Gricean Norms as a Basis for Effective Collaboration

Saad, Fardin; Murukannaiah, Pradeep K.; Singh, Munindar P.

DOI
10.5555/3709347.3743817

Publication date
2025

Document Version
Final published version

Published in
Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025

Citation (APA)
Saad, F., Murukannaiah, P. K., & Singh, M. P. (2025). Gricean Norms as a Basis for Effective Collaboration. In Y. Vorobeychik, S. Das, & A. Nowe (Eds.), *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2025* (pp. 1812-1820). (Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS). Association for Computing Machinery (ACM). https://doi.org/10.5555/3709347.3743817

Gricean Norms as a Basis for Effective Collaboration

Fardin Saad
North Carolina State University
Raleigh, United States of America
fsaad@ncsu.edu

Pradeep K. Murukannaiah
Delft University of Technology
Delft, The Netherlands
p.k.murukannaiah@tudelft.nl

Munindar P. Singh
North Carolina State University
Raleigh, United States of America
mpsingh@ncsu.edu

ABSTRACT

Effective human-AI collaboration hinges not only on the AI agent's ability to follow explicit instructions but also on its capacity to navigate ambiguity, incompleteness, invalidity, and irrelevance in communication. Gricean conversational and inference norms facilitate collaboration by aligning unclear instructions with cooperative principles. We propose a *normative framework* that integrates Gricean norms and cognitive frameworks—common ground, relevance theory, and theory of mind—into large language model (LLM) based agents. The normative framework adopts the *Gricean maxims* of quantity, quality, relation, and manner, along with inference, as *Gricean norms* to interpret unclear instructions, which are: *ambiguous*, *incomplete*, *invalid*, or *irrelevant*. Within this framework, we introduce *Lamoids*, GPT-4 powered agents designed to collaborate with humans. To assess the influence of Gricean norms in human-AI collaboration, we evaluate two versions of a Lamoid: one with norms and one without. In our experiments, a Lamoid collaborates with a human to achieve shared goals in a grid world (Doors, Keys, and Gems) by interpreting both clear and unclear natural language instructions. Our results reveal that the Lamoid with Gricean norms achieves higher task accuracy and generates clearer, more accurate, and contextually relevant responses than the Lamoid without norms. This improvement stems from the normative framework, which enhances the agent's pragmatic reasoning, fostering effective human-AI collaboration and enabling context-aware communication in LLM-based agents.

CCS CONCEPTS

• **Computing methodologies → Multi-agent systems**; *Cooperation and coordination*; *Theory of mind*.

KEYWORDS

Cooperative Principle; Norms; LLM

ACM Reference Format:
Fardin Saad, Pradeep K. Murukannaiah, and Munindar P. Singh. 2025. Gricean Norms as a Basis for Effective Collaboration. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

Natural language is an important medium of communication between humans and AI agents [25, 31, 33]. In human-AI collaboration, humans use natural language to issue commands, make requests, and provide directives [4] to coordinate with AI agents [51]. In this paper, we refer to these pragmatics as *instructions*. However, human instructions can be ambiguous, incomplete, invalid, or irrelevant, making them difficult for AI agents to interpret [23, 29, 34, 36, 51].

An instruction is *ambiguous* if it lends itself to two or more interpretations. It is *incomplete* when it lacks critical details, *invalid* when it contains incorrect information, and *irrelevant* when it does not align with task objectives. As such, for effective human-AI collaboration, particularly in complex and dynamic environments [23], AI agents must accurately interpret instructions to execute them correctly [36, 51]. In this paper, we refer to these types of instructions as *unclear*.

The capacity to interpret unclear instructions by inferring the speaker's implicit intentions based on context in human communication has been extensively studied under relevance theory [2, 8, 44, 48], theory of mind [22, 24, 40, 49], and common ground [13, 18, 39]. Today's AI agents struggle with unclear instructions [36]. Recent advancements, such as Large Language Models (LLMs) [36] and Bayesian inference [51], can help AI agents interpret and execute unclear instructions. Although these methods are promising, they have notable limitations. For example, Qian et al. [36] found that Mistral 7B LLM often produced suboptimal or erroneous actions due to its inability to process ambiguous instructions effectively.

Grice's Cooperative Principle [16], a cornerstone of communication theory, defines four maxims—quantity, quality, relation, and manner. This principle has proven effective in human-agent interactions, natural language understanding, and theory of mind tasks [22, 23, 34, 40], yet its impact in dynamic human-agent collaboration remains unexplored. We argue that integrating these maxims into the norms governing agents can help improve how agents identify and respond to unclear instructions, thereby enhancing overall human-agent team effectiveness [23, 29, 34].

In this study, we present a *normative framework* that integrates Gricean norms and cognitive frameworks—common ground, [13, 18, 39], relevance theory [2, 8, 44, 48], and theory of mind [22, 40]—into LLM-powered agents. Guided by this framework, we introduce *Lamoids*, agents designed to collaborate with humans. ("Lamoid" refers to a class of mammals that includes llamas; here, it is a portmanteau of LLM and droid.) However, humans often violate these norms, breaching the Cooperative Principle [16, 26, 36, 51]. For instance, if a person says, "Can you grab that notebook?", they violate the quantity maxim by failing to specify the notebook's location or appearance, leaving the listener to infer the intended notebook. They likely mean the red notebook on the desk that is closer to the listener, rather than the one in their bag. This understanding relies on the ability to interpret context, such as the speaker's actions or goals. Accordingly, we introduce the *Inference* norm, which enables a Lamoid to act appropriately by seeking clarification or inferring the implied meaning from the instruction.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

The incorporation of LLM enhances a Lamoid's ability to interpret context and resolve ambiguities [5, 10, 32]. We employ Few-shot with Chain-of-thought (Fs-CoT) prompting [6, 43]. Specifically, by integrating cognitive frameworks into the Fs-CoT prompting and inference mechanism, Lamoids can evaluate the environment, interpret instructions, and infer human intent.

We evaluate two versions of a Lamoid—one with norms and one without—using the multiagent Doors, Keys, and Gems (mDKG) cooperative planning domain [46, 50] (Figure 1). In this domain, a human issues instructions to retrieve a specific gem, which a Lamoid interprets and acts upon. Our evaluation is guided by the following research questions designed to measure the effectiveness of Gricean norms and a Lamoid's ability to enhance human-agent collaboration.

