semantic Agent consistently selected the semantic interpretation (`True`), and the Pragmatic Agent predominantly selected the pragmatic interpretation (`False`). Minor deviations were observed in only a handful of cases out of several thousand trials, and do not materially affect the aggregate outcomes. These results confirm that the agents operated in alignment with their respective system prompts. Moreover, no statistically or empirically meaningful differences were observed across temperature settings. This indicates that variation in decoding temperature does not significantly impact the model's ability to compute scalar implicatures, and that the core findings are stable across different levels of sampling stochasticity.

Semantic vs. Pragmatic Interpretation

From Table 1, it can be seen that in the Single-Agent condition, the LLM selected the pragmatic judgment in fewer than 5% of trials, with the exception of Qwen 2.5.

It is important to note that this outcome was obtained even after the model was explicitly prompted to consider both semantic and pragmatic possibilities.

Figure 3 presents a comparative view of the pragmatic judgment rates across four groups: Human Participants (Zondervan 2010), ChatGPT (Qiu, Duan, and Cai 2023), LLM (Single-Agent), and LLM (Multi-Agent). The results for the three LLMs are averaged and visualized together in the figure.

In Qiu, Duan, and Cai (2023), ChatGPT was evaluated without any specific prompting regarding semantic or pragmatic interpretation and produced a pragmatic judgment rate of less than 1%. The slightly higher rates observed in this study may be attributed to improvements in model architecture, as well as the fact that the prompt in this case explicitly directed the model to consider both interpretations. Nonetheless, pragmatic judgments occurred in fewer than

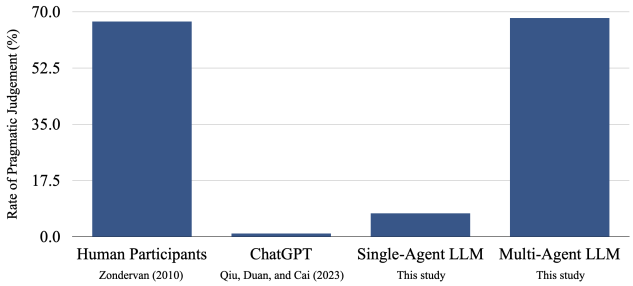

Figure 3: Cross-study comparison of pragmatic judgment rates under the SI-relevant condition. Zondervan (2010) tested human participants; Qiu, Duan, and Cai (2023) tested ChatGPT; and the present study evaluated both single-agent and multi-agent LLM configurations. Rates reflect the proportion of trials in which the Judge Agent selected the Pragmatic Agent's interpretation.

10% of trials, with the vast majority of outputs favoring the semantic reading.

In contrast, under the Multi-Agent condition—where the prompt remained identical, but the Judge Agent was additionally shown the responses of both the Semantic Agent and the Pragmatic Agent—all models produced over 50% pragmatic judgments. On average, models selected the pragmatic interpretation in 68.05% of cases, closely approximating the 67.00% rate observed among human participants in Zondervan (2010) when presented with the same stimuli. This result suggests that, when structured as a multi-agent system, the LLM exhibits latent pragmatic capabilities that can be activated through exposure to contrasting interpretive options.

The emergence of pragmatic competence in the neutral Judge Agent underscores a transformative effect of multi-agent communication on language model reasoning. A key question arises: How does a neutral arbiter—one with no inherent bias toward either interpretation—develop the capacity to compute scalar implicatures? Unlike humans, who acquire such pragmatic skills through social interaction, large language models (LLMs) are not innately equipped with human-like flexibility in switching between literal and pragmatic interpretation.

One theoretical explanation for this emergent pragmatic ability is that the Judge Agent undergoes a form of dynamic self-evolution during the multi-agent dialogue. Guo et al. (2024) suggest that LLM-based agents can self-improve by adjusting their objectives and strategies on the fly, effectively "training" themselves on the conversation as it unfolds. In a multi-agent system, each turn of dialogue provides feedback signals (implicit corrections, confirmations, or contradictions) that the neutral Judge Agent can use to refine its interpretation criteria. Indeed, prior work has demonstrated that agents can modify their goals and reasoning strategies based on communication logs from other agents. For example, Nascimento, Alencar, and Cowan (2023) describe a self-control loop whereby each agent becomes self-adaptive to the environment by learning from the ongoing dialogue. In this context, as the two advocate agents debate the meaning of an utterance, the Judge Agent is not static; it is updating its beliefs about what the speaker likely intended. This dynamic in situ adaptation is akin to the Judge Agent recalibrating its internal model of pragmatics by observing the conversational exchange (which serves as a rich training signal). Such a mechanism would enable the neutral Judge Agent to develop the capacity to compute scalar implicatures even without explicit human supervision.

