Generics are puzzling. Can language models find the missing piece?

Gustavo Cilleruelo Calderón **Emily Allaway** **Barry Haddow** **Alexandra Birch**
School of Informatics, University of Edinburgh
g.cilleruelo-calderon@sms.ed.ac.uk {emily.allaway, bhaddow, a.birch}@ed.ac.uk

Abstract

Generic sentences express generalisations about the world without explicit quantification. Although generics are central to everyday communication, building a precise semantic framework has proven difficult, in part because speakers use generics to generalise properties with widely different statistical prevalence. In this work, we study the implicit quantification and context-sensitivity of generics by leveraging language models as models of language. We create CONGEN, a dataset of 2873 naturally occurring generic and quantified sentences in context, and define *p-acceptability*, a metric based on surprisal that is sensitive to quantification. Our experiments show generics are more context-sensitive than determiner quantifiers and about 20% of naturally occurring generics we analyze express weak generalisations. We also explore how human biases in stereotypes can be observed in language models[1].

1 Introduction

Humans use generalisations to abstract away from particular objects, events or facts and convey regularities about the world.

In this work, our focus is on generic sentences, such as *insects have six legs* or *mosquitoes carry malaria*, which express generalisations without explicit quantification. These two generic sentences are acceptable in many contexts, but the quantifications they convey are widely different: almost all insects have six legs, but fewer than 1% of mosquitoes carry malaria.

One way of expressing generalisations in language is through explicitly quantified sentences, such as *most insects are nocturnal* or *some mosquitoes have white stripes*. Quantified sentences express statistical claims about the members of a kind that share the predicated property: for example *most* if a majority of insects are nocturnal or *some* if a minority of mosquitoes have white stripes.

Even as generics seem to express inconsistent quantifications, they are at the heart of communication and dissemination in science (DeJesus et al., 2019; Bowker, 2022), medical research (Peters et al., 2024), and politics (Novoa et al., 2023). Furthermore, in the social realm generics serve as linguistic vehicle for social essentialism (Rhodes et al., 2012) and stereotyping (Leslie, 2017; Bosse, 2022).

The nature of generics and their importance in communication has led to extensive literature on the semantics of generic sentences (e.g., Carlson, 1977b; Cohen, 1999; Leslie, 2008; Liebesman, 2011; Sterken, 2015a; Nickel, 2016; Tessler and Goodman, 2016; Stovall, 2019; Nguyen, 2020; Bosse, 2021; Kirkpatrick, 2023). However, many open questions remain. These include how they relate to quantifiers and the degree to which generics are context sensitive. In this work, we use language models to explore the implicit quantification and context-sensitivity of generics, and how they are affected by human biases around stereotypes.

Language models have demonstrated unprecedented performance in a variety of linguistic tasks, such as machine translation (Kocmi et al., 2024) or conversational assistance (Chiang et al., 2024). We describe how speakers use generics by studying the surprisal in language models for various naturally occurring generic and quantified sentences.

Most existing datasets of generics are synthetic, often derived from knowledge bases or generated by language models (Bhakthavatsalam et al., 2020; Allaway et al., 2024). Since the examples in these datasets are machine-generated and/or lack a context in which they might be uttered, there is no guarantee that they represent how speakers actually use generics. Therefore, in this work we introduce CONGEN, a dataset of naturally occurring generic

[1]Code and data are available in https://github.com/ilyocoris/generics_are_puzzling

Proceedings of the 31st International Conference on Computational Linguistics, pages 6571–6588
January 19–24, 2025.

and quantified sentences with contexts.

In order to study generics and quantification, we define the p-acceptability metric; given a set of quantifiers, it indentifies the one that best fits a sentence by using the surprisal of a language model. Previous work either fails to use surprisal to describe quantification or focuses on prompting (Collacciani et al., 2024; Allaway et al., 2024). We validate this metric by showing that it recovers the expected dynamics of quantifiers (*all*, *most*, *some*) and the generic on two datasets of generics (CONGEN and GENERICSKB). We then use our metric to study different aspects of generics.

Our contributions are (*i*) CONGEN, a dataset of naturally occurring bare plural generic and quantified sentences in contexts, (*ii*) p-acceptability, a new metric based on language models that is sensitive to quantification and (*iii*) insights into how generics are used, including weak generalisations, context-sensitivity and stereotypes.

2 Background

Semantic theories of generics guide and scaffold our experimental design. In what follows, we present linguistic background on generics (§2.1) and then we sketch two theories of genericity related to our experiments (§2.2): *generics-as-defaults* and *contextualism*. Finally, we introduce two phenomena that involve generics (§2.3): *stereotypes* and *generic overgeneralisation*. These theoretical elements motivate our research questions and frame the interpretation of experimental results.

2.1 Generics

The term *generics* covers multiple linguistic phenomena that abstract away from particular objects, members, or events. In our work, we focus on one specific kind of generics: *bare plural characteristic sentences*. Bare plurals are noun phrases in plural form without a definite or indefinite article[2]. Characteristic sentences are propositions that do not express specific episodes or isolated facts, but rather report a kind of general property or regularity (Carlson and Pelletier, 1995).

In linguistics, generics are traditionally analyzed as quantifiers (Carlson, 1977b), with an unpronounced implicit operator GEN that has a role similar to that of *most* or *generally* in explicit quantification (Lewis, 1975). However, there is no consensus on what the semantic content of GEN is, how it is determined or even if it exists (cf. Carlson and Pelletier, 1995). Despite this, most real-life generics are *majority generics*: they are acceptable when a majority of the members of the kind in question satisfies the predicated property (e.g. *ravens are black*).

However, some generics, such as *mosquitoes carry malaria*, *ducks lay eggs*, or *bees reproduce*, are often used even though they express statistically weak generalisations. We call these *weak generics* (Almotahari, 2022). Weak generics are broadly categorized into two types (Leslie, 2007): those that express properties that are characteristic of the kind, but are only possessed by a minority of its members (*minority generics*) and those that express dangerous, striking or appalling characteristics (*striking generics*).

In addition to the above, Leslie et al. (2011) also distinguish *quasi-definitional* generics. These predicate a property true of all the members of the kind, without exceptions (e.g. *beetles are insects*).

2.2 Generics in Philosophy of language

In this section we introduce two influential accounts from philosophy of language: generics-as-defaults and contextualism. These supply contrasting views on how to explain the diversity in generic use and how they are affected by context; topics we discuss in our experiments (§5.2 and §5.3 respectively).

Generics as defaults. The *generics-as-defaults* theory posits generics as the linguistic manifestation of a default cognitive mechanism of generalisation (Leslie, 2008). In contrast to quantifiers, which express generalisations based on statistical surveying, generics express primitive generalisations based on what we perceive as characteristic, distinctive or striking in the world (Leslie, 2007).

