Quantifying Generalizations: Exploring the Divide Between Human and LLMs' Sensitivity to Quantification

Claudia Collacciani, Giulia Rambelli and **Marianna Bolognesi**
University of Bologna, Italy
{claudia.collacciani2, giulia.rambelli4, m.bolognesi}@unibo.it

Abstract

Generics are expressions used to communicate abstractions about categories. While conveying general truths (e.g., *Birds fly*), generics have the interesting property to admit exceptions (e.g., penguins do not fly). Statements of this type help us organizing our knowledge of the world, and form the basis of how we express it (Hampton, 2012; Leslie, 2014).

This study investigates how Large Language Models (LLMs) interpret generics, drawing upon psycholinguistic experimental methodologies. Understanding how LLMs interpret generic statements serves not only as a measure of their ability to abstract but also arguably plays a role in their encoding of stereotypes. Given that generics interpretation necessitates a comparison with explicitly quantified sentences, we explored i.) whether LLMs can correctly associate a quantifier with the generic structure, and ii.) whether the presence of a generic sentence as context influences the outcomes of quantifiers. We evaluated LLMs using both Surprisal distributions and prompting techniques. The findings indicate that models do not exhibit a strong sensitivity to quantification. Nevertheless, they seem to encode a meaning linked with the generic structure, which leads them to adjust their answers accordingly when a generalization is provided as context.

1 Introduction

Generic generalizations, or simply *generics*, are statements such as *Birds fly* , *Dogs are mammals* or *Clocks are round*, that convey information about categories. They are a powerful way through which language allows us to communicate and learn abstract knowledge that extends beyond present context and direct experience. We use generic sentences to express our knowledge about the world, including stereotypes or prejudices (e.g., *Men are better at math than women*). Generics are fundamental to human cognition because they help us conceptualize the properties we associate with different categories, organizing our experience of the world (Chatzigoga, 2019).

The most notable feature of generics is that they allow for exceptions (Krifka et al., 1995). For example, *Birds fly* is considered true by speakers even if there are birds that cannot fly (e.g., penguins): in this case, therefore, the corresponding universal statement (*All birds fly*) is false. Different generalizations tolerate exceptions to varying degrees, according to their different *semantic content*: *Dogs are mammals* requires for its truth that all dogs be mammals; *Ducks lay eggs* is judged true even if only mature female ducks lay eggs, while *Mosquitoes carry malaria* refers to an even smaller minority (about 1 percent of mosquitoes carry malaria). There are generics that might be better paraphrased with *all*, others with *most*, and others with *some*; however, unlike quantified statements, they do not explicitly convey information about how many category members possess the predicated property.

Given this property, the meaning of generics can be considered "underspecified": humans' correct interpretation is derived through world knowledge and pragmatic abilities (Tessler and Goodman, 2019). The main questions that cognitive and psycholinguistic studies conducted on generics seek to answer are whether generics are a default mechanism, whether there exists a generic bias, and what is the relationship between genericity and prevalence, i.e., to what proportion of category members the property predicted by the generic applies (Cimpian et al., 2010; Leslie et al., 2011; Khemlani et al., 2012; Prasada et al., 2013, among others). These studies investigate the nature of generalizations by contrasting generic and overtly quantified sentences (Chatzigoga, 2019); in this sense, quantifiers are used to make explicit the underspecified meaning of generics.

The present paper investigates the interpretation of generalizations in different Large Language Models (LLMs). Since psycholinguistics experiments conducted on humans involve the compari-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11811–11822
August 11-16, 2024

son with quantified expressions, we also used quantifiers as a means of unraveling generics comprehension to directly compare humans' and models' interpretations.

We aim to explore the capability of LLMs to interpret generalizations that differ in semantic content as humans do. As we mentioned, this ability in people is closely related to world knowledge, which allows us to interpret 'underspecified' and implicitly quantified sentences by making use of our prior information. For this reason, investigating what is encoded in models with regard to generic form and its relation to quantification is crucial to comprehend whether they effectively understand the level of inclusiveness of a conceptual category, based on the context in which it is used, and whether they can leverage the power of quantifiers to replicate human-like distinctions, thereby enhancing their capacity to comprehend and interpret natural language accurately. The capability of LLMs to interpret generic sentences is not only an index of their capacity for abstraction but is also arguably involved in their encoding of stereotypes, since generics are one of the most powerful ways through which language convey them (Beukeboom and Burgers, 2019).

In what follows, we present two experiments that try to answer our research questions (RQs):

> **RQ1: Are LLMs capable of interpreting generic sentences according to their semantic content?** Generalizations can have different implicit quantificational values depending on their semantic content. Humans are able to derive their correct meaning thanks to their world knowledge. In our first experiments (cf. 4), we investigate if LLMs are able to do the same through two different methodologies.
>
> **RQ2: Do LLMs have a linguistic default interpretation associated with the generic form?** People seem to have a default interpretation associated with the *form* of generics (Cimpian et al., 2010). In our second experiment (cf. 5), we conduct an exploratory analysis of how models interpret generalizations aside from their content, i.e., whether they seem to have encoded linguistic knowledge associated with the generic form.

