Understanding Question Effectiveness: A Computational and Pragmatic Approach

Tanmay Mendhey
Email: tanmaytammy1993@gmail.com

***Abstract*—Effective communication relies on well-structured questions that elicit informative responses. This study explores the relationship between question formulation and the informativeness of answers, integrating insights from pragmatics, psycholinguistics, and computational modeling. Using a probabilistic reasoning framework, I analyze how contextual cues influence response quality and develop a predictive model for evaluating question effectiveness. Experimental results demonstrate that certain question structures systematically yield more informative answers. These findings contribute to improving question design in human dialogue, automated systems, and survey methodologies.**

***Index Terms*—pragmatics, computational modeling**

I. Introduction

Q:"Are you gonna eat that apple?"
A:"Oh, go ahead!"

In this dialogue, Q is inquiring if A will eat the apple, but instead of directly requesting the apple, Q cleverly frames the question in a way that avoids being impolite while subtly signaling A about her intentions. A, understanding Q's true purpose, responds in a way that addresses the underlying request rather than the question's surface meaning. This exchange illustrates the complexity of social reasoning within everyday conversations and brings forth two key questions for formal language models: What makes a question useful? And what constitutes an appropriate answer?

Psycholinguistic research has shown that respondents are often aware of the questioner's underlying goals and aim to provide answers that are relevant to these goals. A seminal study by Clark (9) involved a scenario where liquor merchants were contacted by phone. They were first told one of two context-setting phrases: "I want to buy some bourbon" (the *uninformative* condition) or "I've got $5 to spend" (the *five dollar* condition). Following this, the merchants were asked, "Does a fifth of Jim Beam cost more than $5?" The results revealed that merchants tended to give a simple yes/no response much more frequently in the five dollar context, where an exact price was not expected. In contrast, when the context was merely about buying whiskey, more detailed price information was given. The merchants inferred that in the latter context, the questioner's goal was to determine affordability, while in the former, the questioner was more interested in purchasing the whiskey, making precise price information more relevant (9).

The significance of context and the questioner's goals can also be observed in responses to questions that request identification, such as "who is X?" (1), or requests like "Do you have the time?" where answers can vary in terms of precision (2; 3). Similarly, wh-questions like "Who passed the examination?" allow for answers that may be either *exhaustive* or *selective* in nature, and questions like "where are you?" can elicit answers with varying degrees of abstraction (4). These instances highlight the importance of social inference in both asking and answering questions.

Recent developments in Rational Speech Act (RSA) models (5; 6) have mathematically captured the pragmatic understanding of language through recursive Bayesian inference. In this framework, speakers are considered to choose utterances that maximize the information obtained by an imagined listener. In this paper, I extend the RSA framework to apply it to question-answering dialogues.

A challenge within this framework arises from the fact that, unlike statements, questions do not directly provide information. Therefore, we must define what utility a question can have.

I propose, following (15), that the utility of a question lies in how much information it can be expected to generate about the questioner's interests later in the conversation. More specifically, the value of a question for the questioner is defined by the expected information that can be gained about her concerns based on the set of possible answers the question might generate. This differs from the traditional RSA model, in that it emphasizes the information gained by the speaker (rather than the listener), and the relevance of this information is realized later within the brief conversation.

To comprehensively encapsulate this concept, it is essential to formulate a model of the respondent, which functions both as the conceptual framework assumed by the inquirer and as the mechanism underlying response generation. I analyze three progressively sophisticated representations of the respondent. The most fundamental model entails a direct response to the query, where the respondent does not seek to deduce the inquirer's underlying intent. The second model, referred to as the explicit respondent, strives to generate a response that is informative in relation to the literal phrasing of the query, without delving into deeper interpretative assumptions about the inquirer's intent. The third and most advanced model, known as the pragmatic respondent, discerns the true underlying intent of the inquirer and formulates a response that directly aligns with these inferred objectives. This advanced model builds upon RSA to integrate considerations of the dialogue's thematic structure, as originally suggested by (7) to account for the phenomenon of hyperbole. It advances prior research by treating the explicit query as an indirect yet

informative signal regarding the broader subject of discourse.

The structure of this paper is as follows. I begin by detailing the computational models for both questioner and answerer, highlighting the similarities and differences compared to other recent approaches. Following this, I present a series of computational experiments that illustrate how the models effectively account for four classic answerer-sensitivity effects observed in the psycholinguistics literature. Given the limited data on *questioner* behavior, and the fact that the classic effects did not sufficiently consider the questioner component of the model, I devised a communication game paradigm. This allowed me to manipulate private goals, possible questions, and potential answers.

After evaluating the model predictions on data from a simple single-player version of the task, I extend this paradigm to a more natural, real-time, multi-player experiment, introducing a broader range of goal sets, question sets, and answer sets. Particular focus is given to a crucial condition in the task where the questioner models diverge in their predictions, thus facilitating a comparison between them.

The paper concludes with a brief discussion on the methodology, situating language use within the realm of social cognition, and offering perspectives on key future research directions.

II. A Rational Speech Act Model of Questioning and Answering Behavior

How should an inquirer determine which question to pose? I assume that the inquirer aims to *gain information pertinent to a defined objective*. To formulate a question that yields valuable insights, the inquirer evaluates how the respondent would answer under different possible states of the world. The question is then chosen to maximize the likelihood of obtaining a response that assists in fulfilling the objective.

More formally, consider a set of possible world states W, a collection of objectives G, a range of potential questions Q, and a set of anticipated responses A. These sets are assumed to be mutually known between the inquirer and the respondent. An informational objective $g \in G$ acts as a projection function, mapping a world state to a particular feature or subset of features that are relevant to the inquirer's interest. This concept aligns with the idea of a question-under-discussion (8). I denote $P_g(w)$ as the probability $\hat{P}(g(w))$ of the feature pertinent to g under world state w, derived from the projected probability distribution:

$$\hat{P}(v) = \int_W \delta_{v=g(w)}\, P(w)dw.$$

Inquirer Model:

The **inquirer** receives an objective $g \in G$ and returns a probability distribution over possible questions $q \in Q$:

$$P(q|g) \propto e^{\mathsf{E}_{P(w)}[D_{KL}(P_g(w|q,w)\,\|\,P_g(w))]-C(q)}$$

Here, the inquirer weighs the cost $C(q)$ of posing a question against the expected information gain. The cost is likely influenced by factors such as the complexity or length of the question. The information gain is measured via the Kullback-Leibler divergence between the prior probability of g-relevant world states, $P_g(w)$, and the posterior distribution after posing a question q and receiving a response indicating the true world state w:

$$P_g(w \mid q, w) = \sum_{a \in A} P_g(w \mid q, a) P(a \mid q, w)$$

Firstly, it incorporates $P(a|q, w)$, a model of the respondent's answering behavior, which I will elaborate on below. Secondly, it involves $P(w|q, a)$, an "interpreter" that determines the probability of different world states given a question-answer pair.

To formally define the **interpreter** function used by all agents to determine the literal meaning of a question-answer pair, we assume that a question represents an informational goal that maps possible worlds onto an answer set, denoted as A. This approach aligns with the partition semantics of questions introduced by (22). Specifically, considering the pre-image of such a mapping, an answer designates an element of the partition, expressed as $q^{-1}(a)$.

The interpreter constrains the prior world distribution to only those worlds that remain consistent with the semantics of a given question-answer pair:

$$P(w|q, a) \propto P(w)\delta_{q(w)=a} \quad (1)$$

Next, I outline three distinct answerer models, each of which offers a different perspective on how a questioner may interpret responses, thereby leading to different instantiations of the questioner model. These answerers take a question $q \in Q$ and a true world state $w \in W$ as inputs and produce a probability distribution over possible answers $a \in A$.

The extbfliteral answerer generates answers by balancing the prior probability of an answer and how well a given question-answer pair conveys the actual world state to the interpreter:

$$P(a|q, w) \propto P(a)P(w|q, a) \quad (2)$$

For a fixed question, this model closely parallels the speaker component in prior Rational Speech Act (RSA) frameworks, where the question solely determines the literal interpretation of the response.

The extbfexplicit answerer evaluates responses based on how effectively they resolve the explicit question q:

$$P(a|q, w) \propto P(a)P_q(w|q, a) \quad (3)$$

The extbfpragmatic answerer does not take the question's literal meaning at face value but instead considers the broader informational goal it likely serves. By reasoning over the probable goals g that could have motivated the question q, the pragmatic answerer selects answers that optimize communication on average:

$$P(a|q, w) \propto P(a) \sum_{g \in G} P(g|q)P_g(w|q, a) \quad (4)$$

To infer the likely goal g, the pragmatic answerer employs a Bayesian inversion of the explicit questioner model, relying on a prior over goals:

$$P(g|q) \propto P(q|g)P(g) \quad (5)$$

Each questioner and answerer model may differ in their level of optimization—that is, the degree to which they sample from the defined distributions as opposed to deterministically selecting the most probable response. To regulate this selection process, I introduce an optimality parameter α, modifying the probability distributions as follows:

$$P'(x) \propto P(x)^{\alpha} \quad (6)$$

This formulation establishes the model space, comprising three distinct answerer models and their corresponding questioner models. The implementation of these models is carried out using WebPPL, a probabilistic programming language (12).

