Integrating Neural and Symbolic Components in a Model of Pragmatic Question-Answering

Polina Tsvilodub
University of Tübingen
first.last@uni-tuebingen.de

Robert D. Hawkins
Stanford University
hawkrobe@gmail.com

Michael Franke
University of Tübingen
first.last@uni-tuebingen.de

Abstract

Computational models of pragmatic language use have traditionally relied on hand-specified sets of utterances and meanings, limiting their applicability to real-world language use. We propose a neuro-symbolic framework that enhances probabilistic cognitive models by integrating LLM-based modules to propose and evaluate key components in natural language, eliminating the need for manual specification. Through a classic case study of pragmatic question-answering, we systematically examine various approaches to incorporating neural modules into the cognitive model—from evaluating utilities and literal semantics to generating alternative utterances and goals. We find that hybrid models can match or exceed the performance of traditional probabilistic models in predicting human answer patterns. However, the success of the neuro-symbolic model depends critically on how LLMs are integrated: while they are particularly effective for proposing alternatives and transforming abstract goals into utilities, they face challenges with truth-conditional semantic evaluation. This work charts a path toward more flexible and scalable models of pragmatic language use while illuminating crucial design considerations for balancing neural and symbolic components.

1 Introduction

Imagine you are a barista in a café with only three items in stock: iced coffee, soda, and Chardonnay. If a customer asks: "Do you have iced tea?", you might naturally respond "I'm sorry, we don't have iced tea, but I can make you an iced coffee!". This situation exemplifies *pragmatic question answering*, where answerers commonly go beyond the literal question being asked (Clark, 1979). Classical accounts of the semantic meaning of questions and answers (e.g., Hamblin, 1973; Groenendijk and Stokhof, 1984; Hakulinen, 2001), maintain that polar questions like "Do you have iced tea?" are fully resolved by a polar answer {yes, no}. Yet humans routinely provide a *relevant* selection of additional information (e.g., mentioning the iced coffee, but not the Chardonnay).

Understanding what, exactly, makes an answer relevant has been a central question in the field of pragmatics, with extensive work investigating the contextual factors that shape answer selection (e.g. van Rooy, 2003; Stevens et al., 2016; Rothe et al., 2017). One recent framework for modeling these pragmatic choices is the Rational Speech Act framework (Frank and Goodman, 2012; Degen, 2023), which has been successfully applied to both question and answer selection (Hawkins et al., 2015; Hawkins and Goodman, 2017; Hawkins et al., to appear). The probabilistic cognitive models (PCMs) developed within this framework offer significant advantages through their transparent, explicit task decomposition and systematic error analysis (Farrell and Lewandowsky, 2018).

However, these models are typically limited to a small set of predefined examples, restricting their applicability to real-world scenarios. In contrast, Large Language Models (LLMs) offer a complementary set of capabilities. They can process open-ended natural language input and generate flexible responses, but often struggle with subtle pragmatic patterns (Hu et al., 2023; Ruis et al., 2023; Tsvilodub et al., 2024b) and lack the degree of explainability that makes PCMs so valuable for cognitive modeling (Zhao et al., 2023).

To address these complementary strengths and limitations, we explore a family of *neuro-symbolic* models, with different combinations of both approaches to leverage their respective strengths and to overcome known shortcomings.[1] Our ap-

[1]We use the term *neuro-symbolic* in the sense of a model that has *neural* network components (here, LLMs), that are scaffolded by a *symbolic* task analysis, i.e., integrated in a particular computational procedure. Other senses of the term also exist (Bhuyan et al., 2024).

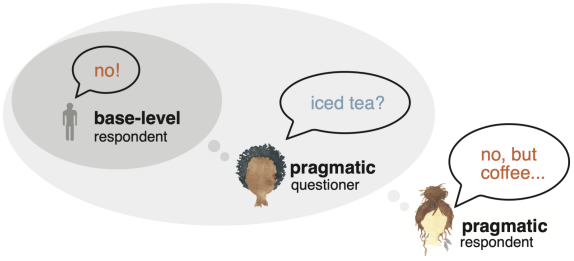

Figure 1: Probabilistic cognitive model (PCM) of pragmatic question answering. The PCM, built in the Rational Speech Act framework, implements recursive back-and-forth reasoning of rational agents. The questioner chooses a question based on their decision problem and an expectation of responses that any question might provoke. The respondent chooses a relevant response based on the decision problem inferred from the question.

proach builds on the task analysis developed in previous work on pragmatic question-answering (Hawkins et al., 2015; Hawkins and Goodman, 2017; Hawkins et al., to appear) in two ways. First, we use it as a *scaffolding structure* that determines the computational steps, with LLMs executing specific subtasks that would traditionally require manual specification in a PCM (Sections 3.2–3.3). Second, we verbalize (parts of) the scaffolding structure in a single prompt, relying on a single LLM call to solve the respective computational task (Section 3.4). This dual approach enables us to systematically investigate the tradeoffs between fine-grained task decomposition and end-to-end neural processing.

Our key contributions are as follows:

- A novel neuro-symbolic framework that extends probabilistic models of pragmatic question answering to more open-ended natural language.
- A systematic investigation of how different integrations of neural and symbolic components affect model behavior.
- Empirical validation against human data, demonstrating that neuro-symbolic models can match or exceed traditional probabilistic approaches in predicting human behavior.

2 A Probabilistic Cognitive Model of Relevant Question-Answering

The probabilistic cognitive model we use for task analysis and scaffolding, which we refer to as the *QA model* (Hawkins et al., to appear), captures a rational *pragmatic respondent* that chooses an answer by reasoning about how a pragmatic questioner chooses a question (see Figure 1 for overview and Appendix A for technical detail). The questioner is grounded in a context-independent *base-level respondent*. The pragmatic questioner selects a question based on the response they expect to get from the base-level respondent, who answers austerely without considering the wider context. The pragmatic respondent, in turn, reasons about the motivation of the speaker for asking the question (i.e., *infers* their goal from the question) and chooses responses that are expected to be relevant to the questioner's goal.

To implement expected relevance of an answer, the QA model builds on decision-theoretic accounts of relevance of questions and answers (van Rooy, 2003; Benz, 2006), which formalizes relevance in terms of a *decision problem (DP)*. The DP includes a real-valued *utility function* of how useful different alternatives (e.g., iced coffee, soda, Chardonnay) are for a given goal (e.g., getting an iced tea). The questioner selects questions that have a high expected relevance (i.e., high *expected utility*) of information from the base-level respondent. The pragmatic respondent uses the questioner's goal-oriented choice of question to infer from the question what kind of DP the questioner likely has. These inferences then guide the respondent's choice of information that will likely increase the expected utility for the questioner, traded off with response costs. We use a probabilistic implementation of the QA model in WebPPL (Goodman and Stuhlmüller, 2014) from Hawkins et al. (to appear) as a starting point and baseline. As commonly done for probabilistic modeling, for these simulations we specified the space of possible answers, possible questions, the literal semantics and the DP utility function specifically for the main experimental materials (see Section 3.1 and Appendix A.1).