2 Related works

Most of the NLP literature dealing with genericity in language has focused on the building of resources geared towards distinguishing generic expressions that refer to whole categories from their non-generic counterparts that refer to specific exemplars (Friedrich et al., 2015; Govindarajan et al., 2019; Uryupina et al., 2020; Collacciani et al., 2024, among others). More recently, the usefulness of generic sentences as a resource to retrieve common sense knowledge, exploitable to boost performance in various NLP applications, has been proposed and demonstrated by (Bhakthavatsalam et al., 2020; Nguyen et al., 2023).

However, to the best of our knowledge, there are no studies investigating the interpretation of linguistic generalizations by LLMs, except for the recent works by Ralethe and Buys (2022), which addresses the generic overgeneralization effect, and Collacciani and Rambelli (2023), which investigates generics interpretation, building on psycholinguistic experimental designs. Both works, however, only focus on Masked Language Models such as BERT and RoBERTa.

We will use quantifiers to investigate generics comprehension, placing them at the beginning of bare generic sentences to explicitly specify their quantificational value. The most relevant studies that have evaluated model predictions following quantifiers are Kalouli et al. (2022), which focus on logical quantifiers such as *all, every*, and *some*, and Michaelov and Bergen (2023) and Gupta (2023), which focus on *few* and *most*-type quantifier; the other few works involving quantifiers focus on predicting the quantifier itself (Pezzelle et al., 2018; Talmor et al., 2020). The present work will contrast LLMs' predictions on generic sentences and sentences quantified by *no, few, some, most*, and *all*, investigating which quantifiers seem to best approximate the meaning of the generic form. Therefore, our work aims not only to understand LLMs' generics interpretation but also to contribute to the exploration of LLMs' knowledge of quantifiers, adding the systematic comparison with generic sentences as a novel element.

3 Materials and Methods

Dataset For this study, we created a dataset in which each generic sentence is paired with the correct quantifier, i.e., the quantifier that humans would prefer to make explicit the implicit quantification value of the generic sentence[1]. From now

[1]The dataset is available at: https://osf.io/ahspu/?view_only=1f789e020b7346338c53b684943dc9f1

on, we will refer to this quantifier as *Human Quantifier*, while we will use the label *LLM Quantifier* to indicate each quantifier when paired with the sentences for the LLMs evaluation.

To assemble our dataset, we employed sentences from different existing resources. In the first place, we looked at the Herbelot and Vecchi (2016) dataset, consisting of concept-feature pairs from McRae et al. (2005), such as *airplane has-engines* or *ant is-black*, labeled by native speakers through quantifiers. For each pair in the norms, annotators were asked to provide a label expressing how many members of the category possess the property in question, choosing among the natural language quantifiers *no, few, some, most, all*. We selected only those pairs on which all three annotators agreed on the same quantifier. From this dataset, we sampled 500 sentences annotated with *some* and *all*, plus 97 sentences annotated with *most*. Sentences annotated by humans with *some, most*, and *all* are those that can be considered as true generalizations, in their generic form. However, in order to better understand whether they are correctly interpreted by LLMs, we decided to add to the dataset also sentences quantifiable with *few* and *no*, that are characterized by implausible or impossible category-property pairs. In this case, the effect of the quantifier is to reverse the truth value of the sentence (from implausible to plausible and from impossible to possible): because of this feature, these sentences will be useful as a touchstone to evaluate the others.

First, we included a sample of 500 concept-property pairs extracted from the COMPS dataset (Misra et al., 2023). We selected 500 cases in which `negative-sample-type` is equal to `random` (i.e., for which the similarity between the acceptable and unacceptable concept for a certain property is equal to 0), and used the unacceptable concept to form our sentences. In these cases, the Human Quantifier would always be *no* because the predicated property is unacceptable, as in *Unicycles clean dishes*. Additionally, we selected 240 stimuli originally constructed by Urbach and Kutas (2010) for a psycholinguistic task and recently used by Michaelov and Bergen (2023) and Gupta (2023) for LLMs evaluation. These stimuli consist of 120 typical subject-verb pairs (called "backbone phrases") completed by both a typical and an atypical object, as in *postmen carry mail* vs. *postmen carry oil*. In the original psycholinguistic experiment of Urbach and Kutas (2010), these sentences were alternatively modified by *most*-type and *few*-type quantifiers in order to collect offline plausibility ratings and record brain activity (using EEG) for the different conditions. Following the plausibility ratings of the original experiment, we included these stimuli by annotating sentences with a typical object with *most* quantifier, while sentences with an atypical object are annotated with *few*.

Human Quantifier	**Generic sentences**	**Examples**
NO	500	Unicycles clean dishes.
FEW	120	Smugglers transport umbrellas.
SOME	500	Oranges are used for juice.
MOST	217	Clocks are round.
ALL	500	Whales are mammals.
Total	1837	

Table 1: Structure of our dataset.

Our final dataset consists of 1,837 sentences. Even if the dataset is not completely balanced, we believe that the stimuli should be sufficient to observe tendencies in intra- and inter-conditions. Table 1 shows the structure of our dataset, along with some examples.