A. Related Models

The probabilistic model described above for question and answer behavior shares similarities with recent decision-theoretic (16) and game-theoretic (17) models of pragmatic reasoning. All of these approaches focus on *reasoning* and *inference* regarding a conversational partner's mental state. In contrast, these theories diverge from the widely recognized interactive alignment model (13) and other dynamical systems-based models of dialogue (10), where coordination is achieved through low-level processes such as priming and adjusting to a partner's syntactic, lexical, and phonological choices.

Although extensive research has been conducted on dialogue models, comparatively little focus has been directed toward the mechanisms governing question-and-answer interactions. The predominant theoretical frameworks in this area stem from formal linguistic principles, notably the concept of informativeness. Groenendijk and Stokhof (22) introduced a foundational perspective on question-answer semantics, suggesting that posing a question segments the space of possible worlds into a partition, where each subset represents a plausible response. An answer gains utility by eliminating subsets from this partition, with the most informative responses ruling out all but the actual world. However, as noted by van Rooy (15) and further discussed by others (11), this model implies that resolving a wh-question such as "Where can I buy an Italian newspaper?" would necessitate an exhaustive enumeration of all locations where such newspapers are available—an impractical expectation. In reality, identifying just one accessible location would be sufficient. Furthermore, these models do not adequately capture the influence of context in determining what qualifies as a meaningful answer.

In response, more recent theories have attempted to resolve these issues by incorporating the questioner's goals. For example, (15) formalizes these goals as a decision problem faced by the questioner. According to this decision-theoretic approach, an answer is useful if it maximizes the expected value of the questioner's decision problem, and a useful question induces a partition fine enough to optimally distinguish the worlds that matter to the decision at hand. While this framework accounts for the context-dependence and relevance-maximization of question and answer interactions, it assumes that the answerer knows the questioner's decision problem in advance. This assumption, however, raises a question: if the answerer already knows the questioner's problem, why would the questioner even need to ask the question?

Our models aim to extend this core concept within a probabilistic framework, which provides an inferential mechanism for the answerer to deduce the "decision problem" rather than assuming it is known. In the following, I will show how these models address four classic case studies of question and answer pragmatics from psycholinguistics.

III. Four Case Studies

A. Clark (1979), Experiment 4

To begin, I illustrate how the model produces varied responses to the same query depending on the surrounding context, sometimes providing excessively detailed or insufficiently informative answers. As an example, I replicate the whiskey-pricing study introduced in the Introduction (9). As highlighted earlier, liquor vendors were more inclined to give detailed responses (e.g., explicitly stating the price) when asked "Does a fifth of Jim Beam cost more than $5?" in a general context (e.g., "I want to purchase some bourbon") rather than in a highly specific context (e.g., "I have exactly $5").

In this framework, the world state represents the whiskey's price, which spans values from $1 to $10. Two potential objectives exist: identifying the precise whiskey price and determining whether it exceeds $5. The available responses include exact price statements and binary answers "yes" or "no," with the latter being less costly. The contextual statement shapes the questioner's goal likelihood: if the context is "I want to purchase whiskey," both goals are equally probable. Conversely, if the context is "I have $5 to spend," the probability of merely wanting to know whether the price surpasses $5 increases to 9:1.

1) Findings: When the question "Does Jim Beam cost more than $5?" is posed, the most frequent response is the correct Boolean answer, occurring with probabilities of 0.44 and 0.49. Crucially, the dependence of the response on context is evident: when the question follows "I want to buy some whiskey," the likelihood of stating the exact price is higher (at 0.18) compared to when the context is "I have $5 to spend" (where it drops to 0.11). In contrast, the literal and explicit models, which lack contextual sensitivity, do not exhibit variation between these scenarios. The literal model assigns equal probability to providing the correct Boolean and numerical answers, while the explicit model consistently predicts the true Boolean response, irrespective of context. This demonstrates that the pragmatic *responder* aligns with human behavioral patterns in psychologically relevant situations, successfully passing an initial qualitative validation test.

B. Groenendijk and Stokhof (1984)

A more complex case is the *mention-some* versus *mention-all* interpretations of wh-questions (14; 22). Some questions, like "Who is coming to dinner tonight?", request an exhaustive list of answers. In contrast, other questions, like "Where can I find a bathroom in this building?", require only a single answer.

The question "Where can one buy an Italian newspaper?" can be interpreted differently depending on the context: a tourist likely wants to know the nearest location, while a businessperson looking to establish a newspaper distribution network would need a comprehensive list. The challenge is determining how the same question can have different interpretations in different contexts, which is handled by inferring the questioner's underlying goal.

In this model, the world state consists of four cafes, each defined by its distance from the speaker and whether it sells Italian newspapers. The two possible goals are determining the identity of the *nearest* cafe selling a newspaper or identifying *all* cafes with newspapers. The possible answers include all combinations of cafes (e.g., "cafe 1 and cafe 3" or "cafe 2, cafe 3, and cafe 4"), as well as the answer "none."

The prior over the answer utterances is set as follows: there is a 10% chance of answering "none," while the agent selects one cafe with a 50

1) Results: The world is set as follows:

```
world = {'cafe1' : [3, false],
         'cafe2' : [1, true],
         'cafe3' : [3, true],
         'cafe4' : [3, true]}
```

After executing the model for both contexts, I find that the most probable response in the "I'm new here" context is the single location "cafe2," with a probability of 0.56. In contrast, in the "I'm a businessperson..." context, the most probable response is the combination "cafe2 and cafe3 and cafe4," with a probability of 0.82. Importantly, cafe 1 was never assigned a probability in either context, as it did not sell Italian newspapers, demonstrating that the pragmatic answerer will not provide false information. Additionally, the nearest cafe was prioritized in the "tourist" context.

The literal and explicit answerers, as in the Clark example, do not account for context, and hence both incorrectly predicted that the context would have no impact on the preferred answer. The literal answerer predicted that all combinations of cafes 2, 3, and 4 would be given equally, while the explicit answerer predicted "cafe2" as the answer in all contexts. The key difference between this scenario and the Clark case lies in the structure of the world and the meaning of the answers, while the questioner and answerer functions remain unchanged.

C. Gibbs Jr. & Bryant (2008): Experiment 3

Previous research has demonstrated that individuals tend to round their responses to the nearest 5- or 10-minute increment when asked, "Do you have the time?", even when using a digital watch (2). Replicating this result, (3) conducted a follow-up study by introducing the context, "I have a meeting at 4:00," before posing the time-related question. Their findings showed a decrease in the tendency to round times as the time approached the deadline. Specifically, when the question was asked at 3:40, respondents were more likely to round to "3:45," but at 3:53, they would give the time to the exact minute. This behavior was explained through the lens of the questioner's goal: when facing an approaching deadline, a precise answer is crucial for decision-making, such as whether to hurry.

In this experiment, I treat the world state as the true current time. Unlike the previous scenarios with only two Question Under Discussion (QUD) elements, here, there is a *family of QUDs* and a fixed context regarding the appointment time. The QUDs are parameterized by a threshold time, where times below this threshold are rounded to the nearest 5 minutes, and times above this threshold are reported exactly. For instance, a QUD with a threshold set at 3:50 reflects the goal of learning the true time if it exceeds 3:50, but only requiring an approximate time if it is below 3:50. The prior over thresholds (τ = 3:30, 3:35, . . . , 3:55) is weighted by its proximity to the appointment time, reflecting the idea that as the appointment time nears, a more precise answer becomes more relevant.

The set of potential responses consists of times from 3:30 to 4:00, each with uniform probability. Importantly, the meaning of a given response is approximate, as it influences the "interpreter" component of the model by giving greater weight to world states close to the response. This allows for the generation of rounded answers in the first place.

1) Results: To evaluate this model, I compare two simulations. In both cases, the context is "I have an appointment at 4:00." In the first, the actual time is 3:34 (the 'appointment far' condition); in the second, it is 3:54 (the 'appointment near' condition). In their study, (3) observed that in the 'appointment far' condition, respondents are more likely to round to 3:35, while in the 'appointment near' condition, they are more likely to give the precise time of 3:54. Indeed, in the 'appointment far' condition, the pragmatic questioner model most frequently replies with the rounded time of '3:35' at a probability of 0.40, compared to the true time of '3:34' at 0.27. In contrast, in the 'appointment near' condition, the model most often provides the exact time '3:54' with a probability of 0.68, and rounds to '3:55' with a probability of 0.16 (see Figure 1).