Contextualism. Sterken (2015a) argues for a contextualist view of generics: generics express widely different generalisations in different contexts. The unpronounced generic operator GEN picks out a generalisation as a function of the context of the utterance, similarly to how the determiner *that* picks out a referent.

[2] *Sharks attack bathers* is a bare plural generic. The same generic could also be expressed in English with the definite (*the shark attacks bathers*) or indefinite (*a shark attacks bathers*) articles.

2.3 Human biases in the usage of generics

Generics are often used in ways that do not follow logical reasoning and highlight human cognitive biases (Leslie, 2017; Neufeld et al., forthcoming). One important example is the connection generics have with how we express stereotypes; we explore this in experiment §5.4.

Generic overgeneralisation. One example of illogical use of generics is *generic overgeneralisation* (Leslie et al., 2011; Lazaridou-Chatzigoga et al., 2017): humans often use universal quantification (*all*) in situations where the generic is acceptable even when exceptions exist. This effect has also been documented in language models (Allaway et al., 2024; Ralethe and Buys, 2022).

Stereotypes. In the social realm, striking generics are linked to stereotyping and the essentialization of social groups (Rhodes et al., 2012; Leslie, 2017). In particular, Cimpian et al. (2010) and Khemlani et al. (2009) demonstrate a psychological connection between striking information and an overestimation of statistical frequencies. This means that humans seem to reason from the quantifier *some* to *most* and even to *all* when striking properties are at play (Leslie, 2008). Additionally, generics are also central in recent NLP studies on preventing and countering stereotypes (Bosse, 2022; Allaway et al., 2023b; Mun et al., 2023).

3 Related work

Recent works that study generics and language models use prompting to test generic overgeneralisation, property inheritance (Allaway et al., 2024) and, more generally, the effect of quantifiers on sentence meaning (Collacciani et al., 2024). However, for studying how language models model quantification and generics, prompting has several shortcomings. In particular, the effect of small variations in the prompt on model behavior is not well understood (Salinas and Morstatter, 2024). Additionally, prompting requires an instruction tuned model, often trained to be a virtual assistant (Zhang et al., 2024), which may skew the underlying language distribution in unaccounted ways.

To avoid the drawbacks of prompting, studies have also looked at the internal states of pre-trained models. Collacciani et al. (2024) compare the surprisals of quantified sentences but fail to find a sensitivity to quantification in language models. As the authors note, this may be due to their metric not being sufficiently expressive. While the work from Gupta (2023) also uses surprisal, in this case of critical words, to draw conclusions about quantifier comprehension in language models, it does not take generics into consideration. In contrast, in this work we develop a new metric that uses the surprisal of the predicated property tokens, and show that it describes rich quantificational dynamics modelled by language models.

Several datasets exist that specifically target generics. GENERICSKB is a dataset (3M samples) that combines naturally occurring generic and quantified statements with synthetic examples derived from knowledge bases (Bhakthavatsalam et al., 2020). The naturally occurring generics are selected with a BERT-based scorer trained on human annotations. The GEN-A-TOMIC corpus contains synthetic generics generated by GPT2-XL (Bhagavatula et al., 2023). Additionally, datasets of synthetic generics exemplars (i.e., cases where the generic does and does not hold) have been constructed (Allaway et al., 2023a, 2024).

All of these datasets contain synthetic examples (either machine generated or derived from knowledge bases) and do not include context, which is key to understanding how speakers use generics. In contrast, our CONGEN dataset contains only naturally occurring human-annotated sentences, each with an associated document as context.

4 Methodology

4.1 Dataset: CONGEN

Theorists emphasize the context-sensitivity of generic sentences (Sterken, 2015a; Nickel, 2016; Almotahari, 2023). The lack of consensus on how context affects the use of generics motivates the construction of CONGEN. To the best of our knowledge, this is the first dataset that targets generics in context.

CONGEN consists of naturally occurring bare plural generics and quantified statements (with *some*, *most* and *all*) in context. These are drawn from a subset of DOLMA (Soldaini et al., 2024) and from 2024 Reddit comments. DOLMA is a cleaner version of Common Crawl and may have been used in the training data for popular language models (e.g., MISTRAL). Therefore, we include recent Reddit comments to validate our findings on data the models have not been trained on.

In order to find bare plural generic sentences in such massive collections of data, we train a binary

DOLMA	REDDIT 2024
The potato variety dictates the color of the flower which for red potatoes can be dark pink to lavender. **Yellow potatoes have white flowers.** Piling our storage potatoes starts in late-September and by mid-October all our potatoes are in the barn.	No. Sprinters typically have long legs. Runners in general have long legs. **Swimmers have long torsos.** Michael Phelps, who is 6'4", has shorter legs than the Olympic runner Hicham El Guerrouj, who is 5'9"
There are several reasons for a high number of repetitive leg movements while sleeping. **Some people with chronic pain at night tend to have poor sleep and frequent repetitive leg movements.** If you have concerns about your sleep, you should discuss them with your doctor.	I have this problem. It's not because I don't like vegetables. I can just taste waaaay too much of the minerals. **Most vegetables taste like iron and dirt.** I'd rather eat actual dirt than a beet. Also it makes water taste weird when you can taste minerals.

Table 1: Examples of generic and quantified sentences in context, extracted from the CONGEN dataset.

Source	Quantifier	# Samples
DOLMA	GEN	559
	ALL	500
	MOST	578
	SOME	551
Reddit (2024)	GEN	411
	ALL	71
	MOST	158
	SOME	45

Table 2: CONGEN dataset: breakdown of the 2873 annotated sentences in context.

classifier to detect generic and quantified sentences. We train a ROBERTA (Liu et al., 2019) classifier on GENERICSKB and GEN-A-TOMIC.

We find candidate sentences in the original data sources by combining the scores of the classifier with linguistic heuristics that filter out sentences in the singular or in past tense. The collected candidate sentences are annotated as irrelevant, bare plural generic or explicitly quantified with *all*, *most* or *some*. The final dataset contains 2873 human-annotated generic and quantified sentences (Table 2). Details on dataset construction are available in Appendix B.

4.2 Metric: p-acceptability

Generic sentences can be used to express generalisations with vastly different quantificational strength: from weak generics (e.g., *mosquitoes carry malaria*) to quasi-definitional ones (e.g., *mosquitoes are insects*). To describe and study these quantificational dynamics in language models, we introduce a criterion to answer the following question: What is the quantifier that best fits the kind-property relation expressed in a sentence?

Consider quantified bare plural generalisations with a simple structure (context + quantifier + bare plural + verb + property), where the quantifier is one of *all*, *most*, *some* or the generic (Ø). We propose a notion of acceptability that selects the *quantifier that makes the property more likely* given the subject, verb and context.