Before diving into neuro-symbolic model evaluation, we first validate whether the task decomposition stipulated in the QA model is actually borne out in human intuitive reasoning. To this end, we conducted an exploratory *answer explanation* experiment. Participants (N=50) were recruited via Prolific and shown four trials with contexts wherein a person asked for a target item while several alternative options were available, similar to the initial café example, which constituted the main materials we describe in more detail in Section 3.1. The

question was followed by a character replying "no" and providing one, most relevant, competitor alternative. Participants were asked to type an explanation of why that response was reasonable and what would justify mentioning the particular option over a different one. We then analyzed the types of provided explanations, distinguishing between explanations that appealed to (1) abstract similarity of options, (2) questioner goals, desires, intentions, or preferences, and (3) features that were functionally relevant for the questioner goal (e.g., being and iced non-alcoholic drink). If participants spontaneously reason about questioner goals and respective relevant option features as formalized in the QA model, we hypothesize that the proportion of (2) and (3) will be higher than (1). We found that 0.43 of responses appealed to goals (2), 0.20 to goal-relevant features (3), and 0.21 to general similarity (1). 0.13 of responses were unclassifiable (e.g., only appealed to respondent politeness). We interpret this as mild *prima facie* support for the task decomposition implemented in the probabilistic QA model. In the next section, we analyze how systematically replacing different components of the QA model with LLM modules affects the fit to human data.

3 Evaluating Neuro-Symbolic QA models

We investigate the neuro-symbolic framework starting with models where only one component of the task is supplied by an LLM. We then incrementally increase the number of LLM-based modules and change their types, while observing the changes of the *fit to human data* and the *qualitative changes in the predictions*. The driving motivation is to make PCMs more generally applicable (open-ended). For that, two steps are necessary. For one, we would like to be able to generate an in principle open-ended set of alternatives over which to reason or which to choose from. Consequently, we test if LLMs can provide plausible sets of responses, questions, and questioner goals for the QA model; we call LLMs in this role **proposers** (cf. Sumers et al., 2023; Tsvilodub et al., 2024a). For another, once we have open-ended sets of alternatives, we need to be able to obtain information about them for downstream computation, i.e., we also use LLMs in the role of **evaluators** for judging literal semantics of answers and for assessing the utility of options.

3.1 Experimental setup

For all reported simulations below, we use `GPT-4o-mini` for the LLM modules, with the sampling temperature $\tau = 0.1$. All simulations are run for five iterations. We report additional results with the open-source LLM `Qwen-2.5-32B-Instruct` in Appendix D. We use experimental materials, human data and the one-shot LLM prompt from Tsvilodub et al. (2023) to investigate what kinds of alternative options (e.g., iced coffee or Chardonnay), if any, different neuro-symbolic QA models mention in the predicted responses, given a polar question (e.g., "Do you have iced tea?") and different options in context.

The materials include 30 commonsense vignettes similar to the initial barista example. The context always included three possible options, but not the requested target (i.e., iced tea). The options always included a best-fitting alternative called the *competitor* (e.g., iced coffee), a conceptually *similar* option that was deemed less relevant for the questioner's goal (e.g., soda), and an *unrelated* option irrelevant for the uttered request (e.g., Chardonnay). Experimental subjects provided answers by freely typing into a text box. Responses were categorized as "target," "similar," and "unrelated." In addition to these three categories, corresponding to mentioning each of the single options, the categorization also distinguished responses that mentioned *all options*, as well as responses that mentioned *no options*.

If a respondent is engaging in pragmatic reasoning, we would expect her to prefer competitor responses over other types. Tsvilodub et al. (2023) found that humans are, in fact, *relevantly overinformative*, strongly preferring competitor responses (0.52 of responses) over exhaustive responses (0.10), no options responses (0.20), similar (0.18) or unrelated responses (0.00). We investigate how well neuro-symbolic models match human behavior, operationalized via Jensen-Shannon divergence between the observed human data and the models' categorical predictions.

3.2 Integrating LLM Evaluators in the PCM

We assess a class of models that, starting from the QA model, systematically incorporate LLM modules into the PCM architecture which take over two functions: (i) the evaluation of utility of an option, and (ii) the evaluation of the truth of a response. Figure 2 (lower panel) shows a schematic overview

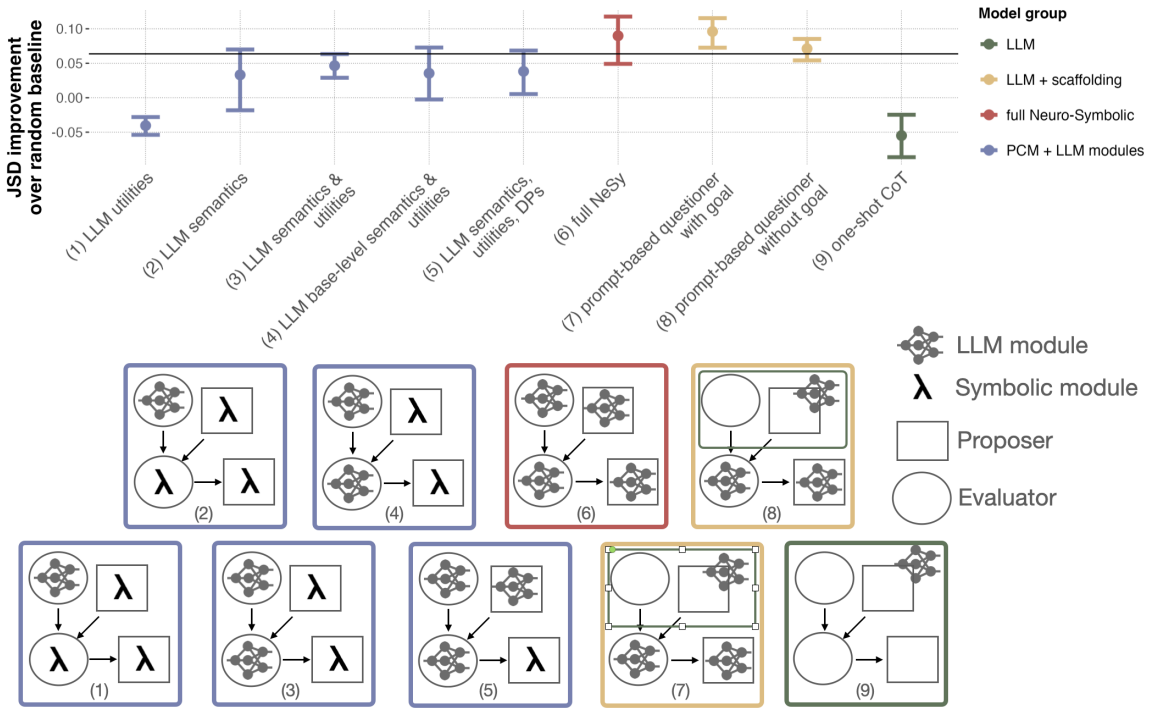

Figure 2: **Upper panel:** Improvement of the model fit to human data in terms of Jensen-Shannon divergence over a uniform response distribution baseline (higher is better, y-axis) of all analyzed models (x-axis). The horizontal line indicates performance of the probabilistic model. Dots indicate the means across simulations, error bars indicate 95% bootstrapped CIs. **Lower panel:** Overview of tested models. Each box shows a schematic of one model, labeled on the x-axis in the plot above it. The models are ordered from closest to the PCM on the left (only one component is LLM-based), to a model only using a single LLM with a single prompt on the right.

of the tested models.