Models We conducted our experiments on BERT-large-uncased (Devlin et al., 2019), a bi-directional masked language model, GPT2-xl (Radford et al., 2019), and 2 open-source pre-trained generative LLMs and their instruction-tuned variants: Llama-2 (Touvron et al., 2023) and Mistral (Jiang et al., 2023) with 7 billion parameters.[2]

4 Are LLMs Capable to Interpret Generic Sentences According to their Semantic Content?

4.1 Surprisal

In our first experiment, we measure the Surprisal of each of the sentences in our dataset modified by each of the five quantifiers considered (*no, few, some, most, all*). We use the Surprisal of the overall sentence (S_s), defined as the sum of the Surprisals of each token (S_t) normalized by the length of the sentence:

$$S_s = \frac{\sum_{t \in S}^{T} S_t}{count(t)}$$

where S_t is the negative log-probability of the occurrence of a token given its context. The Surprisal scores were extracted using the Minicons library, v.

[2]We only focus on open LLMs for reproducibility reasons and because we are interested in comparing the base and the instruct version of the very same models.

0.2.33 (Misra, 2022). The underlying idea is that if LLMs correctly take the meaning of the different quantifiers into account in their decision process, for each sentence, the Surprisal would be lower in the condition modified by the corresponding Human Quantifier than in the others. Let us consider the examples in Table 1: a LLM is considered accurate if

1. (a) S_s (***No*** *unicycles clean dishes*) < S_s (***Few/Some/Most/All*** *unicycles clean dishes.*)
   (b) S_s (***Few*** *smugglers transport umbrellas.*) < S_s (***No/Some/Most/All*** *smugglers transport umbrellas.*)

and so on. In addition to the five quantified conditions, we also extracted the Surprisal of the bare generic sentence, without quantifier. In the case of the generic condition, we expect it to have i) higher Surprisal than the sentences preceded by *no* and *few* for the sentences for which the Human Quantifier is *no* and *few*, whilst ii) having an approximately equivalent Surprisal score to the sentence preceded by *all, some*, and *most*. Sentences annotated with *no* and *few* quantifiers are semantically impossible or implausible sentences and, therefore, should be 'surprising' unless they are preceded by the respective Human Quantifier, which reverses the truth value of the sentences:

2. (a) S_s (*Unicycles clean dishes*) > S_s (***No*** *unicycles clean dishes.*)
   (b) S_s (*Smugglers transport umbrellas*) > S_s (***Few*** *smugglers transport umbrellas*)

In contrast, sentences annotated with *some*, *most*, and *all* refer to semantically plausible events. The generic versions of these sentences are implicitly quantified, that is, semantically equivalent to the respective quantified sentence. Consequently, there should be no difference in the Surprisal scores between the bare generic and the quantified versions:

3. (a) S_s (*Oranges are used for juice*) $\simeq$ S_s (***Some*** *oranges are used for juice*)
   (b) S_s (*Clocks are round*) $\simeq$ S_s (***Most*** *clocks are round*)
   (c) S_s (*Whales are mammals*) $\simeq$ S_s (***All*** *whales are mammals*)

Results Following the above assumptions, we computed the accuracy of each model separately for each Human Quantifier class, reported in Figure 1. On the left (Accuracy QUANT), we report

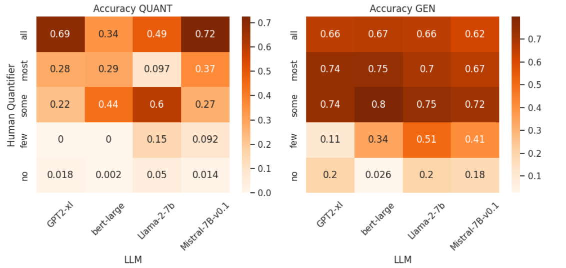

Figure 1: Heatmaps of Accuracy values per Human Quantifier on Surprisals, for each LLM.

accuracy values computed following (1): a model is correct if the sentence with the lowest Surprisal is the one with the same quantifier of the specific Human Quantifier class. As the plot reveals, the highest accuracy is obtained for the Human Quantifier *all* (especially for GPT2 and Mistral models), followed by the Human Quantifier *some*.

On the right (Accuracy GEN), we compare the Surprisal of a generic sentence (without quantifier) with its version modified by the specific Human Quantifier: an LLM is considered accurate if it fulfills the conditions in (2) and (3). For *some, most*, and *all* classes, we considered as approximately equal a Surprisal of $\pm$ 1 std[3]. Similarly, we observe that accuracy scores are higher for *some* and *all* classes, but, in this case, we obtain high accuracy even for the class *most*. On the contrary, the scores are low for *no* and *few* classes.

To inspect the behavior of LLMs in more detail, we examined the distributions of the Surprisal values inside each Human Quantifier class. Figure 2 reports the distributions for GPT2-xl and Mistral, as the other LLMs analyzed (BERT-large and Llama) show the same trends (all boxplots are in Appendix A). For each Human quantifier (x-axis), a boxplot represents the Surprisals of a sentence with a specific quantifier (e.g., "No unicycles clean dishes" vs. "All unicycles clean dishes").We can observe two main trends: by looking at the average mean of each Human Quantifier group, we notice that the *no* and *few* classes tend to have higher Surprisals *in general*, regardless of the LLM Quantifier condition. We can hypothesize that this happens because these sentences contain words that do not usually co-occur with each other precisely because they are meant to identify

[3]For each LLM, we used the standard deviation of its Surprisals on the entire dataset.