Definition 4.1 (p-acceptability)**.** Let Q be a set of candidate quantifiers, s a bare plural generic and θ a language model. We construct $\{q + s \mid q \in Q\}$ the set of variations of s[3]. We call q the p-acceptable quantifier for s if $q + s$ is the sentence with the lowest surprisal of the property tokens (i.e. tokens after the verb):

$$\text{p-acceptable}(s; Q, \theta) := \underset{q \in Q}{\operatorname{argmin}}\, H_p(q + s; \theta) \tag{1}$$

where H_p is the surprisal of the property tokens

$$H_p(s; \theta) := -\frac{1}{|P|} \sum_{i \in P} \log p_\theta(t_i | t_{<i}) \tag{2}$$

with P is the set of indices of the property tokens and t_i the tokens in sentence s.

We build the set of variations of s as $\{s$, 'all'$+s$, 'most'$+s$, 'some'$+s\}$. For sentences that originally had an explicit quantifier, we remove the quantifier to obtain s. To compute the p-acceptability of a sentence s with context c, we build the set of variations $\{c + q + s \mid q \in Q\}$.

For example, consider the sentence s = *tigers have stripes* which can be split into word tokens t_0 = `tigers`, t_1 = `have` t_2 = `stripes`. Recall that the candidate quantifiers are $Q = \{\text{all}, \text{most}, \text{some}, \emptyset\}$. Then, the set of variations will be *all tigers have stripes*, *most tigers have stripes*, *some tigers have stripes* and *tigers have stripes*. The surprisal on the property tokens (in this case t_2) with the quantifier "all" is then calculated as:

[3] $q + s$ refers to string concatenation: if q = most and s = *tigers have stripes*, then $q+s$ = most tigers have stripes.

$$H_p(\texttt{all} + \texttt{tigers have stripes}; \theta) =$$
$$-\frac{1}{|P|}\sum_{i \in P} \log p_\theta(\texttt{stripes}|\texttt{all tigers have})$$

and similarly for the other quantifiers in Q. The p-acceptable quantifier would then be the one with the minimum surprisal.

As we show in the following experiments, p-acceptability is sensitive to the effect of using a quantifier (*all*, *most*, *some*) or the generic (§5.1). We note that previous work found that the surprisal of the whole sequence is not sensitive to the effect of quantifiers (Collacciani et al., 2024) and we replicate this finding (see Appendix C).

5 Experiments

The experiments that follow use p-acceptability (defined in §4.2) to study quantification and generics through language models. First, we validate that p-acceptability describes quantification in CONGEN and GENERICSKB (§5.1). Then, we explore three aspects of generics: their implicit quantificational strength (§5.2), their context-sensitivity (§5.3), and their role in stereotypes (§5.4).

We use three state-of-the-art open-source language models of increasing size: MISTRAL-7B, MISTRAL-8×7B and MISTRAL-8×22B (Jiang et al., 2023). Additional details on the models used are available in Appendix A.

Because our focus is on bare plural generics, we filter out of GENERICSKB those generics that are not bare plural and call this subset GENERICSKB-BP ($N = 570358$). Implementation details can be found in Appendix D.

5.1 Can p-acceptability describe quantification?

Quantifiers specify a prevalence relation between members of a kind and a property. In terms of this relation, we would expect *all* and *most* to be interchangeable in many contexts, likewise for *most* and *some* but never for *all* and *some*. This experiment recovers these commonsense intuitions of quantification with p-acceptability.

Experimental setup. For each sentence in CONGEN and GENERICSKB-BP, we build the set of variations and get the p-acceptable quantifier. We plot these p-acceptability percentages against the original quantifiers of the sentences, that is, how often each quantifier makes the property tokens easiest to predict for the language model.

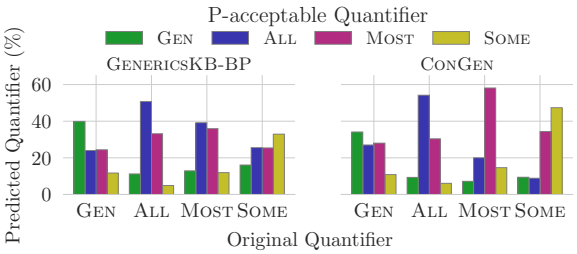

Figure 1: P-acceptable quantifiers on both datasets correspond to semantic intuitions (MISTRAL-7B)

Results. In both datasets, the most prevalent p-acceptable quantifier corresponds to the original quantifier (Figure 1). In GENERICSKB data, the distinctions are less clear, with bigger confusion between, for example, *all* and *most*.

For sentences that were originally generic, *all* and *most* are the most prevalent wrongly p-accepted quantifiers. This agrees with most generics being majority generics.

The prevalence of *all* in originally *some* sentences from GENERICSKB-BP seems counter-intuitive. We believe that this is due to noise in the dataset[4], rather than the metric. On CONGEN, p-acceptability recovers an intuitive profile for *some* sentences: *some* is the most prevalent quantifier and is mostly confused with *most*, rarely with *all* or the generic.

P-acceptability captures semantic intuitions on quantification across both datasets. In what follows, we use p-acceptability to investigate some aspects of how speakers *use* generics.

5.2 What is the implicit quantification of generics?

Although generic sentences present no overt quantification operator, we can investigate which quantifier better describes the kind-property relationship expressed in a generic with p-acceptability. Given a generic sentence, we study its *implicit quantification* by finding the p-acceptable explicit quantifier.

Experimental setup. In this experiment we consider generics from CONGEN and GENERICSKB-BP. We compute the p-acceptability excluding GEN from the candidate quantifiers and only considering *all*, *most* and *some* as possible options for quantification.

[4]For example, the three first *some* sentences in GENERICSKB are *Some aardvarks detect predators*, *Some aardvarks dig holes* and *Some aardvarks dig own burrows*. These feel more appropriate for *most* or *all* quantification.

ALL
Aquatic crustaceans have gills for breathing.
Dead plants contain vital substances beyond just carbon.
Omnivores eat both plants and animals.
MOST
Narcotics cause a good deal of vasodilation.
Banana plants have a lot of root exudates.
People adapt to total blindness.
SOME
Ocean currents carry water over long distances.
Berries are toxic to humans but loved by birds.
Plastics bind heavy metals.

Table 3: Examples of naturally occurring generics from CONGEN with different implicit quantifications. (p-acceptability from MISTRAL-7B).

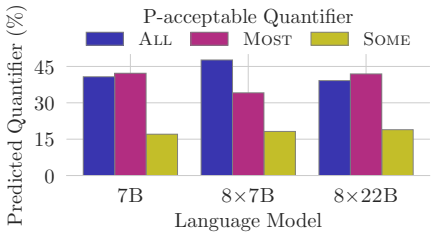

Figure 2: Implicit quantification in CONGEN generics across MISTRAL models.

