Can LLMs Understand Quantifiers Like Humans?

Essam Sleiman (esleiman@g.harvard.edu)
Harvard University

Abstract

This study delves into the significant role of quantifiers such as ”few,” ”some,” and ”many” in human cognition and language, drawing upon Gricean reasoning and the Rational Speech Act (RSA) framework. We explore how these quantifiers facilitate effective information conveyance by framing communication as a cooperative process between speakers and listeners. Leveraging probabilistic models (RSA) framework, we study how quantifiers are semantically understood. Furthermore, we extend our analysis to compare quantifier usage between humans and Large Language Models (LLMs) and investigate the influence of contextual cues on quantifier selection by both groups. Our findings suggest that LLMs closely align with human cognition, particularly when less context is provided, emphasizing the role of context in quantifier usage.

Code: https://github.com/essamsleiman/CocoSciFinal.git
Keywords: Quantifier; RSA; LLM

Introduction

Quantifiers, such as the words: few, some, and many, play a pivotal role in human cognition, serving as essential tools for categorizing, reasoning, and communicating ideas. These words facilitate the ability for humans to sort and measure experiences, creating a bridge between the use of language and the way humans represent quantities.

Building on the role of quantifiers in cognition, the theory of pragmatics as proposed by (Grice, 1975) (Gricean reasoning) explains their use. This theory posits that speakers, being rational beings, tailor their language to achieve specific conversational objectives, choosing expressions that most directly convey their knowledge and intentions. For instance, a well-informed speaker would more likely state a specific fact (like ”The person is at work”) rather than use a broader, less informative expression (like ”The person is at work or school”). Furthermore, the Rational Speech Act (RSA) model was introduced to explain various psycho-linguistic outcomes in the realm of pragmatic reasoning with a probabilistic model. (Goodman & Stuhlmüller, 2013) (Goodman & Frank, 2016) (Frank, 2016) (Kao, Bergen, & Goodman, 2014).

The RSA framework betters the understanding of human communication by framing it as a recursive conversation between a speaker and a listener. At its core, RSA models communication as a cooperative process where a speaker intentionally conveys maximal information to a listener, who then interprets this information through reasoning about the speaker’s intentions. This model uses Bayesian inference to deduce the most likely state of the world based on the utterance, while simultaneously considering the speaker’s choice of words to effectively communicate with a naive listener. (Yuan, Monroe, Bai, & Kushman, 2018)

Recently, the principles of Gricean reasoning and the RSA framework have been utilized to better understand and model human communication using quantifiers in language. (van Tiel, Franke, & Sauerland, 2021) In this study, we employ these probabilistic pragmatic models as tools to evaluate theories of semantics, particularly concentrating on quantifiers like ”few”, ”some” and ”most”. This focus allows us to delve into how these quantity words function within the framework of semantic understanding.

Expanding upon this approach, we aim not only to utilize these probabilistic models for a deeper insight into human usage of quantifiers but also to apply the same probabilistic framework to the outputs of Large Language Models (LLMs). This allows us to directly compare and contrast the nuances of quantifier usage between human cognition in the form of communcation and artificial linguistic systems, to better understand how similar LLMs understand quantifiers in comparison to humans.

Furthermore, our investigation extends to examining how varying levels of context in prompts influence the use of quantifiers by both humans and LLMs. By altering the context provided in these prompts, we aim to explore how context dependency affects the interpretation and selection of quantifiers. This study not only highlights the differences and similarities in quantifier usage between human cognition and LLMs but also sheds light on the role of contextual nuances.

Our results suggest that LLMs more closely match human cognition when it comes to using quantifiers in the case that less context is provided.

Related Work

Rational Speech Act

In the Rational Speech Act (RSA) framework, researchers explore the role of quantifiers in communication by analyzing the interaction between a speaker and listener. The RSA framework categorizes listeners into two types: the Pragmatic Listener (L1), who delves beyond literal meanings to infer deeper intentions, and the Naive, Literal Listener (L0), who interprets language at face value. Meanwhile, the Pragmatic Speaker (S1) carefully selects quantifiers to communicate effectively. This framework can be used to better understand quantifier use, in an effort for better human language and communication interpretation. Researchers leverage RSA to examine how different contexts and speaker intentions influence quantifier selection and interpretation.(Goodman & Stuhlmüller, 2013) (Goodman & Frank, 2016) (Frank, 2016) (Kao et al., 2014).

Quantifiers in Large Language Models

With the rising popularity of Large Language Models (LLMs), recent work has studied how well LLMs can understand quantifiers.