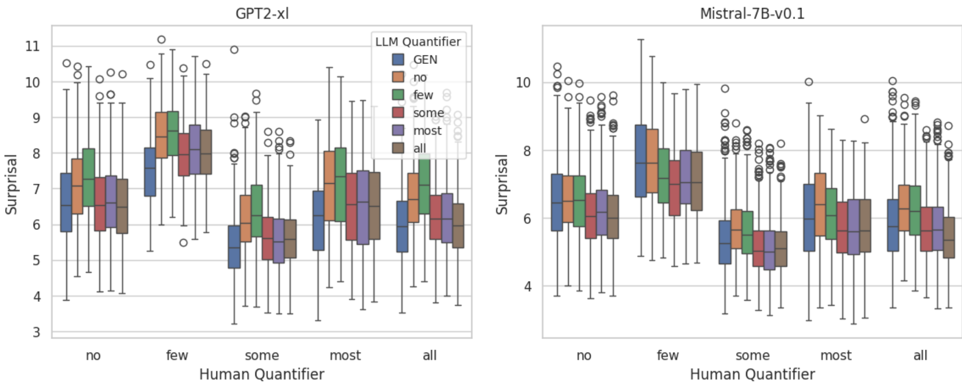

Figure 2: Sentence surprisal distributions per Human Quantifier and LLM Quantifier, for GPT2-xl and Mistral.

properties that are impossible or implausible for the categories in question (Kauf et al., 2023). However, the *overall meaning* of these sentences should become less surprising when introduced by the appropriate Human Quantifier (*no* or *few*), since it has the effect of reversing the truth value of the sentence, as illustrated in (2).

Nevertheless, the presence of the quantifier does not model the Surprisal scores as theoretically expected. Looking inside each LLM Quantifier group, we notice that the Surprisal distribution is the same across the five groups, and we do not see the reversal of the ratios among the distributions that should occur if the quantifier meaning was properly taken into account. In other words, if the quantifiers' meaning were correctly taken into consideration, the sentences with the lowest score should be the ones with the same quantifier of the target class (cf. (1)). Conversely, the average surprisal of generic sentences (GEN) should be similar to the Surprisal of sentences quantified with *some, most* and *all* in their respective classes (cf. (3)), while they should be higher than *no* and *few* in their respective classes (cf. (2)).

However, sentences quantified with *some, most* and *all* tend to have lower Surprisals in all five conditions across all LLMs (with the partial exception of Llama, in which the subgroups are roughly all at the same level). This inspection can help us interpret the accuracy values: the fact that the models perform better on the *some, most* and *all* classes seems to be due to a general preference for these quantifiers over the others in all cases, rather than a real grasp of the meaning of the quantifiers, also with respect to the generic sentences.

What we just observed leads us to point out that the recent results reported by Gupta (2023) on quantifiers comprehension in LLMs may be misleading. In their experimental paradigm, the accuracy of Surprisal is calculated on sets of minimal pairs, such as *S (Most postmen carry mail) < S (Few postmen carry mail)* and *S (Most postmen carry oil) > S (Few postmen carry oil)*. In this task, the two complementary conditions are satisfied by the two opposite outcomes. The accuracy values they report are consistently around 0.5, which means the model satisfies the conditions in about half of the cases. In light of our results, this outcome seems to be due to a general agnostic preference of LLMs for a quantifier on the other: *most* has a tendency to always have a lower Surprisal than *few*, regardless of what would be the correct Human Quantifier, as well as the other quantifiers to maintain their position in the reciprocal distribution.

Our results align with those of Michaelov and Bergen (2023), as well as with previous studies on the sensitivity of LLMs probability values to negation and logical quantifiers (Ettinger, 2020; Kassner and Schütze, 2020; Kalouli et al., 2022). In the next section, we will discuss a possible explanation for these outcomes and propose an alternative method for investigating the LLMs' interpretation of generic sentences through quantifiers.

4.2 Prompting

From the analysis of Surprisals, it emerged that LLMs are unable to correctly interpret generic sentences through quantifiers with respect to their semantic content (i.e., their Human Quantifier). However, we want to point out that the Surprisals, as well as the probability values produced by LLMs, are an index of the LLM's 'online' decision-making process: in this sense, they are somewhat comparable to human brain activity in response to linguistic stimuli, and they have indeed been used in works comparing them to brain responses such as the

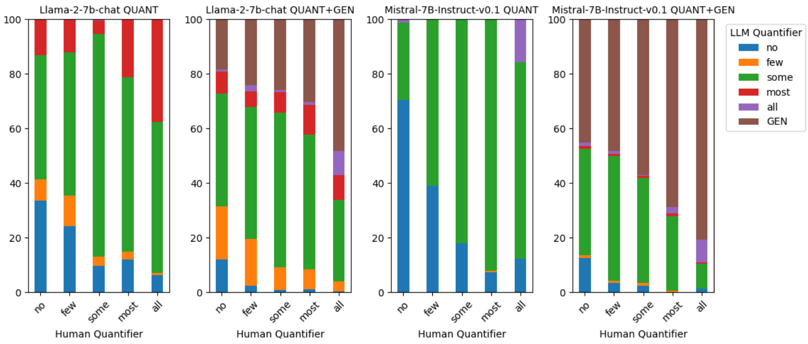

Figure 3: Percentages of occurrence for each options (LLM Quantifier) per Human Quantifier class in the LLMs responses when prompted. For each LLM, we show the responses to both QUANT and QUANT+GEN prompting.