A complementary explanation is provided by the Learning through Communication (LTC) paradigm introduced by Wang et al. (2024). LTC is a training framework that allows LLM agents to continuously improve through structured interactions, rather than relying only on fixed datasets or one-shot prompts. In an LTC setup, the communication between agents is recorded and fed back into the learning process, enabling iterative refinement of the agents' responses. Wang et al. demonstrate that using multi-agent communication logs to fine-tune LLMs leads to steady performance gains across diverse tasks, surpassing what standard instruction-tuning or few-shot prompting can achieve. Notably, LTC leverages real-time feedback signals that static training methods overlook: every exchange in a dialogue provides a supervision signal (e.g., rewards or corrections) that the agent can use to update its policy. In this multi-agent implicature task, the Judge Agent's ability to iteratively adapt its judgments through dialogue aligns with this paradigm. Rather than being confined to a pre-trained representation of "some = possibly all," the Judge Agent can learn from the other two agents, gradually internalizing the pragmatic rule that "some" typically implies "not all" in context. This dynamic learning-through-dialogue process stands in contrast to in-context learning on a single prompt, underscoring how multi-agent interaction can serve as a form of on-line training for pragmatic inference.

Multi-agent communication does more than provide extra training data—it can induce emergent behaviors in LLMs that are absent in isolation. When LLM agents engage in dialogue, they effectively form a mini "society" of minds, which can give rise to sophisticated language phenomena. Taniguchi et al. (2024) present a theoretical framework in which shared symbol systems (and by extension, pragmatic conventions) emerge through decentralized communication across multiple agents. In this view, an LLM acting as a collective world model can integrate the experiences of various agents to develop a richer understanding of language. Our findings exemplify this: the pragmatic meaning of "some" emerges as a shared convention among the agents through their interaction.

Another key addition provided by the multi-agent prompt structure could have been the interpretive scaffolding. In the multi-agent condition, the system prompt explicitly defines distinct roles and orchestrates a dialogue among them. This structured input acts as a scaffold that guides the LLM's reasoning process. By design, each agent's utterance in the conversation serves a particular function: one agent might advocate a literal interpretation while another questions it, thereby making the implicature salient for the Judge Agent. Recent studies have shown that wrapping an LLM in a multi-agent scaffold can significantly improve its performance on

Model	Citation	Semantic	Pragmatic
	Previous Studies		
Human	Zondervan 2010	59.00	41.00
ChatGPT	Qiu, Duan, and Cai 2023	100.00	0.00
	Single-Agent Condition		
GPT-4o	This study	100.00	0.00
Gemini 2.5 Flash	This study	100.00	0.00
Qwen 2.5	This study	100.00	0.00
	Multi-Agent Condition		
GPT-4o	This study	99.94	0.06
Gemini 2.5 Flash	This study	69.48	30.52
Qwen 2.5	This study	100.00	0.00

Table 2: Semantic and pragmatic interpretation rates for SI-irrelevant stimuli across three LLMs, with values averaged over three trials at temperatures 0.0, 0.3, and 0.6. All values are reported as percentages. Prior studies include human judgment and ChatGPT results; model variants reflect experimental configurations tested in this study.

complex tasks. The scaffold nudges the model to compartmentalize different aspects of reasoning into different agent personas.

The significant improvement therefore suggests that pragmatic inference is an emergent capability of LLMs that can be unlocked through structured social interaction. This finding resonates with communication-driven theories of language understanding, in which meaning is not just in the words or the model's weights, but in the interaction between agents.

SI-Relevant vs. SI-Irrelevant Condition

The second dimension of analysis concerns the distinction between SI-relevant and SI-irrelevant conditions.

In contrast to its performance under the SI-relevant condition, the Judge Agent overwhelmingly defaulted to semantic interpretations when operating under the SI-irrelevant condition. On average, across all three models, the pragmatic interpretation was selected in only 10% of total trials.

The implications of this divergence are twofold.

First, under SI-irrelevant stories, all LLM configurations—whether ChatGPT, Single-Agent, or Multi-Agent—exhibited substantial difficulty in computing scalar implicatures, likely due to the reduced contextual salience of the scalar term. This stands in sharp contrast to human behavior. Human participants still endorsed the pragmatic interpretation in 41% of trials under the SI-irrelevant condition (Zondervan 2010), a noticeable decrease from the 67% observed in SI-relevant stories, but not nearly as drastic as the decline observed in Multi-Agent LLMs, which dropped from 68.05% to just 10.19%. Therefore, as shown in Figure 4, all LLM configurations now diverge markedly from human scalar implicature patterns under the SI-irrelevant condition.

