Generics are not quantificational: A new path from language models to semantic theory

Gustavo Cilleruelo Calderón[1λ] **Mahrad Almotahari**[2λ]
Emily Allaway[1] **Barry Haddow**[1] **Alexandra Birch**[1]
[1]School of Informatics, University of Edinburgh
[2]School of Philosophy, Psychology and Language Sciences, University of Edinburgh
g.cilleruelo-calderon@sms.ed.ac.uk malmotah@ed.ac.uk
{emily.allaway, bhaddow, a.birch}@ed.ac.uk

Abstract

Generic sentences express generalizations that tolerate exceptions without *explicitly* communicating information about quantities. For example, the sentence *Ravens are black* is true even though there are albino ravens. The sentence doesn't explicitly communicate the number or frequency of black ravens. Whether generics semantically encode information about quantities *implicitly* is controversial. This work takes a large-scale distributional approach to the semantic debate. It compares thousands of naturally occurring generics and quantificational sentences using language-model probabilities. It shows that language models recover many semantic facts about quantifiers. It also shows that they recover semantic facts about surface distributional differences between generics and their "quantificational counterparts". Accordingly, and contrary to dominant views in other fields, we formulate an empirical argument to the effect that *generics are not quantificational*.[1]

1 Introduction

There are many different kinds of generalizations. Some are universal: *All ravens are black*. Some are existential: *There is a black raven*. Some are statistical: *Typically, ravens are black*. And some have a logical character that can't be traced back to an explicit quantifier: *Ravens are black*. Generics are sentences that convey exception-permitting category-wide generalizations of this last kind. And they're the subject of ongoing interdisciplinary debate. Linguists, psychologists, and philosophers of language continue to disagree about the proper way to semantically analyze their meaning (Carlson and Pelletier, 1995; Reuter et al., 2025; Neufeld et al., 2025).

One of the major fault lines in the debate is whether generics should be analyzed in quantificational terms. According to this view, which we call *Quantificationalism*, generics are basically synonymous with a quantificational sentence. For example, the meaning of *Ravens are black* is roughly the same as the meaning of *Usually, ravens are black* (Sterken, 2015; Nguyen, 2020; Lee and Nguyen, 2022) or *All normal ravens are black* (relative to some technical definition of normality) (Nickel, 2016). In contrast, the most influential *non*-Quantificationalist approach to generics maintains that they give voice to a default mode of generalizing information that reflects the quirkiness in our natural propensity to reason inductively (Leslie, 2007, 2008; Leslie et al., 2011).

The debate about Quantificationalism focuses on the analysis of a small set of paradigm examples. In this work, we bring a new body of evidence to bear on the discussion: the millions of generics in the MGEN dataset (Cilleruelo et al., 2025b), which, along with the contexts they inhabit, are drawn from real-world sources. Our large-scale computational approach identifies distributional patterns within a representative corpus of naturally occurring text. If generics are quantificational, we should expect their patterns to coincide with their explicitly quantificational counterparts.

Language models are effective tools for the detection and analysis of large-scale distributional patterns in language (Baroni, 2021; Grindrod, 2024; Boleda, 2025). Our investigation motivates their use in the debate about Quantificationalism by verifying that they recover many well-known facts about the semantics of quantifiers. The results strongly indicate that language models can provide new insight on the semantic relationship between quantity and genericity.

Our experiments contrast generics of the form *Ks are F* with their corresponding quantificational

[λ]Equal contribution and corresponding authors.
[1]Code and data available at gustavocilleruelo.com/not_quantificational.

Findings of the Association for Computational Linguistics: ACL 2026, pages 21875–21892

counterparts: QUANTIFIER⌢*Ks are F*. We use language models to derive probability distributions over post-verb tokens (those that occur within the predicate *F*) and measure the effect of interventions on the sentence-initial quantifier or generic noun phrase with KL-divergence. If, for a particular sentence and context, the addition of some quantifier preserves the meaning of the generic (as Quantificationalism maintains), then the probabilities of tokens in F shouldn't change much after that intervention. However, contrary to this prediction, our main empirical finding is that none of the 11 quantifiers we consider preserve the meaning on many of the generics we test. This motivates our core claim: *generics are not quantificational*.

The main contribution of this paper is thus an argument that challenges the dominant view in linguistics and the philosophy of language, which maintains that generic sentences are quantificational (§2.1). Our argument is empirically motivated by a novel large-scale distributional analysis of over $600,000$ sentences (§5.2) and theoretically grounded in both the concepts of information theory and the platitude that meaning is reflected in patterns of use (§2.5). Additional contributions are methodological: we build on previous work to derive facts about synonymy from language-model probability distributions by using KL-divergence (§4.1). With this method, we show that *language models reliably model quantifier meanings* (§5.1). Finally, we identify major distributional differences between generics and their quantificational counterparts. We package these results in the form of a new argument challenging the empirical adequacy of Quantificationalism (§6).