Results. Across all three models (Figure 2), *all* and *most* are the most p-acceptable implicit quantifiers at a prevalence of 40% each. Around 18% of sentences are consistently quantified as *some*.

For the 7B and 8X22B models, *most* is the most prevalent quantifier, which mirrors the fact that generic sentences often express properties shared by a majority of members of a kind. Nevertheless, *all* has comparable or even bigger prevalence for MISTRAL-8×7B. Although we expect *all* to be the implicit quantifier for quasi-definitional generics, the observed high prevalence of *all* suggests that language models also model the generic overgeneralisation effect (§2.3), as found in other studies (Allaway et al., 2024). A more fine-grained annotation is needed to verify this on CONGEN data.

We take those generics with *some* as the p-acceptable quantifier to be weak generics. We observe close to 20% of weak generics across all models both in CONGEN (Figure 2) and GENERICSKB-BP (Figure E.1). To the best of our knowledge, this is the first estimation of the prevalence of weak generics in natural language. Table 3 shows examples of generics with different implicit quantifications (also in Appendix I, Table I.5).

5.3 Are generics context-sensitive?

Semantic theories in philosophy of language hypothesize that the context of generic sentences determines the semantic content of GEN (§2.2). We quantify the effect of different context windows on implicit quantification using p-acceptablility and the multi-sentence contexts in CONGEN.

Experimental setup. For each sentence in CONGEN, we compute the p-acceptable quantifier at increasing sizes of left-side context. We increase the context size in chunks of 4 tokens, irrespective of word or sentence boundaries (Table I.6).

We measure the percentage of correct predictions by p-acceptability as instances where it recovers the original quantifier. For sentences that are originally generics, we also replicate this setup (excluding GEN from the candidate quantifiers) for different left-context windows.

Results. Figure 3 shows the percentage of correct predictions for each original quantifier (e.g. green corresponds to the percentage of times *gen* is p-accepted on generic sentences at each context length). In originally generic sentences, we have a 20% increase in accuracy across the first 20 tokens of context, which roughly correspond to the preceding sentence. For explicitly quantified expressions, context does not improve the accuracy of p-acceptability as much as for generics.

As control, we replicate the experiment with random context sampled from other documents with the same original source (DOLMA or Reddit) and find no improvement on any quantifier, including GEN. Details are available in Appendix F.

We investigate if the increase in accuracy that context has on generic sentences is related to their implicit quantification. In Figure 4 the relative percentages of each quantifier are mostly unaffected by context, with a slight increase of *all* and decrease of *most*.

For those samples where context is needed for p-acceptability to predict the correct quantifier, we define the *minimal context* as the smallest context needed for the correct prediction. We find a very low presence of quantifiers in the minimal contexts of generic sentences. A preliminary analysis of the linguistic characteristics of these contexts is available in Appendix G.

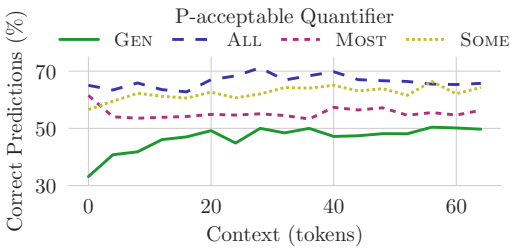

Figure 3: Percentage of correct p-acceptable quantifiers with different contexts. (MISTRAL-8×22B)

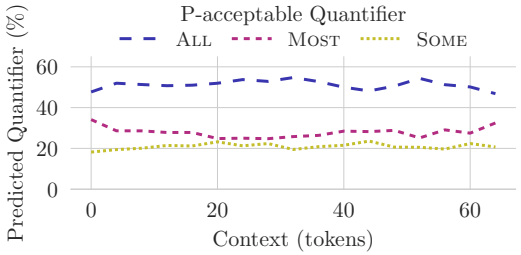

Figure 4: Implicit quantification with different left-side contexts on generic sentences from CONGEN. (MISTRAL-8×22B)

5.4 Are stereotyping generics different?

Stereotypes are often expressed linguistically through striking generics where the subject is a social group. This is partly because, when dangerous properties are predicated, humans perceive them as more prevalent than they really are (Cimpian et al., 2010; Leslie, 2017). In the following experiment, we study the implicit quantification in language models of negative and positive stereotypes.

Experimental setup. To study the implicit quantification in this subset of generics, we collect a small dataset of stereotypes ($N = 504$) divided into *real* (the subject-property is a real-world stereotype) and *invented* sentences (the subject is an invented word that morphologically resembles a demonym).

We extract real negative stereotypes from the Social Bias Frames dataset (Sap et al., 2020), a collection of offensive texts annotated with implied stereotypes. For real positive stereotypes, we generate samples based on tradition and culture for different social groups. The invented sentences are built by combining invented demonyms with a list of negative and positive predicates (e.g., *craguils are murderers* or *corriards are warm and hospitable*). Details are available in Appendix H and Table I.7.

To further explore how effective purely linguistic strategies are at mitigating the bias in striking generics (Leslie, 2017; Carnaghi et al., 2008; Gelman and Heyman, 1999), we generate the following three paraphrases for each social group in a stereotype: bare plural (*catalans are lovely*), singular + 'people' (*catalan people are lovely*) and 'people who are' + singular (*people who are catalan are lovely*).

We compare the results between the pre-trained and instruction tuned versions of MISTRAL-7B, as one objective of language model designers when instruction-tuning models is to mitigate social biases (Zhang et al., 2024).

Results. Figure 5 reports the percentage of p-acceptable quantifiers for each paraphrase and type of stereotyping generic. For negative stereotypes, *all* is the predominant quantifier. This aligns with the theoretical and empirical observation that speakers use universal quantification with this subset of striking generics (Cimpian et al., 2010). Note that with the *people who are* paraphrase this is not the case; we observe a stark contrast, where *some* is the most prevalent p-acceptable quantifier. The instruction-tuned model predicts more *some* and less *all*. Interestingly, for invented negative cases even in the *ppl who* paraphrase, *all* is the most prevalent quantifier.

In contrast to the negative stereotypes, the predominant quantifier is *most* for positive stereotypes. This further supports hypotheses that the implicit universal quantification of negative stereotypes is due to the strikingness of the predicate.

6 Discussion

In this work, we study different aspects of generics and quantified sentences through language models. We now discuss our results in relation to existing theories of generics.