First, we implement an LLM *utility evaluator* for instantiating the utility function in the questioner's decision problem (resulting in the **"LLM utilities" model**). The utility function defines real-valued utilities for the different alternatives (e.g., the iced coffee, soda), conditioned on a target object (e.g., iced tea). In the original QA model, the utilities were elicited in a human rating experiment wherein participants were asked to provide slider ratings for each possible option (e.g., iced tea, iced coffee, soda, Chardonnay), given another option as the goal (see Appendix A.1). To replace the human input with an LLM, we prompted the utility evaluator in a way identical to the instructions of the human elicitation experiment, namely to predict the full space of utilities via ratings on a scale with range 0–100 instead of slider ratings. Importantly, the prompt (and the original human experiment) only asked for abstract ratings, independent of the functional context in which the options occurred in the question answering scenario (see Appendix B for all full prompts). The remaining model components (e.g., the set of alternative utterances, the semantics) remained symbolic in this model.

Beyond replacing the utility component, another function-based component to replace with LLMs for open-ending the PCM is *semantic evaluation*. Semantic evaluation is necessary for the base-level and for the pragmatic respondent and assesses whether a response is true in a particular context. While base-level and pragmatic respondent have slightly different responses at their disposition owing to the fact that the base-level responder is not reasoning about the context (see Appendix A), the semantic evaluation is essentially the same. For an answer like "No, but we have iced coffee." the module has to check whether the polar answer part (e.g., "yes", "no") is true for a context (e.g., the café has soda and iced coffee), given the question (e.g., "Do you have iced tea?"). It also has to evaluate whether the added information (e.g., "We have iced coffee.") is actually correct. We explored models with different combinations of these evaluators. The **"LLM semantics"** model uses an LLM-based semantic evaluator for both the base-level and the pragmatic respondent, while using the same utility component as the original QA model (based on

the human experimental data). The **"LLM semantics & utilities"** model employs all described LLM evaluators. The **"LLM base-level semantics & utilities"** only uses an LLM-based base-level respondent, a rule-based pragmatic respondent, and the LLM utility evaluator. The predictions of all models are compared in Section 4.

3.3 Integrating LLM Proposers in the PCM

Next, we integrate LLMs as *proposers* for sets of alternatives required by the QA model. We start with sampling the possible questioner goals with a *goal proposer*. The LLM was prompted to generate plausible text-based goals, given the context and question (see Figure 11). While the set of possible goals in the PCM only contained four DPs (each defining a preference for one of the options: target, competitor, similar, unrelated option), the proposer may sample any text-based questioner goal description. These sampled text-based goals are connected to a DP representation via the *utility evaluator* (Section 3.2). The evaluator was prompted to generate the utilities for the available options, conditioned on each proposed goal. The **"LLM semantics, utilities, DPs"** model uses the goal proposer together with the evaluators from Section 3.2, while the sets of possible utterances and questions are symbolic (i.e., pre-specified manually).

Further open-ending the QA model, we introduce a *response proposer* and a *question proposer* which provide the set of alternative questions and pragmatic answers that the respective pragmatic agents reason over. In both cases, the LLM was concisely prompted to generate n alternatives to an observed utterance or question given the context vignette (see Figure 9, Figure 10). We set $n = 10$ for the response proposer, and $n = 3$ for the question proposer. Here, we address the empirical question whether LLMs, out of the box, can be (easily) prompted to produce the expected types of alternative pragmatic responses in the context of the QA model (no options, competitor, similar, unrelated, all options). Based on exploratory qualitative analyses described in Section 4 in more detail, we append "no-options" and "all-options" responses constructed in a rule-based manner to the set of sampled alternatives. The observed question was always added to the set of sampled alternatives provided by question proposer.

The question and response proposers were tested as part of the fully neuro-symbolic replication of the PCM (**"full NeSy"** model). This model implements the full task decomposition of the QA model, capturing the pragmatic respondent's recursive reasoning (Figure 1) fully via the modules described above. The base-level respondent uses an LLM-based semantic evaluator to (symbolically) select an informative, true response to a given question (assuming that the decision problem is known). For the pragmatic interpreter, the different possible questions are supplied by an LLM-based question proposer. An LLM-based utility evaluator rates the usefulness of potential options to (symbolically) compute the questioner's expected utility of each question (based on the expected behavior of the base-level respondent). Finally, the pragmatic respondent estimates likely DPs among the neurally sampled alternatives, given the question, symbolically via Bayes rule (where the likelihood term is approximated via samples of generated questions given a DP). Given her posterior beliefs about the DPs, the respondent chooses a response from the set provided by the response proposer that maximizes her utility function. The respondent's utility function combines the expected utility of a response with informativeness, formalized as a KL divergence term (see Appendix A for details). We assume flat priors and no utterance costs throughout the model.

3.4 Scaffolding Prompted LLMs with Cognitive Modules

All previous models have implemented computational components suggested by the original QA model with LLM-based proposers and evaluators. These LLM-based components implemented rather "local", smaller computational elements of the task analysis suggested by the QA model. Alternatively, we may also use LLMs to replace larger chunks of computation, such as the full pragmatic question answering agent, or even the full task analysis captured by the QA model. In the following, we introduce three models that instantiate this general strategy.

We first consider a model called **prompt-based questioner**, of which we consider two versions, one prompted with questioner goals, and one prompted without goals. This model decomposes the pragmatic respondent's task into its two high-level components suggested by the PCM: inferring the questioner's goal based on the observed question, and selecting a response that optimizes the questioner's utility given the inferred DP. We implement a purely prompt-based pragmatic questioner

module that supplies the first component. This prompt-based questioner is used by the pragmatic respondent of the "full NeSy" model for inferring the distribution over DPs sampled with an LLM-based goal proposer. The prompt-based questioner takes a questioner goal, the context, and prompts the LLM to provide a likelihood of someone asking the given question (see Fig. 12). The elicited likelihoods for all questions and DPs are then renormalized and used by the pragmatic respondent. We then compare the role of conditioning this module on the goal, and also use a goal-free prompt where the LLM is asked to assess the question likelihood based on the context only (**prompt-based questioner without goal**, see Fig. 13).

For comparison, we also consider a purely *monolithic* prompting of the LLM. In particular, the **one-shot chain-of-thought model** has a chain-of-thought prompt which *verbalizes* the reasoning steps suggested by the QA model in the chain-of-thought for a single example item (see Figure 14). That is, this model is fully LLM-based, using only one call to one neural module (i.e., the LLM).

4 Results

Quantitative results We used the human answer proportions reported in Section 3 as reference and quantitatively compared models in terms of fit to the human data by calculating the Jensen-Shannon divergence (JSD) between the human and the models' predictions. Specifically, we calculated the score Δ_i of model M_i in comparison to the performance of a baseline B given by a flat distribution over all answer categories:

$$\Delta_i = JSD(B, \text{humans}) - JSD(M_i, \text{humans})$$

where $JSD(B, \text{humans}) = 0.154$. We report Δ_i-s in Figure 2 (upper panel; higher JSD differences are better, indicating closer fit to human data). The figure additionally shows the reference value provided by the PCM (solid line).