2 Background

2.1 Generics

Generics can appear in different shapes and sizes. We restrict our attention to sentences of the form *Ks are F*.[2] Sentences of this form express kind-wide generalizations that often ascribe a characteristic property (*Ravens are black*), a distinctive function (*Toasters toast bread*), or a stereotypical feature (*Bachelors are slobs*). They do so without giving explicit quantitative information. Although many aspects of generic generality deserve attention, such as its resilience to exceptions or its role in social-group marginalization (Leslie, 2017; Allaway et al., 2023), we focus on their relation to quantification. There seems to be a consensus among specialists that generics typically express some kind of quantity akin to most (majority) or many (prevalence); the main source of disagreement seems to concern how to refine this core idea (Carlson, 1977b; Cohen, 1999a; Carlson and Pelletier, 1995).

Quantificationalists interpret *Ks are F* in terms of a contextually relevant quantitative relation between the set of *K*s and the set of things that are *F*. Influential accounts of this sort include Cohen (1999a), which analyzes generics in terms of probability; Sterken (2015), which posits an unpronounced indexical operator that takes distinct quantificational meanings as values (universality, typicality, etc.) depending on the context; Nickel (2008, 2016), which argues that generics quantify over normal sets of individuals; and many others (e.g., Tessler and Goodman, 2016; Bosse, 2021; Kirkpatrick, 2024; Hoorens et al., 2026).

The most influential theory of generics that doesn't rely on quantification is the one due to Leslie (2007, 2008) and Leslie et al. (2011). This account takes *Ks are F* to be the language faculty's way of articulating the default generalizations of a primitive cognitive propensity to make inductive inferences. Other approaches that don't rely on quantificational meaning analyze generics in terms of reference to kinds (Carlson, 1977b; Liebesman, 2011; Teichman, 2023) or as defeasible permissions to engage in certain kinds of inference (Stovall, 2019). These theories have one thing in common: they don't define truth or acceptability conditions for generics in terms a quantitative relationship between the set of *K*s and the set of things that are *F*, as Quantificationalists do.

Overall, the semantic debate about generics is wide open, with recent works arguing both for (Neufeld et al., 2025) and against (Almotahari, 2026) Quantificationalism. Discussion often centers on a small number of paradigmatic examples, analyzed with the tools of formal semantics or experimental psychology (Hermans et al., 2026). Here, we bring another method to bear on the debate, one that takes advantage of language models as a powerful tool for identifying large-scale distributional patterns.

[2] Following the convention in the literature, the verb *to be* is just for readability. Bare-plural generics with some other verb are understood to be instances of this schema.

2.2 Quantifiers

Maybe the most influential theory of quantification is *Generalized Quantifier Theory*, which maintains that quantifiers denote relations between sets (Barwise and Cooper, 1981). For example, *All Ks are F* is true if and only if the set of *Ks* is a subset of the set of things that are *F* ($K \subseteq F$) and *Some Ks are F* is true if and only if the intersection between the two sets is non-empty ($K \cap F \neq \emptyset$). Other quantifiers, such as *most*, *few*, or *many*, would depend on some contextually determined threshold (e.g., *Many Ks are F* would be true if and only if $|K \cap F| > t_c$).

More recent theories take into account pragmatic effects that can occur when quantifiers interact. For example, in cases where both *some* and *all* would be true, speakers use the latter to be more informative (Grodner et al., 2010; Lassiter and Goodman, 2017).

The quantifiers we consider in this work are classified according to their strength as follows: universality (*all*), majority (*most*, *generally*, *typically*, *usually*, *normally*), prevalence (the majority quantifiers plus *often* and *many*), minority or weak quantification (*some*, *few*) and universal negation (*no*).

2.3 The MGEN dataset

The MGEN dataset (Cilleruelo et al., 2025b) is a massive collection of over 4.1 million naturally occurring generic and quantificational sentences, extracted from the ZYDA dataset (Tokpanov et al., 2024). The contexts for these sentences are long—averaging over 5000 words—and they appear out in the wild: on websites and in academic papers.

In MGEN, generics always have a bare-plural noun in subject position, which can be followed by a predicate that's either simple or complex. Six quantifiers (*all*, *most*, *many*, *some*, *few*, *no*) always appear as the first word in the sentence. Another five (*often*, *generally*, *typically*, *usually*, *normally*) can be the first word of the sentence or appear alongside the main verb.