D. Clark (1979): Experiment 5

This experiment offers a critical test for the questioner component of the model. In earlier computational experiments, the answerer's inferences were based solely on the context and the QUD prior, as the question space contained only a single element. However, one of the most innovative predictions of the RSA model is that the questioner's utterance choice should influence the pragmatic answerer's inferences about the underlying goals. While there is limited experimental work directly testing questioner behavior as a dependent variable, there are

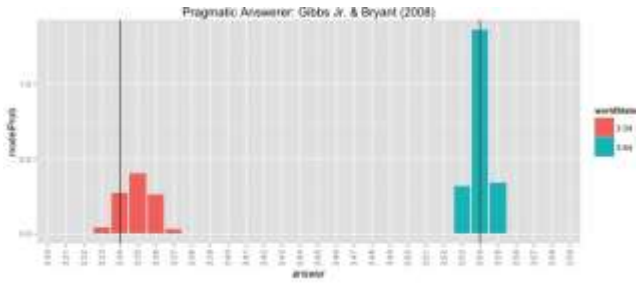

Fig. 1. Simulation results from the replication of Gibbs Jr. and Bryant (2008), showing the likelihood of different time answers. The vertical lines indicate the actual time in each scenario.

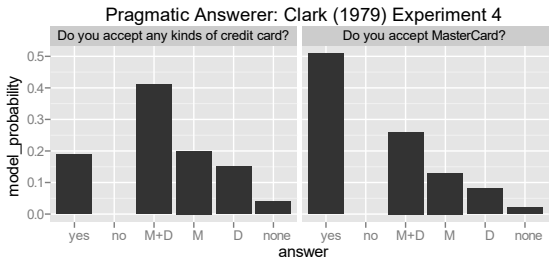

Fig. 2. Results of the computational experiment replicating Clark (1979), Experiment 4. The M label indicates responses regarding MasterCard, D for Diner's Club, and $M + D$ for answers that mention both cards. In the left panel, an over-informative answer is provided for the question, "Do you accept any kinds of credit card?" while the right panel shows a more literal yes/no answer to "Do you accept MasterCard?"

studies where the question asked serves as the independent variable, and the focus is on its impact on responses.

One such study is a follow-up to the first experiment modeled in this paper, where instead of calling liquor merchants, (9) contacted restaurants and posed one of four yes/no questions about the types of *credit cards* they accepted:

1) "Do you accept MasterCard?"
2) "Do you accept American Express?"
3) "Do you accept credit cards?"
4) "Do you accept any kinds of credit cards?"

The responses were analyzed to determine the likelihood of receiving a simple yes/no answer versus the more informative response listing all accepted cards. It was found that (1) and (2) almost always received a yes/no answer, (3) was equally likely to receive a yes/no or full list, and (4) nearly always prompted the full list.

I model this scenario as follows: the set of possible worlds is defined by the status of five different credit cards ('Visa', 'MasterCard', 'American Express', 'Diner's Club', and 'Carte Blanche'), with a Boolean indicating whether each card is accepted. The true world state is known by the answerer but not the questioner. The four questions form the question space, and the answer space includes 'yes', 'no', and all possible combinations of the cards accepted.

There are four possible goals depending on the questioner's situation. The first goal is the "MasterCard" goal, where the questioner only has a MasterCard and wants to know if it is accepted. The second goal, "Master + Diner's", arises when the questioner has both MasterCard and Diner's Club cards and wants to know if either is accepted. The third goal, "Master + Diner's + American", occurs when the questioner possesses MasterCard, Diner's Club, and American Express, and only needs to know if any of these cards are accepted. Finally, the "names" goal involves the questioner seeking the full list of accepted cards.

1) Results: The world state for testing the pragmatic answerer is as follows:

```
var world = {
        'Visa' : false,
        'MasterCard' : true,
        'AmericanExpress' : false,
        'Diners' : true,
        'CarteBlanche' : false
};
```

In this case, when the questioner asks, "Do you accept MasterCard?", the pragmatic answerer most likely responds with 'yes'. However, when the question is, "Do you accept any kinds of credit cards?", the most likely answer is the full list of accepted cards, as seen in Figure 2.

Upon examining the inferred QUD in each scenario, these results become clearer. For the "MasterCard" question, the pragmatic answerer deduces that the goal is to determine whether MasterCard is accepted, so a simple 'yes' is sufficient. For the "any kind" question, the answerer infers that the goal is to obtain the full list of cards, making a 'yes' response inadequate. As a result, the answerer chooses to provide the complete list instead. These inferences are based purely on the questioner's behavior within the model, rather than on broader contextual cues as in previous experiments.

E. Discussion

It is noteworthy that the same questioner-answerer model was able to replicate patterns of question-answer behavior across four distinct scenarios. This included both explicit and implicit contextual effects, as well as situations where the question itself acted as a signal for the underlying goals. All of these investigations primarily focused on answerer behavior, allowing us to demonstrate that the questioner model is consistently integrated as a submodule within the answerer. However, because questioner behavior was always manipulated as an independent variable, testing the questioner model as an independent predictor of human questioning behavior was not possible. This reflects the general oversight of questioner behavior in psycholinguistic studies, and additional experiments are necessary to assess the model's predictions.

Several complex questions remain regarding the model's behavior in these scenarios, especially the last one. Recall that Clark observed a difference between the questions "Do you accept any kinds of credit cards?" and "Do you accept credit cards?" whereas the model treats these two as identical. It

interprets both questions with literal semantics that yield true if any of the Boolean conditions in the object are true, and false otherwise, which subsequently influences the pragmatic answerer. The source of this asymmetry is not immediately clear. Another issue is that the answer distribution resulting from this model is not consistently stable across all parameter settings and world scenarios. The answer prior relies on a parameter to determine the likelihood of 'yes' or 'no' answers compared to lists of card types, and changing this parameter causes significant shifts in the answer distributions (though all distributions trend toward fewer 'yes'/'no' responses and more list-type answers when responding to the "any kinds" question). A more thorough exploration of the model's parameter space, alongside sensitivity tests to Question Under Discussion (QUD) alternatives, would be valuable.

Although the questioner component was pivotal in shaping the pragmatic answerer's responses in the final simulation, none of these experiments specifically tested the questioner component. In fact, much of the empirical literature treats the question or context as an independent variable, with the answer being the dependent one. The model, as suggested by (15), however, proposes that the question itself plays a significant role in eliciting a relevant answer. To evaluate these predictions, I devised a series of experiments using a communication game designed to collect data on both question-asking and question-answering behaviors.

IV. Exp. 1: Hierarchical Inquiry and Response

I created a guessing-game experiment with two participants—a questioner and an answerer—to examine how questioners formulate inquiries based on a specific objective and how responders respond under ambiguity regarding this objective. Four hidden animals—a dalmatian, a poodle, a cat, and a whale—were part of the game setup, and they were all placed behind different gates. According to a class hierarchy, these animals held different levels of positions (see Fig. 3). While the answerer knew where every animal was, they were not aware of the questioner's private objective, which was to identify a particular animal (e.g., "locate the poodle"). The questioner had to ask one question from a small list of premade questions before choosing a gate, and the answerer would then reveal the animal behind it. A key feature of the original scenario served as the inspiration for the question choice restriction: the questioner must use an indirect query when the most direct question (such as "Can I eat your food?") cannot be asked because of social norms, complexity, or other limitations.

The limitation on available questions played a crucial role in distinguishing between the pragmatic and explicit versions of the model. If every possible question were accessible, both models would invariably select the most direct inquiry. To explore the divergence in behavior when only a subset of questions was permitted, consider a case where the question "poodle?" was not an option. If the questioner instead asked about a "dog?", both a dalmatian and a poodle would be plausible responses for an explicit answerer, given their shared

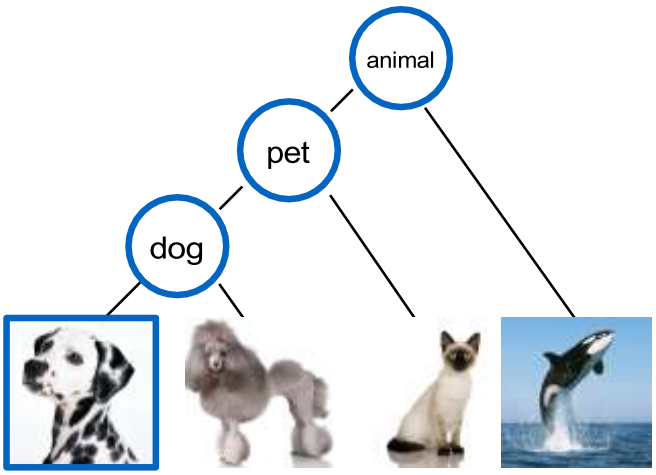

Fig. 3. Hierarchical structure in Exp. 1. The goal and response elements consisted of the four terminal nodes (animals) concealed behind gates (tree nodes). The set of available questions, however, was limited to the highlighted nodes within the hierarchy, promoting indirect questioning.

classification as dogs. However, a pragmatic answerer might infer that if the questioner were genuinely searching for the dalmatian, they would have directly asked about it. Since the questioner did not do so, it is likely that their interest lies in the poodle, the only other relevant option given the constraints on their questioning ability.