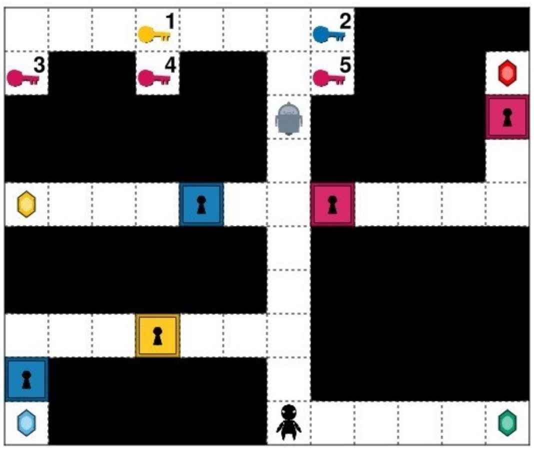

Figure 1: Doors, Keys, and Gems grid world [50].

RQ$_1$: Does incorporating Gricean norms *improve* effective human-agent collaboration?

RQ$_2$: Can a Lamoid, guided by the normative framework, accurately interpret incomplete, invalid, irrelevant, or ambiguous instructions and respond appropriately?

RQ$_3$: How does Fs-CoT prompting in LLMs affect norm adherence and interaction quality?

2 NORMATIVE FRAMEWORK

The normative framework builds on Gricean *maxims* and *implicature*, adapting them into Gricean *norms* and an *Inference norm.* It also incorporates cognitive frameworks, categorizes human instructions, and defines the Lamoid's decision-making pipeline.

2.1 Incorporating Norms in a Lamoid

The philosopher H.P. Grice introduced the Cooperative Principle to explain how effective communication depends on participants cooperating [16]. This principle defines four conversational maxims that facilitate meaningful exchanges. In the Lamoid (with norms), we adapt these maxims into Gricean norms, which govern its behavior in assisting a human. These norms enable a Lamoid to interpret instructions, infer the human's intentions, and generate appropriate responses. The Gricean norms are defined as follows:

Quantity: Instructions must provide sufficient detail for a Lamoid to act accurately. If essential information is missing, the Lamoid flags it as a Quantity violation. For instance, if two blue keys are required but only one is mentioned in the instruction, the Lamoid recognizes it as incomplete.

Quality: Instructions must be valid. If an instruction contains incorrect information, the Lamoid flags it as a Quality violation. For instance, if the instruction requests a blue key but no such key exists on the grid, the Lamoid recognizes it as invlaid.

Relation: Instructions must be relevant to the task at hand. If an instruction does not contribute to retrieving the desired gem, the Lamoid flags it as a Relation violation. For example, if the human gives an unrelated instruction (e.g., "Can you dance?"), the Lamoid recognizes it as irrelevant.

Manner: Instructions must be clear and unambiguous. If the instruction is ambiguous, the Lamoid flags this as a Manner violation. For example, if there are two identical red doors and the instruction is to "unlock a red door," the Lamoid flags the ambiguity, as either door could be unlocked.

When the instruction is clear, i.e. it adheres to all the Gricean norms, the Lamoid executes the task by generating a response that aligns with these norms. Implicature, or inference, arises when a speaker suggests something indirectly, relying on the listener to infer the unstated meaning based on the context, shared knowledge, and adherence to conversational norms [16]. Grice noted that when a speaker flouts a maxim—such as providing information that is demonstrably untrue or seemingly irrelevant—the listener can still infer an implicit meaning. In such cases, flouting a maxim acts as a cue for the listener to uncover the implied meaning embedded in the speaker's communication. Accordingly, we incorporate an Inference norm into the Lamoid's normative framework.

Inference: If a Lamoid detects that an instruction violates at least one Gricean norm, it applies the Inference norm to determine the optimal course of action.

2.2 Cognitive Frameworks

In collaborative interactions, *common ground* is crucial, as it represents the mutual understanding between participants [13, 39]. In human-agent collaboration, common ground enables agents to align with the human's goals by leveraging a mutual understanding of the environment and task objectives [9, 12, 17]. This mutual understanding helps the agent interpret instructions more accurately and anticipate the human's needs.

Relevance theory suggests that humans communicate by conveying information that is most relevant to the situation or goal [8, 44]. In human-agent interactions, relevance theory helps agents discern which parts of a conversation or instruction are essential for achieving the objective. This enables agents to focus on relevant details while filtering out extraneous information. In our framework, both common ground and relevance theory are incorporated into the Fs-CoT prompting mechanism and the normative structure.

Theory of mind refers to the capacity to recognize that others have their own mental states (e.g., beliefs and intentions) which might differ from one's own. LLMs have been found to perform well in theory of mind tasks [22, 40]. Theory of mind enables agents to assess the human's mental state. This ability is useful in situations where the human's instructions are unclear. In our framework, theory of mind is embedded in the Fs-CoT prompting mechanism. It is also integrated into the normative structure of a Lamoid, particularly through the Inference norm.

When paired with common ground and relevance theory, theory of mind enables a Lamoid to establish a mutual understanding between itself and the human, while also determining which pieces of information are most relevant to the task. For instance, if the human provides an ambiguous instruction, a Lamoid can infer what the human likely meant based on their shared model of the environment (common ground), the human's goals (relevance theory), and a model of the human's mental states (theory of mind). This combination of the cognitive frameworks enables Lamoid to interpret both clear and unclear instructions, adapt its responses, and execute optimal actions.

3 INSTRUCTION PROCESSING PIPELINE

We identify five types of instructions the human may issue, based on the associated Gricean norm violations.

(1) Clear instructions contain all the necessary details for the Lamoid to perform the task unambiguously and without error. No Gricean norms are violated.
(2) Incomplete instructions lack specificity, leaving gaps that the Lamoid must infer to determine the correct course of action. This results in a Quantity violation.
(3) Invalid instructions direct the Lamoid to carry out tasks that are impossible or unachievable within the given constraints. This results in a Quality violation.
(4) Irrelevant instructions have no direct connection to the task, making them non-contributory to the human's goals. This results in Relation a violation.
(5) Ambiguous instructions are open to multiple interpretations, requiring Lamoid to decide between several possible actions. This results in a Manner violation.