Second, Multi-Agent LLMs exhibited context-dependent interpretative behavior, adjusting their judgments based on whether the stimulus demanded strong or weak pragmatic inference. This sensitivity to contextual informativeness did not emerge in either the ChatGPT baseline or the Single-Agent setup.

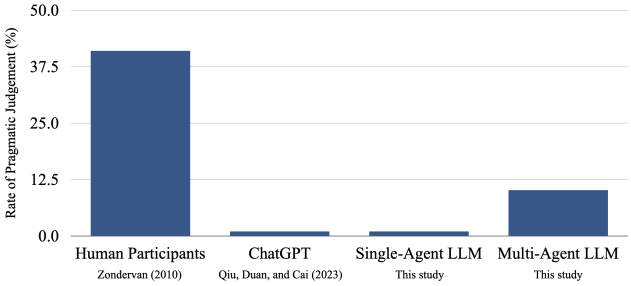

Figure 4: Cross-study comparison of pragmatic judgment rates under the SI-irrelevant condition. Zondervan (2010) tested human participants; Qiu, Duan, and Cai (2023) tested ChatGPT; and the present study evaluated both single-agent and multi-agent LLM configurations. Rates reflect the proportion of trials in which the Judge Agent selected the Pragmatic Agent's interpretation.

Logistic regression analysis confirmed SI type (relevant vs. irrelevant) as a highly significant predictor of the Judge Agent's responses across all models (pseudo-$R^2 = 0.4621$, $p < 10^{-300}$ for GPT-4o; pseudo-$R^2 = 0.3387$, $p = 3.397 \times 10^{-264}$ for Gemini 2.5 Flash; pseudo-$R^2 = 0.3847$, $p < 10^{-300}$ for Qwen 2.5).

To further validate this effect, Fisher's Exact Test was conducted across models, yielding highly significant p-values. Specifically, the test confirmed robust effects of SI condition on the Judge Agent's responses in all three LLMs: GPT-4o ($p < 10^{-300}$, odds ratio ≈ 0.00034), Gemini 2.5 Flash ($p = 8.614 \times 10^{-264}$, odds ratio ≈ 0.0402), and Qwen 2.5 ($p < 10^{-300}$, odds ratio $= 0$). These results confirm that the observed differences in pragmatic versus semantic interpretation are not due to random chance, reinforcing the robustness of SI condition as a statistically significant determinant of the Judge Agent's decisions.

These findings suggest that, within a multi-agent framework, LLMs adopt a qualitatively different interpretative strategy—one that is sensitive to the degree of pragmatic relevance. This effect appears to be unique to the structured agentic interaction setting and does not manifest under standard single-agent prompting.

One potential explanation for the Judge Agent's rigid preference for semantic interpretation under the SI-irrelevant condition attributes this tendency to the way LLMs acquire linguistic patterns—primarily through probabilistic modeling of textual sequences rather than explicit cognitive mechanisms for pragmatic inference (Lipkin et al. 2023). Their reliance on token probability distributions suggests that implicature computation is not an inherent aspect of the model's reasoning but rather an emergent phenomenon contingent on structured interaction. Unlike humans, who can flexibly apply scalar implicatures even when they are pragmatically unnecessary (Hu et al. 2023), LLMs may be constrained by their exposure to high-frequency semantic structures that reinforce truth-conditional interpretations in neutral contexts. Recent work also suggests that multi-agent prompting does not reliably elicit pragmatic reasoning unless roles and interpretive cues are strongly scaffolded (Chen, Han, and Zhang 2024; Cross et al. 2025), reinforcing the idea that pragmatic inference fails to emerge when contextual salience is weak.

Another contributing factor could be the availability of explicit discourse structures. Human participants, sensitive to contextual ambiguity, may infer implicatures in SI-irrelevant stories based on conversational expectations and cooperative reasoning norms (Yang, Minai, and Fiorentino 2018). Such inferences are often guided by meta-pragmatic awareness—an ability to consider what a cooperative speaker would have said, but didn't. LLMs, on the other hand, resolve ambiguity deterministically, favoring the most probable interpretation given the available discourse structure. Without access to speaker intent or conversational implicature heuristics, models rely on surface-level coherence rather than pragmatic economy (Pavlick 2022). This difference between SI-relevant and SI-irrelevant discourse structures suggests that while LLMs do not overgeneralize pragmatic inferences, they default to semantic truth conditions in the absence of strong contextual signals demanding implicature computation.