We found that most tested models with intermediate or high degrees of task decomposition came close to the original PCM (the CIs overlap with the PCM reference line or lie above it), indicating that the neuro-symbolic framework provides a potentially viable method for explaining human data. Visually, the "full NeSy" model and the "prompt-based questioner with goals" fit human data best in terms of Δ. The PCM + LLM models tended to improve with a higher number of LLM modules, but generally provided a somewhat worse fit than the PCM (the means are below the line). Supporting LLMs with a theoretically motivated task decomposition led to significant improvement within the LLM + scaffolding models: the "prompt-based questioner" models showed a better fit than the "one-shot CoT" model. Therefore, overall we found that the neuro-symbolic approach to open-ending pragmatic PCMs showed quantitative fit to human data on par with established cognitive modeling, while offering a more realistic interface to natural language inputs and outputs.

Qualitative results Next to the quantitative analyses, we analyzed qualitatively the differences between model predictions and the performance of the single modules. Figure 3 shows the proportions of different response categories (e.g., competitor, no-options responses etc.) predicted by the different models, next to PCM predictions and human data from Tsvilodub et al. (2023). The figure reveals that although many neuro-symbolic models have similar fit to human data in terms of Δ, there are qualitative differences in the predicted response proportions. The two models with "LLM semantics" overpredicted the proportion of unrelated responses, while the "LLM base-level semantics & utilities" model overpredicted the all-options response rate and slightly underpredicted the competitor rate.

Comparisons of the base-level and pragmatic respondent semantic modules revealed that the base-level semantics module performed reliably, while the pragmatic respondent semantic module made mistakes more frequently, including when evaluating unrelated responses. This may have led to the overprediction of the unrelated responses, as shown by the comparison of the "LLM semantics & utilities" and the "LLM base-level semantics & utilities" models because the former only differs from the latter by using an LLM-based pragmatic respondent semantics evaluator. We correlated the utility evaluator predictions with data elicited from humans for the PCM (see Figure 5) and found a very high correlation ($R = 0.92$), so we can likely rule out the utility evaluator as the source of overprediction of the unrelated category.

The comparison of the PCM + LLM models to the "full NeSy" model highlights the difference in response proportions that is driven by adding LLM proposers for the set of available responses and questions. The addition of response and question proposers decreased the rate of unrelated re-

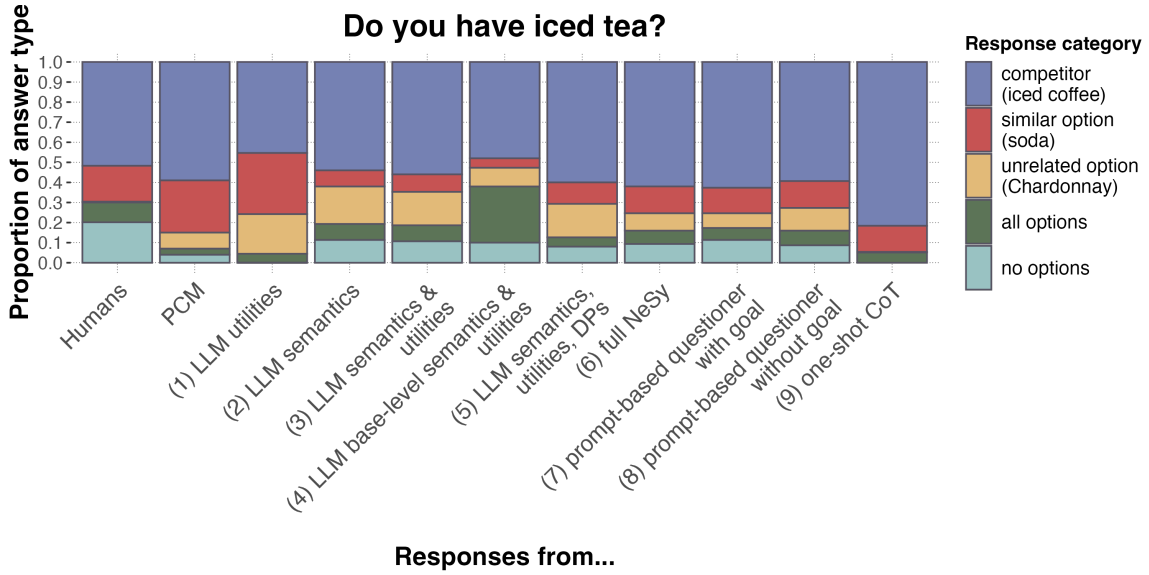

Figure 3: Proportions of different response categories produced by humans (left column) and predicted by different models. The categories are based on which options are mentioned in the response.

sponses and slightly increased the rate of similar and exhaustive responses. Since the "full NeSy" model included the pragmatic respondent semantic evaluator module, we can conclude that semantic evaluations might work more reliably with the LLM's own proposals than with the pre-specified sets of responses and questions. These observations are in line with one of the well-known challenges of neuro-symbolic modeling concerning difficulty of converting between neural and symbolic representations that is required in order to reliably compute truth values for open-ended sentences and contexts (Bader et al., 2004), as well as with debates around LLMs' ability to provide reliable evaluations (Bavaresco et al., 2024).

We also explored decreasing and increasing the n of alternative responses proposed by the LLM. We found that results with $n < 10$ proposals were unlikely to contain the "all options" or "no options" responses. For $n = 10$ this was more often the case, but we appended these two response types to set of alternatives manually nonetheless, to ensure availability of all conceptually meaningful response types. Sampling $n = 50$ responses ensured full coverage of response types but became computationally expensive. Generally the proposals often contained multiple instances of one response type (e.g., multiple competitor responses), an observation we return to in the discussion. However, this is unlikely the sole driving force beyond the fit of the framework, as the "LLM semantics, utilities, DP" model showed a similar competitor response proportion, while operating on a fully prespecified set of responses.

We qualitatively assessed the samples of the goal proposer module that generates possible text-based questioner goals, given the vignette. We compared the samples to human data from a web-based experiment wherein participants were asked to write three plausible goals of the questioner, given the vignette context (see Appendix C for details and human results). We focused on analyzing whether the LLM-proposed goal focused on getting the *target* mentioned in the question, on a more *general* information gain, or on *specific* situation aspects. We observed that, while LLM proposals were plausible, they focused on the target and specific goals around the target more, while humans showed more diversity in their specific goals, e.g., often involving social aspects of the described situation.

Turning to the LLM + scaffolding model type, comparing the "prompt-based questioner model without goals" and the "prompt-based questioner model with goals" revealed a trend towards predicting unrelated and similar responses more uniformly in the goal-free model, which is expected given that the distinction between these types of answers is based on reasoning about the questioner's goal. However, these differences are small and indicate that, even under certain (ablating, from a theoretical perspective) prompt variation, LLMs may be able to approximate pragmatic behavior.