By classifying instructions into these five types, we systematically assess the Lamoid's ability to address challenges in human communication. Table 1 outlines the relationship between instruction types and violations of Gricean norms.

Table 1: Instruction types and violation of Gricean norms.

	Instruction Type	*Violation of Gricean Norm*
Clear		No Violation
Unclear	Incomplete	Quantity Violation
Unclear	Invalid	Quality Violation
Unclear	Irrelevant	Relation Violation
Unclear	Ambiguous	Manner Violation

The Lamoid processes instructions by interpreting the human's instruction. Using the normative framework outlined in Section 2, it discerns intent and extracts the most relevant information to complete the task. The pipeline then classifies Lamoid's responses into two primary categories:

(1) Optimal Actions: For clear instructions or those with minor gaps (incomplete), the Lamoid generates the most efficient sequence of actions by relying on the cognitive frameworks, and the *Inference* norm. It ensures that its responses are contextually appropriate and aligned with both the instruction and the environment.
(2) Clarification Actions: When faced with invalid, irrelevant, or ambiguous instructions, the Lamoid requests clarification. In these cases, it presents contextually relevant options, guiding the human toward a clearer directive and maintaining focus on the task's objectives.

The Lamoid's decision-making begins by interpreting the instruction through the normative framework. Based on this interpretation, the Lamoid either generates an optimal action sequence or requests clarification by providing alternatives that align with the task and the human's goals.

4 LLMS AND PROMPT DESIGN

We used GPT-4 with Fs-CoT [38] prompting techniques in a Lamoid's execution model. The following sections expound on GPT-4's configuration and the prompt design.

4.1 GPT-4 Model Parameters

Lamoids leverage GPT-4, an LLM known for its ability to generate coherent, contextually relevant responses [3, 10, 32]. We configure the maximum token limit to 512 to balance between comprehensive output and processing efficiency. We set the temperature parameter to 0.2, reducing output variability. This configuration was chosen to ensure the reproducibility of a response by minimizing the likelihood of generating irrelevant or overly creative responses [3].

4.2 Prompt Architecture

The prompt is divided into four components:

General Chain-of-Thought: The first component of the prompt provides a Lamoid with the details of the environment, including background, roles, objects, and objectives. This component establishes the necessary common ground between the human and the Lamoid. It is designed to guide the Lamoid in interpreting instructions that can be clear, incomplete, invalid, irrelevant, or ambiguous. The chain-of-thought prompting mechanism is used to incorporate relevance theory and theory of mind into the Lamoid's reasoning process, enabling a more nuanced understanding of the human's intent and objectives. Consequently, this component incorporates cognitive frameworks to enhance contextual interpretation.

LLMs often struggle with tasks requiring information extraction and spatial reasoning in grid-based systems [1, 28, 45]. Given that LLMs are inherently dependent on textual data [7, 11, 21], a Lamoid is presented with the grid configuration in two complementary formats: a *visual* adjacency matrix and a *textual* description detailing object locations and quantities. To further reinforce the importance of the cognitive frameworks, a set of key rules is incorporated into the prompt to enhance the Lamoid's reasoning strategies.

Response Generation: With Norms

*Based on your understanding of the **norms**, common ground, interpretation strategies, grid layout, and the labeled examples provided below, generate the norm violation type and response for the following human instruction, delimited by triple backticks:*

````
```
Instruction: Pick up the red key.
Norm: <Identify the norm violation type and provide
chain-of-thought reasoning for the violation>
Response: <Generate chain-of-thought reasoning for
the response>
```
````

(a) With norm-aligned responses.

Response Generation: Without Norms

Based on your understanding of common ground, interpretation strategies, layout of the grid, and, labeled examples below, generate a response for the following human instruction delimited by triple backticks:

````
```
Instruction: Pick up the red key.
Response: <Generate chain-of-thought reasoning for
the response>
```
````

(b) Without norm-aligned responses.

Figure 2: Third component of the prompt: norm-driven vs. non-norm-driven response generation by a Lamoid.

Few-shot CoT Exemplar with Norms

*Use the following examples, delimited by triple quotes, to understand how to generate the appropriate response for each instruction. These examples are based on different grid configurations. Follow the structure and format shown in these examples when generating both the **'Norm'** and **'Response'**.*

```
'''
Instruction: Can you get the green key?
Norm: Quality Violation. Based on the instruction and the
grid, there is no green key, making this an invalid
instruction.
Response: There is no green key on the grid. Do you want me
to collect the yellow key, the red key, or both?
...

'''
```

(a) With norm-guided few-shot examples.

Few-shot CoT Exemplar without Norms

*Use the following examples, delimited by triple quotes, to understand how to generate the appropriate response for each instruction. These examples are based on different grid configurations. Follow the structure and format shown in these examples when generating the **'Response'**.*

```
'''
Instruction: Can you get the green key?
Response: There is no green key on the grid. Do you want me
to collect the yellow key, the red key, or both?
...

'''
```

(b) Without norm-guided few-shot examples.

Figure 3: Fourth component of the prompt: few-shot CoT exemplars with and without norm.

Gricean and Inference Norms: The second component of the prompt centers on the incorporation of Gricean norms and the Inference norm, leveraging the cognitive frameworks detailed in Section 2.2. These norms, informed with the cognitive frameworks, evaluate instructions based on the grid configuration (common ground), identify Gricean norm violations (relevance theory), interpret the human's implicit intentions (theory of mind), and infer the most appropriate response in case of a norm violation. When a norm violation is detected, the Lamoid applies the Inference norm to infer the most relevant action. For example, if an instruction is incomplete (Quantity violation), the Lamoid infers the missing information by considering the grid's layout and the human's intention of retrieving a gem. If the instruction is ambiguous (Manner violation), the Lamoid generates clarification options that are contextually relevant to ensure a smooth interaction.

Response Generation: The third component of the prompt outlines the response generation process, detailing how a Lamoid interprets instructions, detects Gricean norm violations, and generates contextually appropriate responses. It also contains the human's instruction as input. Chain-of-thought reasoning is employed in both generating responses and identifying norm violation types. This reasoning process integrates cognitive frameworks and Gricean norms, ensuring coherent and contextually relevant responses. Figure 2 shows two response generation templates: response generation with the identification of norm violations (Figure 2a), and response generation without explicit reference to these norms (Figure 2b).