1) Participants: This experiment involved 125 participants recruited through Amazon Mechanical Turk. Eleven participants were excluded because they did not understand the instructions or were not native English speakers.

A. Stimuli & Procedure

The model's design included a world space W with 24 possible combinations of four objects assigned to four gates, resulting in $4! = 24$ different configurations. The goal space G contained the set of four objects the questioner may be attempting to identify, represented as the leaves of the decision tree in Fig. 3. The answer space A consisted of four gates from which the answerer could select. The restricted question space Q included the highlighted nodes of the decision tree: 'dalmatian?', 'dog?', 'pet?', and 'animal?'.

Each participant took part in eight trials—four as questioners and four as answerers—one trial for each goal and question, respectively. In the questioner block, participants were assigned a private goal from G, such as "find the poodle!" and were tasked with selecting the most useful question from a drop-down menu containing elements of Q. During the answerer block, participants were shown the assignment of objects to gates and were informed of the question posed by the other player. They were then asked to choose the gate that would give the most useful information to the questioner. To avoid any learning effects, questioners were not given answers, and neither role was able to see the results of the game. The order of the questioner and answerer blocks was randomly

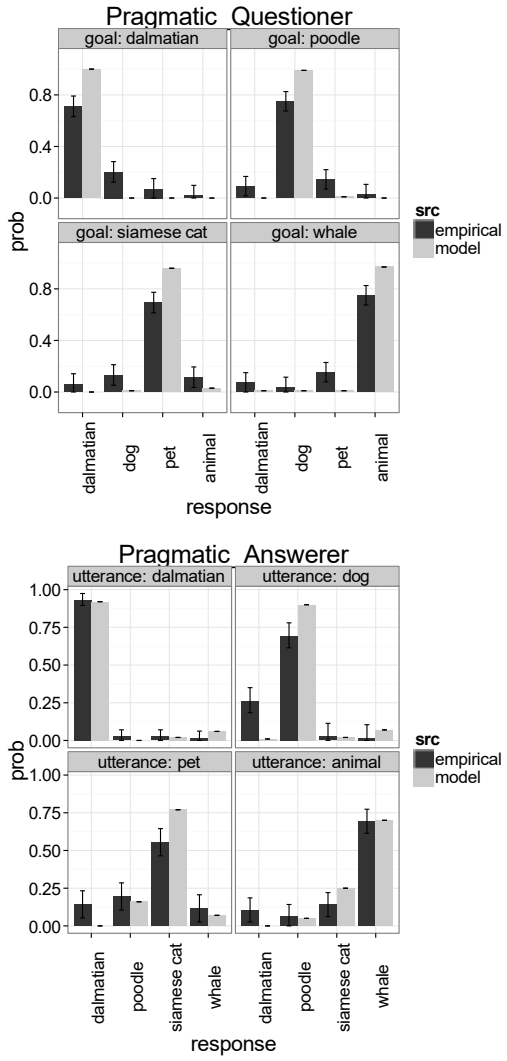

Fig. 4. Experiment 1's findings are contrasted with the most accurate model's predictions for questioners (left) and answerers (right). The pragmatic answerer model more closely matches the qualitative patterns found in the response data than the explicit answerer model, even though the two questioner models produce predictions that are comparable.

determined for each participant, and within each block, the stimuli were also presented in a random order.

1) Results: The outcomes for the questioner role are illustrated in Fig. 4 (left). The analysis indicates that questioners consistently ask distinct types of questions based on the objective, with a noticeable trend toward more indirect questioning as the goals change. Statistical evaluations revealed significant deviations from uniform distributions across the four response categories. Specifically, questioners showed a greater tendency to inquire about the 'dalmatian' when the objective was to locate the dalmatian, $\chi^2(3) = 137$, $p < .001$, about the 'dog' when the aim was to identify the poodle, $\chi^2(3) = 152$, $p < .001$, about the 'pet' when the goal was the cat, $\chi^2(3) = 120$, $p < .001$, and about the 'animal' when the target was the whale, $\chi^2(3) = 150$, $p < .001$.

The results for the answerer role are presented in Fig. 4 (right). Answerers exhibit high sensitivity to the constraints posed by the questioner, offering details about the dalmatian when asked about a 'dalmatian', $\chi^2(3) = 281$, $p < .001$, the poodle when inquired about a 'dog', $\chi^2(3) = 137$, $p < .001$, the cat when asked about a 'pet', $\chi^2(3) = 57$, $p < .001$, and the whale when queried about an 'animal', $\chi^2(3) = 121$, $p < .001$. Interestingly, while a literal interpretation of the question could allow revealing either the dalmatian or the poodle in response to a question about a 'dog', answerers consistently preferred disclosing the location of the poodle.

2) Model Comparison: Each model was evaluated with uniform prior probabilities across the worlds, goals, questions, and answers, assigning equal costs to all utterances. The optimality parameter for each model was fine-tuned to maximize its correlation with the empirical data.

The *literal answerer* and *literal questioner* models can be dismissed. The *literal answerer* model predicts an equal likelihood for all four gates, which is inconsistent with the observed patterns for the answerer role. In the same vein, the *literal questioner* model, which determines the question that would elicit the most informative answer from the literal answerer, does not exhibit any preference for specific questions. The predictions made by these models, when compared to the empirical results, are shown in the left column of Fig. 5.

Among the two remaining questioner models—the explicit and pragmatic models—their predictions are quite similar, which makes it difficult to distinguish between them using the current data. The pragmatic questioner model results in a model-data correlation of $r = 0.99$, while the explicit questioner model achieves $r = 0.96$. The difference in these correlations is statistically significant, with a d-value of 0.03, and the 95% confidence interval for Hou's estimate being [-0.097, -0.004]. Despite the pragmatic model fitting the data slightly better, both models yield similar qualitative patterns.

The predictions of the pragmatic questioner model for each response distribution are displayed in Fig. 4 (left). Although the model's predictions do not precisely align with the empirical data in terms of magnitude, it effectively captures the key qualitative trends, particularly the modal responses.

In comparison, the pragmatic answerer model offers a much closer fit to the data than the explicit answerer model, with a model-data correlation of $r = 0.95$ for the pragmatic answerer and $r = 0.8$ for the explicit answerer. This difference in correlations is statistically significant, as evidenced by a d-value of -0.15 and Zou's 95% confidence interval ranging from [-0.43, -0.02].

3) Discussion: The pragmatic answerer model is the only one that effectively accounts for the key qualitative features of the response data. For example, while the explicit answerer would predict that participants are equally likely to choose between the 'dalmatian,' 'poodle,' and 'cat' when asked about a pet, the actual data reveal a distinct preference for selecting the cat, with 'dalmatian' and 'poodle' showing similar levels of preference to the other option. This preference pattern is accurately captured by the pragmatic answerer model (see

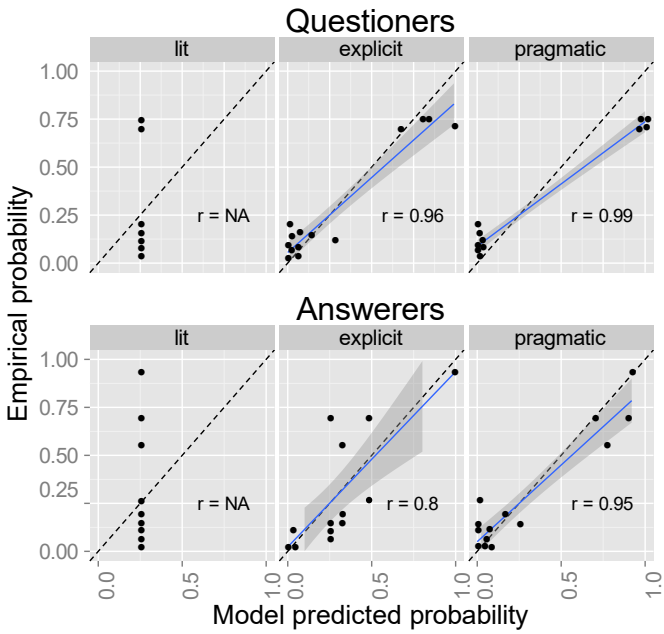

Fig. 5. The space of models and their association with the empirical data from Experiment 1 is illustrated. The row corresponding to the questioner models highlights the reasoning employed by each model in relation to the answerers below, with the pragmatic answerer considering the direct questioner.

Fig. 4 (right)). A more pronounced discrepancy arises in response to the 'animal?' question, where the explicit answerer anticipates a uniform distribution across all answers (as all four responses are animals). However, the observed distribution deviates notably from uniform, reinforcing the necessity of employing a pragmatic answerer to explain these findings.

While the data strongly advocate for the pragmatic answerer model, the results concerning the explicit and pragmatic questioner models are less definitive. The predictions made by both models did not significantly differ, and both demonstrated strong performance. To further investigate this, a follow-up study was conducted focusing on a special case within the guessing-game framework, where the explicit and pragmatic questioners diverge in their predictions.