Weak generics. Weak generics are central to discussions of generics in philosophy of language. On the one hand, Leslie (2008) uses the prevalence of striking generics to support the idea that generics express primitive psychological generalisations. In contrast, Sterken (2015b) proposes an error theory where striking generics are false. Additionally, Gustafsson (2023) examines weak (striking) generics and argues that generics are more heterogeneous than what the previous theories take them to be. Although these works use striking generics as their running examples, they pay little attention to *how many* generics are actually weak nor what

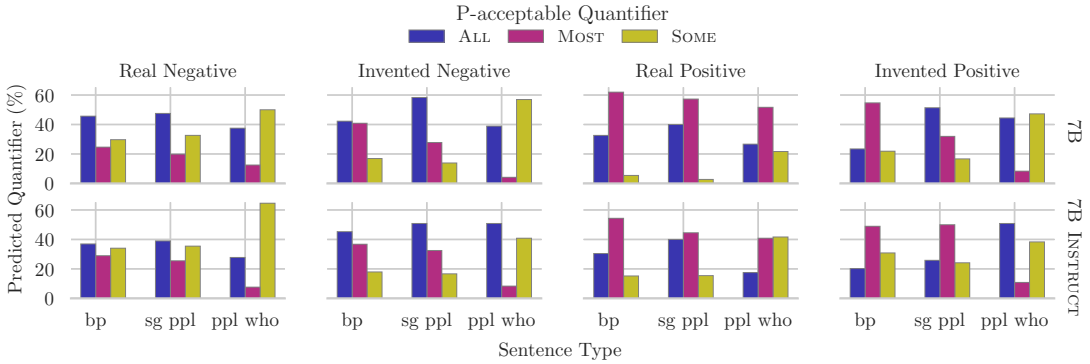

Figure 5: Different p-acceptability rates for each paraphrase of stereotyping generic sentences for MISTRAL-7B and MISTRAL-7B INSTRUCT. Paraphrases are indicated as *bp* (bare plural), *sg ppl* (singular + 'people') and *ppl who* ('People who are' + singular).

the non-striking weak generics *look like*.

In our experiments on implicit quantification (§5.2) we characterize weak generics as those generics implicitly quantified with *some*. If we take the 437414 generics in CONGEN and GENERICSKB-BP combined to be a representative sample of language, between 18% and 23%[5] of generics are weak generics. Manual examination of these weak generics reveals that they are mostly non-striking (Tables 3 and I.5).

These results suggest that *non-striking weak generics are common in language use*: almost 1 in 5 generics we analyze expresses a weak generalisation. Our notion of implicit quantification not only can give an estimation of *how many* weak generics are there, but, when applied to CONGEN, yields explicit examples of weak generics in-context. This is a new resource for theorists to draw from.

Context-sensitivity. The degree of context-sensitivity of the semantic content of GEN is a controversial topic in the literature. The *generics-as-defaults* view posits stable, non-context-sensitive content whereas the *contextualist* view claims a distinct and strong context-sensitivity. Recently, Almotahari (2023) argued that both views are compatible, by attributing some context-sensitivity to psychologically salient features in *generics-as-default*.

In our experiments (§5.3), increasing context improves the accuracy of p-acceptability for generic sentences much more than for explicitly quantified sentences. Additionally, when considering the implicit quantification of generic sentences, the prevalence of each quantifier is not affected by the context. These results suggest that *generics are context-sensitive in a way that determiner quantifiers are not*.

In this work, we limit our contribution to showing that some dynamics of context-sensitivity in generics can be revealed with language models, and leave further inquiry, both empirical and theoretical, for future work.

Stereotypes. The use of generics to express stereotypes is a much explored topic both in philosophy of language and experimental psychology (Cimpian et al., 2010; Leslie, 2017). The following two hypotheses are central to the discussion: (*i*) humans interpret negative stereotypes as universally quantified and (*ii*) how speakers express a stereotype changes its the perceived quantificational force.

In the context of stereotypes, we take implicit quantification as a measure of bias: interpreting stereotypes as universally quantified is a sign of essentialization and prejudice. An unbiased system should quantify stereotypes as *some*, sometimes *most*, but never *all*.

Our results (§5.4) are congruent with both hypotheses. We find that negative stereotypes are overwhelmingly implicitly quantified as universals (*all*), whereas for positive stereotypes the most p-accepted quantifier is *most*. For those same negative stereotypes, when paraphrased as *ppl who*, *some* becomes increasingly prevalent, suggesting an existential perception of quantification.

The study of social bias in language models is important in order to make them fair and safe to use. Bias mitigation is often a priority of language model designers when instruction tuning. For real stereotypes, the instruction-tuned model predicts

[5]These are the percentages of generic sentences with the p-acceptable quantifier *some* with MISTRAL-8×22B on CONGEN and GENERICSKB-BP respectively.

less *all* and more *some* than the base model, which suggests a success in bias mitigation. Nevertheless, on invented negative stereotypes, the instruction model predicts a large amount of *all*, even for the *ppl who* case. Although our experiments are not on generation, this raises doubts on the effectiveness of instruction tuning for bias mitigation, especially when new social kinds are combined with striking properties.

7 Conclusion

Generics are similar to quantifiers, yet speakers use them in logically inconsistent ways. In this work, we study the dynamics of quantification on generic and quantified sentences through language models.

To do so, we introduce a new dataset (CONGEN) and metric (p-acceptability). With these tools we estimate the prevalence of weak generics, identify a distinct context-sensitivity in generics and show how linguistic strategies can help mitigate stereotypes.

We believe our findings and methodology open new doors for research on generics and quantification in language.

8 Limitations

Language models. Even though we use SOTA open-source language models for our experiments, currently all competitive language models are trained for profit rather than research; this inevitably hinders any research effort.

Compute limitations mean we run MISTRAL-8×7B 8-quantized and MISTRAL-8×22B 4-quantized. There is empirical evidence that quantization does not have a big impact on performance for MISTRAL models across a wide range of tasks (Badshah and Sajjad, 2024).

We test on a family of models with the autoregressive transformer architecture. Exploration of how other autoregressive families or architectures (such as MAMBA or diffusion) model quantification and generics is left for future work.

Metric. The p-acceptability metric, as defined, is specific to English sentences with a simple structure, as the bare plural does not exist in many other languages.

Implicit quantification The implicit quantification in generic sentences could be more closely related to adverbial quantifiers than to the determiner quantifiers we study (Kirkpatrick, 2024). Future work will need to expand both the metric and dataset to include other quantifiers like *many*, *every* or *few*, in order to get a more comprehensive picture of quantification and its relation to generics.

Data. Sentences in CONGEN are first collected from DOLMA and Reddit by a classifier trained to identify generic and quantified sentences. Performance and biases of this classifier are not well explored in this work and could affect the significance of the sampling. Another source of bias in CONGEN is that the first author annotates most of the data. We plan on adressing these issues and expanding CONGEN in future work.