Taken together, our key results are:

- the neuro-symbolic modeling approach fits human data quite closely, potentially making it a framework for computational modeling of pragmatic question answering performing on par with the PCM;

- at least some level of task decomposition when using LLM modules is required for a good fit to human data;

- LLM modules are generally good proposers, although attention should be paid to *types* of proposals that are expected for explanatory purposes;

- LLMs are good evaluators for functions based on abstract world knowledge like the utility evaluator;

- LLMs may struggle with truth-conditional semantics of certain utterances, but perform well when evaluating yes/no responses to polar questions.

5 Related work

Our work is situated at the intersection of several strands of like-minded work in different areas, in addition to the work we build on directly (Hawkins et al., 2015; Tsvilodub et al., 2023). The idea and promise of neuro-symbolic models has been studied in artificial intelligence for many years (Bhuyan et al., 2024). Further, our framework is closely related to recent work outlining various approaches to combining scaffolding structures, computational modeling or cognitive architectures with LLMs (e.g., Nye et al., 2021; Collins et al., 2022; Sumers et al., 2023; Wong et al., 2023; Kambhampati et al., 2024). Combining LLMs with PCMs specifically in the context of computational pragmatics has received some attention in recent work (e.g., Lew et al., 2020; Franke et al., 2024; Tsvilodub et al., 2024a) but the present work focuses specifically on systematically comparing and evaluating families of related models with varying degrees of neural or symbolic computation.

On an algorithmic level, our models combine several LLM calls in a particular architecture, which has been widely used in recent prompt techniques (Nye et al., 2021; Prystawski et al., 2023; Yao et al., 2023), and systems that use LLM calls to retrieve information (e.g., Lewis et al., 2020), to access different tools (e.g., Schick et al., 2023) or to solve complex reasoning tasks (e.g., Creswell et al., 2022; He-Yueya et al., 2023).

Systems with multiple LLM calls per input have also been specifically applied to question answering (Wang et al., 2023), mainly with a focus on improving factual accuracy of responses, or on training systems to improve their question asking capabilities (Andukuri et al., 2024). Therefore, our case study addresses a highly relevant task, with a novel focus on modeling *pragmatic, human-like* answering behavior.

6 Discussion

Taken together, in this case study we outlined and systematically assessed a neuro-symbolic framework for computational pragmatic modeling that uses probabilistic cognitive models as scaffolding structure that integrates LLM components for more flexible interfaces with language and background knowledge. The experiments on a case study of pragmatic question answering revealed that such modeling can be a viable candidate in the toolbox for more flexible models of human behavior in question answering. The systematic comparison of neuro-symbolic models with different degrees of task decomposition suggests fine-grained differences in how LLMs perform on different subtasks common to PCMs.

Our case study has several limitations, but also opens up paths for future work. For one, the full neuro-symbolic models implement Bayesian inference via enumeration, which results in computational bottlenecks when scaling the number of proposals and options in context. Related work connecting LLMs and Bayesian inference might be a promising avenue for improvements (Lew et al., 2023). Additionally, the current main results are based only on one closed-source LLM (but see Appendix D for exploratory results with an open-source LLM), and only use zero-shot prompting (except the CoT model). In this initial case study, we prioritized using relatively simple, non-engineered prompts, but nonetheless LLM prompting comes with potential risks of hallucination, errors and biases (e.g., Bender et al., 2021; Ji et al., 2023; Liu et al., 2023).

Finally, the use of LLMs as proposers and evaluators opens up interesting questions. For instance, response proposals supplied by the LLM might contain a trend towards certain response types, which can arguably be seen as a learned prior over human

preferences reflected in the training data. Additionally, cognitive models usually assume utterance costs for human language production and comprehension, but such online processing costs might not have a clear counterpart in LLMs. Further, varying performance of LLM evaluators might suggest that some aspects of semantics might be amortized in training data (White et al., 2020). Our results suggest that LLMs might not approximate different aspects of human intuitive knowledge equally well, touching upon important considerations of replacing human judgements with LLMs (Shiffrin and Mitchell, 2023; Löhn et al., 2024). For the LLMs + PCM models, one other potential source of improved performance with scaffolding of the LLM could be due to higher inference time compute budget that comes with decomposing the task into several LLM calls (Yu et al., 2024).

In sum, we presented a detailed case study as a starting point for exploring neuro-symbolic models of human language use, showing that task decomposition supplied by a cognitive model can be leveraged in synergy with recent LLMs, working towards open-ending pragmatic computational modeling.

Acknowledgments

We would like to thank Fausto Carcassi for his contributions to developing the framework, and the anonymous reviewers for insightful comments. MF is a member of the Machine Learning Cluster of Excellence at University of Tübingen, EXC number 2064/1 – Project number 39072764. PT and MF gratefully acknowledge the support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

References

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. 2024. STar-GATE: Teaching language models to ask clarifying questions. In *First Conference on Language Modeling*.

Sebastian Bader, Pascal Hitzler, and Steffen Hoelldobler. 2004. The integration of connectionism and first-order knowledge representation and reasoning as a challenge for artificial intelligence. *Preprint*, arXiv:cs/0408069.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2024. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. *URL https://arxiv.org/abs/2406.18403*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Anton Benz. 2006. *Utility and relevance of answers*. Springer.

Bikram Pratim Bhuyan, Amar Ramdane-Cherif, Ravi Tomar, and TP Singh. 2024. Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications*, pages 1–36.

Herbert H Clark. 1979. Responding to indirect speech acts. *Cognitive psychology*, 11(4):430–477.

Katherine M. Collins, Catherine Wong, Jiahai Feng, Megan Wei, and Joshua B. Tenenbaum. 2022. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. *Preprint*, arXiv:2205.05718.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Judith Degen. 2023. The rational speech act framework. *Annual Review of Linguistics*, 9(1):519–540.

Simon Farrell and Stephan Lewandowsky. 2018. *Computational modeling of cognition and behavior*. Cambridge University Press.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Michael Franke, Polina Tsvilodub, and Fausto Carcassi. 2024. Bayesian statistical modeling with predictors from LLMs. *arXiv preprint arXiv:2406.09012*.

Noah D Goodman and Andreas Stuhlmüller. 2014. The Design and Implementation of Probabilistic Programming Languages. `http://dippl.org`. Accessed: 2025-1-30.

Jeroen Antonius Gerardus Groenendijk and Martin Johan Bastiaan Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, Univ. Amsterdam.

Auli Hakulinen. 2001. Minimal and non-minimal answers to yes-no questions. *Pragmatics*, 11(1):1–15.

CL Hamblin. 1973. Questions in Montague English. *Foundations of Language*, 10(1):41–53.

Robert D. Hawkins and Noah D. Goodman. 2017. Why do you ask? The informational dynamics of questions and answers. PsyArXiv.

Robert D. Hawkins, Andreas Stuhlmüller, Judith Degen, and Noah D. Goodman. 2015. Why do you ask? Good questions provoke informative answers. *Cognitive Science*.

Robert D. Hawkins, Polina Tsvilodub, Claire Augusta Bergey, Noah D. Goodman, and Michael Franke. to appear. Relevant answers to polar questions. *Philosophical Transactions B*.

