Can LLMs Understand Quantifiers Like Humans?

Essam Sleiman[1] **Ced Zhang**[2] **Joshua Tenenbaum**[2]
[1]Harvard University [2]Massachusetts Institute of Technology (MIT)

Abstract

Quantifiers such as "few," "some," and "many" express imprecise numerical information, and their use depends on context. We test whether humans and Large Language Models (LLMs) select quantifiers similarly and whether specifying the reference-set size changes their responses. We collected 20 human quantifier judgments and 20 responses from GPT-3.5 and GPT-4 across two prompt conditions. One prompt specified that 5 of 30 students scored above 90%; the other omitted the class size. We used these response distributions in Rational Speech Act models and compared the human and LLM pragmatic-speaker distributions using Kullback–Leibler divergence. Average divergence was 0.51570 when the class size was specified and 0.40038 when it was unspecified. Thus, within this study, the modeled distributions were closer when the reference-set size was not provided. These preliminary results motivate broader study of how contextual specificity shapes human and LLM quantifier use.

Code: `Quantifiers-In-LLMs`

Keywords: Quantifier; RSA; LLM

Introduction

Quantifiers such as "few," "some," and "many" allow speakers to describe quantities without providing exact numerical values. Their meanings overlap and can change with context. For example, five students may be described differently when the total class size is known than when it is unspecified. Quantifier selection therefore provides a useful setting for examining how contextual information shapes language use (van Tiel et al., 2021).

Gricean pragmatics treats communication as cooperative and proposes that speakers select expressions that are informative for their conversational goals (Grice, 1975). The Rational Speech Act (RSA) framework formalizes this idea as probabilistic reasoning between speakers and listeners (Goodman and Stuhlmüller, 2013; Goodman and Frank, 2016). A speaker selects an utterance based on a state of the world, while a listener interprets the utterance by reasoning about why the speaker selected it. This Bayesian formulation provides quantitative predictions about pragmatic language use.

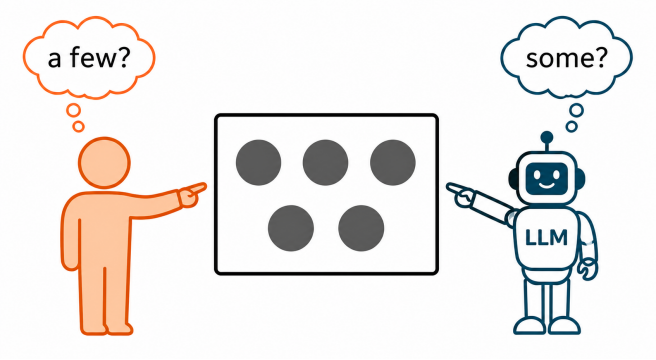

Figure 1: Illustration of the central task: humans and LLMs may map the same observed quantity to different linguistic quantifiers.

Quantifiers are particularly suited to this framework because several expressions can describe the same numerical value. A speaker might describe five objects as "a few," "some," or "several," depending on the reference set and the intended message. Previous work has used probabilistic pragmatic models to explain graded patterns in human quantifier production (van Tiel et al., 2021). These models offer a structured account of how numerical information, semantic meanings, and conversational expectations influence quantifier choice.

Large language models also produce quantifiers in natural language. However, fluent output alone does not establish that their quantifier-selection patterns match those of humans. We make a narrower comparison by presenting humans and LLMs with matched prompts, constructing parallel RSA models from their responses, and comparing the resulting probability distributions. This approach evaluates similarity in observable quantifier selection without making broader assumptions about human or machine cognition.

We examine two prompt conditions. Both state that five students scored above 90%. The first specifies that the class contains 30 students, while the second leaves the class size unspecified. We use human judgments and responses from GPT-3.5 and GPT-4 as priors in RSA models, then compare the resulting pragmatic-speaker distributions using Kullback-Leibler divergence. Within this study, the human and LLM distributions were closer when the class size was unspecified. This preliminary result shows how a shared probabilistic framework can be used to study the effect of contextual specificity on human and LLM quantifier selection.

Related Work

Rational Speech Act

The Rational Speech Act (RSA) framework models communication as recursive reasoning between speakers and listeners (Goodman and Stuhlmüller, 2013; Goodman and Frank, 2016). A **Literal Listener (L0)** interprets an utterance according to its literal meaning. A **Pragmatic Speaker (S1)** selects an utterance based on how effectively it communicates an intended state. A **Pragmatic Listener (L1)** then reasons about why the speaker selected that utterance. This probabilistic framework has been used to explain graded patterns in natural language quantification, where choices among expressions such as 'few', 'some', and 'many' depend on numerical information, context, and the available alternatives (van Tiel et al., 2021).