N400 amplitude (Ettinger, 2020; Michaelov and Bergen, 2023; Gupta, 2023).

It is interesting to note that there is experimental evidence in psycholinguistic works that shows that the manipulation of both *no-some-all* (Fischler et al., 1984; Kounios and Holcomb, 1992), and *few-most* (Urbach and Kutas, 2010), which leads to a reversal of the truth values of the modified sentences[4], while well taken into account in offline plausibility judgments, does not similarly reverse the N400 amplitudes in incremental sentence comprehension. This is considered by Urbach and Kutas (2010) as a dissociation between the patterns of quantifier and typicality effects for the offline and online measures. Given these considerations, the LLMs' online processing (measured through Surprisal), which reveals low sensitivity to quantifiers but good sensitivity to typicality (as shown in the previous paragraph and similar to Michaelov and Bergen (2023)) is not that dissimilar to that of humans. In other words, both humans and LLMs, in online processing, are more sensitive to the plausibility of the predicated property on a given category (i.e., the fact that *Postmen carry **mail*** is more plausible than *Postmen carry **oil***), rather than to the presence of the correct quantifier.

Therefore, we decided to test our dataset through another methodology that is possibly more comparable to offline plausibility judgments (as are the Human Quantifier classes annotation on our dataset): metalinguistic prompting. For this task, we tested the instruction-tuned variants of Llama-2 and Mistral, using the same hyperparameters[5]. We used two different versions of prompting strategies, both in zero-shot settings, since we are interested in eliciting the knowledge already encoded in each model (examples of each prompt are reported in Appendix B). In the first condition, models were asked to choose the *most truthful* sentence from the list of its quantified versions for each of the sentences in our dataset. In the second condition, the only difference is that the generic form of the sentence is also presented among the options; we will call the first version QUANT and the second QUANT+GEN.

Results Figure 3 reports the percentages in which each different option (LLM Quantifier condition) occurs in the LLMs responses per Human Quantifier class, in both QUANT and QUANT+GEN prompting versions. Both language models show the same trends. When they are given only the quantified sentences as options (QUANT prompting), they take a 'conservative' stance, overextending the existential quantifier *some* over all classes. Interestingly, we can observe a trend for which from the Human Quantifier class *no* to *all* there is a progressive extension of the quantifiers *most* and *all*, and a simultaneous reduction of *no* and *few* in the responses. When the generic form is also provided among the options (QUANT+GEN prompting), this is often preferred, especially by Mistral. Even in this case, we can observe a progression in the extension of the generic form from left to right. This progressive trend seems to suggest that the instruction-tuned models are able to partially discriminate between different classes on the basis of their semantic content and have encoded some kind of meaning associated with quantifiers and the generic form, although not particularly refined.

However, the accuracy (Table 2) remains overall not satisfying. In this case, the accuracy val-

[4]E.g., from the cited studies: *[All/some/no] gems are rubies. - [All/some/no] rubies are gems.*; *[Most/Few] farmers grow [crops/worms.]*

[5]`Temperature=0, do_sample=False, top-k=10, max-tokens=50, frequency and presence penalty=0.`

ues were computed considering the model accurate if its choice matched with the Human Quantifier class:

$$Accuracy = \frac{N_{LLMQUANT=HUMANQUANT}}{N_{TOT}}$$

Furthermore, in the QUANT+GEN version, the choice of the generic form was considered accurate if the Human Quantifier class was *some, most* or *all*, given that these are the cases for which the generic expression is acceptable (cf. (3)). As in the previous experiment, the accuracy is not good on all classes consistently, but only on which there is a strong general preference, i.e., when there is overextension (e.g., *some*).

	Llama		Mistral	
Human Quantifier	QUANT	QUANT+GEN	QUANT	QUANT+GEN
no	.278	.082	.690	.124
few	.092	.125	.000	.008
some	.746	.498	.818	.384
most	.198	.097	.000	.009
all	.000	.076	.158	.080

Table 2: Prompting Accuracy per Human Quantifier class on QUANT and QUANT+GEN versions.

5 Do LLMs Have a Linguistic Default Interpretation of the Generic Form?

Our second study is more exploratory and aims to investigate the relationship between generalizations and quantification from a more formal point of view. We draw inspiration from the work of Cimpian et al. (2010), who found that people, when presented with a generic sentence about a novel category (*Morseths have silver fur.*) and asked to estimate how many members of the category possess the characteristic predicated by the generic, tend to assign very high percentages (on average,

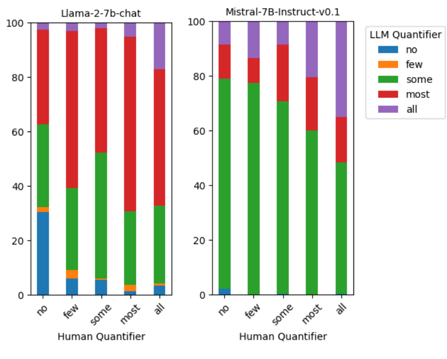

Figure 4: Percentages of occurrence for each option (LLM Quantifier) per Human Quantifier in the LLMs responses when prompted on the entailment condition.

very close to 100 percent). From that, we can infer that people have a default interpretation of the generic form: if informed about a made-up category, that lacks associations to properties in their minds, through a generic form, humans tend to extend by default the predicated property on all members of the category. Since models do not seem to encode the world knowledge necessary to interpret generics on account of their semantic content as humans, we decided to test them on a similar paradigm. Our aim is to comprehend whether and how LLMs encode a default interpretation associated with a generic form.