V. Experiment 2: A Comprehensive Evaluation of Questioner Models

1) Participants: A total of 50 individuals were enlisted to participate solely in the questioner phase of the guessing game outlined earlier. Of these, ten participants were excluded from the analysis due to being non-native English speakers or expressing uncertainty regarding the instructions.

2) Stimuli and Procedure: The procedure adhered to the format established previously, with some modifications to the stimuli. The world space X encompassed possible configurations of three pets across three gates. The two designated goals T were the dalmatian and the poodle (with the cat excluded). The questions Q were phrased as either 'dalmatian?' or 'cat?'. The available responses R corresponded to the three

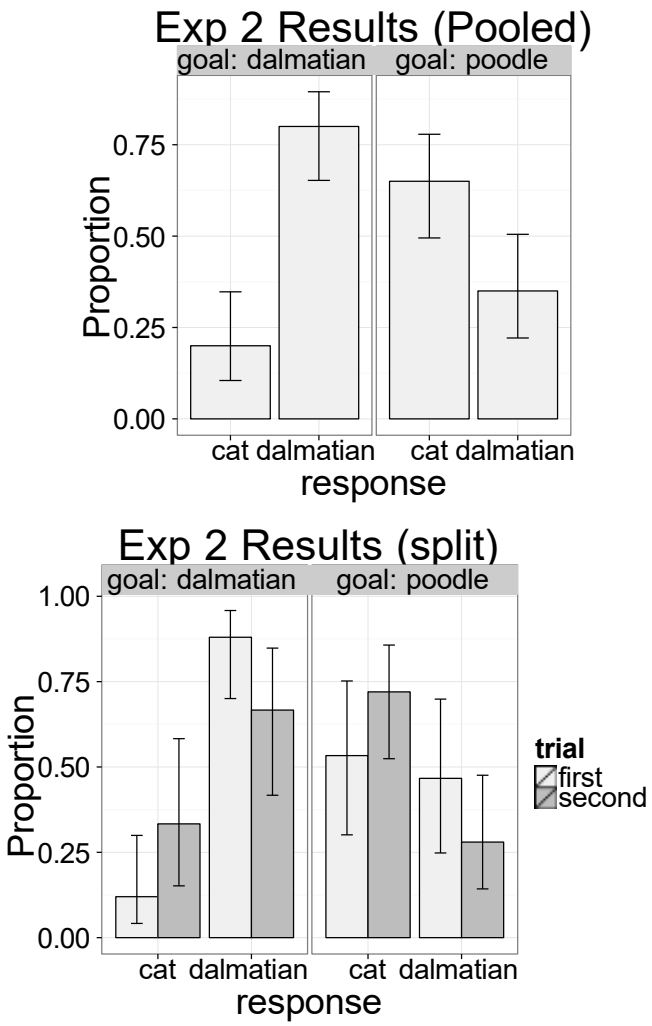

Fig. 6. Distribution of responses across all trials in Exp. 2 (left) and the response distribution segmented by first- and second-trial data (right).

gates. Each participant was presented with the two goals in a randomized order.

3) Results: When tasked with identifying the dalmatian, participants exhibited a notable tendency to inquire about the dalmatian more frequently than the cat, $\chi^2(1) = 12, p < 0.001$. Conversely, when the goal was to locate the poodle, participants showed a marginal preference for asking about the cat over the dalmatian, $\chi^2(1) = 3.6, p = 0.058$. However, focusing solely on the first trial, the preference for the dalmatian remained strong, $\chi^2(1) = 14.4, p < 0.001$, while the tendency to inquire about the cat diminished significantly, $\chi^2(1) = 0.07, p = 0.79$. These results are depicted in Fig. 6.

4) Comparison of Questioner Models: The explicit questioner model predicts that when the goal is to find the 'poodle,' there will be no clear preference for a specific question because the explicit answerer is equally unlikely to provide the desired answer for both targets. In contrast, the pragmatic questioner model predicts that participants will inquire about the cat. This is based on the assumption that the (internal) pragmatic answerer would deduce that if the questioner was interested

in the dalmatian, they would ask about it directly; thus, not asking about it indicates a focus on the alternative goal.

The question of which questioner model best captures the data has yet to be fully addressed. Overall, the response distribution confirms the pragmatic model's predictions, with participants preferring questions about the cat. However, this trend is not observed when only the first trial data are considered. This difference could be attributed to a variety of factors. Intriguingly, the pragmatic model suggests that if the questioner ignores the constraints on possible goals, a more explicit-like distribution of responses may emerge. If participants mistakenly believed the poodle was the only goal (against the instructions), asking about the dog would be consistent with the pragmatic model. It's possible that participants didn't fully consider the alternative goal (dalmatian) until after the initial trial, when it was actually the goal.[1]

5) Interpretation: Experiment 1 demonstrated that the pragmatic answerer model is essential for explaining the behavior of answerers in the basic guessing-game task, complementing the findings from computational simulations. Furthermore, it showed that both the explicit and pragmatic questioner models were able to capture key features of the questioner data, such as the tendency to prefer under-informative questions when the explicit label for an object is not available. However, neither experiment 1 nor experiment 2 succeeded in distinguishing between the explicit and pragmatic questioner models. In experiment 1, both models made identical predictions; in experiment 2, the response data were influenced by participants' assumptions about alternative goals.

There are several methodological issues worth considering in this task. For example, participants only provided hypothetical judgments about what they *would* say under different goal or question scenarios, rather than engaging in an actual game. Moreover, participants did not perceive that they were interacting with a real partner. These issues contributed to a general sense of confusion about the task and may have encouraged abstract, logic-based reasoning instead of the social and linguistic intuitions the experiment aimed to investigate.

To address these concerns, I modified the task in the subsequent experiments. In experiment 3, I replicated experiment 1 in a real-time, multiplayer setting where participants were assigned to either the "questioner" or "answerer" role and interacted with each other while playing the full game. In experiment 4, I expanded the range of items beyond the singular animal hierarchy used in earlier experiments. This new set of items covered four domains (animals, plants, places, and artifacts) and incorporated three different hierarchy structures, yielding a broader range of predictions from the models. One of these hierarchical structures was designed specifically to provide a critical test for distinguishing between the explicit and pragmatic questioner models. Unlike experiment 2, this condition did not require participants to reason about unconventional question alternatives that do not refer to any existing real-world items.

[1]This pattern was also seen in a simpler stimulus set involving red and blue squares and circles.

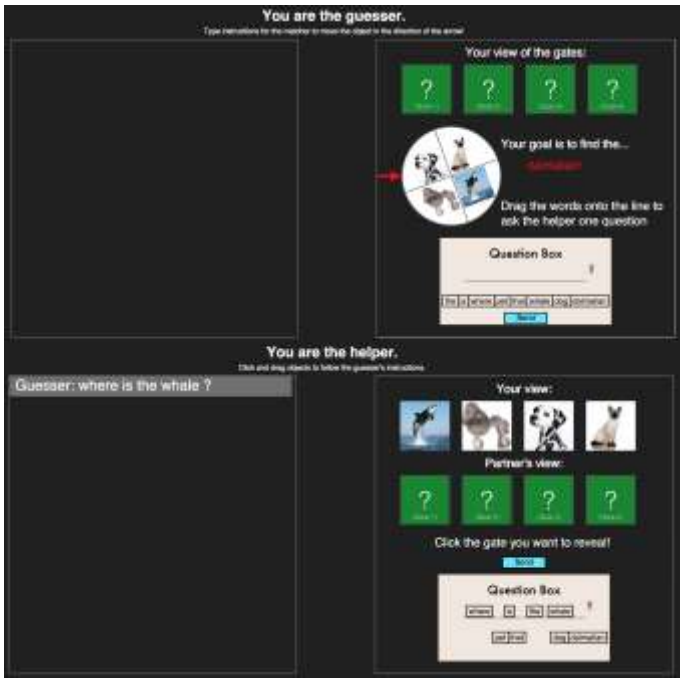

Fig. 7. Interface views for Experiment 3: the questioner's screen (top) and the answerer's screen (bottom).

VI. Experiment 3: Interactive Q&A Task

1) Participants: A total of 50 participants were recruited via Amazon's Mechanical Turk for this task. Among them, 25 participants were assigned to the questioner role and the remaining 25 to the answerer role, resulting in 25 completed games.

2) Stimuli & Procedure: The world space W, goal space G, question space Q, and answer space A remained consistent with those used in Experiment 1 (refer to Figure 3). The procedure was adjusted to incorporate real-time interaction between the players, following the methodology of (23). Participants first completed a brief quiz on the game instructions before being directed to the main interface. The first player to join was designated as the "questioner" (or "guesser" in the cover story) and instructed to wait for a second player. Once the second player joined, they were assigned the role of "answerer" (or "helper"), and the game commenced.