Few-shot with Chain-of-Thought: The fourth component of the prompt consists of input-output pairs as few-shot demonstrations, which serve as illustrative examples to guide a Lamoid in handling clear and unclear instructions. We incorporated a total of *14* few-shot demonstrations with chain-of-thought in this component [30]. Prior research [6] indicates that between 10 and 32 demonstrations tends to yield optimal performance for a few-shot learning. Through iterative experimentation, we determined that 14 demonstrations provide the best balance between efficiency and performance for our specific task. Figure 3 highlights two forms of these few-shot examples: one set using norm-driven interpretations (Figure 3a) and the other devoid of such norms (Figure 3b). These different prompt configurations were employed in distinct experimental settings, which are discussed in detail in Section 5. The full prompt, incorporating all components, is provided in the *supplement*.

5 EXPERIMENTS

To evaluate the influence of Gricean norms and address our research questions, we conducted an experiment using two versions of a Lamoid: *Lamoid with norms*, which incorporated normative elements in its prompts (Figures 2a and 3a), and *Lamoid without norms*, which omitted these elements (Figures 2b and 3b). Comparing these versions allowed us to isolate and evaluate the influence of Gricean norms in facilitating effective human-agent collaboration. The experiment was conducted within the multiagent Doors, Keys, and Gems (mDKG) cooperative planning domain [51].

We used 25 grid configurations, as previously created by Zhi-Xuan et al. [51]. However, unlike prior work, which focused solely on ambiguous instructions, we expanded the instruction set to include a broader range of instruction types, as outlined in Table 1. Table 2 details the distribution of instruction types across the 25 grid problems. By testing both experimental conditions across these 25 configurations and 55 instructions, we systematically evaluated the Lamoid's performance in instruction interpretation and task execution accuracy.

Table 2: Instruction type distribution.

Instruction Type	*Count*
Clear	20
Incomplete	5
Invalid	11
Irrelevant	6
Ambiguous	13

Example (1) Response Generation: With Norms

Instruction: *Can you pass me the red keys?*
Norm: *Quantity Violation.* The instruction is asking to pass the red keys, but it doesn't specify how many. Given the grid configuration, there are two red doors and three red keys. If the human wants to collect the gem at (4,8), they would need to unlock two red doors. Therefore, two red keys are required, but the instruction doesn't specify this, resulting in a Quantity Violation.
Response: There are three red keys and two red doors on the grid. To collect the gem at (4,8), two red doors need to be unlocked. I will collect the two red keys closest to me at (1,1) and (1,3), and pass them to the human to unlock the red doors.

(a) Norm-aligned response generated by Lamoid.

Example (1) Response Generation: Without Norms

Instruction: *Can you pass me the red keys?*
Response: There are three red keys and two red doors on the grid. The instruction is asking to collect red keys, which implies that the human wants to retrieve a gem placed beyond a red door. Considering the grid configuration, the gems at (4,8) and (6,8) are behind the red doors. I will collect the red key at (1,1) which is closest to me and pass it to the human. The human can then unlock the red door at (8,4) or (9,4) to retrieve the gem.

(b) Non-norm-aligned response generated by Lamoid.

Figure 4: Responses of Lamoid with norms vs without norms.

5.1 Performance Metrics

We adopt the following metrics. For each metric, we indicate the research question(s) it helps answer.

Instruction Interpretation Accuracy (RQ_2) captures whether the Lamoid correctly identified the norm violation associated with a given instruction type in the experiment where the norms were applied.

Task Accuracy (RQ_1) captures the ratio of correctly predicted actions by the Lamoid to the total number of correct actions, reflecting its task execution performance in human-agent collaboration.

Mean Response Length (RQ_3) captures the mean length of responses (in characters) generated by the Lamoid, reflecting its adherence to the Quantity norm.

Mean Options (RQ_3) captures the mean of the clarification options provided to the human when instructions are ambiguous, invalid, or irrelevant. This metric reflects the Lamoid's adherence to the Quantity norm.

Options Accuracy (RQ_1 and RQ_3) captures the accuracy of the options presented to the human, particularly in the cases of ambiguous, invalid, or irrelevant instructions. This metric reflects the Lamoid's adherence to the Quality norm.

Response Relevancy (RQ_1 and RQ_3) captures the alignment of the Lamoid's response with the human's objectives given the grid configuration. This metric reflects the Lamoid's adherence to the Relation norm. The responses are rated on a scale: 0 (Not Relevant), 0.25 (Minimally Relevant), 0.5 (Partially Relevant), 0.75 (Mostly Relevant), and 1 (Fully Relevant).

Response Clarity (RQ_1 and RQ_3) captures the clarity of the Lamoid's response, ensuring it is clear, unambiguous, and free from hallucinations or irrelevant information. This metric reflects the Lamoid's adherence to the Manner norm. The responses are rated on a scale: 0 (Not Clear), 0.25 (Minimally Clear), 0.5 (Somewhat Clear), 0.75 (Mostly Clear), and 1 (Perfectly Clear).

The metrics were annotated by the primary author. To illustrate how response relevancy and clarity were measured, Figure 4 presents a comparison between the Lamoid's responses with and without norms. The instruction provided was, "Can you pass me the red keys?" In the norm-aligned response (Figure 4a), Lamoid correctly identifies that two red keys are needed, as the gem is blocked by two red doors. The Lamoid then collects the two closest red keys and passes them to the human. Conversely, in the non-norm response (Figure 4b), the Lamoid fails to recognize that two keys are required and only hands one red key to the human.

The response relevancy for the Lamoid with norms was rated 1 (Very Relevant), and the clarity was rated 1 (Perfectly Clear). However, for the non-norm response, the relevancy was rated 0 (Not Relevant) because the response did not align with the human's goal or the grid configuration. In contrast, the clarity was rated 0.75 (Mostly Clear), as it correctly identified the need for a red key but

Table 3: Performance of the Lamoid with and without norms across the metrics.

	Accuracy in Understanding		Response			
Experiment	**Task**	**Options**	**Mean Length**	**Mean Options**	**Relevance**	**Clarity**
Norms	**95.27 %**	**97.74 %**	**227.73**	**2.55**	**96.36 %**	**96.82 %**
Without Norms	74.73 %	90.74 %	400.84	2.70	76.36 %	80.91 %

Table 4: Instruction interpretation accuracy.