In the stereotypes collection we build, the stereotypes are derived from a dataset based on American Twitter, which means they are centered around American and Western culture and prejudice.

Context. The concept of context can broadly mean three things in the philosophy of language literature: the spatial and temporal context of the utterance, the subjective context of the speaker (such as intentions) or the linguistic context (previous utterances). In this work, we assume linguistic context as the only source of context-sensitivity, as we study how language models model the context. If generics were context-sensitive in ways that are not expressed or conveyed in language, our methodology could not capture it.

Acknowledgements

This work was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10039436].

The first author would like to thank Mahrad Almotahari for introducing him to the topic of generics and Dan Lassiter, Annie Bosse, Nicolas Navarre, Tristan Baujault-Borresen, Guillem Ramírez and Aina Centelles for their proof-reading, corrections and discussions.

We also thank the anonymous reviewers for their comments on the manuscript.

References

Emily Allaway, Chandra Bhagavatula, Jena D. Hwang, Kathleen McKeown, and Sarah-Jane Leslie. 2024. Exceptions, Instantiations, and Overgeneralization: Insights into How Language Models Process Generics. *Computational Linguistics*, pages 1–60.

Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen McKeown, Doug Downey, and Yejin Choi. 2023a. Penguins don't fly: Reasoning about generics through instantiations and exceptions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2618–2635, Dubrovnik, Croatia. Association for Computational Linguistics.

Emily Allaway, Nina Taneja, Sarah-Jane Leslie, and Maarten Sap. 2023b. Towards countering essentialism through social bias reasoning. *Preprint*, arXiv:2303.16173.

Mahrad Almotahari. 2022. Weak generics. *Analysis*, 82(3):405–409.

Mahrad Almotahari. 2023. Generic cognition: A neglected source of context sensitivity. *Mind and Language*.

Sher Badshah and Hassan Sajjad. 2024. Quantifying the capabilities of llms across scale and precision. *Preprint*, arXiv:2405.03146.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Preprint*, arXiv:2001.08435.

Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Lianhui Qin, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. 2023. I2d2: Inductive knowledge distillation with neurologic and self-imitation. *Preprint*, arXiv:2212.09246.

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

Anne Bosse. 2021. Generics: Some (non) specifics. *Synthese*, (5-6):14383–14401.

Anne Bosse. 2022. Stereotyping and generics. *Inquiry: An Interdisciplinary Journal of Philosophy*, pages 1–17.

Mark Bowker. 2022. A problem for generic generalisations in scientific communication. *Journal of Applied Philosophy*, 39(5):1002–1017.

Robert Brandom. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press, Cambridge, Mass.

Greg N. Carlson, editor. 1977b. *Reference to Kinds in English*.

Greg N. Carlson and Francis Jeffry Pelletier, editors. 1995. *The Generic Book*. University of Chicago Press.

Andrea Carnaghi, Anne Maass, Stefania Gresta, Mauro Bianchi, Mara Cadinu, and Luciano Arcuri. 2008. Nomina sunt omina: On the inductive potential of nouns and adjectives in person perception. *Journal of Personality and Social Psychology*, 94(5):839–859.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

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.

Ariel Cohen. 1999. Generics, frequency adverbs, and probability. *Linguistics and Philosophy*, 22(3):221–253.

Claudia Collacciani, Giulia Rambelli, and Marianna Bolognesi. 2024. Quantifying generalizations: Exploring the divide between human and llms' sensitivity to quantification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11811–11822. Association for Computational Linguistics.

Jasmine M. DeJesus, Maureen A. Callanan, Graciela Solis, and Susan A. Gelman. 2019. Generic language in scientific communication. *Proceedings of the National Academy of Sciences*, 116(37):18370–18377. Contributed by Susan A. Gelman, July 20, 2019 (sent for review October 15, 2018; reviewed by Ellen M. Markman and Douglas L. Medin).

John D. Garrett. 2021. garrettj403/SciencePlots.

Susan A. Gelman and Geoffrey D. Heyman. 1999. Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological Science*, 10(6):489–493.

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, Singapore. Association for Computational Linguistics.

Matti Gustafsson. 2023. Taking truth seriously: The case of generics. *Synthese*, 202(3).

Martin Heidegger. 1927. *Being & Time*. Max Niemeyer.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

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. *Preprint*, arXiv:2310.06825.

Sangeet Khemlani, Sarah-Jane Leslie, and Sam Glucksberg. 2009. Generics, prevalence, and default inferences.

James Ravi Kirkpatrick. 2023. The dynamics of generics. *Journal of Semantics*, 40(4):523–548.

James Ravi Kirkpatrick. 2024. Are generics quantificational? *Synthese*, 204(17).

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era is Here but MT is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, United States.

Dimitra Lazaridou-Chatzigoga, Linnaea Stockall, and Napoleon Katsos. 2017. A new look at the 'generic overgeneralisation' effect. *Inquiry*, 66(9):1655–1681.

Sarah-Jane Leslie. 2008. Generics: Cognition and acquisition. *Philosophical Review*, 117(1).

Sarah-Jane Leslie. 2017. The original sin of cognition: Fear prejudice, and generalization. *Journal of Philosophy*, 114(8):393–421.

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.

Sarah-Jane Leslie. 2007. Generics and the structure of the mind. *Philosophical Perspectives*, 21:375 – 403.

David Lewis. 1975. Adverbs of quantification. pages 5–20.

David Liebesman. 2011. Simple generics. *Noûs*, 45(3):409–442.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777, Singapore. Association for Computational Linguistics.

Eleonore Neufeld, Annie Bosse, Guillermo Del Pinal, and Rachel Sterken. forthcoming. Giving generic language another thought. *WIREs Cognitive Science*.

Anthony Nguyen. 2020. The radical account of bare plural generics. *Philosophical Studies*, 177(5):1303–1331.

Bernhard Nickel. 2016. *Between Logic and the World: An Integrated Theory of Generics*. Oxford University Press UK, Oxford, GB.

Gustavo Novoa, Margaret Echelbarger, Andrew Gelman, and Susan A. Gelman. 2023. Generically partisan: Polarization in political communication. *Proceedings of the National Academy of Sciences*, 120(47):e2309361120. Contributed by Susan A. Gelman; received June 3, 2023; accepted September 25, 2023; reviewed by Yphtach Lelkes and Gregory L. Murphy.

Uwe Peters, Henrik Sherling, and Benjamin Chin-Yee. 2024. Hasty generalizations and generics in medical research: A systematic review. *PLOS ONE*, 19.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *CoRR*, abs/2003.07082.

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.

Marjorie Rhodes, Sarah-Jane Leslie, and Christina M. Tworek. 2012. Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, 109(34):13526–13531. Edited by Douglas L. Medin, Northwestern University, Evanston, IL, and approved July 2, 2012 (received for review May 25, 2012).