The interfaces for the questioner and answerer are shown in Figure 7. Chat messages were displayed on the left side of the screen, while the right side served as the workspace for players to view goals, pose questions, and provide answers. At the start of each trial, a wheel on the questioner's screen (see Figure 7, top) would spin and randomly select one of the four goals. The questioner then clicked and dragged words into the "Question box" to formulate a query to assist in identifying the goal. The answerer could observe the words being dragged in real time. Upon clicking the 'send' button, the question appeared in the chat log, and control shifted to the answerer, who selected one of the four gates to respond with the location

of the chosen object. The questioner then attempted to guess which gate they believed the goal object was behind.

Each participant completed four trials, with the object locations randomized and a new goal randomly assigned for each trial. This design prevented questioners from using "process of elimination" reasoning when multiple trials included the same goal.

3) Findings: The results from the questioner role are illustrated alongside the model's predictions in Figure 8 (left). It was noted that questioners consistently tailored their questions based on the specific goal, with a trend toward asking more indirect questions as the experiment progressed. χ^2 tests across the four different response distributions revealed significant departures from uniformity. In particular, questioners were more inclined to inquire about the 'dalmatian' when the goal was the dalmatian, $\chi^2(3) = 77$, $p < .001$, the 'dog' when the goal was the poodle, $\chi^2(3) = 50$, $p < .001$, the 'pet' when the goal was the cat, $\chi^2(3) = 47$, $p < .001$, and the 'animal' when the goal was the whale, $\chi^2(3) = 39$, $p < .001$.

The findings for the answerer role are presented in Figure 8 (right). Answerers showed a high degree of responsiveness to the constraints posed by the questioners, offering information related to the dalmatian when asked about a 'dalmatian', $\chi^2(3) = 102$, $p < .001$, the poodle when asked about a 'dog', $\chi^2(3) = 47$, $p < .001$, the cat when asked about a 'pet', $\chi^2(3) = 45$, $p < .001$, and the whale when asked about an 'animal', $\chi^2(3) = 31$, $p < .001$. A comparison of these results with the predictions from the model is provided in the following section.

4) Model Comparison: Model comparison was conducted similarly to Experiment 1: a single optimality parameter was fitted for each of the six models to maximize correlation with the empirical data.

Once again, both the literal questioner and literal answerer models were ruled out due to their prediction of a uniform distribution across the four questions and answers. Among the remaining models, the predictions of the questioner models were found to be nearly identical. Specifically, I obtained a correlation of $r = 0.971$ for the explicit questioner model and $r = 0.996$ for the pragmatic questioner model. The difference in these correlations was statistically significant, with Zou's confidence interval of $[-0.079, -0.009]$ suggesting that the pragmatic questioner model provides a better fit. However, both models made nearly identical qualitative predictions, as shown for the pragmatic questioner model's predictions in Figure 8 (left).

For the answerer models, the pragmatic answerer provided a substantially better fit to the data than the explicit answerer, with correlations of $r = 0.7$ for the explicit answerer and $r = 0.99$ for the pragmatic answerer. Taking into account the overlap of empirical data used for these correlations, the difference was statistically significant (Zou's confidence interval = $[-0.676, -0.107]$).

5) Discussion: In this real-time, interactive setting, I replicated the results from Experiment 1. As in the previous experiment, both the explicit and pragmatic questioner models

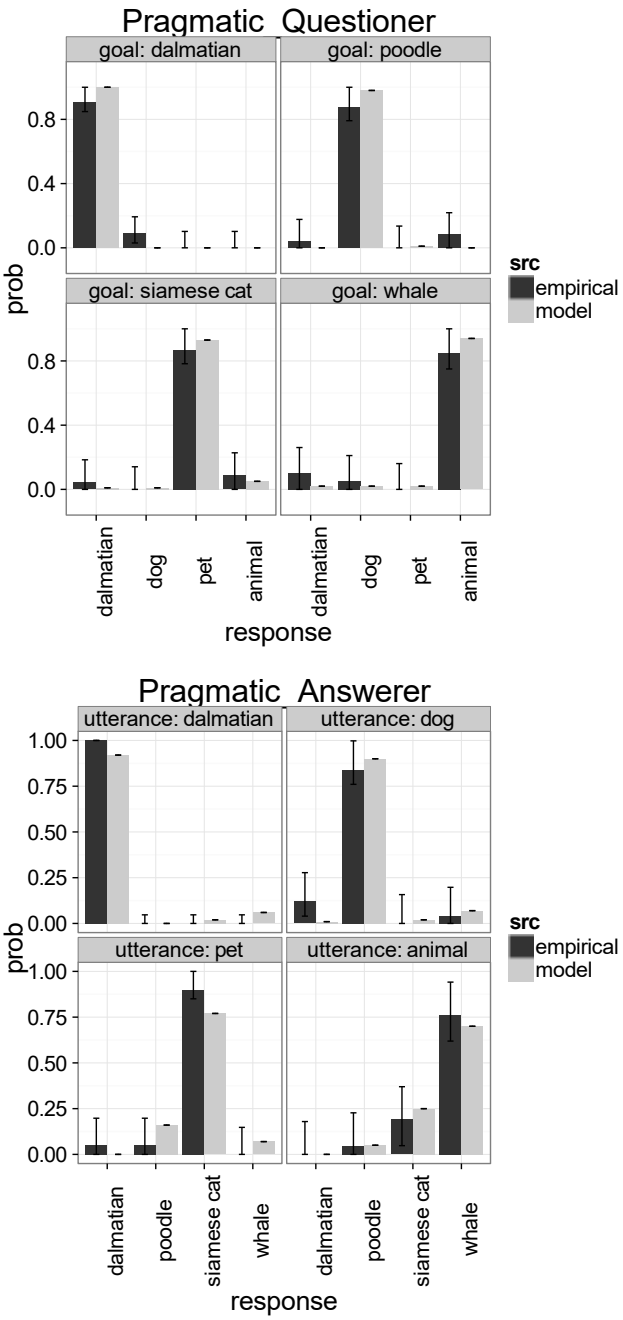

Fig. 8. Experiment 3 results and model fits for the pragmatic questioner (top) and answerer (bottom) models.

fit the data for questioners extremely well, while the pragmatic answerer model provided a superior fit compared to the explicit answerer model, both quantitatively and qualitatively. Furthermore, the interactive nature of the game appeared to enhance the experience, making it more engaging and less confusing for participants compared to the dropdown menu design used in the first two experiments. The continuous feedback from the answerer created a more social atmosphere, and the real-time visibility of the questioner's actions made it seem like participants were interacting with another human,

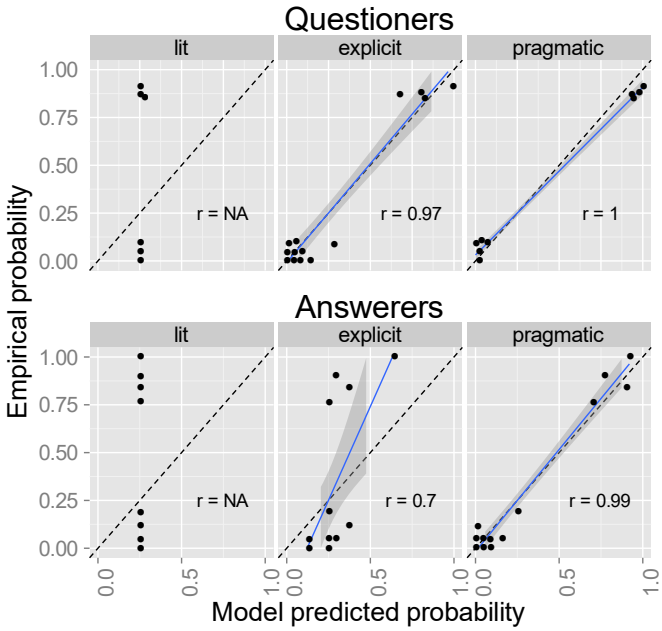

Fig. 9. The full model space, showing the correlations with Experiment 1 data. The questioner models are presented in the first row, each reasoning about the corresponding answerer models in the row beneath. The pragmatic answerer model reasons about the explicit questioner.

addressing concerns from previous experiments.

It is important to note that the stimulus set used in all experiments so far—based on a simple animal hierarchy—provided only 16 points of comparison between the models and the data. Additionally, there exists a heuristic strategy for selecting questions based on the hierarchy structure that leads to the same response patterns without requiring social inference. For example, if the questioner saw their goal, they could eliminate labels that do not apply (e.g., a 'cat' is neither a 'dalmatian' nor a 'dog') and then choose the most specific remaining label (e.g., 'pet' is more specific than 'animal').

In Experiment 4, I aim to test the generality of the model by expanding the stimulus set to include multiple domains and hierarchy structures. This will address the concern that the observed behavioral patterns might be specific to the animal set and tree-like hierarchy. I will also include a critical hierarchy structure where the explicit and pragmatic questioner models make distinct predictions, and the heuristic strategy fails to account for these differences.

