Are LLMs good pragmatic speakers?

Mingyue Jian
University of Edinburgh
s2531301@ed.ac.uk

N. Siddharth
University of Edinburgh
n.siddharth@ed.ac.uk

Abstract

Large language models (LLMs) are trained on data assumed to include natural language pragmatics, but do they actually behave like pragmatic speakers? We attempt to answer this question using the Rational Speech Act (RSA) framework [9, 10], which models pragmatic reasoning in human communication. Using the paradigm of a reference game constructed from the TUNA [27] corpus, we score candidate referential utterances in both a state-of-the-art LLM (Llama3-8B-Instruct) and in the RSA model, comparing and contrasting these scores. Given that RSA requires defining *alternative* utterances and a truth-conditional *meaning function*, we explore such comparison for different choices of each of these requirements. We find that while scores from the LLM have some positive correlation with those from RSA, there isn't sufficient evidence to claim that it behaves like a pragmatic speaker. This initial study paves way for further targeted efforts exploring different models and settings, including human-subject evaluation, to see if LLMs truly can, or be made to, behave like pragmatic speakers.

1 Introduction

With the emergence of large language models (LLMs) [1, 4, 6, 14, 15, 25, 26], a key question arises: can these models, trained on data presumed to include natural language pragmatics, exhibit pragmatic reasoning akin to humans? While LLMs have demonstrated formal linguistic competence (adherence to linguistic rules), their functional competence in pragmatic language use remains uncertain and warrants further investigation [17]. Although much research has focused on evaluating LLMs' pragmatic abilities as listeners, particularly their comprehension of non-literal language [12, 16, 19, 23, 24], less attention has been given to their pragmatic capabilities as speakers. Specifically, it remains unclear whether LLMs can effectively use context to generate non-literal utterances; i.e., being *informative* beyond being simply *true*. Investigating this aspect is crucial for deepening our understanding of LLMs' generative processes and enhancing their reliability.

Humans are typically pragmatic agents in communication. Imagine a room with three pieces of furniture: a small red desk, a small yellow desk, and a large red chair. If a friend directs your attention to "the red one", you would first eliminate the yellow desk as a possibility. Given the remaining red objects, your reasoning would then likely discount the chair, on account of it being distinct in the room, and that if it were the intended referent, the simpler utterance would have been just "the chair". This finally leads you to the intended referent being the small red desk. This recursive reasoning process that takes *informativity* into account beyond simply being *literally* true, encapsulates pragmatic communication.

A particularly influential model of pragmatic communication is the Rational Speech Act (RSA) framework [7, 10] which quantitatively models theory of mind [8, 21], formalising how speakers and listeners use context, shared knowledge, and probabilistic reasoning to communicate effectively. This framework operates language understanding as a recursive process, where both speakers and listeners in a conversation behave rationally to reason each other's intention.

NeurIPS 2024 Workshop on Behavioral Machine Learning.

In the earlier example, when a friend refers to "the red one", you exclude the yellow desk as a literal listener, guided by a "meaning function" that checks alignment with the words. The pragmatic speaker, your friend, chooses words with the expectation that you, as the listener, will interpret them correctly. As a pragmatic listener, you refine your interpretation, using both the message and context to infer intent. This interaction between speaker and listener, each considering the other's perspective, is key to effective communication, as modelled by the RSA framework.

Nguyen [20] applies the RSA framework to view RLHF-fine-tuned language models as bounded pragmatic speakers, where RLHF equips the LLM with a "theory of mind" listener model. This enables the LLM to anticipate listener interpretation when computing the distribution of the pragmatic speaker. However, this study does not examine the pragmatic generation ability of unmodified LLM. Carenini et al. [5] examines vanilla LLMs' pragmatic reasoning using RSA to find that GPT-2 XL's reasoning aligns with RSA in a metaphor task structured as "X is Y", with a restricted meaning and utterance space. However, this focuses on the LLM as a *listener*; we instead propose using a reference game to compare LLM scores as a *speaker* against RSA using a natural language format aligned with LLM training data. Additionally, we explore how the alignment varies with different sources of alternative utterances and RSA models that employ distinct meaning functions.

Our results indicate that our vanilla LLM model (Llama3-8B-Instruct [26]) has a positive correlation with the two RSA models that employ different meaning functions in the context of the reference game task, the correlation is stronger when scoring the logic-constructed utterances, and when the RSA model is with a rule-based meaning function. However, we do not see a clear alignment of the LLM's pragmatic scoring with that from the RSA models.

2 Rational Speech Act Model (RSA)

The RSA model iteratively refines a heterogeneous relation between alternative utterances U and intended meanings O, such that the relations begin being purely literal, and is refined using pragmatic reasoning: $U \times O \to [0, 1]$. The framework begins with a literal listener L_{lit}:

$$P_{L_{lit}}(o|u) \propto M(u, o) \cdot P(o), \tag{1}$$

where, in the context of our reference game, each object is equally likely to be selected, resulting in a uniform prior $P(o)$. Thus, the literal listener's interpretation relies entirely on the meaning function $M()$. The pragmatic speaker P_{S_p} is constructed from a literal listener L_{lit}:

$$P_{S_p}(u|o) \propto e^{\alpha(\ln P_{L_{lit}}(o|u) - \ln |u|)} = \left(\frac{P_{L_{lit}}(o|u)}{|u|} \right)^{\alpha}. \tag{2}$$

Here, $|u|$ is the utterance length imposing a cost on longer productions. This cost function aligns with the maxims of a pragmatic speaker [11], favouring the use of fewer attributes to convey the intended meaning within a controlled attribute space. Additionally, this approach ensures that comparisons between the RSA models and LLM remain valid. The preset prompt to the LLM (Appendix B) instructs it to describe the object using as few words as possible, effectively serving as a cost function that indirectly penalises longer outputs. α is a parameter that scale the rational level of S_p. A higher α will sharpen the probability distribution and vice verse.