Joy He-Yueya, Gabriel Poesia, Rose E Wang, and Noah D. Goodman. 2023. Solving math word problems by combining language models with symbolic solvers. *arXiv preprint arXiv:2304.09102*.

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)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

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. *ACM Comput. Surv.*, 55(12).

Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. 2024. LLMs can't plan, but can help planning in LLM-Modulo frameworks. *Preprint*, arXiv:2402.01817.

Alexander K Lew, Michael Henry Tessler, Vikash K Mansinghka, and Joshua B Tenenbaum. 2020. Leveraging unstructured statistical knowledge in a probabilistic language of thought. In *Proceedings of the annual conference of the cognitive science society*.

Alexander K. Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash K. Mansinghka. 2023. Sequential Monte Carlo steering of large language models using probabilistic programs. *Preprint*, arXiv:2306.03081.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Lea Löhn, Niklas Kiehne, Alexander Ljapunov, and Wolf-Tilo Balke. 2024. Is machine psychology here? On requirements for using human psychological tests on large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 230–242, Tokyo, Japan. Association for Computational Linguistics.

Maxwell Nye, Michael Tessler, Josh Tenenbaum, and Brenden M Lake. 2021. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning. *Advances in Neural Information Processing Systems*, 34:25192–25204.

Kathryn Pruitt and Floris Roelofsen. 2011. Disjuntive questions: Prosody, syntax, and semantics. Handout, Göttingen.

Ben Prystawski, Paul Thibodeau, Christopher Potts, and Noah Goodman. 2023. Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.

Anselm Rothe, Brenden M Lake, and Todd Gureckis. 2017. Question asking as program generation. *Advances in neural information processing systems*, 30.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Richard Shiffrin and Melanie Mitchell. 2023. Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, 120(10):e2300963120.

Jon Scott Stevens, Anton Benz, Sebastian Reuße, and Ralf Klabunde. 2016. Pragmatic question answering: A game-theoretic approach. *Data & Knowledge Engineering*, 106:52–69.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. 2023. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Polina Tsvilodub, Michael Franke, and Fausto Carcassi. 2024a. Cognitive modeling with scaffolded LLMs: A case study of referential expression generation. In *ICML 2024 Workshop on LLMs and Cognition*.

Polina Tsvilodub, Michael Franke, Robert Hawkins, and Noah D. Goodman. 2023. Overinformative question answering by humans and machines. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Polina Tsvilodub, Paul Marty, Sonia Ramotowska, Jacopo Romoli, and Michael Franke. 2024b. Experimental pragmatics with machines: Testing LLM predictions for the inferences of plain and embedded disjunctions. In *Proceedings of CogSci*, pages 3960–3967.

Robert van Rooy. 2003. Questioning to resolve decision problems. *Linguistics and Philosophy*, 26(6):727–763.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Julia White, Jesse Mu, and Noah D. Goodman. 2020. Learning to refer informatively by amortizing pragmatic reasoning. *Preprint*, arXiv:2006.00418.

Li Siang Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought. *ArXiv*, abs/2306.12672.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. 2024. Distilling system 2 into system 1. *Preprint*, arXiv:2407.06023.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*.

A QA model

Below, we report the QA model by Hawkins et al. (to appear), described in Section 2, in more formal detail.

The base-level respondent that provides literal responses r to a question q given the world w is defined as follows:

$$R_0(r \mid w, q) \propto \begin{cases} 1 & \text{if } r \text{ is true in } w \text{ \& safe for } q \\ 0 & \text{otherwise.} \end{cases}$$

The notion of safety is couched in prior work on semantics of questions and answers (Pruitt and Roelofsen, 2011) and entails that, for the tested vignettes, only the literal answers $r \in \{\text{'yes'}, \text{'no'}\}$ are evaluated here.

The pragmatic questioner selects a question given their decision problem, based on the responses they expect from the base-level respondent R_0. Formally, a *decision problem* (DP) is a tuple $\mathcal{D} = \left\langle \mathcal{W}, \mathcal{A}, \mathcal{U}, \pi_Q^{\mathcal{W}} \right\rangle$, consisting of a set of world states $\mathcal{W}$, a set of options $\mathcal{A}$, a utility function $\mathcal{U} : \mathcal{W} \times \mathcal{A} \to \mathbb{R}$, and a probability distribution $\pi_Q^{\mathcal{W}} \in \Delta(\mathcal{W})$ capturing the questioner's prior beliefs about the world states. Then, the *value of a decision problem* $\mathcal{D}$ is the expected utility under a policy $\aleph^{\mathcal{D}}$ that chooses options according to their expected utility:

$$V(\mathcal{D}) = \mathop{\mathbb{E}}_{a \sim \aleph^{\mathcal{D}}} \left[\mathop{\mathbb{E}}_{w \sim \pi_Q^{\mathcal{W}}} \left[\mathcal{U}(w, a) \right] \right]$$

The pragmatic questioner then selects a question by soft-maximizing the expectation over the values of the decision problems $\mathcal{D}^{|r,q}$ given likely responses from the base-level respondent, resulting in $Q(q \mid D)$ (see Figure 4), where C(r) and C(q) are the production costs associated with the response and question, respectively.

The pragmatic respondent then reasons about the pragmatic questioner's choice of question in order to infer their likely decision problem:

$$\pi_{R_1}^{\mathfrak{D}|q}(\mathcal{D}) \propto Q(q \mid \mathcal{D})\, \pi_{R_1}^{\mathfrak{D}}(\mathcal{D})$$

Finally, the pragmatic respondent chooses a response by soft-maximizing the expected utility of the response given their posterior beliefs about the questioner DP. Utility is defined as a (parameterized) combination of informativity (defined via KL divergence) and action-relevance (defined via the decision problem value), resulting in $R_1(r \mid q)$ (see Figure 4).

A.1 Parameterization of the QA model

As commonly done for probabilistic modeling, in order to run simulations with the QA model parameters of the model were specified by the modelers or with elicited human data (Hawkins et al., to appear). For each vignette, the set of alternative questions included polar questions about the availability of each of the possible options individually, and a wh-question inquiring about all possible options.

$$Q(q \mid \mathcal{D}) = \underset{\alpha_Q}{\mathrm{SM}} \left(\mathop{\mathbb{E}}_{w \sim \pi_Q^{\mathcal{W}}} \left[\mathop{\mathbb{E}}_{r \sim R_0(\cdot \mid w, q)} \left[V(\mathcal{D}^{\mid r,q}) - \mathrm{C}(r) \right] \right] - \mathrm{C}(q) \right)$$

$$R_1(r \mid q) = \underset{\alpha_R}{\mathrm{SM}} \left(\mathop{\mathbb{E}}_{\mathcal{D} \sim \pi_{R_1}^{\mathcal{D} \mid q}} \left[(1-\beta) \left(-\mathrm{KL} \left(\pi_Q^{\mathcal{W}^{\mid r,q}} \mid\mid \pi_{R_1}^{\mathcal{W}} \right) \right) \; + \beta \, V \left(\mathcal{D}^{\mid r,q} \right) - \mathrm{C}(r) \right) \right] \right)$$

Figure 4: Formal definitions of the pragmatic questioner $Q(q \mid D)$ and respondent $R_1(r \mid q)$.