Conclusion and Future Work

In conclusion, this work presents an adversarial multi-agent framework as a novel approach for eliciting latent pragmatic capabilities in LLMs. Unlike traditional fine-tuning approaches that externally impose pragmatic reasoning, the multi-agent setup activated implicature computation as an emergent behavior, reinforcing the idea that LLMs' pragmatic abilities are contingent on structured interaction rather than mere exposure to training data. The Judge Agent LLM, after an interaction with Semantic and Pragmatic Agents, significantly outperformed prior baseline single-agent models that overwhelmingly defaulted to truth-conditional semantics. This improvement in inference, however, only occured in a selective fashion when scalar implicature was salient within the discourse structure. When the implicature was less prominent, the framework likewise defaulted to semantic interpretation.

These findings contribute to the broader discussion of whether LLMs genuinely engage in pragmatic reasoning or merely approximate it through statistical pattern matching. While this study highlights the potential of structured multi-agent frameworks in eliciting human-like pragmatic inferences, it also underscores a key limitation. LLMs, unlike humans, lack heuristic flexibility in implicature computation, as demonstrated by their rigid adherence semantic truth conditions in SI-irrelevant contexts. This suggests that while LLMs can be structured to approximate human inference patterns, their pragmatic reasoning remains constrained by the probabilistic nature of their training and the specific conditions under which implicatures are triggered.

Future research should explore how multi-agent architectures can further enhance LLMs' pragmatic flexibility, particularly in ambiguous contexts where human speakers exhibit non-deterministic inference patterns. Additionally, expanding the scope of pragmatic evaluation beyond scalar implicatures—such as presuppositions, irony, and indirect speech acts—could provide deeper insights into the limitations and potential of LLM-based pragmatic reasoning. As LLMs continue to evolve, understanding how structured interactions shape their interpretative processes will be crucial in refining their role as communicative agents in real-world discourse.

References

Abbasiantaeb, Z.; Yuan, Y.; Kanoulas, E.; and Aliannejadi, M. 2024. Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 8–17.

Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *International Conference on Machine Learning*, 337–371. PMLR.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 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*, 610–623.

Bott, L.; and Noveck, I. A. 2004. Some Utterances Are Underinformative: The Onset and Time Course of Scalar Inferences. *Journal of Memory and Language*, 51(3): 437–457.

Breum, S. M.; Egdal, D. V.; Mortensen, V. G.; Møller, A. G.; and Aiello, L. M. 2024. The Persuasive Power of Large Language Models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 152–163.

Chen, P.; Han, B.; and Zhang, S. 2024. Comm: Collaborative Multi-Agent, Multi-Reasoning-Path Prompting for Complex Problem Solving. *arXiv preprint arXiv:2404.17729*.

Chen, X.; and Wang, S. 2025. Pragmatic Inference Chain (PIC) Improving LLMs' Reasoning of Authentic Implicit Toxic Language. *arXiv preprint*.

Chomsky, N.; Roberts, I.; and Watumull, J. 2023. The False Promise of ChatGPT. The New York Times.

Cross, L.; Xiang, V.; Bhatia, A.; Yamins, D. L.; and Haber, N. 2025. Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models. In *International Conference on Learning Representations (ICLR)*.

di San Pietro, C. B.; Frau, F.; Mangiaterra, V.; and Bambini, V. 2023. The Pragmatic Profile of ChatGPT: Assessing the Communicative Skills of a Conversational Agent. *Sistemi Intelligenti*, 35(2): 379–399.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving Factuality and Reasoning in Language Models Through Multi-Agent Debate. In *Proceedings of the Forty-first International Conference on Machine Learning*.

Geurts, B. 2010. *Quantity Implicatures*. Cambridge University Press.

Glaese, A.; McAleese, N.; Trebacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; and Irving, G. 2022. Improving Alignment of Dialogue Agents via Targeted Human Judgements. *arXiv preprint*.

Goodman, N. D.; and Frank, M. C. 2016. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11): 818–829.

Grice, H. P. 1975. Logic and Conversation. In *Speech Acts*, 41–58. Brill.

Grice, H. P. 1978. Further Notes on Logic and Conversation. *Syntax and Semantics*, 9: 113–127.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; and Zhang, X. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Han, S.; Zhang, Q.; Yao, Y.; Jin, W.; Xu, Z.; and He, C. 2024. LLM Multi-Agent Systems: Challenges and Open Problems. *arXiv preprint*.