2.1 RSA model with different meaning functions

We construct two RSA models with different meaning functions (MFs) for further investigation. They take the form $U \to [0, 1]$, indicating whether the utterance u literally describes the object o.

Prompt-based MF: This leverages the natural language understanding capabilities of LLMs for scoring. We use prompt engineering to generate numeric scores from the LLM, employing 3-shot prompting to guide the model with input-output examples that establish a fixed output template (Figure 6 in Appendix F). The prompt-based meaning function is defined as:

$$M_{\text{p}}(u, o) = \frac{P(\text{Yes} \mid \texttt{LLM}(o, u))}{P(\text{Yes} \cup \text{No} \mid \texttt{LLM}(o, u))}, \tag{3}$$

representing the probability of the model answering "Yes". The prompt template is refined through a process of trial and error (Appendix F).

Rule-based MF: This is based on feature exclusion: an utterance u that includes a feature contradicting those of the object o does not describe o. For example, if o is "a large, grey chair facing forwards", then the utterance u as "a green thing" does not describe o as the colour feature in u contradicts that of o. We define the rule-based meaning function as:

$$o = \{f_1, f_2, ...f_n\} \subsetneq F, \tag{4}$$

$$M_{\text{r}}(u, o) := \nexists w.(\exists f.f \in F \setminus o \wedge D(w, f)) \wedge (w \in u), \tag{5}$$

where $f_1, f_2, ...$ are the specific features of the object o, F is a full set of predefined features in the dataset, and D is a relation containing (w, f) iff word w describes feature f.

We evaluate the two MFs against human-labelled ground truth data and find that the rule-based function consistently identifies literal relationships for the logical-constructed sequences, and most of the top-k generated sequences (mean Acc = 99.9%), while the prompt-based method occasionally falls short in both constructed sequences (mean Acc = 97.3% for the 3-shot prompt-based method) (Appendix F).

3 Evaluation Pipeline

We use a reference game [13, 22] as the task, where a set of objects O includes target object o_t. The speaker selects an utterance $u_t \in U$ to convey o_t to the listener, who must identify it based on O and utterance u_t. For this task, we employ the TUNA dataset (furniture domain) [27], which organises each reference game around 7 objects with predefined attributes and features (Appendix A).

We propose a pipeline for the evaluation (Appendix C), which is organised into three stages: first, constructing alternative utterance U and meaning O spaces within the context of the reference game; second, getting scores from the vanilla `LLM` and the `RSA` models with different meaning functions for each alternative sequence; and finally, evaluating the output distribution across various metrics.

3.1 Construction of the meaning and utterance space

Meaning space: For each reference game, we map object attributes from the TUNA dataset into a noun phrase template to generate descriptions: `a <SIZE>, <COLOUR> <TYPE> facing <ORIENTATION>`. This results in the meaning space of a set of 7 object descriptions for each game.

Utterance space: In a reference game, the utterance space includes all alternative utterances within the restricted world that could describe any object in the meaning space. An optimal utterance space would encompass both literal and pragmatic expressions, enabling a thorough assessment of communicative effectiveness. However, even in a restricted setting, the construction of such an utterance space, accounting for the wide variety of sentence structures and connotations found in natural language, can be difficult. Previous research has predominantly concentrated on producing sentences that are pragmatic, rather than exploring the full spectrum of meaning generation [28]. In pragmatic referring expression generation, this typically involves sampling from a learned model during inference [2, 18, 28]. However, this method inherently produces only pragmatic sequences, as `LLM`s are presumed to be trained on pragmatic data. We present two approaches for constructing the utterance space U.

Top-k alternatives: This samples the top-k utterances from the `LLM` using beam search, generating pragmatic sequences with flexible phrasing. The `LLM` receives a prompt (Appendix B) containing the world context and a concise task description to identify the target object o_t, ensuring minimal instruction to test the `LLM`'s inherent pragmatic reasoning. To maintain consistency, generation begins with both "a" and "the" to ensure a noun phrase format. We generate sentences then deduplicate semantically identical ones that differ only in minor details like punctuation.

Logical rule alternatives: This approach constructs both pragmatic and literal utterances based on logical rules. This method is particularly effective in a text-based reference game setting, where a literal utterance includes all relevant features, while a pragmatic utterance may involve omitting some features. In particular, since the generated sequence should follow a noun-phrase format, the omission

for the object feature would be rephrased as 'thing'. The formalisation of the logical construction for the utterance space is as follows:

$$F_* = \bigtimes_{A \in \text{Attributes}} (F_A \cup \{\varepsilon\}), \tag{6}$$

$$U_{\text{logic}}(O) = \{u(\mathbf{f}) | \mathbf{f} \in F_* | \exists\, o \in O . \mathbf{f} \subseteq o\}, \tag{7}$$

where F_* represents all possible combinations of features in the reference game setting, and O is a set of objects in a particular game. $U_{\text{logic}}(O)$ is a set of possible utterances that describe objects in O. $u()$ is a function we define to transform the feature set $\mathbf{f}$ to an utterance. Appendix D displays the formalisation of the logical construction process with examples.

3.2 Getting scores from the two models

We score alternatives in the LLM and in RSA given the same reference game world.