As in Section 4.2, we used prompting to test the instruction-tuned variants of Llama-2 and Mistral, using the same parameter configurations presented in the previous experiment. Our prompting strategy for this task is inspired by the experimental design of (Cimpian et al., 2010); an example is shown in B. This prompting strategy is analogous to the QUANT condition used in the previous section, since the options are exactly the same; the only difference is that, in this case, the generic sentence is given as a premise in an entailment condition (e.g., *Birds fly, therefore...*), with the aim of exploring whether this leads the models to reshape their response accordingly.

Results Figure 4 reports the percentages in which each different option (LLM Quantifier condition) occurs in the LLMs responses per Human Quantifier class in the entailment condition. If we compare them with the results in Figure 3, for the QUANT condition - in which the options were exactly the same, we can observe that in Llama there is a strong reduction of *no* and *few* and a large overextension of the quantifier *most*, as well as the emergence of *all* on the last classes; in Mistral, *no* and *few* have practically disappeared and, although there is still an overextension of *some*, there is a strong increase of *most* and *all*. Overall, the generic sentence provided as a premise seems to lead both models to skew toward "strong" positive quantifiers (*most* and *all*), to the expense of negative ones.

6 General Discussion

This paper offers both quantitative and qualitative insights into how LLMs interpret generics, employing experimental designs that utilized quantified expressions to probe the comprehension of generic statements. Our two experiments were conceived to evaluate two related but separate abilities: first, the models' capacity to accurately recognize the common knowledge implied in generic statements (i.e., they can generalize a property to the right level of inclusiveness of categories); secondly, their ability to comprehend generalizations irrespective of their content, specifically, whether they incorporate any linguistic cues linked to the generic form. To the best of our knowledge, we are the first to perform this investigation with recent LLMs, including their instruction-tuned variant, and testing them with prompting methodologies.

The experiment illustrated in 4 was designed to investigate whether LLMs are capable of interpreting generic sentences according to their semantic content through quantifiers (RQ1). We observed that Surprisals do not seem to be particularly sensitive to the effect of quantifiers on sentence meaning, thus preventing us from using them as an explicit marker of the interpretation of generic sentences that differ in semantic content. However, it is possible that this outcome is not due to a complete insensitivity of the models to the meaning of quantifiers as much as to the method employed. In fact, the measurement of Surprisals could be more akin to measurements of human online processing (such as recording of brain activity) rather than offline judgments (such as the annotations we have on our dataset). Interestingly, the Surprisal of the models with respect to the effect of quantification does indeed seem to follow a similar pattern to that emerging from comparable studies on human N400 potentials (Fischler et al., 1984; Kounios and Holcomb, 1992; Urbach and Kutas, 2010). Therefore, we investigated the behavior of the models through prompting, which mirrors offline human judgments. The analyzed outcomes suggest that the instruction-tuned models have encoded some kind of meaning associated with quantifiers and the generic form, although not particularly refined. LLMs judge the choice of *most* and *all*, as well as of the generic form over the others, as more suitable as the semantic content of the sentence goes from impossible/implausible category-property pairs (*no/few* classes) to plausible category-property pairs (*some/most* classes), to necessary category-property pairs (*all* class).

However, the comprehension of the meaning of generic and quantified sentences with respect to their semantic content does not seem to be particularly accurate. LLMs tend to take a very 'conservative' stance, preferring the intermediate quantifier *some* when given only quantified sentences as options, and the generic sentence itself (inherently vague) when this is added among the options[6]. This could be due to the fact that explicit quantification is actually a relatively rare phenomenon in naturally occurring text, on which LLMs are exclusively trained, while underspecified constructions like generic sentences are much more frequent (Herbelot and Copestake, 2011; Herbelot and Vecchi, 2016). Moreover, the different quantifiers all appear in the same syntactic positions and in superficially very similar contexts; the choice of one or the other is inextricably linked to our extralinguistic knowledge of the categories and the properties predicated on them, something LLMs do not possess.

For this reason, we conducted a last study to explore the models' interpretation of generalization and quantification aside from the semantic content of the predications, i.e., whether they seem to have encoded linguistic knowledge associated with the generic form (RQ2). We found that, when a generic sentence is provided as a premise in an entailment condition, instruction-tuned models tend to reshape the distributions of the different quantifiers in their responses (cf. Figure 3 vs. Figure 4), skewing their preferences toward "strong" positive quantifiers (*most* and *all*). People behave similarly (interpreting a property predicated by a generic as applicable to virtually all members of the category) when tested on novel categories, for which they have no prior understanding. However, in a real language setting humans undeniably modulate their interpretations of generalizations through their knowledge of real categories.