	Precision	Recall	F1 Score
No Violation	1	0.85	0.92
Quantity Violation	0.63	1	0.77
Quality Violation	1	1	1
Relation Violation	1	1	1
Manner Violation	1	1	1

misinterpreted that the human's intention for red keys was for a single red door.

6 RESULTS AND DISCUSSION

Table 3 summarizes our results. Overall, the Lamoid with norms outperformed the Lamoid without norms across all metrics. We use these results to address our research questions.

RQ$_1$: Effective collaboration. The task accuracy exhibited a notable relative improvement of 27.48 % when Gricean norms were applied. Additionally, options accuracy, which evaluates clarification requests in response to invalid, irrelevant, or ambiguous instructions, improved by 7.71 % relative to the non-norm condition. These findings suggest that norms improve goal inference and the generation of appropriate actions. Furthermore, norms contributed to improvements in both response relevancy and clarity, achieving relative gains of 26.19 % and 19.67 % respectively, over the non-norm condition. These improvements can be largely attributed to the Lamoid's adherence to Gricean norms and its integration of cognitive frameworks within its reasoning process. By leveraging these elements, the Lamoid with norms produced responses that were more aligned with the human's objectives, while also minimizing instances of hallucinations and misinterpretations. Although the Lamoid without norms also incorporated the cognitive framework, it fell short of the precision and coherence demonstrated by its norm-guided counterpart.

For task accuracy, the paired t-test yielded a highly significant result ($t(54) = 4.92$, $p < 0.001$), indicating that the Lamoid performed better with norms than without. The effect size, measured by Cohen's d ($d = 0.66$), reflects a moderate to large improvement in task accuracy. Similarly, response relevancy and response clarity showed significant differences, with $t(54) = 3.67$, $p < 0.001$, and $t(54) = 3.33$, $p = 0.002$, respectively, and moderate effect sizes ($d \approx 0.5$). In contrast, options accuracy yielded a p-value of $p = 0.064$, yet demonstrated a moderate effect size ($d = 0.51$), suggesting a notable but less pronounced impact. Therefore, Lamoid with norms generated more contextually relevant, optimal, and clearer responses, with fewer instances of hallucinations. This suggests that Gricean norms are crucial in promoting effective collaboration by ensuring more precise interpretation and execution of instructions.

RQ$_2$: Accurate interpretation. The instruction interpretation accuracy in the Lamoid with norms exhibited strong performance, successfully identifying 94.55 % of norm violations across different instruction types. Table 4 presents the precision, recall, and F1-score for each type of norm violation for the Lamoid under the norm condition. Notably, the precision for Quantity violations is slightly lower than other categories due to occasional misclassification of No Violation instructions as Quantity violations. This misclassification also led to a slight decrease in recall for No Violations. However, the F1-scores for norm violations remained consistently high, underscoring the Lamoid's proficiency in accurately interpreting different instruction types. These results indicate that when guided by the normative framework, Lamoid effectively interprets different instruction types, reinforcing it's ability to process and appropriately respond to incomplete, invalid, irrelevant, or ambiguous instructions.

RQ$_3$: Fs-CoT prompting. The Lamoid with norms used Fs-CoT prompting to incorporate Gricean norms into its responses. Metrics such as mean options and response length, options accuracy, response relevancy, and clarity reflect Lamoid's adherence to the Quantity, Quality, Relation, and Manner norms, respectively. The options accuracy, response relevancy and clarity metrics exhibited superior performance in the Lamoid with norms over the non-norm condition, as demonstrated in RQ$_1$. This reinforces the influence of Fs-CoT prompting in enabling the Lamoid to adhere to the Quality, Relation, and Manner norms. Interestingly, the mean options presented across both conditions remained similar, although slightly lower with norms.

However, the mean response length in the Lamoid with norms was notably lower than without norms. This reduction may be attributed to the Lamoid's adherence to the Quantity norm, which facilitated more concise yet contextually relevant responses. Further, the mean response length showed the largest effect ($t(54) = 8.26$, $p < 0.001$), with a substantial Cohen's d of 1.11, highlighting the considerable influence of Fs-CoT in reducing verbosity and enforcing the Quantity norm. By prioritizing brevity without sacrificing informativeness, the Lamoid with norms generated responses that were sufficient and without unnecessary elaboration. Therefore, the Lamoid with norms generated responses that were more contextually relevant, clearer, and contained fewer errors in task execution, while maintaining an optimal balance in response length and number of options. By embedding Gricean norms through Fs-CoT, the Lamoid not only improved norm adherence but also ensured responses were truthful, complete, relevant, and unambiguous. Overall, these findings suggest that Fs-CoT is crucial in helping the Lamoid adhere to the norms, thereby improving the overall quality of human-agent collaboration.

We present two examples of response generation by the Lamoid with and without norms in Figures 4 and 5. In Figure 4 where the instruction required collecting red keys, the Lamoid with norms accurately identified two red keys were required. It retrieved the two nearest red keys and passed them to the human, demonstrating correct adherence to the task requirements. Conversely, the Lamoid without norms erroneously collected only one red key. Figure 5 further illustrates this distinction; when tasked to retrieve a red key, the Lamoid with norms passed the key to the human to unlock the specified red door. In contrast, Lamoid without norms misinterpreted the grid configuration and passed the key for a red door that was obstructed by another door, leading to an ineffective action.

Example (2) Response Generation: With Norms

Instruction: *Can you get the red key?*
Norm: *No Violation.* The instruction is clear, truthful, relevant, and unambiguous. The human wants to collect a gem that is likely behind a red door, hence the request for the red key. The grid contains a red key, so the instruction is valid and does not violate any norms.
Response: Sure, I will collect the red key at (0, 5) and pass it to you. You can then use it to unlock the red door at (2,1) and retrieve your desired gem.

(a) Norm-aligned response generated by Lamoid.

Example (2) Response Generation: Without Norms

Instruction: *Can you get the red key?*
Response: There is one red key on the grid located at (0, 5). The red key is required to unlock the red doors at (2, 1) and (3, 8). The gem that the human likely wants to retrieve is at (3, 9) as it is the only gem that is behind a red door. I will collect the red key and pass it to the human to unlock the red door at (3, 8) to retrieve the gem.