Vanilla LLM: The logit probabilities, i.e., raw output values from the LLM before they are transformed into a probability distribution, directly indicate the model's preferences. These are used as scores to assess the LLM's behaviour:

$$p(u \mid O, o_t) = p(\boldsymbol{u} \mid \boldsymbol{c}(O, o_t)) = \prod_{i=1}^{N} p(u_i \mid \boldsymbol{c}(O, o_t), \boldsymbol{u}_{1:i-1}),$$

where $\boldsymbol{c}(\cdot)$ is the prompt template (Appendix B) and N is token length. Scores for top-k alternatives are generated along with the sequence using beam search. For the logic-constructed alternatives, we compute the probability retroactively for each utterance. Many popular pre-trained LLMs excel in downstream tasks, but few are open-source and provide logit probabilities. For this project, we use the open-source Meta-Llama3-8B-Instruct model [3, 26]. We access logit probabilities using the `python-llama-cpp` library.

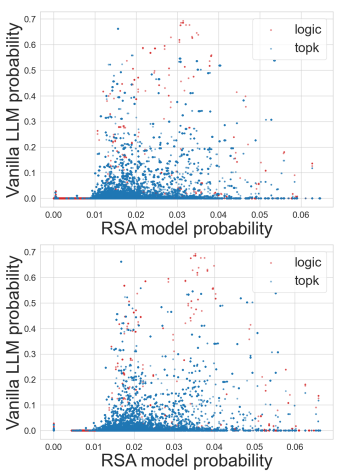
RSA models: The scores are calculated using Eq.1 and Eq.2 with the two constructed meaning functions.

4 Results

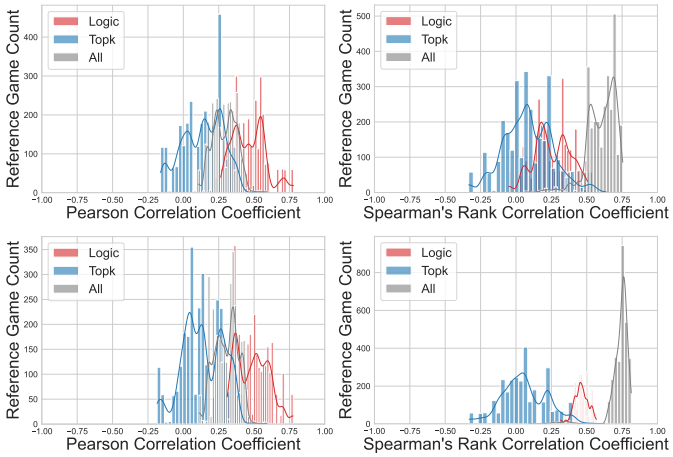
Data Overview: We have 2,940 reference games, each with a set of utterances describing one object, and we generate 386,510 utterance instances in total. Of these, 88,310 are generated by top-k sampling, and 298,200 by logic-based rules.

Evaluation metrics: We test models using probability outputs from the vanilla LLM and two RSA models: one with a prompt-based MF and the other with a rule-based MF. We assess the correlation between scores from the vanilla LLM and RSA models using Pearson Correlation Coefficient (PCC) for linear relationships and Spearman's Rank Correlation Coefficient (SRCC) for ranking similarity.

Experiment 1: We analyse the overall correlation across all reference games by comparing the scoring of each utterance instance related to any object for both models. Figure 1a illustrates the correlation between the vanilla LLM and the RSA models using different meaning functions. Both scatter plots show no clear linear relationship between the scores. The performance of the two meaning functions is comparable, as evidenced by the similar patterns observed across both plots. The RSA model's scoring can be interpreted as a spectrum from incorrect to literal to pragmatic utterances. The vanilla LLM scoring reveals that literal utterances are favoured by the LLM more than pragmatic ones. Particularly notable is the right-hand tail of the graph, where many pragmatic utterances are either minimally acknowledged or largely overlooked by the LLM. This suggests that while the LLM is able to correctly make factual judgement, they are unlikely to rank the utterances pragmatically. The RSA models in Figure 1 are configured with $\alpha = 1.0$. To further investigate the effect of α on correlation, we experiment with different values of $\alpha \in \{0.2, 0.6, 1.0, 1.4, 1.8, 3.0\}$. Correlation plots for each α are provided in Appendix E. Our results indicate that varying α does not alter our primary finding: the scores from LLM do not exhibit strong alignment with those of the RSA models, and the general distribution pattern remains consistent. Furthermore, we observe that increasing α sharpens the probability distribution, as reflected by the increase of the upper-bound on the x-axis.



(a) Overall correlation.



(b) Correlation histograms by condition with (l) PCC and (r) SRCC.

Figure 1: Correlation analysis for scores in the LLM and RSA models using (top) prompt-based MF and (bottom) rule-based MF. Red indicates logic-based alternatives, blue indicates the top-k alternatives and grey indicates all alternatives regardless of construction method.

Experiment 2: We then compare models by assessing their scoring preferences for each set of utterances describing each object, the result shows positive correlation on PCC and SRCC scores all model evaluations (Appendix G). Figure 1b displays the histogram for the correlation of each experimented instance on both metrics. The vanilla LLM aligns more closely with the RSA model using a rule-based MF (PCC =0.303, SRCC =0.736) than with the one using a prompt-based function (PCC =0.291, SRCC =0.606). We can see that the correlation between the LLM and RSA models is stronger for the logic-constructed utterances than for the top-k ones, likely due to the predictability of the former. In contrast, top-k utterances, prone to hallucinations, show more variability, highlighting the challenges LLMs face in maintaining coherence and accuracy in less structured tasks.