The set of available pragmatic answers included answers of all categories described in Section 3.1.

In order to specify the utility functions of the questioner DPs, a web-based experiment was run with human participants. Participants ($N = 453$) were asked to provide slider ratings for each possible option (e.g., iced tea, iced coffee, soda, Chardonnay), given another option as the goal. The full space of possible combinations was elicited. The slider ratings were on a scale of 0–100. Importantly, participants were asked to rate how happy they think a person would be to receive an option, given the target, resulting in *abstract* conditional preferences. The DP utilities for each vignette were bootstrapped from human preferences in the QA model simulations. Human results for ratings of the alternatives, given the option used as the target in the free production experiments as the goal (e.g., the iced tea) are shown in Figure 5 (left) together with respective LLM module predictions. Human and `GPT-4o-mini` ratings correlated highly, and supported the intuitive ordering of the relevance of alternatives (e.g., the competitor received higher ratings than the unrelated option for a given target).

B Prompts

Prompts for all LLM modules are presented below in Figures 6–14.

B.1 Semantic Evaluators

The base-level semantic evaluator only evaluates the set of literal responses {'yes', 'no'}. The pragmatic respondent semantic evaluator evaluates the set of possible overinformative responses. In models where the set of pragmatic responses is prespecified, the possible responses are of the form "I'm sorry, we don't have {target}. {continuation}", where the continuation was constructed for all response types (no-options, competitor, similar, unrelated, all-options responses).

C Human Experiment on Goal Inference

In an exploratory *goal inference* study, participants (N=35) were shown vignette contexts without the available options, followed by the question asked by a speaker. Participants were asked to name three plausible goals in three separate text fields that the questioner might have in mind when asking the question. We focused on distinguishing whether participants named goals focused on acquiring the *target* mentioned in the question, on acquiring more *general* information, or on goals related to more *specific* aspects of the situation.

Participants were most likely to infer *specific* goals (0.42 of the responses), followed by *target-related* goals (0.35 of the responses). More *general* information-seeking goals were less likely (0.17 of the responses), and some responses were non-classifiable (0.06).

We then manually analyzed the proposals of the LLM goal proposer module. Qualitatively, the target-related goals mostly were about acquiring the target or an item with the same functional features (e.g., when the target was veggie pizza, the functional feature would be being a vegetarian option), both for humans and LLMs. The specific goals produced by humans often involved more details than just acquiring the target, e.g., acquiring the target for a friend, or mentioned different specific preferences participants came up with. In contrast, the specific goals produced by LLMs were less likely to mention social aspects like acquiring something for a friend, and more likely to produce possible more specific questioner preferences (e.g., "asking about certain dietary restrictions"). The more general goals produced by humans and LLMs often mentioned learning about the set of available alternatives.

Figure 5: **Left**: GPT-4o-mini utilities plotted against human utilities, $R = 0.92$. **Right**: Qwen-2.5-32B-Instruct utilities plotted against human utilities, $R = 0.93$.

	1-shot CoT	1-shot example	1-shot explanation	0-shot
Qwen-2.5-32B-Instruct	0.21	0.15	0.25	0.28
Qwen-2.5-14B-Instruct	0.16	0.24	0.22	0.39
Qwen-2.5-7B-Instruct	0.33	0.19	0.50	0.17

Table 1: Jensen-Shannon divergence between human response proportions and the proportions of different response categories predicted by Qwen models of different sizes under various prompting (lower is better).

Utility Evaluator Prompt

```
In this study we are interested in how you think about other people.
On each trial, you will be given some information about a person: 'Suppose someone wants to have Italian food.'

Then we'll ask how happy you think this person would be about other things, given this information. For instance, we might ask: 'How happy do you think they would be if they had French food instead?'
You'll use ratings from 0-100 to answer the questions. Return the rating only.

Suppose someone wants {goal}. How happy do you think they would be if they got {option}?
```

Figure 6: **Utility Evaluator Prompt**

Base-level Evaluator Prompt

```
Safe answers to questions only provide information that the questioner genuinely does not know, given what they asked.
True answers to questions only provide information that is true given the context.

Here is an everyday situation where someone asks a question: {context + question}
Here is a potential answer to the question: {utterance}

Is the answer safe and true in this context, according to the definition above?
Return 'yes' or 'no' only.
```

Figure 7: **Base-level Evaluator Prompt**

Pragmatic Respondent Semantic Evaluator Prompt

```
True answers to questions only provide information that is true given the context.

Here is an everyday situation where someone asks a question: {state}
Here is a potential answer to the question: {utterance}

Is the answer true in this context, according to the definition above?
Return 'yes' or 'no' only.
```

Figure 8: **Pragmatic Respondent Semantic Evaluator Prompt**

Response Proposer Prompt

```
Safe answers to questions only provide information that the questioner genuinely does not know, given what they asked.
True answers to questions only provide information that is true given the context.

Here is a question someone could ask in an every day situation: {question}
Here are the available options: {options}

Generate {num_samples} literal answers to the question.
Return them as a numbered list.
```

Figure 9: **Response Proposer Prompt**

Question Proposer Prompt

```
Suppose a person has the following goal: {goal}
The person is in the following everyday situation: {context}
Generate {num_samples} well formed short questions(s) the person might naturally ask in the context to achieve their goal.
```

Figure 10: **Question Proposer Prompt**

Goal Proposer Prompt

```
You will be given a context in which a person asks a question.
What plausible different goals might the person be interested in, given what they asked?
Your task is to generate {num_samples} alternatives in a comma separated list.
```

Figure 11: **Goal Proposer Prompt**

Prompt-based questioner with goals

```
We are interested in how likely a person would be to ask the following question in a simple context, given their goal.
Please return only the likelihood, provided on a scale between 0 and 1.
Goal: {goal}
Context: {state}
{utterance}
```

Figure 12: **Prompt-based questioner with goals**

Prompt-based questioner without goals

```
We are interested in how likely a person would be to ask the following question in a simple context.
Please return only the likelihood, provided on a scale between 0 and 1.
Context: {state}
{utterance}
```

Figure 13: **Prompt-based questioner without goals**

One-shot chain-of-thought prompt

```
You are hosting a barbecue party. You are standing behind the barbecue. You have the following goods to offer: pork sausages, vegan burgers, grilled potatoes and beef burgers.
Someone asks: Do you have grilled zucchini?

Let's think step by step. You reason about what that person most likely wanted to have. That they asked for grilled zucchini suggests that they might want vegetarian food. From the items you have pork sausages and beef burgers are least likely to satisfy the persons desires. Vegan burgers and grilled potatoes come much closer. Grilled potatoes are most similar to grilled zucchini. You reply:

I'm sorry, I don't have any grilled zucchini. But I do have some grilled potatoes.
```

Figure 14: **One-shot chain-of-thought prompt**

D Simulation Results with an Open-Source LLM

Additionally to the main experiments performed with `GPT-4o-mini`, we ran all experiments with an open-source LLM — `Qwen-2.5-32B-Instruct` (Team, 2024), providing insights about advantages and open questions for our neuro-symbolic modeling framework when it is based on LLMs that can be run locally.