Studies by (Pezzelle, Steinert-Threlkeld, Bernardi, & Szymanik, 2018) and (Talmor, Elazar, Goldberg, & Berant, 2020) highlight LLMs' difficulties with understanding quantifiers, revealing a notable gap in their ability to use them accurately and contextually. (Kalouli, Sevastjanova, Beck, & Romero, 2022) furthered this exploration by focusing on logical quantifiers such as 'all', 'every', and 'some', offering detailed analysis on how LLMs process these quantifiers. The concept of inverse scaling, where larger LLMs may exhibit decreased performance in quantifier tasks, was introduced by (Michaelov & Bergen, 2022), presenting a significant scalability challenge for LLMs. A study by (Jang, Ye, & Seo, 2023) investigated how LLMs handle negation, which can be extended to their lack of understanding of the negation of quantifiers. These collective efforts provide a better understanding of the state of LLMs in terms of quantifier processing, indicating both the current challenges and potential directions for future research in this area of language understanding.

Human vs LLM

To to best of our knowledge, prior work has not investigated how humans and Large Language Models (LLMs) differ in understanding quantifiers using the Rational Speech Act (RSA) method. Secondly, we have not seen work investigating how LLMs handle quantifiers in varying contexts and levels of specificity in comparison to humans.

This work works to answer the following two questions:

(1) Can we quantify how humans understand language quantifiers in comparison to State of the Art, Large Language models?

(2) When more or less context if provided about a quantifier, how does the comparison between human understanding of quantifiers and Large Language Models differ?

Methods

In this section, we present the survey questions we used and the experiments we ran. Our first step was to survey both humans and large language models (our respondents) with questions that helped us get a better understanding of how the two groups represent quantifiers. This can be done by asking a survey question that has a hypothetical scenario, and asking what quantifier would be used in the scenario. We investigate different questions that has various levels of context within the scenario. We then build a model using the RSA framework around the responses from humans and another for the responses of the LLMs. Lastly, we compare the two distributions obtained from these two models.

Survey Questions

We considered a couple of different ”prompts” to use as survey questions. We first considered the following type of question:

Question: Given the following scenario, there are some X? Please choose a number that best represents 'some' in this context.

This type of prompt requests the respondents to assign a number that corresponds to a quantifier in a hypothetical scenario. After careful consideration, we decided against this prompt as it would require us to survey each question the number of times equal to the number of quantifiers. We then decided to flip the prompt such that it follows the following format:

Question: X Number out of Y? Which quantifier would you use to describe the number of X out of Y? Quantifier Options: [option1, option2, option3,..]

This prompt allows us to condition on the quantifiers without surveying the number of times equal to the number of quantifiers times.

We developed the following 2 survey questions to be asked to both humans and our LLM. These were crafted where the first question provides more context than the second one.

1. 5 students in scored above 90% out of a class of 30 students. Question: ”Which quantifier would you use to describe the number of students who scored above 90%? Options: [“a couple”, 'few', 'some', “several” 'most', “a majority”, “almost all” 'all']

2. 5 students scored above 90% out of a class of students. Question: ”Which quantifier would you use to describe the number of students who scored above 90%? Options: [“a couple”, 'few', 'some', “several” 'most', “a majority”, “almost all” 'all']

The first question includes the total number of students in a class while the second does not, leaving that up to the respondent's intuition.

Experiment

Given the limited time, we tried our best to survey a diverse population of humans. To decide which LLMs to use, the easiest and most available Large Language Models to prompt are OpenAI's GPT3.5 and GPT4 models, accessed through the https://chat.openai.com/ interface.

We ask 10 individuals each of the two questions. Then we prompt each of GPT3.5 and GPT4 5 times each with both questions and note the responses. We then use the RSA framework to model the human responses and the LLMs' responses.

After modeling each of our respondent groups, the RSA framework allows us to run two experiments. Firstly, given a quantifier, the model returns a probability distribution across all possible numbers that correspond to that quantifier. Second, given a number, the model returns a probability distribution across all possible quantifiers corresponding to that value. We provide figures for both distributions in both question scenarios.

To compare the two models, we compare the difference in the probability distributions assigned to each quantifier given a value. We average the difference over distributions over multiple values. In our paper, we test the two values: 1 and 5, however, given more time in a future iteration of this work, we would compare all possible values for a better comparison of the two models.

Results

For each of the survey questions, we've provided two sets of figures. The first compares the distributions assigned to each quantifier given a number. The second compares the distributions assigned to each value given a quantifier.