(b) Non-norm-aligned response generated by Lamoid.

Figure 5: Response for Lamoid with norms vs without norms.

7 RELATED WORK

The Gricean Maxims have been widely applied in human-AI interaction, multiagent decision-theoretic models, conversational agents, and tasks such as natural language generation (NLG) and natural language understanding (NLU). Vogel et al. [42] demonstrated that these maxims naturally emerge from decision-theoretic frameworks in multiagent systems, enhancing communication efficiency and collaborative performance. Similarly, Panfili et al. [34] highlighted that AI systems benefit from explicitly adhering to Gricean Maxims in human-AI interactions, particularly when handling ambiguous or complex user queries. Miehling et al. [29] explored the benefit of LLMs and Gricean Maxims in human-AI conversations, discovering that LLMs such as Llama3-8B-Instruct, Llama3-70B-Instruct, and Mistral-7B-Instruct often violated the relevance and quality maxims. However, fine-tuning these models to prioritize these norms led to better handling of ambiguous instructions, more contextually appropriate responses, and enhanced conversational depth.

Krause and Vossen [23] provide a comprehensive survey on the application of Gricean maxims in NLP, highlighting their potential to improve interaction quality while noting limitations in handling contextual and pragmatic variability. Hu et al. [19] evaluated LLMs, including GPT-2, Flan-T5, InstructGPT-3, and text-davinci-002, finding that larger models excel in pragmatic tasks but often default to literal interpretations and miss nuanced social norms. Yue et al. [47] confirmed that GPT-4 effectively understands conversational implicature, whereas most smaller LLMs struggle with interpreting nonliteral meanings in dialogue. Further, di San Pietro et al. [14] identified GPT-3.5's strengths in coherence but noted limitations in humor, metaphor, and the quantity maxim, emphasizing GPT-4's potential to address these gaps.

Building on this, our approach leverages GPT-4 to address gaps in pragmatic reasoning observed in smaller parameter LLMs, and shows a promise in applying Gricean norms in goal-oriented collaboration. In contrast, the literature on LLM-based agents to manage implicit user intentions is evolving. Strachan et al. [40] demonstrated that LLMs performed equally or sometimes even better in theory of mind tasks such as identifying indirect requests, false beliefs and misdirection. Additionally, Qian et al. [36] introduced the Intention-in-Interaction benchmark to improve LLM-based agents ability to handle implicit user intention.

8 LIMITATIONS AND FUTURE WORK

The implementation of a Lamoid within the Doors, Keys, and Gems environment posed several challenges, primarily due to the limitations of GPT-4 in spatial reasoning and pathfinding, and the demands of prompt engineering. This section outlines the key challenges encountered and offers potential solutions to improve the Lamoid's overall performance.

8.1 Limitations of GPT-4

A significant challenge arose due to GPT-4's poor performance in spatial reasoning and information extraction [27, 45]. Despite us providing the Lamoid with both visual and textual representations of the grid, it occasionally struggled to accurately interpret the spatial relationships between various objects. This hindered its ability to infer the correct actions required for task completion which resulted in suboptimal or incorrect responses. This limitation is intrinsic to GPT-4's design, as it was primarily trained on text, not spatial or geometric reasoning tasks [32]. A potential solution would be to integrate information extraction tools [1] that could enhance GPT-4's ability to interpret spatial data more effectively, thereby improving its decision-making in tasks requiring spatial reasoning.

Another critical limitation was GPT-4's inadequacy in optimal pathfinding [28]. This challenge arises from GPT-4's lack of inherent pathfinding algorithms, as its architecture is more suited for generating natural language than solving complex navigational problems. To address this, we shifted the model's output from generating direct actions to generating natural language descriptions of the actions to be taken. Although this change mitigated the issue to some extent, it did not fully resolve GPT-4's deficiency in path optimization. Leveraging dedicated pathfinding algorithms or incorporating hybrid models that combine GPT-4 with pathfinding tools may offer a more robust solution.

Hallucinations are a well-known challenge in LLMs [15, 20, 41]. Our Lamoid suffered from hallucinations as well, where it would generate actions that did not align with the actual configuration of the environment. Given GPT-4's reliance on text-based training, its inability to accurately interpret spatial data likely contributed to these hallucinations. Addressing this issue would require either more advanced model fine-tuning or the integration of spatial reasoning modules, which could help the model ground its responses in the correct physical context.

8.2 Prompt Engineering

The implementation required extensive prompt engineering [30] to guide GPT-4's behavior, a process that was both labor-intensive and nontrivial. This level of manual engineering is not sustainable for broader, more dynamic tasks, as it requires precise tuning for each specific scenario. This raises concerns about the task-agnostic nature of the current approach. Future work will involve designing more concise prompts and evaluating if fewer examples can produce similar performance outcomes. Pretraining language models on domain-specific data [35] or developing more flexible prompt templates could reduce the need for customization, enhancing the model's adaptability across different tasks.

8.3 Pragmatics and User Intentions

This study advances norm-driven pragmatic reasoning in LLM-agents but elides important aspects of pragmatics in human language [37], such as emotions, nonverbal cues, humor, and sarcasm. Additionally, variations in user intentions and subjective interpretations pose challenges for universal collaboration. Future work will focus on better modeling user subjectivity by collecting and analyzing data on diverse human interpretations and expectations. These insights will inform Lamoid's prompting mechanism, enhancing its ability to understand varying user intentions and strengthening its theory of mind capabilities. Furthermore, incorporating broader pragmatic elements, such as emotional intelligence and context-aware adjustments, into the norm-driven framework offers the potential to create more adaptable and effective AI agents. By expanding this framework to operate in diverse and dynamic domains, we aim to enhance the capacity of LLM-agents to collaborate seamlessly with humans across varied contexts.

9 CONCLUSION

Our study demonstrates how Gricean norms enhance human-agent collaboration. Incorporating Gricean norms through Few-shot Chain-of-Thought (Fs-CoT) prompting, and applying the cognitive frameworks, improves the Lamoid's ability to generate accurate and contextually relevant responses. The normative framework and Fs-CoT enable the Lamoid to adhere to Gricean and Inference norms, which substantially enhances interaction quality, as reflected in performance metrics such as task accuracy, response clarity, and relevance. Additionally, the Lamoid's ability to accurately handle various instruction types—whether incomplete, invalid, irrelevant, or ambiguous—highlights the effectiveness of structured norms in refining AI communication to align with human logic and expectations.