The experimental settings were the same as reported in 3.1. Quantitative results comparing the predictions of the different models to human results in terms of JSD improvement over a random baseline Δ, introduced in 4, are shown in Figure 15. The results indicate that some models with LLM evaluators (i.e., semantics and utility evaluators, models (1) and (3)) perform on par with the models based on a powerful closed-source LLM, as well as close to the original probabilistic model. The high correlation between DP utilities predicted by Qwen and human results (Figure 5, right) corroborates that such evaluations can also be reliably elicited from an open-source model. Similarly to GPT-based models, the performance of the utility evaluator was more robust than for the literal semantic evaluators, as indicated by the better fit to human data for model (1). However, for model (2) and for models introducing a proposer (models (4)–(5)) the fit of the models decreased. Manual evaluations of the single modules in these models indicated that, qualitatively, the generated evaluations and proposals were adequate for the respective modules. However, this LLM struggled more to follow formatting instructions, so that processing the proposals for passing them to the neural evaluator modules was more brittle. Simulation runs which resulted in unrecoverable parsing errors were excluded form analysis.[2] Models which use a Qwen-based prompted questioner module ((6)–(7)) improved the fit to human data over the random baseline, although the role of conditioning the questioner prompt on the goal was opposite to the GPT-based models.

Qualitative results comparing the proportions of different response types under different models are shown in Figure 16. The qualitative patterns suggest that Qwen-based models preferred responses mentioning a relevant alternative (i.e., competitor responses) over no options or exhaustive responses.

[2]For this reason, no results of the full neuro-symbolic model are reported.

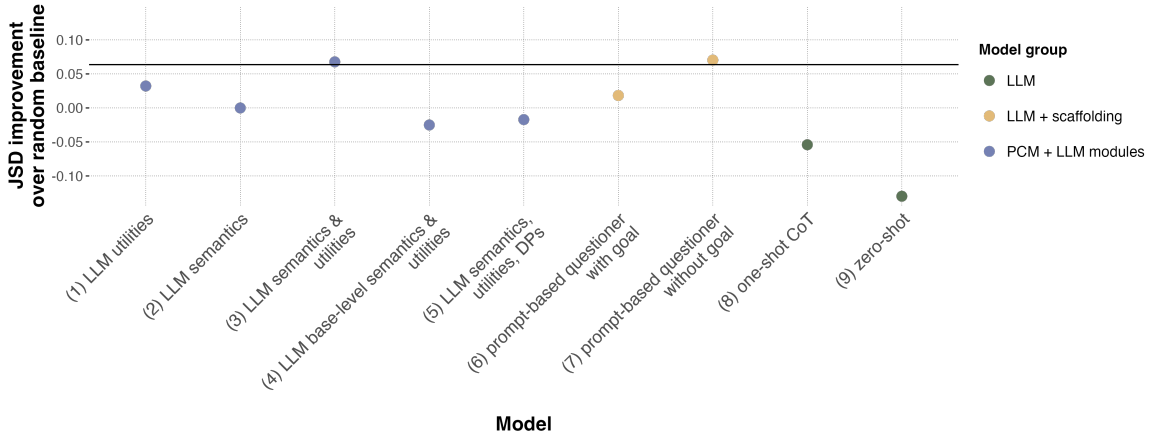

Figure 15: Improvement of the fit to human data of a model with an open-source Qwen-2.5-32B-Instruct backbone over a uniform response distribution baseline (higher is better). The horizontal line indicates the performance of the symbolic probabilistic model. The points indicate averages over simulations.

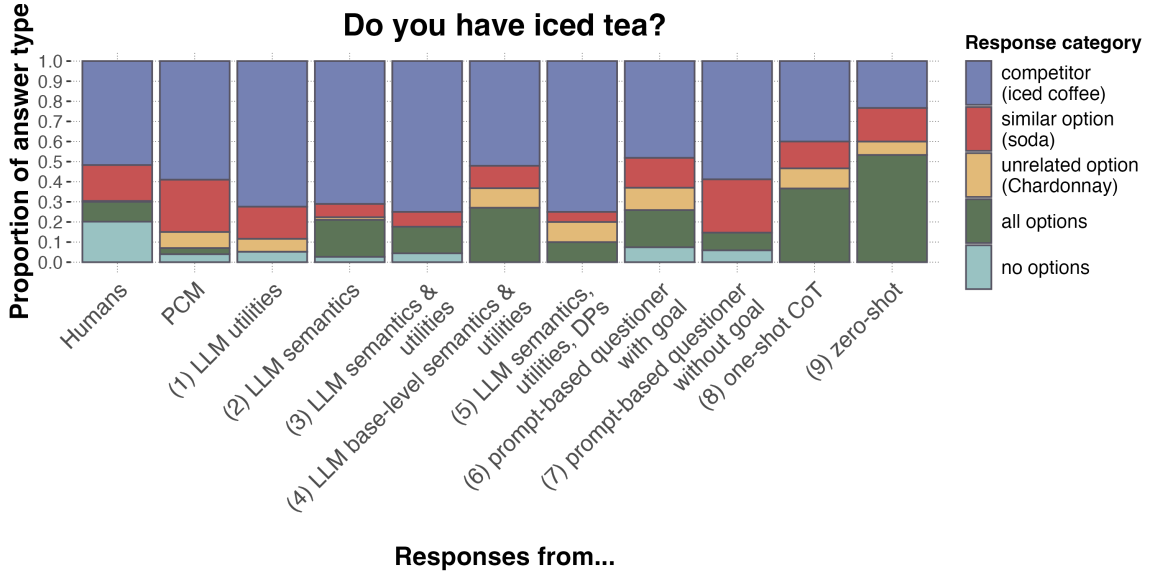

Figure 16: Proportions of different response categories predicted by Qwen-2.5-32B-Instruct used in different models (1–7), and with different prompting strategies (8–9).

LLM-only predictions, both in the one-shot chain-of-thought and the zero-shot prompting conditions, on the other hand, showed a larger proportion of exhaustive responses. We also report the JSD values for predictions from different sizes of Qwen under different prompting strategies from Tsvilodub et al. (2023) and human results in Table 1. These results suggest variation in the effectiveness of such prompting for different model sizes. For the two larger models, prompts that verbalize the PCM improve results over zero-shot prompting, although for the 32B model, ablated prompts further improve the fit to human data, suggesting substantial variation of human-likeness of the predictions when using only neural modules.

In sum, most neuro-symbolic Qwen-based models scaffolded with the PCM showed a better fit to human data than the random baseline, while the predictions of the LLM alone, even under one-shot chain-of-thought prompting, showed worse fit than the baseline. Additionally, given the open availability of the LLM, light-weight fine-tuning for better formatting instruction-following might offer a promising avenue for more robust neuro-symbolic modeling with open-source LLMs. Therefore, we can cautiously conclude that, given sufficient instruction-following capabilities for formatting, the neuro-symbolic framework might allow open-source LLMs to produce more human-like response patterns.