Interestingly, when evaluating performance on SRCC we find that correlation scores across the entire sequence space for each reference game set are positive regardless of meaning function method, with most exceeding 0.5. This suggests that the LLM can effectively distinguish top-k sequences from logically constructed sequences and rank them correctly to some extent. However, it lacks the ability to accurately rank the top-k sequences within the group, as indicated by SRCC scores ranging from -0.25 to 0.50 for both SRCC graphs.

5 Discussion

Based on our experiments and analyses, we see no clear evidence to suggest that LLMs are good pragmatic speakers. When comparing its scoring with the RSA models using different meaning functions, the LLM aligns more with the rule-based MF, especially for logic-constructed utterances. In our meaning function evaluation, the rule-based MF outperforms the prompt-based approach in factual judgement tasks, highlighting the LLM's strength in structured reasoning.

While these results highlight the LLM's pragmatic abilities in a controlled setting, their generalisability to everyday language remains uncertain. The structured nature of the reference games may not fully reflect the complexities of real-world communication. However, this research offers a framework for evaluating LLMs' pragmatic abilities and could be extended to more natural language use.

Future work should explore more diverse datasets to reflect a wider range of communication settings and natural language use. Testing on other LLMs, especially those with advanced pragmatic reasoning like GPT models trained on large datasets, would provide deeper insights into handling pragmatic tasks. Further research could also compare LLM alignment with the RSA model when iterated multiple times, rather than a single interaction, and examine the effects of scaling parameters and cost functions on alignment.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] J. Andreas and D. Klein. Reasoning about pragmatics with neural listeners and speakers. *arXiv preprint arXiv:1604.00562*, 2016.

[3] bartowski. bartowski/meta-llama-3.1-8b-instruct-gguf, 2024. URL `https://huggingface.co/bartowski/Meta-Llama-3.1-8B-Instruct-GGUF`. Accessed: 2024-10-29.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] G. Carenini, L. Bodot, L. Bischetti, W. Schaeken, and V. Bambini. Large language models behave (almost) as rational speech actors: Insights from metaphor understanding. In *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*, 2023.

[6] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways, 2022.

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

[8] D. C. Dennett. True Believers:The Intentional Strategy and Why It Works. In *Mind Design II: Philosophy, Psychology, and Artificial Intelligence*. The MIT Press, 03 1997. ISBN 9780262275071. doi: 10.7551/mitpress/4626.003.0003. URL `https://doi.org/10.7551/mitpress/4626.003.0003`.

[9] M. C. Frank and N. D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012. doi: 10.1126/science.1218633. URL `https://www.science.org/doi/abs/10.1126/science.1218633`.

[10] N. D. Goodman and M. C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.

[11] H. P. Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

[12] J. Hu, S. Floyd, O. Jouravlev, E. Fedorenko, and E. Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. arxiv. *arXiv preprint arXiv:2212.06801*, 2022.

[13] R. M. Krauss and S. Weinheimer. Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1:113–114, 1964.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[15] O. Lieber, O. Sharir, B. Lenz, and Y. Shoham. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*, 1, 2021.

[16] A. Louis, D. Roth, and F. Radlinski. " i'd rather just go to bed": Understanding indirect answers. *arXiv preprint arXiv:2010.03450*, 2020.

[17] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 2024.

[18] W. Monroe, R. X. Hawkins, N. D. Goodman, and C. Potts. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338, 2017.

[19] A. Nematzadeh, K. Burns, E. Grant, A. Gopnik, and T. L. Griffiths. Evaluating theory of mind in question answering. *arXiv preprint arXiv:1808.09352*, 2018.

[20] K. Nguyen. Language models are bounded pragmatic speakers: Understanding rlhf from a bayesian cognitive modeling perspective. *arXiv preprint arXiv:2305.17760*, 2023.

[21] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[22] S. Rosenberg and B. D. Cohen. Speakers' and listeners' processes in a word-communication task. *Science*, 145(3637):1201–1203, 1964.

[23] L. Ruis, A. Khan, S. Biderman, S. Hooker, T. Rocktäschel, and E. Grefenstette. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36, 2024.

[24] S. L. Sravanthi, M. Doshi, T. P. Kalyan, R. Murthy, P. Bhattacharyya, and R. Dabre. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. *arXiv preprint arXiv:2401.07078*, 2024.

[25] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[26] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[27] K. van Deemter, I. van der Sluis, and A. Gatt. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 130–132, 2006.

[28] J. White, J. Mu, and N. D. Goodman. Learning to refer informatively by amortizing pragmatic reasoning. *arXiv preprint arXiv:2006.00418*, 2020.

A TUNA dataset attributes and features

Attribute	Possible features
Type	chair, sofa, desk, fan
Colour	blue, red, green, grey
Size	large, small
Orientation	left, right, front, back

Table 1: Preset attributes and corresponding possible features for the 'furniture' domain in the TUNA dataset.