Quantifiers in Large Language Models

Previous work has studied how neural language models predict and represent quantifiers. Pezzelle et al. (2018) compared humans and neural models on predicting quantifiers from local and global linguistic context. Human performance improved when more context was available, while model performance decreased. This result suggests that humans and models may use contextual information differently when selecting quantifiers.

Kalouli et al. (2022) examined whether contextualized language models capture the semantic constraints associated with function words, including logical quantifiers such as 'all', 'every', and 'some'. Michaelov and Bergen (2022) studied predictions following 'few'-type quantifiers across language models of different sizes. They reported inverse scaling, in which larger models performed worse on the tested predictions. Together, these studies show that quantifier behavior provides a useful test of how language models represent meaning and use context.

Human vs LLM

Prior work has compared human and model performance on quantifier prediction (Pezzelle et al., 2018). However, the task and model setting differ from the present study. At the time of this study, we were not aware of prior work that compared distributions of human and GPT-3.5 or GPT-4 quantifier choices using parallel RSA models. We were also not aware of work testing whether specifying or omitting the reference-set size changes the similarity between those distributions.

This work addresses the following two questions:

1. How similar are the distributions over quantifier choices produced by humans and state-of-the-art Large Language Models?
2. Does specifying the reference-set size change the similarity between the human and LLM distributions?

Methods

We presented humans and Large Language Models (LLMs) with two matched forced-choice prompts. Both prompts stated that five students scored above 90%. The first specified that the class contained 30 students, while the second omitted the class size. We converted the resulting human and LLM response frequencies into utterance priors for parallel Rational Speech Act (RSA) models. We then compared the models' pragmatic-speaker distributions using Kullback–Leibler divergence.

Survey Questions

The prompts differed only in the reference-set information provided:

1. **Specified class size:** "5 students scored above 90% out of a class of 30 students."
2. **Unspecified class size:** "5 students scored above 90% out of a class of students."

Both scenarios were followed by the same question and response options:

> **Question:** Which quantifier would you use to describe the number of students who scored above 90%? **Options:** ['a couple', 'a few', 'some', 'several', 'most', 'a majority', 'almost all', 'all']

The unspecified condition left the total class size to the respondent's interpretation.

Participants and Model Responses

Ten human participants answered both prompts, producing 20 human quantifier judgments. GPT-3.5 and GPT-4 were accessed through the ChatGPT interface. Each model was prompted five times with each question, producing 20 model responses. Responses from GPT-3.5 and GPT-4 were pooled to construct one LLM response distribution for each prompt condition.

RSA Model

The RSA model represented possible quantities as the integers $n \in \{0, \ldots, 30\}$. A uniform state prior assigned equal probability to each quantity. For each respondent group and prompt condition, the observed frequencies of the eight quantifier choices determined the utterance prior. Unobserved options were assigned probability 0.01, and the remaining probabilities were normalized. This avoided undefined KL divergence when an option was absent from one group's responses.

The literal meaning function mapped each utterance to an inclusive numerical interval:

'a couple'	$[0, 14]$	'a few'	$[0, 10]$
'some'	$[4, 10]$	'several'	$[5, 13]$
'most'	$[22, 30]$	'a majority'	$[20, 30]$
'almost all'	$[25, 29]$	'all'	$[28, 30]$

Numerical perception used a Weber fraction of 0.1, with speaker optimality $\alpha = 1$. The pragmatic speaker

mapped quantities to quantifier distributions, and the pragmatic listener performed the inverse.

The fixed, overlapping intervals and all other numerical parameters were held constant across groups and conditions, so only the response-derived utterance priors differed. A shared 0–30 state space enabled direct comparison even when the prompt omitted the class size.

Distribution Comparison

We evaluated the pragmatic-speaker distributions at $n = 5$, matching the survey scenario, and at $n = 1$, a second low-count point. We calculated $D_{\mathrm{KL}}(P_{\mathrm{LLM}} \| P_{\mathrm{human}})$ at each state and averaged the two values for each condition. Lower divergence indicates greater similarity.

Results

Figures 2 and 4 show the pragmatic-speaker distributions over quantifiers for the state $n = 5$. Figures 3 and 5 show the corresponding pragmatic-listener distributions over states for the utterance 'a few'. In both prompt conditions, the speaker distributions placed nearly all probability on 'a couple', 'a few', 'some', and 'several', but the human and LLM models allocated that probability differently.