We then calculate the difference between the distributions using the kl-divergence (Kullback–Leibler divergence). Results are found in the table. The results indicate that when less context is provided in a prompt, understanding of quantifiers is better matched between humans and LLMs in comparison to when context is provided.

Table 1: KL-divergence between distributions of RSA Model on Humans and LLMs.

Experiment type	avg divergence
Question 1	0.27019
Question 2	0.04736

Discussion/Conclusion

In conclusion, we formulate a method to compare quantifier understanding using the RSA framework between humans and advanced Large Language Models. Our results suggest that Language Models behave more similarly to Humans when less context is provided. With more time, we hope to survey more people and provide a more detailed comparison between the two models, specifically in the distribution comparison. With these two, we believe we can produce a more informed and accurate comparison to validate the results we have so far.

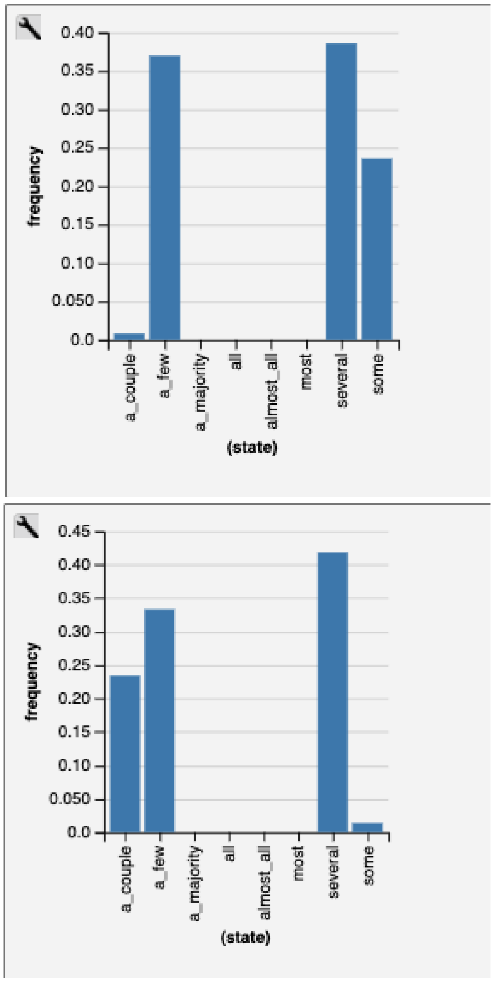

Figure 1: Question 1: A comparison of the probability distribution across quantifiers assigned to the value of 10 out of 30. Top: LLM. Bottom: Human.

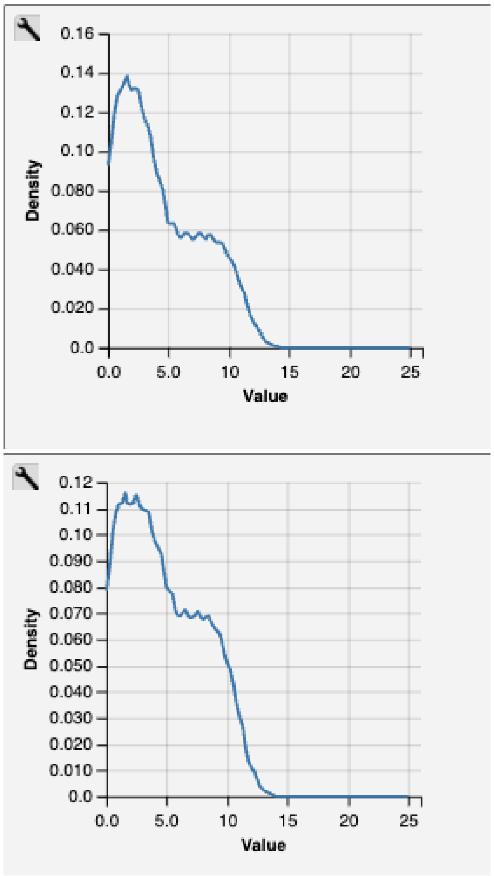

Figure 2: Question 1: A comparison of the probability distribution of the value from 0-30 of the quantifier: a few. Top: LLM. Bottom: Human.