B Prompt template for getting top-k sequences and the scores from LLM

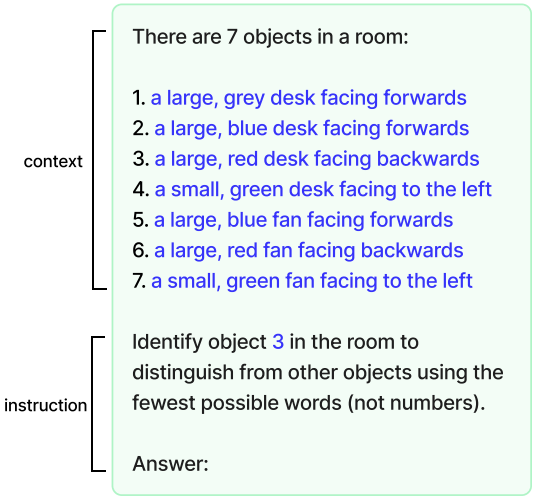


Figure 2: Example of the prompt used for generating top-k sequences with the LLM. The blue text indicates variable elements specific to each reference game instance.

C Evaluation pipeline

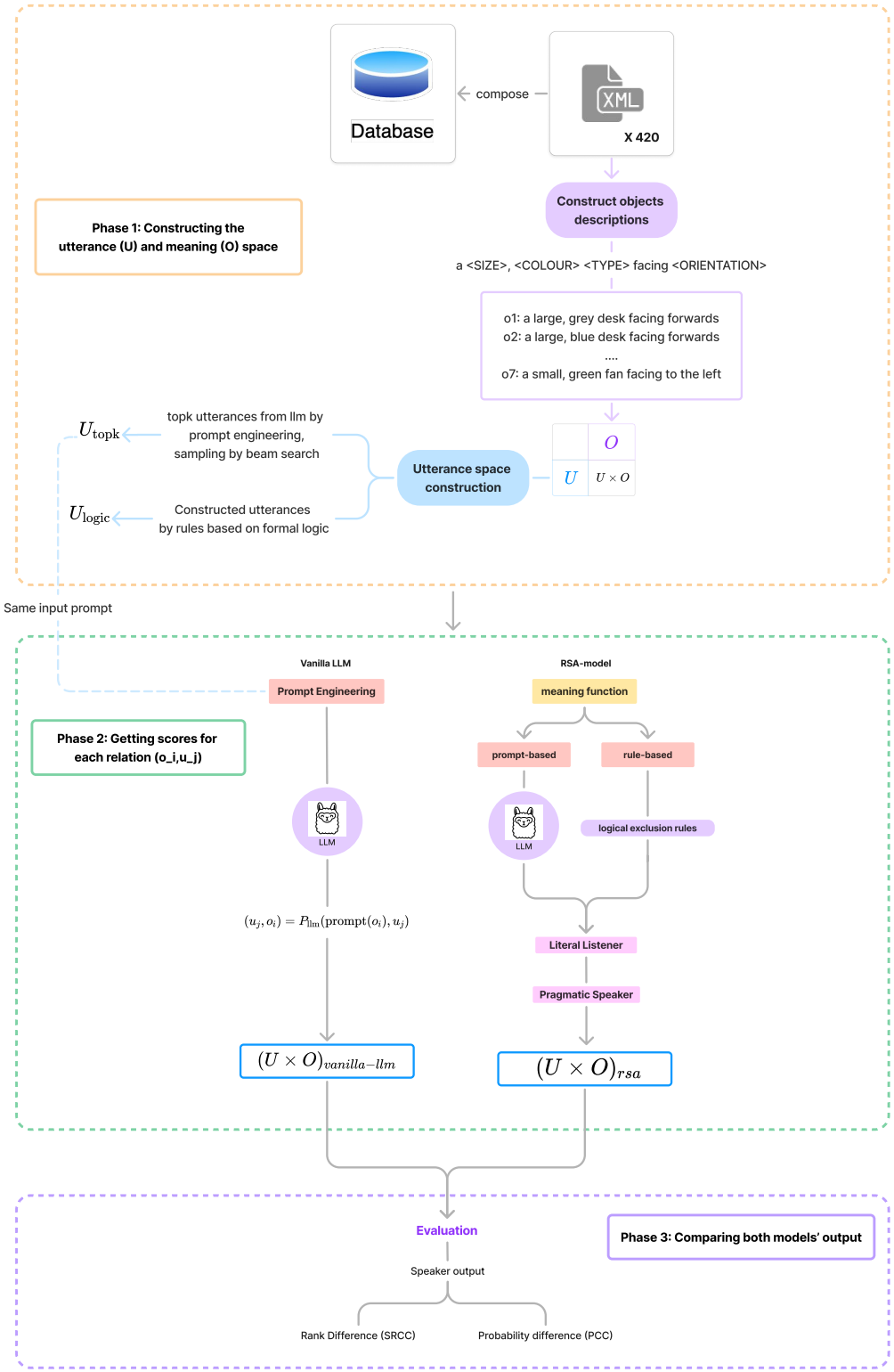


Figure 3: Project methodology pipeline

D Logical construction process

Figure 4 is a concrete example of a logical construction process.