VII. Experiment 4: Generalizing Predictions

1) Participants: A total of 199 participants were recruited from Amazon's Mechanical Turk for this experiment. However, fifty participants were excluded due to a server crash that interrupted the task before completion. Additionally, two participants were excluded because they were not non-native English speakers. This resulted in 74 completed games from unique participants.

2) Stimuli & Procedure: Twelve distinct items were created by combining four domains (animals, plants, places,

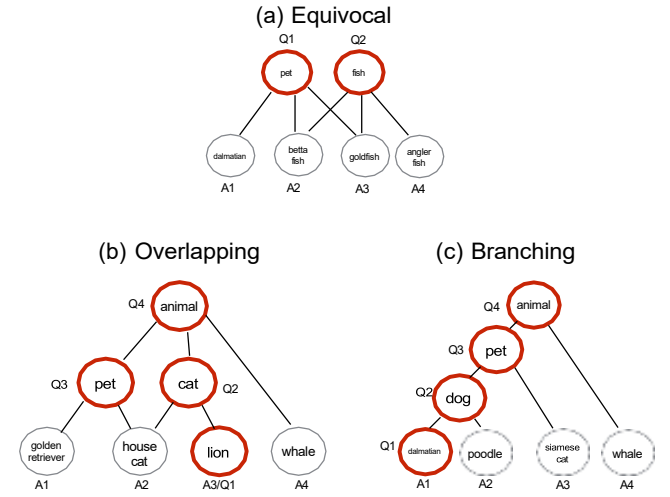

Fig. 10. Illustrations of the three hierarchy types used in Experiment 4. The bottom row represents the answer space, while the labels in the questioner space are highlighted in red.

and artifacts) with three hierarchy structures ("branching," "overlapping," and "equivocal"; see Figure 10). For each item, there was a set of four goal objects in G, four question labels in Q that the questioner could use to inquire about their assigned goal, and four answers in A corresponding to the four items in G displayed on the wheel.

The procedure followed the same format as Experiment 3 with two key changes. First, I addressed an imbalance in the task where the answerer could observe the questioner's movements, but the questioner only saw template text responses. In pilot testing, it was found that answerers were more likely to believe they were playing with a human partner than questioners. To align the beliefs of both participants, I modified the way the answerer responds. Instead of selecting a gate and clicking a 'send' button, a "Reveal Box" was introduced. When the answerer was ready to reveal a gate, they would drag the object into the box. The questioner could then observe the gate's outline moving in real time and view the image once the answerer released it.

The second adjustment was made in response to the fact that some participants exploited the flexibility of the "question box" drag-and-drop procedure after several rounds. On challenging items, particularly those with an equivocal hierarchy where the model predicts no preferred question for certain goals, participants began using non-grammatical signals. While this behavior is interesting, the goal of the experiment was to test models that operate within a defined set of question utterances, so I introduced a predefined frame for the question box. Participants were only allowed to drag one question label into the box.

Each participant responded to one trial for each of the twelve items, with the items presented in random order.

3) Results: First, it was observed that there was a high correlation in response probabilities across domains for both questioners and answerers (see Table I). For simplicity, I will collapse the data across domains in the analysis. However, it

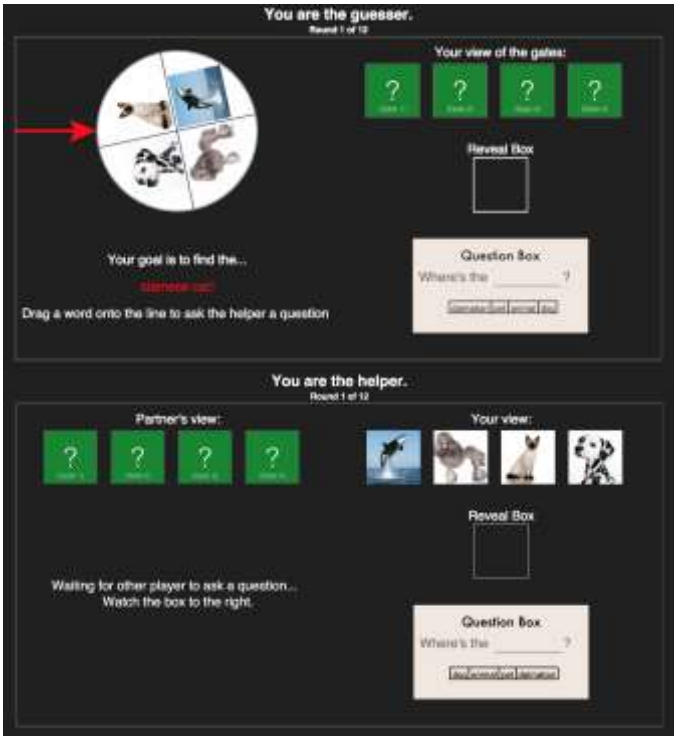

Fig. 11. Interfaces for participants in Experiment 4: the questioner (top) and answerer (bottom).

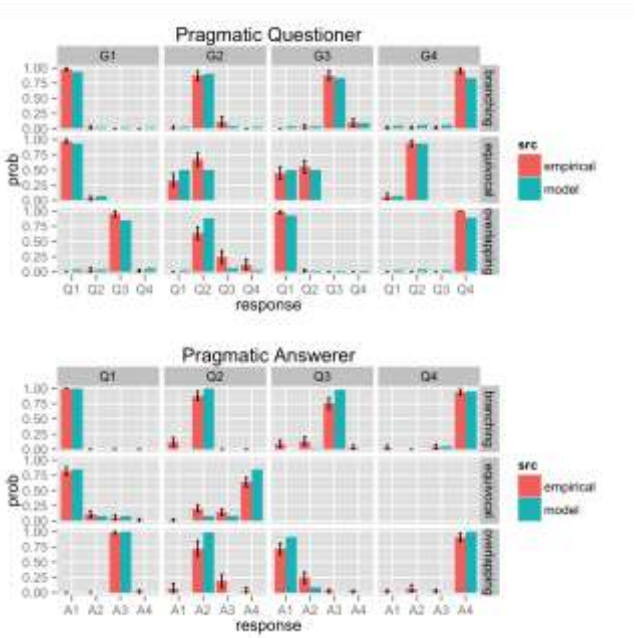

Fig. 12. Results and model fits for Experiment 4, showing the best-performing questioner (left) and answerer (right) models across all domains.

is worth noting that the ”place” domain showed the lowest inter-domain correlations for both questioners and answerers, which suggests it may be an outlier — I will revisit this in the discussion section.

Next, I will examine the response patterns for each hierarchy type, using the ‘animal’ domain for clarity. In the ‘branching’ hierarchy, I find that . . .

4) Model Comparison: Model comparison followed the same method as in Experiments 1 and 3. To prevent overfitting and minimize the number of free parameters, I pooled the data from all hierarchy structures and domains and fitted a single rationality parameter for each model to optimize model-data correlations across the entire pooled dataset.

The fits for these models, broken down by hierarchy structure, are shown in Figure 13. I can immediately rule out both the literal answerer and literal questioner models, as they predict a uniform response distribution. Additionally, I find that the explicit answerer models perform significantly worse than the pragmatic answerer model across all hierarchy structures, with $r = 0.67$ versus $r = 0.96$ in the ‘branching’ hierarchy, $r = 0.44$ versus $r = 0.92$ in the ‘equivocal’ hierarchy, and $r = 0.68$ versus $r = 0.86$ in the ‘overlapping’

	Questioners				Answerers			
	animal	place	plant	artifact	animal	place	plant	artifact
animal	1.00	0.91	0.94	0.94	1.00	0.78	0.92	0.97
place	0.91	1.00	0.96	0.95	0.78	1.00	0.78	0.78
plant	0.94	0.96	1.00	0.97	0.92	0.78	1.00	0.91
artifact	0.94	0.95	0.97	1.00	0.97	0.78	0.91	1.00

TABLE I
CORRELATIONS BETWEEN DOMAINS IN EXPERIMENT 4

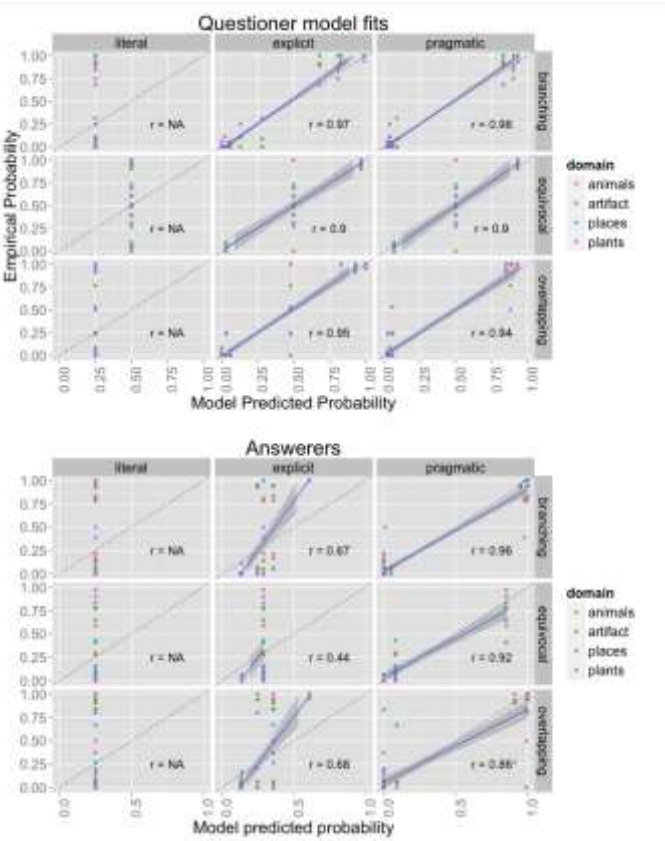

Fig. 13. Full range of models and their correlations with the data from Experiment 4, categorized by item and type.

hierarchy, respectively.