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *Preprint*, arXiv:2401.03729.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Amir Sepehri, Mitra Sadat Mirshafiee, and David M. Markowitz. 2023. Passivepy: A tool to automatically identify passive voice in big text data. *Journal of Consumer Psychology*, 33(4):714–727.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha

Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. *Preprint*, arXiv:2402.00159.

Rachel Sterken. 2015a. Generics in context. *Philosophers' Imprint*, 15:1–30.

Rachel Katharine Sterken. 2015b. Generics, content and cognitive bias. *Analytic Philosophy*, 56(1):75–93.

Preston Stovall. 2019. Characterizing generics are material inference tickets: A proof-theoretic analysis. *Inquiry: An Interdisciplinary Journal of Philosophy*, (5):668–704.

Michael Henry Tessler and Noah D. Goodman. 2016. The language of generalization. *CoRR*, abs/1608.02926.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Blackwell, Oxford.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

A Language Models

In our experiments we use autoregressive transformers from the MISTRAL family (Jiang et al., 2023). Table A.1 compares their size and performance in the MMLU benchmark (Hendrycks et al., 2020), which test the performance of language models across a wide range of topics and tasks.

Model name	Active params.	MMLU (↑)
MISTRAL-7B	7 B	62.5%
MIXTRAL-8×7B	12.9 B	70.6%
MIXTRAL-8×22B	39 B	77.8%
GPT-3.5 TURBO	undisclosed	70.0%

Table A.1: Comparison of size and performance of the language models used, with GPT-3.5 TURBO (ChatGPT) for comparison.

B CONGEN dataset

Data sources. We use a sample of 1.45 million documents (over 100 million sentences) from DOLMA (`https://huggingface.co/datasets/andersonbcdefg/dolma_sample`). We scrape the comments written in June and July 2024 from Reddit using the `pullpush.io` API (Baumgartner et al., 2020) from a list of popular subreddits. The context for the Reddit comments is the comment itself, not a thread with other responses.

Generics classifier training. For the GEN-A-TONIC dataset, we label as negative examples sentences with a i2d2 score < 0.6 or that do not match the following regex pattern: `can | may | should | you |before`. We also remove adverbial quantifiers like `typically` and `generally`. For GENERICSKB we use the scores of the original dataset as labels. The resulting training dataset has $N = 9414001$ samples.

We train a ROBERTA-base model (Liu et al., 2019) with a sequence classification head on this dataset for one epoch with learning rate 2×10^{-5}.

Candidate sentences and annotation. To reduce candidate sentences from millions to few thousands, we filter the sentences with the ROBERTA classifier (score > 0.7) and the regex pattern below, which includes words that are often incompatible with the generic and quantified bare plurals we study:

```
is | may | can | should | would | must | have to
↪  | will | you |^i | were | was | many | we |
↪  they | ought | your |^[^ ]+ of | us | \? |
↪  this | that | those | these | all in all
↪  |,|^the |^a |than
```

We also filter out sentences in the passive voice (Sepehri et al., 2023).

The candidate sentences are annotated by the authors using Label Studio (Tkachenko et al., 2020-2022).

C P-acceptabilty with H and H_p

We compare the notion of acceptability in equation 1 using the surprisal over the whole sequence (H) or the surprisal on the property tokens (H_p). We observe that the surprisal over the whole sequence is not able to correctly predict generic sentences unless we provide big context windows.

Context (tokens)	H	H_p
0	0.007	0.34
32	0.03	0.43
128	0.32	0.45

Table C.2: Accuracy on originally generic sentences of p-acceptability using the entropy over the whole sequence (H) or on the property tokens (H_p); H needs a big context window to be sensitive to generics (MISTRAL-7B).

D Experimental setup

To use GENERICSKB in our setting, we select only bare and quantified plurals. The quantified samples are already labeled in the original dataset (*all*, *most* and *some*). We select the bare plurals from the generic samples by finding sentences with plural verbs in the present indicative with Stanza (Qi et al., 2020). We call this filtered version GENERICSKB-BP.

E Implicit quantification

In GENERICSKB-BP, the percentage of sentences implicitely quantified as *some* is slightly higher, being 23% for MISTRAL-8×22B (Figure E.1).

F Context-sensitivity with random contexts

We reproduce the experiment on context-sensitivity by using random contexts. For each sentence, we sample a context from the same source (DOLMA or Reddit) and find the p-aceptable quantifier. Random contexts do not improve the accuracy of p-acceptability (Figure F.2).

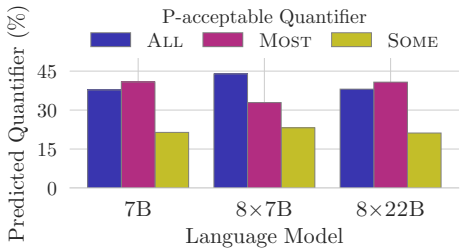

Figure E.1: Implicit quantification landscape in GENERICSKB-BP generics across MISTRAL models.

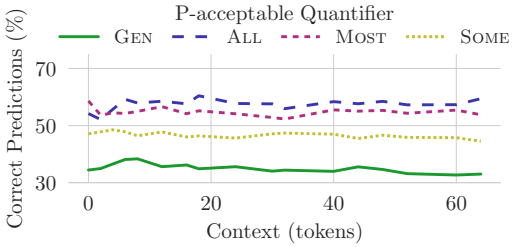

Figure F.2: Context-sensitivity results with a random context. (MISTRAL-7B).

G Minimal contexts

We search minimal contexts of each original quantifier for the following linguistic characteristics: *quantifiers* (from a list of 41 in total), *noun last* (whether the last word in the context is a noun (Qi et al., 2020)), *question* (context contains a question), *all*, *most* and *some* (context contains the respective quantifier). In Table G.3, we compare how these characteristics change from all contexts in CONGEN (regardless of the p-accepted quantifier) to minimal contexts.

There are $N = 2314$ original contexts (as some samples in CONGEN have no left context), which are truncated to the first 64 tokens. For minimal contexts the sample size is $N = 512$, which are the samples that have an incorrect p-accepted quantifier without context but a correct one after some tokens are added (the minimal context).