Our experiments use a subset of MGEN (Table 1) filtered so the key sentence occurs roughly in the middle of the text (with at least five context sentences at either side). To filter out long clauses, we take sentences with between three and 60 words. Finally, all of the adverbial quantifiers we tested appear as the first word in the sentence. Find examples from MGEN in Appendix C.

Quantifier	Sentences	Quantifier	Sentences
Generic	499,653	No	2,511
All	37,569	Often	5,462
Most	36,963	Generally	7,053
Many	36,850	Typically	5,706
Some	37,014	Usually	3,983
Few	4,974	Normally	1,499

Table 1: Sentence counts in our split of MGEN.

2.4 Language models

Auto-regressive language models are the state-of-the-art probabilistic models of language. Our experimental results were obtained with OLMO3-32B (Olmo et al., 2025), a competitive fully-open language model, but we also reproduce our findings in a wide array of open-weight pretrained language models: MISTRAL-7B (Jiang et al., 2023), LLAMA3.1-8B (Grattafiori et al., 2024), QWEN3-8B (Yang et al., 2025), GEMMA-7B (Team et al., 2024) and MIXTRAL-8×22B (Jiang et al., 2024).

2.5 The correlation principle

Bringing language models and information theory to bear on the debate over Quantificationalism requires an assumption about the relation between linguistic meaning and the distribution of text in a corpus. This assumption posits a default connection between similarity of distribution within a large enough corpus and sameness of meaning: all else equal, similarity of distribution correlates with sameness of meaning (Harris, 1982; Boleda, 2020).

There are stronger principles in the vicinity: that similarity of distribution *entails* sameness of meaning, or that the distribution of a term *just is* its meaning. The correlation we rely on is compatible with representationalist theories that foreground the notions of reference and truth. From the representationalist's perspective, the correlation can be defeated by the presence of opacity-inducing elements (paradigmatically, quotation marks and attitude verbs), shifts in syntax, discourse-level effects, and so on. For discussion, see Harris (1991). Grindrod (2023) provides half a dozen considerations in favor of the correlation principle.

3 Related work

The information-theoretic properties (Shannon, 1948; Mudireddy et al., 2025) of token probability distributions derived from language models are the starting point of new and exciting investigations on language. Several recent studies apply surprisal to investigate the structure of human communication, finding hierarchical (Tsipidi et al., 2024), harmonic (Tsipidi et al., 2025), and fractal (Alabdulmohsin et al., 2024) patterns in naturally occurring texts. These investigations are often motivated by a broader framework, namely, surprisal theory (Hale, 2001; Levy, 2008; Giulianelli et al., 2023, 2024; Staub, 2025).

On the subject of generics, quantification, and language models, recent work uses theoretical insights to probe, benchmark, and evaluate model performance, often through prompting (Madusanka et al., 2023; Allaway et al., 2024; Allaway and McKeown, 2025; Kirkpatrick and Sterken, 2025). Gupta (2023) uses surprisal on synthetic sentences to evaluate whether language models reliably represent quantifier-meanings. Cilleruelo et al. (2025a) also use a surprisal-based method to compare quantifiers and generics, but do so on a small manually annotated set of sentences (less than 3000), only for three quantifiers and don't discuss the implications of their work for the wide range of views in linguistics and philosophy of language.

Some recent works argue that modern language models don't reliably model quantifier meanings. (Collacciani et al., 2024; Enyan et al., 2024; Montero et al., 2025). However, they fail to consider the crucial role of context in quantification. Context sensitivity is central to linguistic theories of quantification (Lewis, 1975; Barwise and Cooper, 1981; Lappin, 2000; Lassiter and Goodman, 2017) and has been empirically verified in cognitive science (Urbach et al., 2015; Heim et al., 2015; Macuch Silva et al., 2024). Nevertheless, these previous works evaluate language models mostly on single sentences without any context or with minimal context, and they use synthetic data, which may not be in-distribution for the language model. In contrast, we experiment on real-world sentences, extracted from the internet and scientific publications, along with three left-side context sentences. In §5.1 we show that semantic facts about quantifiers can be reliably extracted from the language models we use, thus vindicating our method and its application to Quantificationalism.

4 Methodology

This work presents an empirical argument about the meaning of a particular linguistic phenomenon through a distributional analysis enabled by an autoregressive language model. Since there's no firmly established methodological framework for this kind of argument, we will first briefly discuss existing approaches, then present our improved method based on KL-divergence (§4.1), and finally detail the experimental setup (§4.2). We use the term *operator* for both quantifiers and quantifier-free generic noun phrases.