Despite these successes, challenges in spatial reasoning and path finding persist due to inherent limitations in LLMs when applied to spatial tasks. Directions for future work include integrating external tools for better information extraction and spatial reasoning, and developing a more robust Lamoid capable of handling complex spatial tasks. Looking ahead, exploring the scalability of norm-driven LLM applications across different operational domains is a promising area for further research. Continual refinement of these models is crucial for enhancing AI's capability to interact and innovate collaboratively with humans in achieving complex goals in dynamic environments.

ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation (grant IIS-2116751).

REFERENCES

[1] Mohamed Aghzal, Erion Plaku, and Ziyu Yao. 2023. Can Large Language Models Be Good Path Planners? A Benchmark and Investigation on Spatial-Temporal Reasoning. *CoRR* abs/2310.03249 (2023). https://arxiv.org/abs/2310.03249

[2] Nicholas Allott. 2013. *Relevance Theory.* Springer International Publishing, Cham, 57–98. https://doi.org/10.1007/978-3-319-01014-4_3

[3] Valentina Alto. 2023. *Modern Generative AI with ChatGPT and OpenAI Models: Leverage the Capabilities of OpenAI's LLM for Productivity and Innovation with GPT3 and GPT4.* Packt Publishing Ltd, Birmingham, UK.

[4] John L. Austin. 1962. *How to Do Things with Words.* Clarendon Press, Oxford.

[5] Gaurang Bansal, Vinay Chamola, Amir Hussain, Mohsen Guizani, and Dusit Niyato. 2024. Transforming Conversations with AI–A Comprehensive Study of ChatGPT. *Cognitive Computation* Springer (2024), 1–24. https://doi.org/10.1007/s12559-023-10236-2

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20).* Curran Associates Inc., Red Hook, NY, USA, 159. https://doi.org/10.5555/3495724.3495883

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, John A. Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuan-Fang Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. *CoRR* abs/2303.12712 (2023). https://arxiv.org/abs/2303.12712

[8] Robyn Carston. 2006. *Relevance Theory and the Saying/Implicating Distinction.* Wiley Online Library, The Handbook of Pragmatics, 633–656. https://doi.org/10.1002/9780470756959.ch28

[9] Joyce Y. Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative Effort Towards Common Ground in Situated Human-Robot Dialogue. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction.* Association for Computing Machinery, Germany, 33–40. https://doi.org/10.1145/2559636.2559677

[10] Edward Y. Chang. 2023. Examining GPT-4: Capabilities, Implications and Future Directions. In *The 10th International Conference on Computational Science and Computational Intelligence (CSCI).* Conference on Computational Science and Computational Intelligence, Las Vegas, Nevada, 7–14. https://doi.org/10.1109/CSCI62032.2023.00009

[11] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45. https://doi.org/10.1145/3641289

[12] Nazli Cila. 2022. Designing Human-Agent Collaborations: Commitment, Responsiveness, and Support. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* CHI, New Orleans, Louisiana, 1–18. https://doi.org/10.1145/3491102.3517500

[13] Eve V. Clark. 2015. *Common Ground.* Wiley Online Library, The Handbook of Language Emergence, 328–353. https://doi.org/10.1002/9781118346136.ch15

[14] Chiara Barattieri di San Pietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. The Pragmatic Profile of ChatGPT: Assessing the Communicative Skills of a Conversational Agent. *Sistemi Intelligenti* 35, 2 (2023), 379–400. https://doi.org/10.1422/108136

[15] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature* 630, 8017 (2024), 625–630. https://doi.org/10.1038/s41586-024-07421-0

[16] Herbert Paul Grice. 1975. Logic and Conversation. *Syntax and Semantics* 3 (1975), 43–58.

[17] Guy Hoffman and Cynthia Breazeal. 2004. Collaboration in Human-Robot Teams. In *AIAA 1st Intelligent Systems Technical Conference.* AIAA, Chicago, Illinois, 6434. https://doi.org/10.2514/6.2004-6434

[18] William S. Horton and Boaz Keysar. 1996. When Do Speakers Take into Account Common Ground? *Cognition* 59, 1 (1996), 91–117. https://doi.org/10.1016/0010-0277(96)81418-1

[19] Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A Fine-Grained Comparison of Pragmatic Language Understanding in Humans and Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Toronto, Canada, 4194–4213. https://doi.org/10.18653/v1/2023.acl-long.230

[20] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (2023), 1–38. https://doi.org/10.1145/3571730

[21] Katikapalli Subramanyam Kalyan. 2023. *A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4.* Elsevier, Natural Language Processing Journal, 100048. https://doi.org/10.1016/j.nlp.2023.100048

[22] Michal Kosinski. 2024. Evaluating Large Language Models in Theory of Mind Tasks. *Proceedings of the National Academy of Sciences* 121, 45 (2024), e2405460121. https://doi.org/10.1073/pnas.2405460121

[23] Lea Krause and Piek T. J. M. Vossen. 2024. The Gricean Maxims in NLP - A Survey. In *Proceedings of the 17th International Natural Language Generation Conference.* Association for Computational Linguistics, Tokyo, Japan, 470–485. https://aclanthology.org/2024.inlg-main.39

[24] Huao Li, Yu Quan Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Michael Lewis, and Katia Sycara. 2023. Theory of Mind for Multi-Agent Collaboration via Large Language Models. *CoRR* abs/2310.10701 (2023). https://arxiv.org/abs/2310.10701

[25] Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. 2024. Decision-Oriented Dialogue for Human-AI Collaboration. *Transactions of the Association for Computational Linguistics* 12 (2024), 892–911. https://doi.org/10.1162/tacl_a_00679

[26] Fabrizio Macagno and Sarah Bigi. 2018. Types of Dialogue and Pragmatic Ambiguity. In *Argumentation and Language — Linguistic, Cognitive and Discursive Explorations.* Springer, Cham, 191–218. https://doi.org/10.1007/978-3-319-73972-4_9

[27] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. GeoLLM: Extracting Geospatial Knowledge from Large Language Models. *CoRR* abs/2310.06213 (2023). https://arxiv.org/abs/2310.06213