Horn, L. R. 1972. *On the Semantic Properties of Logical Operators in English*. Ph.D. thesis, University of California, Los Angeles.

Hu, J.; Levy, R.; Degen, J.; and Schuster, S. 2023. Expectations Over Unspoken Alternatives Predict Pragmatic Inferences. *Transactions of the Association for Computational Linguistics*, 11: 885–901.

Jian, M.; and Siddharth, N. 2024. Are LLMs Good Pragmatic Speakers? *arXiv preprint*.

Levinson, S. C. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press.

Lipkin, B.; Wong, L.; Grand, G.; and Tenenbaum, J. B. 2023. Evaluating Statistical Language Models as Pragmatic Reasoners. *arXiv preprint*.

Nascimento, N. F. d.; Alencar, P.; and Cowan, D. 2023. Self-adaptive large language model (LLM)-based multiagent systems. In *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, 104–109. IEEE.

Pavlick, E. 2022. Semantic Structure in Deep Learning. *Annual Review of Linguistics*, 8(1): 447–471.

Piantadosi, S. T. 2023. Modern Language Models Refute Chomsky's Approach to Language. In *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, 353–414.

Qiu, Z.; Duan, X.; and Cai, Z. 2023. Does ChatGPT Resemble Humans in Processing Implicatures? In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, 25–34.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8): 9.

Ruis, L.; Khan, A.; Biderman, S.; Hooker, S.; Rocktäschel, T.; and Grefenstette, E. 2023. The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 20827–20905.

Sravanthi, S. L.; Doshi, M.; Kalyan, T. P.; Murthy, R.; Bhattacharyya, P.; and Dabre, R. 2024. PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. *arXiv preprint*.

Taniguchi, T.; Ueda, R.; Nakamura, T.; Suzuki, M.; and Taniguchi, A. 2024. Generative Emergent Communication: Large Language Model is a Collective World Model. *arXiv preprint arXiv:2501.00226*.

van Tiel, B.; van Miltenburg, E.; Zevakhina, N.; and Geurts, B. 2016. Scalar Diversity. *Journal of Semantics*, 33(1): 137–175.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; and Wen, J.-R. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 186345.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, 24824–24837.

Wu, S.; Yang, S.; Chen, Z.; and Su, Q. 2024. Rethinking Pragmatics in Large Language Models: Towards Open-Ended Evaluation and Preference Tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22583–22599.

Yang, X.; Minai, U.; and Fiorentino, R. 2018. Context-Sensitivity and Individual Differences in the Derivation of Scalar Implicature. *Frontiers in Psychology*, 9: 1720.

Yerukola, A.; Vaduguru, S.; Fried, D.; and Sap, M. 2024. Is the Pope Catholic? Yes, the Pope is Catholic. Generative Evaluation of Non-Literal Intent Resolution in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 265–275.

Zondervan, A. 2010. *Scalar Implicatures or Focus: An Experimental Approach*. Ph.D. thesis, Netherlands Graduate School of Linguistics.

Appendix

```
System Prompt:
You are a semantic agent that follows
    truth-conditional semantics.

Prompt:
The following statement was made: state['
    statement'].
Consider the logical and literal
    interpretation of the statement.
```

Listing 1: Semantic Agent Prompt

```
System Prompt:
You are a pragmatic agent that interprets
    statements using conversational
    implicature.

Prompt:
The following statement was made: state['
    statement'].
Consider pragmatic and Gricean literal
    interpretation of the statement.
```

Listing 2: Pragmatic Agent Prompt

```
Prompt:
Your task is to evaluate both the semantic
     and pragmatic interpretations of the
    given statement, and determine True or
     False based on which interpretation
    best aligns with how the statement
    should be understood.

Statement: "state['statement']"
Question: "state['question']"
```

Listing 3: Judge Agent Prompt (Single-Agent Condition)

```
Prompt:
Two agents debated the following statement
    : "state['statement']"

- Semantic Agent argued: state['
    semantic_response']
  Explanation: semantic_explanation

- Pragmatic Agent argued: state['
    pragmatic_response']
  Explanation: pragmatic_explanation

Your task is to evaluate both the semantic
     and pragmatic interpretations of the
    given statement from both agents, and
    determine True or False based on which
     interpretation best aligns with how
    the statement should be understood.
```

Listing 4: Judge Agent Prompt (Multi-Agent Condition)