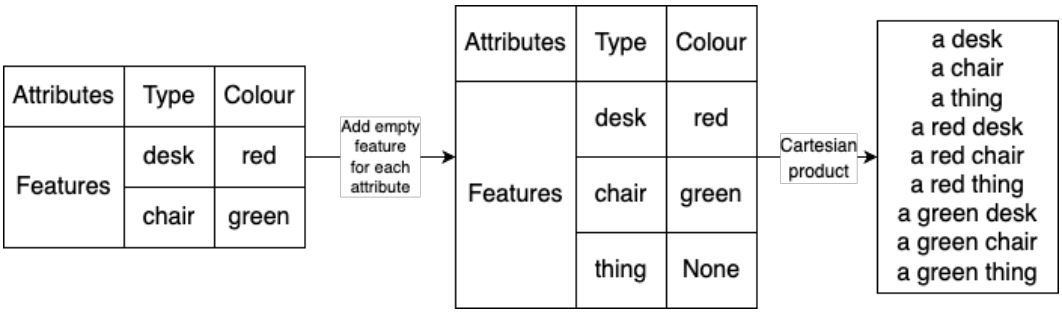
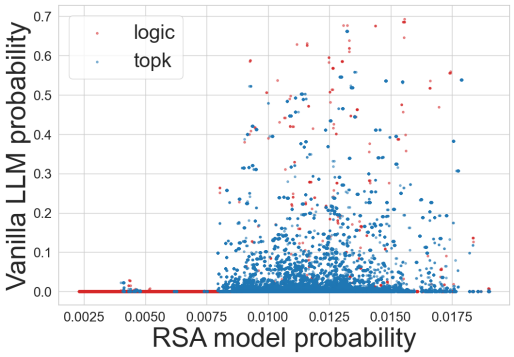
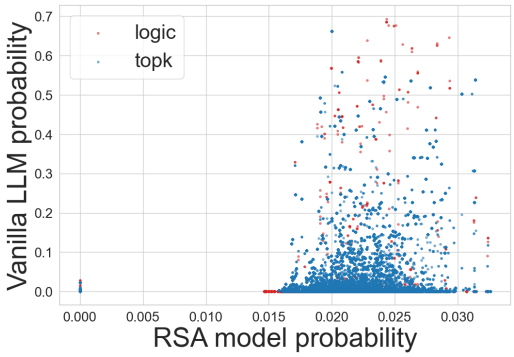


Figure 4: Example of logical construction process, given the attribute sets in the world.

E Results of the correlation scores of LLM and RSA model using different α values

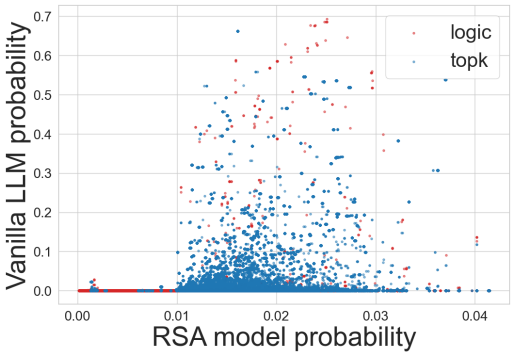


(a) PCC_{topk} = 0.153, PCC_{logic} = 0.372, PCC_{all} = 0.259

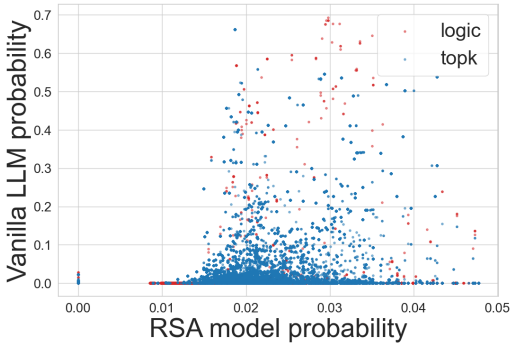


(b) PCC_{topk} = 0.149, PCC_{logic} = 0.395, PCC_{all} = 0.288

$\alpha = 0.2$

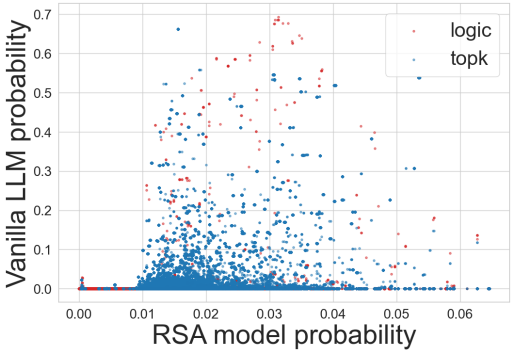


(c) PCC_{topk} = 0.147, PCC_{logic} = 0.417, PCC_{all} = 0.288

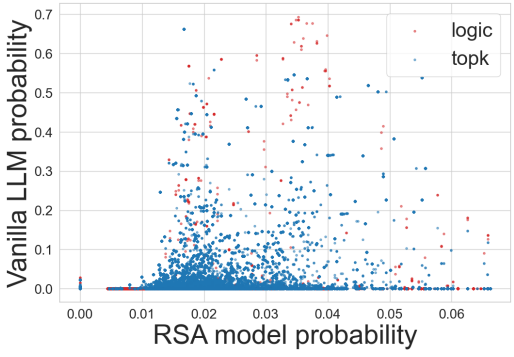


(d) PCC_{topk} = 0.146, PCC_{logic} = 0.459, PCC_{all} = 0.304

$\alpha = 0.6$



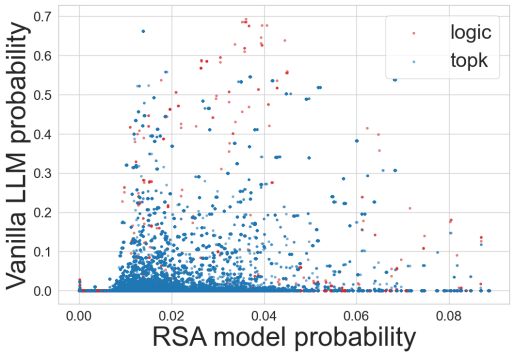
(e) PCC_{topk} = 0.139 , PCC_{logic} = 0.465, PCC_{all} = 0.291