Table 1: Raw counts before smoothing and RSA modeling. 'Most', 'a majority', 'almost all', and 'all' were never selected.

Condition	Group	Couple	Few	Some	Several
Specified	Human	3	4	0	3
	LLM	0	5	2	3
Unspecified	Human	3	3	1	3
	LLM	0	4	3	3

We quantified the difference between the speaker distributions using Kullback–Leibler divergence:

$$D_{\mathrm{KL}}(P_{\mathrm{LLM}} \| P_{\mathrm{human}}) = \sum_x P_{\mathrm{LLM}}(x) \times \log \frac{P_{\mathrm{LLM}}(x)}{P_{\mathrm{human}}(x)}.$$

Table 2: Average KL divergence between the LLM and human RSA speaker distributions over the tested states 1 and 5.

Prompt condition	Average KL divergence
Specified class size	0.51570
Unspecified class size	**0.40038**

Average divergence was 0.51570 with the class size specified and 0.40038 with it unspecified. Thus, the modeled distributions were closer in the unspecified condition. This comparison is descriptive.

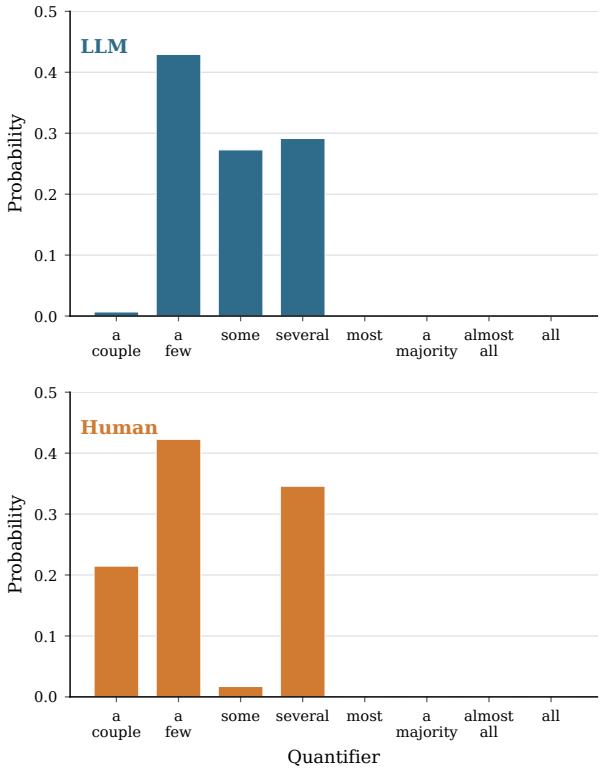

Figure 2: Specified class-size condition: pragmatic-speaker distributions over quantifiers for the state $n = 5$. Top: LLM; bottom: human.

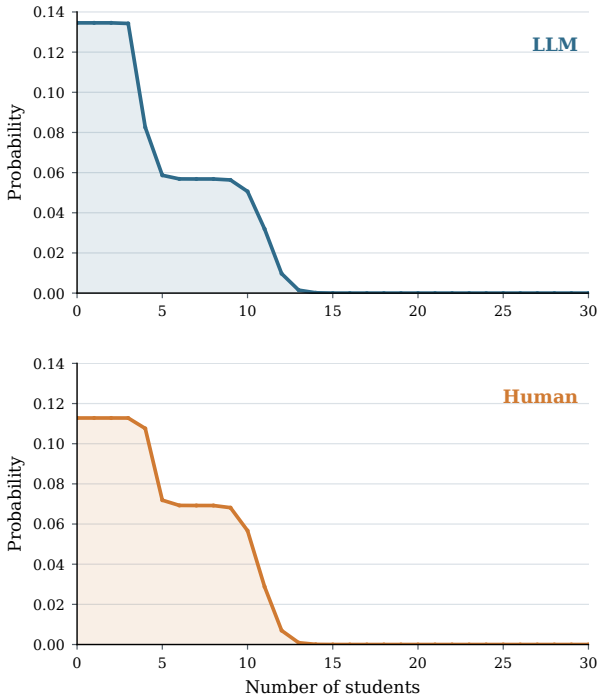

Figure 3: Specified class-size condition: pragmatic-listener distributions over states 0–30 for the utterance 'a few'. Top: LLM; bottom: human.