In examining the questioner predictions, I find that both

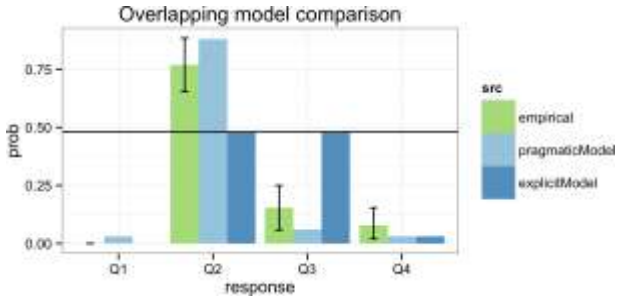

Fig. 14. Detailed comparison of model fits with the data from Experiment 4, highlighting the overlapping hierarchy condition.

the explicit and pragmatic models fit the data exceptionally well, with $r = 0.97$ compared with 0.98 in the 'branching' hierarchy, $r = 0.90$ compared with $r = 0.90$ in the 'equivocal' hierarchy, and $r = 0.95$ compared with 0.94 in the 'overlapping' hierarchy, respectively.

However, because the overlapping structure was designed to provide a critical test for the questioner models, the primary concern is not the overall correlation but the distribution of responses to the goal "G2" (the "house cat" in the animal domain). Since the goal "house cat" is not directly available as a question, the explicit model predicts equal probabilities for the two parent categories, "pet" and "cat," both of which have two children.

The pragmatic model, however, predicts that "cat" will be favored over "pet" because the other child of "cat," the "lion," is pragmatically blocked by the answerer. When the answerer hears "cat," they reason that if the questioner's goal were the "lion," they would have mentioned "lion"; hence, the goal must be the other cat.

Figure 14 zooms in on this critical condition and compares the predictions of both models to the empirical data, with 95% confidence intervals. The pragmatic model makes the correct qualitative prediction, showing that the mean probability of choosing "Q2" ($p_{Q2} = 0.77$, $n = 52$, 95% bootstrapped CI $= [0.65, 0.88]$) is significantly different from the mean probability of choosing "Q3" ($p_{Q3} = 0.15$, $n = 52$, 05% bootstrapped CI $= [0.06, 0.25]$), as predicted by the pragmatic model. The predicted probabilities also closely match the observed probabilities within the 95% confidence interval, providing additional confidence that the model is not overfitting.

5) Discussion: This experiment replicates the results of Experiments 1 and 3, demonstrating that they generalize across a variety of domains and hierarchy structures. Furthermore, the results help distinguish between the pragmatic and explicit questioner models, showing that the pragmatic model is necessary to explain the data. The "place" domain, which showed an outlier pattern, was found to be due to the particular image used, which biased participants toward a "bar" response instead of the "restaurant" response predicted by the pragmatic model.

VIII. General Discussion

A significant contribution of this study is the extension of the Rational Speech Act framework to analyze not only individual utterances in context but also the dynamics of simple dialogues, specifically involving a single question and its corresponding answer. This shift in focus changes the goal from merely conveying true information to extracting useful information from the interlocutor. On the answerer's side, this necessitates a nuanced understanding of the questioner's intentions. This connects closely with game-theoretic and decision-theoretic models (15; 16), which stress the importance of goals and beliefs in communication, while emphasizing the inferential dynamics between conversational partners.

The results presented in this study suggest that answerer behavior is most effectively modeled using a pragmatic approach that reasons about the questioner's intentions, interpreting the question as a signal. This builds on earlier intention-based models by providing a detailed mechanism for how the answerer could infer these intentions through Bayesian conditioning. The superiority of this pragmatic model was consistent across all experimental settings.

However, questioner behavior in Experiment 2 appeared more dependent on prior experience. In an initial version of Experiment 1, which did not emphasize certain aspects of the task, participants displayed a mixture of explicit and pragmatic questioners and answerers. The multiplayer, interactive framework employed in Experiments 3 and 4 proved to be more robust against such variations. Moreover, in conditions such as the overlapping hierarchy in Experiment 4, questioners exhibited higher-order pragmatic reasoning regarding how the answerers would deduce their goals. This behavior could not be explained by a simple heuristic strategy; it indicates that questioners actively reason about the answerers' goals when formulating their questions.

Finally, none of the models fully captured the questioner's behavior in the "equivocal" hierarchy. Although the models predicted no preference between the terms "fish" and "pet," participants overwhelmingly chose "fish," likely due to pre-existing beliefs about label suitability. In future studies, I aim to extend the model to incorporate mixtures of Questions Under Discussion (QUDs) and include label fitness when interpreting question words.

While the artificial nature of the task may limit its naturalness, it offers distinct advantages in controlling the precise set of questions, goals, and answers. Future work will focus on expanding the model to handle a broader range of conversational contexts, scaling it to more complex interactions.

References

[1] Boer, S. E., & Lycan, W. G. (1975). Knowing who. Philosophical Studies, 28, 299-344.
[2] Van Der Henst, J. B., Carles, L., & Sperber, D. (2002). Truthfulness and relevance in telling the time. Mind & Language, 17(5), 457-466. Wiley.

[3] Gibbs, R. W., & Bryant, G. A. (2008). Striving for optimal relevance when telling time. Discourse Processes, 45(1), 1-20. Taylor & Francis.
[4] Potts, C. (2012). Goal-driven answers in the Cards dialogue corpus. In Proceedings of the 30th West Coast Conference on Formal Linguistics (WCCFL).
[5] Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. Science, 336(6084), 998. American Association for the Advancement of Science.
[6] Goodman, N. D., & Stuhlmu¨ller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. In Proceedings of the 34th Annual Conference of the Cognitive Science Society, 1560-1565.
[7] Kao, J. T., Wu, J., Bergen, B., & Goodman, N. D. (2014). Nonliteral number words: Learning the meaning of hyperbole. In Proceedings of the 36th Annual Conference of the Cognitive Science Society.
[8] Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. PhD thesis, Ohio State University.
[9] Clark, H. H. (1979). Responding to indirect speech acts. Cognitive Psychology, 11(4), 430-477.
[10] Fusaroli, R., & Tyle´n, K. (2015). Investigating conversational dynamics: Interactive alignment, intersubjectivity, and the interactive brain. Frontiers in Psychology, 6, 181.
[11] Ginzburg, J. (1995). Resolving questions, I. Linguistics and Philosophy, 18(5), 459-527.
[12] Goodman, N. D., & Stuhlmu¨ller, A. (2014). The Design and Implementation of Probabilistic Programming Languages. Retrieved from http://dippl.org.
[13] Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27(2), 169-190.
[14] Schulz, K., & van Rooij, R. (2006). Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. Linguistics and Philosophy, 29(2), 205-250.
[15] Van Rooy, R. (2003). Questioning to resolve decision problems. Linguistics and Philosophy, 26, 727-763. Springer.
[16] Vogel, A., Bodoia, T., Potts, C., & Jurafsky, D. (2013). Emergence of Gricean maxims from multi-agent decision theory. Proceedings of NAACL-HLT, 1072-1081.
[17] Franke, M. (2013). Game theoretic pragmatics. Philosophical Perspectives, 27(1), 333-375.
[18] Diedenhofen, B., & Musch, J. (2014). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. PLoS ONE, 9(4), e90759. https://doi.org/10.1371/journal.pone.0090759
[19] Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences, 28(5), 675–691. https://doi.org/10.1017/S0140525X05000129
[20] Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. Evolution and Human Behavior, 26(4), 313–331. https://doi.org/10.1016/j.evolhumbehav.2004.08.015
[21] Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. American Journal of Psychology, 57(2), 243–259. https://doi.org/10.2307/1416950
[22] M. Groenendijk and M. Stokhof, ”Studies in the Semantics of Questions and the Pragmatics of Answers,” *PhD Thesis*, University of Amsterdam, 1984.
[23] Hawkins, L. (2014). Real-Time Web Experiments: Methodology and Applications. *Journal of Web Studies*, 5(2), 125–140.