7 Conclusions

In conclusion, this study observed that i) LLMs seem not to have the world knowledge necessary for the comprehension of the meaning of generic and quantified sentences with respect to their semantic content in a human-like way; ii) LLMs overextend the truth of a generic sentence when this is presented as an assumption, on *most/all* members of real-world categories, regardless of the meaning of the predication. This behavior could play a role in their encoding of stereotypes, which could be a potentially harmful bias. Overall, we believe that

[6]In this regard, it should be kept in mind that for our experiments we used temperature=0, which makes models' responses more focused and deterministic.

further investigations are needed to clarify the interpretation of generics in Language Models and, more generally, the role that this phenomenon has in their behavior.

Acknowledgements

Funded by the European Research Council, Abstraction project (attributed to Marianna Bolognesi. Grant agreement: ERC-2021-STG-101039777). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Limitations

Prompting strategies In this study, we assessed models under a conservative condition by employing a low temperature. Future research could explore the responses of the same models under higher temperatures, investigating how enhancing the linguistic creativity of LLMs impacts their performance in the presented tasks.

Another limitation pertains to the prompts utilized. We evaluated all LLMs using the query "Tell me which of the following is the most truthful sentence" on the first prompting task, and "What is the correct completion?" for the second one, in each case followed by a list of the options. While we experimented with different prompts before choosing this format, we did not quantitatively investigate whether alternative queries could enhance the accuracy of the models, nor did we explore whether different examples within the prompt could yield different results.

Study on English The current dataset and research are exclusively centered on English. Extending the dataset to include other languages would be advantageous. However, we currently face a scarcity of resources for other languages annotated with comparable linguistic information.

Ethics Statement

The resources used to build our dataset (Herbelot and Vecchi, 2016; Misra et al., 2023; Urbach and Kutas, 2010) are publicly available. We will release the dataset used in the present experiments and the obtained results in the OSF project of this study[7]. For reasons of replicability, we chose to use only LLMs freely available through huggingface. Given a limited GPU, we relied on 7 billion parameter models and used quantization techniques to reduce memory and computational costs, using `bitsandbytes` library. However, the experiments presented require a considerable memory and computational cost, especially for the prompting tasks. In addition, there is still a significant ethical concern regarding Language Models (LLMs). These models have been demonstrated to produce inaccurate information, potentially generating offensive material when prompted with certain inputs. However, it appears that LLMs fine-tuned with specific instructions have undergone training to mitigate the harmful nature of their responses. Nevertheless, some responses may still contain objectionable content. Any showcases of LLMs' linguistic capabilities should not suggest their safety or alignment with human preferences and values.

References

Camiel J Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication (scsc) framework. *Review of Communication Research*, 7:1–37.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *arXiv preprint arXiv:2005.00660*.

Dimitra Lazaridou Chatzigoga. 2019. Genericity. In *The Oxford Handbook of Experimental Semantics and Pragmatics*, pages 156–177. Oxford University Press.

Andrei Cimpian, Amanda C Brandone, and Susan A Gelman. 2010. Generic statements require little evidence for acceptance but have powerful implications. *Cognitive science*, 34(8):1452–1482.

Claudia Collacciani and Giulia Rambelli. 2023. Interpretation of generalization in masked language models: An investigation straddling quantifiers and generics. *Proceedings of the 9th Italian Conference on Computational Linguistics - CLiC-it 2023*.

Claudia Collacciani, Andrea Amelio Ravelli, and Marianna Bolognesi. 2024. Specifying genericity through inclusiveness and abstractness continuous scales. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15126–15136.

[7] https://osf.io/ahspu/?view_only=1f789e020b7346338c53b684943dc9f1

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Ira Fischler, Paul A Bloom, Donald G Childers, A Antonio Arroyo, and Nathan W Perry Jr. 1984. Brain potentials during sentence verification: Late negativity and long-term memory strength. *Neuropsychologia*, 22(5):559–568.

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. 2015. Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 21–30.

Venkata Govindarajan, Benjamin Van Durme, and Aaron Steven White. 2019. Decomposing generalization: Models of generic, habitual, and episodic statements. *Transactions of the Association for Computational Linguistics*, 7:501–517.

Akshat Gupta. 2023. Probing quantifier comprehension in large language models: Another example of inverse scaling. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 56–64.

James A Hampton. 2012. Generics as reflecting conceptual knowledge. *Recherches linguistiques de Vincennes*, (41):9–24.

Aurelie Herbelot and Ann Copestake. 2011. Formalising and specifying underquantification. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Aurélie Herbelot and Eva Maria Vecchi. 2016. Many speakers, many worlds. *LiLT (Linguistic Issues in Language Technology)*, 13.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Aikaterini-Lida Kalouli, Rita Sevastjanova, Christin Beck, and Maribel Romero. 2022. Negation, coordination, and quantifiers in contextualized language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3074–3085, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event Knowledge in Large Language Models: The Gap Between the Impossible and the Unlikely. *Cognitive Science*, 47(11):e13386.

Sangeet Khemlani, Sarah-Jane Leslie, and Sam Glucksberg. 2012. Inferences about members of kinds: The generics hypothesis. *Language and Cognitive Processes*, 27(6):887–900.