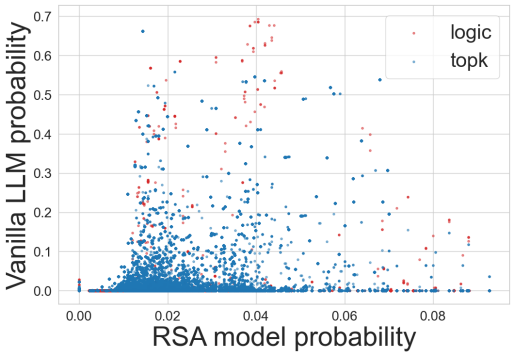
(f) PCC_{topk} =0.138 , PCC_{logic} =0.491, PCC_{all} = 0.303

$\alpha = 1$

Figure 5: Results of the correlation scores of LLM and RSA model using different α values. Each subplot group shows the overall correlation for scores in the LLM and RSA models using (left) prompt-based MF and (right) rule-based MF. We report the PCC scores for each utterances type.

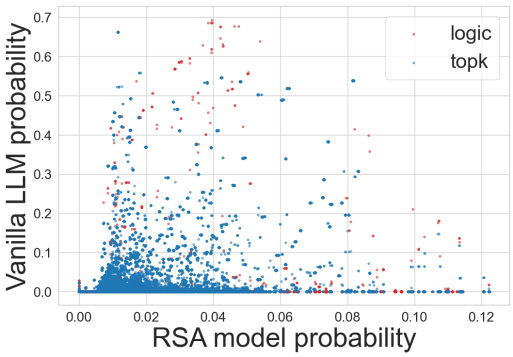


(g) $\texttt{PCC}_{topk}$ = 0.129, $\texttt{PCC}_{logic}$ = 0.476, $\texttt{PCC}_{all}$ = 0.278

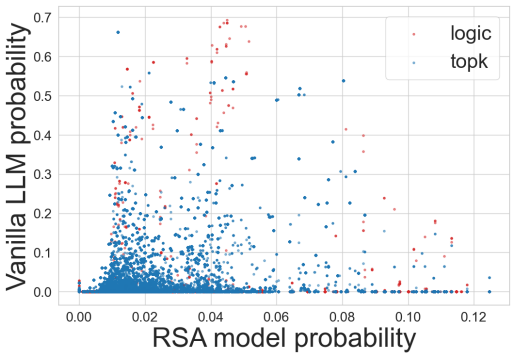


(h) $\texttt{PCC}_{topk}$ = 0.130, $\texttt{PCC}_{logic}$ = 0.496, $\texttt{PCC}_{all}$ = 0.290

$\alpha = 1.4$

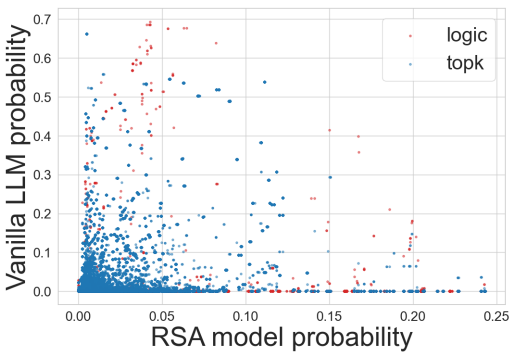


(i) $\texttt{PCC}_{topk}$ = 0.120, $\texttt{PCC}_{logic}$ = 0.465, $\texttt{PCC}_{all}$ = 0.257

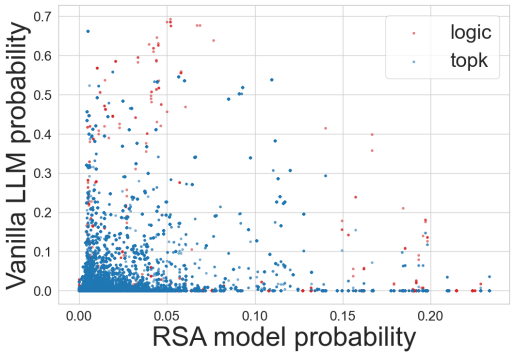


(j) $\texttt{PCC}_{topk}$ = 0.121, $\texttt{PCC}_{logic}$ = 0.482, $\texttt{PCC}_{all}$ = 0.269

$\alpha = 1.8$



(k) $\texttt{PCC}_{topk}$ = 0.089, $\texttt{PCC}_{logic}$ = 0.394, $\texttt{PCC}_{all}$ = 0.186



(l) $\texttt{PCC}_{topk}$ = 0.094, $\texttt{PCC}_{logic}$ = 0.412, $\texttt{PCC}_{all}$ = 0.202

$\alpha = 3$

Figure 5: Results of the correlation scores of LLM and RSA model using different α values. Each subplot group shows the overall correlation for scores in the LLM and RSA models using (left) prompt-based MF and (right) rule-based MF. We report the PCC scores for each utterances type.

F Meaning function evaluation

We evaluate the two proposed meaning functions (prompt-based MF and rule-based MF) by comparing their results on a set of test cases against human-labelled data. This evaluation is essential for fine-tuning parameters such as the number of examples given in the prompt, as well as for assessing the relative performance of the two meaning functions. We selected four constructed worlds for this purpose: "topk1" and "topk2", each comprising 2,056 (o, u) pairs generated from top-k sequences, and "logic1" and "logic2", containing 497 and 602 pairs respectively, with utterances derived from rule-based logical constructions.