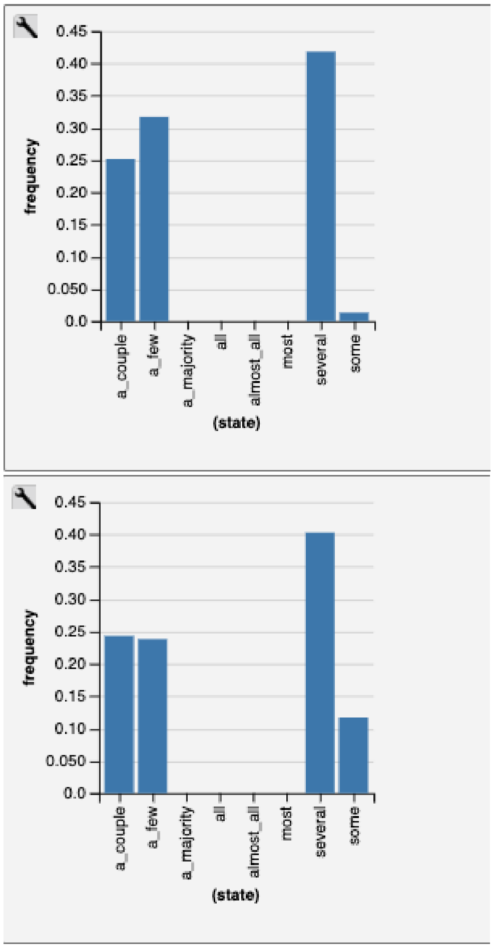

Figure 3: Question 2: A comparison of the probability distribution across quantifiers assigned to the value of 10 out of 30. Top: LLM. Bottom: Human.

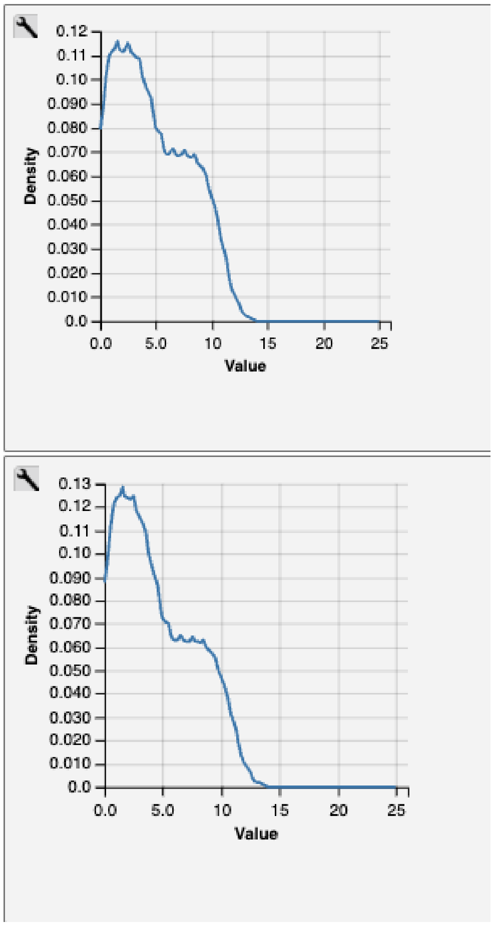

Figure 4: Question 2: A comparison of the probability distribution of the value from 0-30 of the quantifier a few. Top: LLM. Bottom: Human.

Author Contribution

I worked on project alone, all contributions are from me, with some advice from Ced Zhang.

References

Frank, M. C. (2016). Rational speech act models of pragmatic reasoning in reference games.

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

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Jang, J., Ye, S., & Seo, M. (2023). Can large language models truly understand prompts? a case study with negated prompts. In *Transfer learning for natural language processing workshop* (pp. 52–62).

Kalouli, A.-L., Sevastjanova, R., Beck, C., & Romero, M. (2022). Negation, coordination, and quantifiers in contextualized language models. *arXiv preprint arXiv:2209.07836*.

Kao, J., Bergen, L., & Goodman, N. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Michaelov, J. A., & Bergen, B. K. (2022). ’rarely’a problem? language models exhibit inverse scaling in their predictions following’few’-type quantifiers. *arXiv preprint arXiv:2212.08700*.

Pezzelle, S., Steinert-Threlkeld, S., Bernardi, R., & Szymanik, J. (2018). Some of them can be guessed! exploring the effect of linguistic context in predicting quantifiers. *arXiv preprint arXiv:1806.00354*.

Talmor, A., Elazar, Y., Goldberg, Y., & Berant, J. (2020). olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, *8*, 743–758.

van Tiel, B., Franke, M., & Sauerland, U. (2021). Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences*, *118*(9), e2005453118.

Yuan, A., Monroe, W., Bai, Y., & Kushman, N. (2018). Understanding the rational speech act model. In *Cogsci*.