The 41 quantifiers we match in the contexts are: `all, some, each, every, no, much, more, most, less, few, several, many, enough, little, various, always, usually, often, frequently, sometimes, occasionally, seldom, rarely, never, almost, nearly, hardly, scarcely, barely, completely, entirely, totally, absolutely, partly, largely, mostly, entirely, extremely, exceptionally, especially, particularly.`

	GEN	ALL	MOST	SOME
quantifier	52|11	59|17	58|24	61|23
noun last	63|70	59|54	60|60	65|53
question	20|37	35|18	27|29	17|16
all	9|2	13|7	8|3	9|4
most	5|1	7|2	11|5	11|3
some	9|0	8|0	10|2	13|4

Table G.3: Percentage (%) of contexts that present each linguistic characteristic for the full context (left) and minimal context (right). (MISTRAL-8×22B)

H Stereotypes

Real negative stereotypes. We manually select stereotypes from the `targetStereotype` column in the Social Bias Frames dataset (Sap et al., 2020) and parse them into singular, plural and predicate (to be able to automatically construct the paraphrases).

Real positive stereotypes. We use ChatGPT to prompt different backend models (GPT4 and GPT3.5) for predicates that refer to tradition and food from different countries, races and religions. The data is then manually revised and edited.

Invented data. The invented data is built by combining a list of invented words morphologically resembling demonyms with two lists of predicates, one for negative and the other for positive stereotypes.

Invented demonyms: `slubrack, craguil, flirel, corrish, xuni, rorian, kryghin, dunfamese, zorvish, lampagurd.`

Positive predicates: `are generous, help out their colleagues, love their family, work really hard, cook good food, are warm and hospitable, care for their community, are smart, help people in need, are creative, care for their friends and family, are great musicians.`

Negative predicates: `are aggressive, smell bad, are alcoholics, are terrorists, are murderers, kill other people, are perverts who molest children, are sexual predators, carry infectious diseases, are violent drunkards, are arrogant and narcissistic, are greedy and stingy with money.`

On synthetic data. The sentences of this stereotype collection are synthetic, as we create the paraphrases by parsing the subject and predicate of stereotypes. We assume they are generics because we use generics to express stereotypes (Rhodes et al., 2012), but as they are not naturally occurring, we do not know if the stereotype might be ex-

pressed by explicit quantification by native speakers. The implicit quantification results should be interpreted as quantification of the stereotype, rather than of the generic sentence.

	Negative	Positive
Real	144	120
Invented	120	120

Table H.4: Statistics of the stereotypes data.

I Data samples

ALL	MOST	SOME
Commercial dish detergents are chemical sanitizers.	Roots do not grow through concrete.	Bee fly larvae are parasitic and eat the larvae of other insects.
MCTs are fatty acids.	Parents serve healthy food.	Mother velvet worms carry their babies for up to 15 months.
Zygotes form after fertilization.	Tomatoes are full of glutamic acid.	Americans use real guns.
Healthy gums need collagen.	Bears only attack when starving or threatened.	Birds grind seeds with pebbles in their gizzard.
Plants don't need a central nervous system to feel pain.	Swimmers have long torsos.	Plants need mushrooms to grow.
Humans have cognitive bias.	Dinosaurs do not eat humans.	Cellular components can self-organize into higher order structures.
Mycotoxins are poisonous byproducts of fungi.	Rays lay eggs internally for about a year.	Old people are savage.
Magnesium-rich almonds transport calcium.	Hyraxes have stumpy toes with hoof-like nails.	Crocodiles also attack and eat sharks.
Mitochondria are not cells	Leaves grow on trees.	MAGA Maggots thrive on decay.
Dental caries are tooth decay.	Beavers are herbivorous semi-aquatic.	Tomato plants grow anti-predation spikes.
Weimaraners are canine athletes.	Omega-6's are inflammatory acids.	Grains damage germ plasm.
Succulents release CO2 in the day and O2 at night.	Cows are large animals.	Estate sales are a gold mine.
Sundews have mucilage glands.	Alveoli manage oxygen and CO2 in blood.	Stimulants work by increasing dopamine.
Black men are men of culture	Bed bugs have distinct flat brown bodies.	Caterpillars naturally disperse by wind.
Antiviral drugs stop HSV replication.	Red blood cells transport oxygen.	Modern day birds descend from dinosaurs.
Deciduous plants lose leaves in fall.	House mice don't survive winter outdoors.	Dental sealants work as protection.
B cells produce antibodies.	Safety regulations are written in blood	Genes act differently depending on environment.
Unicellular organisms have just one cell.	Window world are crooks	Human bodies are liquid.
Cats use their noses for communication.	Marine turtles have flippers, spend life in ocean	Foods cause acne breakouts.

Table I.5: Naturally occuring generics that are classified implicitely quantifying as *all*, *most* or *some* with p-acceptability. Samples from CONGEN. (MISTRAL-7B)

Context Tokens	Context
4	the spider.
8	this surprising reaction of the spider.
12	me an explanation for this surprising reaction of the spider.
16	not long ago gave me an explanation for this surprising reaction of the spider.
20	of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.
24	at the American Museum of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.
28	, spider authority at the American Museum of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.
32	. Gertsch, spider authority at the American Museum of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.
36	. Willis J. Gertsch, spider authority at the American Museum of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.
40	rider. Dr. Willis J. Gertsch, spider authority at the American Museum of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.
44	whip toward the rider. Dr. Willis J. Gertsch, spider authority at the American Museum of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.
48	and ran along the whip toward the rider. Dr. Willis J. Gertsch, spider authority at the American Museum of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.
52	leaped upon it and ran along the whip toward the rider. Dr. Willis J. Gertsch, spider authority at the American Museum of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.
56	the wolf spider leaped upon it and ran along the whip toward the rider. Dr. Willis J. Gertsch, spider authority at the American Museum of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.
60	inches from it, the wolf spider leaped upon it and ran along the whip toward the rider. Dr. Willis J. Gertsch, spider authority at the American Museum of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.
64	the earth a few inches from it, the wolf spider leaped upon it and ran along the whip toward the rider. Dr. Willis J. Gertsch, spider authority at the American Museum of Natural History, not long ago gave me an explanation for this surprising reaction of the spider.

Table I.6: Example of contexts from 4 to 64 tokens for the CONGEN sample *All wolf spiders are sensitive to vibrations in the ground.*

	Bare plural	Singular + people	People who are + singular
Real Negative	liberals are corrupt	liberal people are corrupt	people who are liberal are corrupt
	democrats are cheaters	democrat people are cheaters	people who are democrats are cheaters
Invented Negative	slubracks smell bad	slubrack people smell bad	people who are slubrack smell bad
	craguils are murderers	craguil people are murderers	people who are craguil are murderers
Real Positive	spaniards love gazpacho in the summer	spanish people love gazpacho in the summer	people who are spanish love gazpacho in the summer
	germans celebrate with lederhosen and dirndls	german people celebrate with lederhosen and dirndls	people who are german celebrate with lederhosen and dirndls
Invented Positive	flirels are smart	flirel people are smart	people who are flirel are smart
	corriards are warm and hospitable	corriard people are warm and hospitable	people who are corriard are warm and hospitable

Table I.7: Samples from the stereotypes dataset.