For the prompt-based meaning function, we test with 3-shot prompts (Figure 6) and 6-shot prompts (Figure 7), and calculate the threshold T that would give the best performance for each n-shot prompt setting. The threshold allows us to compare the prompt-based meaning function to our ground-truth annotations, by considering values of at least T as 1 and other values as 0.

```
Does the description apply to the object?

1.
Description: That's a green sofa facing left.
Object: small, green sofa facing to the left
Yes

2.
Description: That's a green sofa facing left.
Object: small, grey sofa facing to the left
No

3.
Description: A fan.
Object:  large, grey fan facing backwards
Yes

4.
Description: {u}
Object: {o}\n
```

Figure 6: 3-shot template for the prompt-based meaning function.

```
Does the description apply to the object?

1.
Description: That's a green sofa facing left.
Object: small, green sofa facing to the left
Yes

2.
Description: That's a green sofa facing left.
Object: small, grey sofa facing to the left
No

3.
Description: A fan.
Object:  large, grey fan facing backwards
Yes

4. Description: A thing
Object: a large, grey chair facing forward
Yes

5.
Description: a small, grey desk
Object: a small, grey desk facing backwards
Yes

6.
Description: a grey chair
Object: a green desk
No

7.
Description: {u}
Object: {o}\n
```

Figure 7: 6-shot template for the prompt-based meaning function.

Table 2 displays the performance of the prompt-based meaning function across different metrics using n-shot ($n = \{3, 6\}$) with optimised thresholds for best performance, as well as the performance of the rule-based meaning function. Overall, the rule-based meaning function consistently identifies

the literal relationships of all logic-constructed pairs, but falls short on top-k generated pairs that require natural language understanding ability for correct interpretation. For example, for the pair (u_{topk} ="the small grey chair that is not facing forwards", o ="a small, grey chair facing backwards"), the rule-based meaning function scored it wrongly as it fails to interpret negation correctly.

The prompt-based method occasionally falls short as well. It generally performs better on the top-k generated sequences comparatively to the logical constructed ones. Errors are more frequent when the sentence object is "thing" within the group of logically constructed sequences. Notably, the prompt-based meaning function achieves higher performance with our constructed 3-shot prompt compared to the 6-shot prompt.

We consider the rule-based meaning function particularly well-suited for our reference game setting due to the restricted attribute set for each object. Although the generated top-k sequences may exhibit more diverse phrasings — such as when u_{topk} is "a tiny green table" and o is "a small green desk", where "tiny" and "table" are not present in the world vocabulary, the variations still fall within the attribute space of the furniture domain ($A_{\text{furniture}}$ = {'Type', 'Colour', 'Orientation', 'Size'}). Consequently, the rule-based meaning function can effectively capture these variations through synonym mapping.

We anticipate that a more carefully crafted prompt or a more advanced language model could improve the performance of the prompt-based meaning function, although this would necessitate additional human and time resources. Nonetheless, this type of meaning function may be more suitable in a more flexible task setting, where the relationships between o and u extend beyond literal templates and require natural language understanding ability.

	3-shot (T = 0.6)			6-shot (T = 0.8)			rule-based		
	Acc	P	R	Acc	P	R	Acc	P	R
topk1	0.997	0.992	0.988	0.991	0.973	0.976	0.999	0.996	0.996
topk2	0.987	0.925	1.000	0.972	0.875	0.966	0.999	1.000	0.996
logic1	0.956	1.000	0.804	0.966	0.970	0.875	1.000	1.000	1.000
logic2	0.952	0.966	0.768	0.938	0.830	0.830	1.000	1.000	1.000

Table 2: Performance of the prompt-based and rule-based meaning function across different metrics (Accuracy, Precision and Recall), the prompt-based meaning functions are using n-shot ($n = \{3, 6\}$) with optimised thresholds T for best performance.

G Experiment 1: Evaluation between each reference game

Table 3 presents the mean scores and standard deviations of PCC and SRCC across all reference games, when the utterance space is composed of the two different utterance types, and the RSA model is calculated with two different meaning functions.

Overall, the six model evaluations all show a positive correlation between the scoring of the vanilla LLM and the RSA model. The results indicate a preference for utterance type in the meaning function. The PCC and SRCC scores for logic-constructed utterances are higher when the RSA model is configured with a rule-based meaning function. However, the LLM shows stronger alignment with the top-k utterances compared to the RSA configured with a prompt-based meaning function. This aligns with our evaluation of the meaning function: the rule-based meaning function achieves full accuracy when judging pairs in a controlled setting but falls short in flexible settings requiring advanced natural language understanding for interpretation.

Focusing on the SRCC scores, an interesting observation is that while the SRCC score for top-k utterances is very low, the SRCC score for all utterances, regardless of type, is high (0.606 and 0.736, respectively). This suggests that the LLM can rank sequences of different constructed types correctly to some extent, but lacks the advanced pragmatic ability to rank pragmatic sequences accurately, as indicated by the low SRCC scores for the top-k cases (0.086 and 0.059, respectively).

		PCC		**SRCC**	
Utt. Type	**RSA MF**	**Mean**	σ	**Mean**	σ
Logic	Prompt-based	0.465	0.118	0.253	0.128
	Rule-based	**0.491**	0.114	**0.460**	0.051
Top-k	Prompt-based	**0.139**	0.145	**0.086**	0.173
	Rule-based	0.138	0.144	0.059	0.163
All	Prompt-based	0.291	0.082	0.606	0.103
	Rule-based	**0.303**	0.083	**0.736**	0.078

Table 3: Comparison of the mean and standard deviation (σ) for PCC and SRCC metrics across different reference games, highlighting the correlation between the LLM model and two RSA models using different meaning functions (MFs), and with different utterance types.