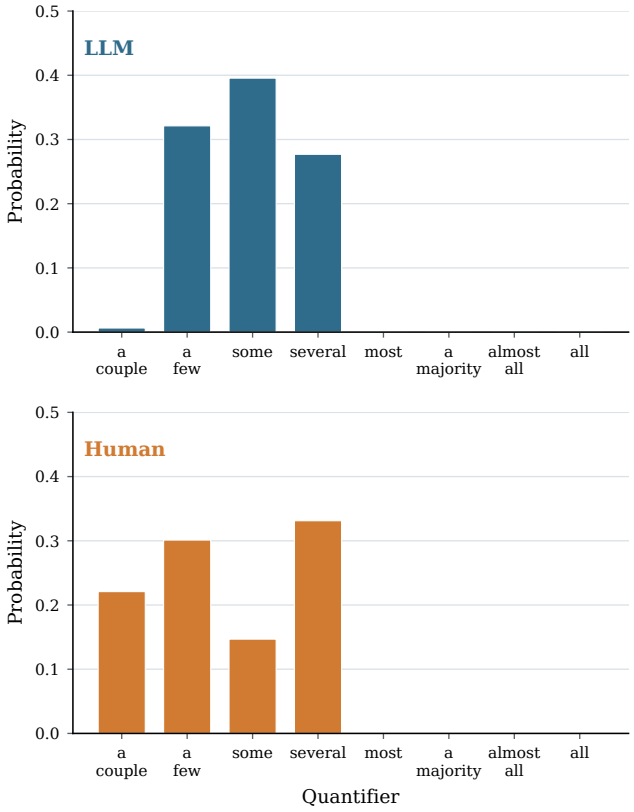

Figure 4: Unspecified class-size condition: pragmatic-speaker distributions over quantifiers for the state $n = 5$. Top: LLM; bottom: human.

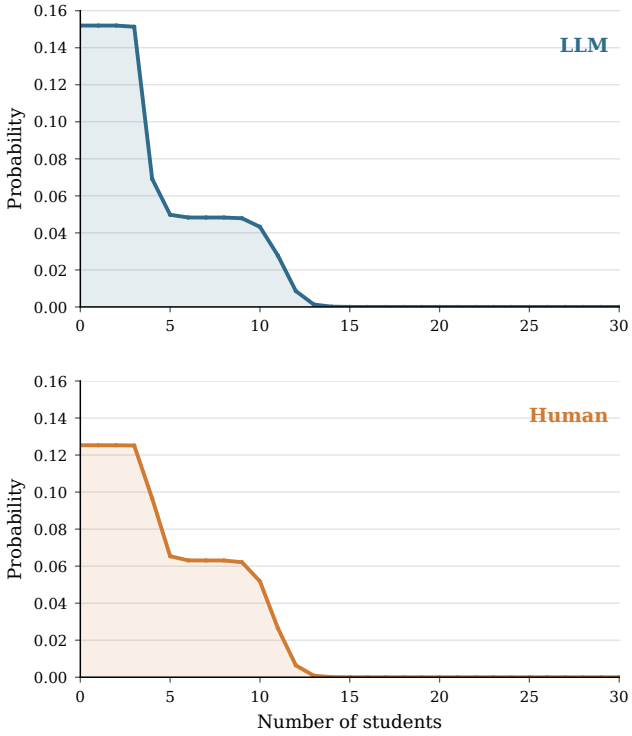

Figure 5: Unspecified class-size condition: pragmatic-listener distributions over states 0–30 for the utterance 'a few'. Top: LLM; bottom: human.

Discussion/Conclusion

Using parallel RSA models, we found that the average KL divergence between the human and LLM pragmatic-speaker distributions was 0.51570 when the class size was specified and 0.40038 when it was unspecified. Because lower divergence indicates greater similarity, the modeled distributions were closer when the class size was omitted. This result is narrower than the claim that LLMs generally understand quantifiers more like humans when less context is available. It shows that, for these prompts and modeling assumptions, the human and LLM response patterns were more similar in the unspecified condition.

One possible explanation is that when the class size is unspecified, both humans and LLMs rely more heavily on general expectations about what counts as 'a few', 'some', or 'several'. Specifying a class size of 30 provides a more precise reference point, which the two groups may use differently. However, the present study cannot determine the mechanism responsible for the observed difference.

The study is limited by its small sample, two prompt conditions, pooled GPT-3.5 and GPT-4 responses, fixed semantic intervals, smoothing choices, and evaluation at only two states. Future work should include more participants, analyze language models separately, test additional quantities and reference-set sizes, and examine whether the result is robust to different RSA assumptions. Despite these limitations, the study demonstrates a quantitative framework for comparing human and LLM quantifier choices across contextual conditions.

Author Contributions

Essam Sleiman conducted the project and analysis, with guidance from Ced Zhang and Joshua Tenenbaum.

References

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.

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

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*.

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.