John Kounios and Phillip J Holcomb. 1992. Structure and process in semantic memory: evidence from event-related brain potentials and reaction times. *Journal of experimental psychology: General*, 121(4):459.

Manfred Krifka, Francis Jeffry Pelletier, Gregory Carlson, Alice ter Meulen, Gennaro Chierchia, and Godehard Link. 1995. Genericity: An introduction. In Greg N. Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, pages 1–124. University of Chicago Press.

Sarah-Jane Leslie. 2014. Carving up the social world with generics. *Oxford studies in experimental philosophy*, 1.

Sarah-Jane Leslie, Sangeet Khemlani, and Sam Glucksberg. 2011. Do all ducks lay eggs? the generic overgeneralization effect. *Journal of Memory and Language*, 65(1):15–31.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

James Michaelov and Benjamin Bergen. 2023. Rarely a problem? language models exhibit inverse scaling in their predictions following few-type quantifiers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14162–14174, Toronto, Canada. Association for Computational Linguistics.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, pages 1907–1917.

Sandro Pezzelle, Shane Steinert-Threlkeld, Raffaella Bernardi, and Jakub Szymanik. 2018. Some of them can be guessed! exploring the effect of linguistic context in predicting quantifiers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 114–119, Melbourne, Australia. Association for Computational Linguistics.

Sandeep Prasada, Sangeet Khemlani, Sarah-Jane Leslie, and Sam Glucksberg. 2013. Conceptual distinctions amongst generics. *Cognition*, 126(3):405–422.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Sello Ralethe and Jan Buys. 2022. Generic overgeneralization in pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3187–3196, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Michael Henry Tessler and Noah D Goodman. 2019. The language of generalization. *Psychological review*, 126(3):395.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Thomas P Urbach and Marta Kutas. 2010. Quantifiers more or less quantify on-line: Erp evidence for partial incremental interpretation. *Journal of Memory and Language*, 63(2):158–179.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26(1):95–128.

A Analysis of LLMs Surprisals

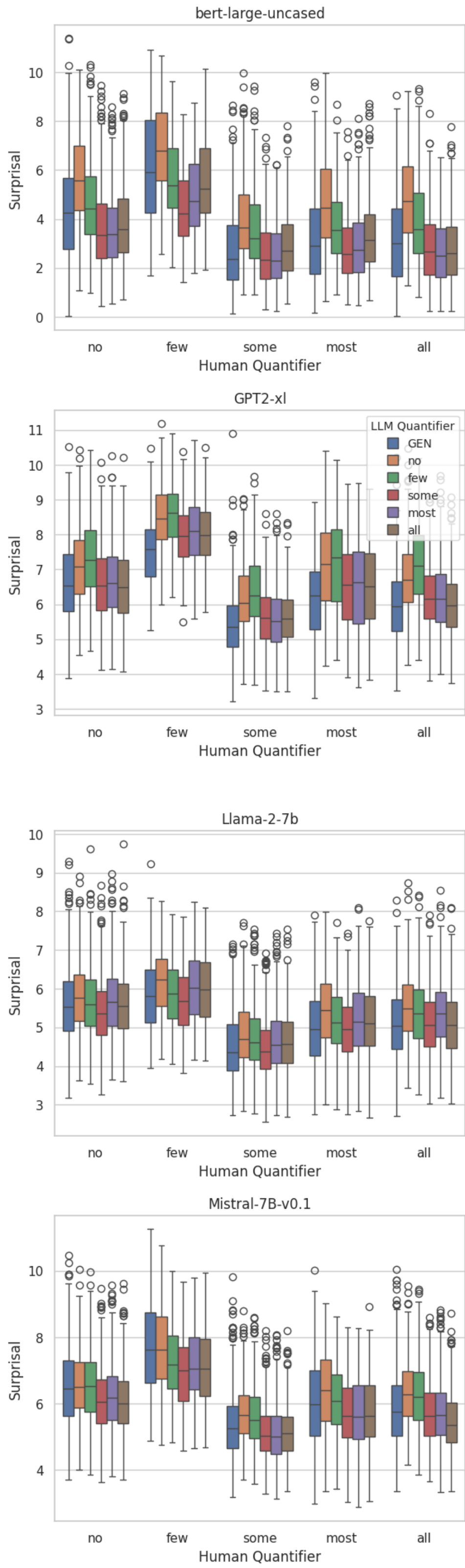

Figure 5: Sentence surprisal distributions per Human Quantifier and LLM Quantifier, for each LLM analyzed.

B Prompting strategies

We report an example for each of the three prompting strategies used. For each of them, the options were randomized for each iteration.

- **Section 4.2**

QUANT version

```
Tell me which of the following is the
most truthful sentence:
    No birds fly.
    Few birds fly.
    Some birds fly.
    Most birds fly.
    All birds fly.
```

QUANT+GEN version

```
Tell me which of the following is the
most truthful sentence:
    Birds fly.
    No birds fly.
    Few birds fly.
    Some birds fly.
    Most birds fly.
    All birds fly.
```

- **Section 5**

```
What is the correct completion? Birds
fly, therefore...
    no birds fly.
    few birds fly.
    some birds fly.
    most birds fly.
    all birds fly.
```