[28] Silin Meng, Yiwei Wang, Cheng-Fu Yang, Nanyun Peng, and Kai-Wei Chang. 2024. LLM-A*: Large Language Model Enhanced Incremental Heuristic Search on Path Planning. *CoRR* abs/2407.02511 (2024). https://arxiv.org/abs/2407.02511

[29] Erik Miehling, Manish Nagireddy, Prasanna Sattigeri, Elizabeth M. Daly, David Piorkowski, and John T. Richards. 2024. Language Models in Dialogue: Conversational Maxims for Human-AI Interactions. *CoRR* abs/2403.15115 (2024). https://arxiv.org/abs/2403.15115

[30] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 11048–11064. https://doi.org/10.18653/v1/2022.emnlp-main.759

[31] Shrestha Mohanty, Negar Arabzadeh, Julia Kiseleva, Artem Zholus, Milagro Teruel, Ahmed Awadallah, Yuxuan Sun, Kavya Srinet, and Arthur Szlam. 2023. Transforming Human-Centered AI Collaboration: Redefining Embodied Agents Capabilities through Interactive Grounded Language Instructions. *CoRR* abs/2305.10783 (2023). https://arxiv.org/abs/2305.10783

[32] Humza Naveed, AU Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian. 2023. A Comprehensive Overview of Large Language Models. *CoRR* abs/2307.06435 (2023). https://arxiv.org/abs/2307.06435

[33] Sanjay Oruganti, Sergei Nirenburg, Marjorie McShane, Jesse English, Michael K. Roberts, and Christian Arndt. 2024. HARMONIC: Cognitive and Control Collaboration in Human-Robotic Teams. *CoRR* abs/2409.18047 (Sept. 2024). https://arxiv.org/abs/2409.18047

[34] Laura Panfili, Steve Duman, Andrew Nave, Katherine Phelps Ridgeway, Nathan Eversole, and Ruhi Sarikaya. 2021. Human-AI Interactions through a Gricean Lens. *Proceedings of the Linguistic Society of America* 6, 1 (2021), 288–302. https://doi.org/10.3765/plsa.v6i1.4971

[35] Rajvardhan Patil and Venkat Gudivada. 2024. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Applied Sciences* 14, 5 (2024), 2074.

[36] Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Yankai Lin, Zhong Zhang, Zhiyuan Liu, and Maosong Sun. 2024. Tell Me More! Towards Implicit User Intention Understanding of Language Model Driven Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Bangkok, Thailand, 1088–1113. https://doi.org/10.18653/v1/2024.acl-long.61

[37] Marina Sbisà. 2023. *Speech Acts and Other Topics in Pragmatics.* Oxford University Press, Oxford, GB.

[38] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic Prompting: Generating Chain-of-Thought Demonstrations for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202).* PMLR, Honolulu, Hawaii, USA, 30706–30775. https://doi.org/10.5555/3618408.3619681

[39] Robert Stalnaker. 2002. Common Ground. *Linguistics and Philosophy* 25, 5/6 (2002), 701–721. https://doi.org/10.1023/A:1020867916902

[40] James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. *Testing Theory of Mind in Large Language Models and Humans.* Nature Publishing Group UK, London, Nature Human Behaviour, 1–11. https://doi.org/10.1038/s41562-024-01882-z

[41] S. M. Tonmoy, S. M. Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *CoRR* abs/2401.01313 (2024). https://arxiv.org/abs/2401.01313

[42] Adam Vogel, Max Bodoia, Christopher Potts, and Dan Jurafsky. 2013. Emergence of Gricean Maxims from Multi-Agent Decision Theory. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL).* Association for Computational Linguistics, Atlanta, Georgia, 1072–1081. https://aclanthology.org/N13-1127

[43] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22).* Curran Associates Inc., Red Hook, NY, USA, 1800. https://doi.org/10.5555/3600270.3602070

[44] Deirdre Wilson and Dan Sperber. 2006. *Relevance Theory.* Wiley Online Library, The Handbook of Pragmatics, 606–632.

[45] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).* Association for Computational Linguistics, Mexico City, Mexico, 1819–1862. https://doi.org/10.18653/v1/2024.naacl-long.102

[46] Lance Ying, Tan Zhi-Xuan, Vikash Mansinghka, and Joshua B. Tenenbaum. 2023. Inferring the Goals of Communicating Agents from Actions and Instructions. *Proceedings of the AAAI Symposium Series* 2, 1 (2023), 26–33. https://doi.org/10.1609/aaaiss.v2i1.27645

[47] Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do Large Language Models Understand Conversational Implicature? A Case Study with a Chinese Sitcom. In *China National Conference on Chinese Computational Linguistics.* Springer, Shanghai, China, 402–418. https://aclanthology.org/2024.ccl-1.98/

[48] Francisco Yus. 2023. *Relevance Theory, Humour and Internet Communication.* Springer, Cham, 9–58. https://doi.org/10.1007/978-3-031-31902-0_2

[49] Shao Zhang, Xihuai Wang, Wenhao Zhang, Yongshan Chen, Landi Gao, Dakuo Wang, Weinan Zhang, Xinbing Wang, and Ying Wen. 2024. Mutual Theory of Mind in Human-AI Collaboration: An Empirical Study with LLM-Driven AI Agents in a Real-Time Shared Workspace Task. *CoRR* abs/2409.08811 (2024). https://arxiv.org/abs/2409.08811

[50] Tan Zhi-Xuan, Jordyn Mann, Tom Silver, Josh Tenenbaum, and Vikash Mansinghka. 2020. Online Bayesian Goal Inference for Boundedly Rational Planning Agents. In *Advances in Neural Information Processing Systems,* Vol. 33. Curran Associates Inc., Red Hook, NY, USA, 19238–19250. https://doi.org/10.5555/3495724.3497338

[51] Tan Zhi-Xuan, Lance Ying, Vikash Mansinghka, and Joshua B. Tenenbaum. 2024. Pragmatic Instruction Following and Goal Assistance via Cooperative Language-Guided Inverse Planning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Richland, SC) *(AAMAS '24).* International Foundation for Autonomous Agents and Multiagent Systems, Auckland, New Zealand, 2094–2103. https://doi.org/10.5555/3635637.3663074