4.1 Meaning in models of language

Computational approaches to the study of language fundamentally changed with the introduction of systems that represent words (or tokens) with dense vectors, notably *word2vec* (Mikolov et al., 2013). They changed again with the first transformer-based bi-directional text encoders, such as BERT (Devlin et al., 2018). Much of this literature focused on vector representations of words, reproducing semantic phenomena by operating on them (Rogers et al., 2020; Grindrod and Grindrod, 2025). Adapting these methods to modern autoregressive language models isn't straightforward, as their next-token prediction objective makes it less clear what should be the representation of a token, and their autoregressive nature means those representations don't have information from their right context. This is exacerbated in the case of generics because there's no overt genericity-expressing token to be represented.

To overcome these issues, thereby making the use of autoregressive language models a sound strategy, researchers are shifting the focus from how a word is represented in an embedding space to how interventions on that position affect the token probability distributions in subsequent tokens of the sentence. This idea has been operationalized by looking at how interventions on a given token affect the surprisal of tokens that follow.[3] For example, Cilleruelo et al. (2025a) and Gupta (2023) modify the quantifier and measure surprisal of the predicate and Liu et al. (2025) studies discourse

[3] Given a language model θ and a context c, the *surprisal* (or *self-information*) of a token t is the negative log-probability θ assigns to that token: $-\log p_\theta(t|c)$.

connectives by averaging the surprisal of text after the expression (i.e., *otherwise*).

Although these prior works successfully recover semantic facts about the phenomena they study, we identify a weakness in their approach. The tail of the token probability distribution of an autoregressive language model is often driven by spurious and noisy correlations on the data (Holtzman et al., 2020). When comparing surprisal, if the tokens in the sentence are assigned a low-probability by the language model, the comparison ceases to be meaningful.

In this work, we modify sentences like QUANTIFIER⌢*Ks*⌢VERB⌢*F* by changing or removing the quantifier and measuring which modifications are synonymous with the original sentence. To overcome the previous drawback, we propose that two quantifiers are synonymous for a given sentence if the maximum Kullback-Leibler divergence in all post-verb tokens is bounded by a threshold t. Given two discrete probability distributions, p and q, their KL-divergence is

$$D(p||q) = \sum_{x \in \mathcal{V}} p(x) \log \frac{p(x)}{q(x)}. \quad (1)$$

This measures how much information is lost by assuming q given that the true distribution is p. Let p_i be the probability distribution over tokens at position i in the original sentence and let q_i be the probability distribution when the word at the quantifier position has been modified. Then:

$$\max_{i \in \mathcal{P}} D(p_i||q_i) < \tau, \quad (2)$$

where the members of $\mathcal{P}$ are the positions occupied by the post-verb tokens. Two expressions will be close (relative to τ) only when every single KL-divergence on the post-verb tokens is below the threshold. If this threshold is low enough, then modifying the sentence-initial phrase doesn't affect token probability distributions later on. Equivalently, *the language model treats pre- and post-modification phrases as interchangeable*. Because we are taking a maximum, notable differences in a single token will be enough to separate interventions. More importantly, KL-divergence is asymmetrical: it compares the probability distributions on the modified sentence q over tokens that are high probability in the same position in the original sentence p. Even if the token is itself low-probability given the language model, the comparison is made over the high-probability tokens at that position in the original sentence (not only the actual token itself, as with surprisal).

4.2 Experimental setup

For each sentence in the subset of MGEN that we test (Table 1), we construct modified counterparts by adding or subtracting a quantifier for each of the other 11 operators considered (*generic*, *all*, *most*, *many*, *some*, *few*, *no*, *often*, *generally*, *typically*, *usually*, *normally*). For each pair consisting of the original sentence and a modified counterpart, we compute the KL-divergence at each token following the verb,[4] with three additional sentences providing subsequent context. We incoroporate this additional context because it's well known that both generics and quantificational sentences can be highly context sensitive (Sterken, 2015; Almotahari, 2024). We then take the *maximum* KL-divergence and, for a threshold τ, measure whether the operator in the modified sentence is synonymous with that of the original, that is, the new operator doesn't disturb the post-verb token distributions beyond τ. There is no obvious criterion for selecting the value of τ. We set $\tau = 0.15$, as that surfaces many of the semantic features associated with quantifiers, but we also present results across thresholds in Figure 2.

5 Results

In this section, we argue that our method, applied to current language models, successfully recovers a wide range of semantic facts and intuitions (§5.1). Then, we show how generics differ from their quantificational counterparts (§5.2). Results are similar across models. For a comparison, see Appendix B.

5.1 Quantifier semantics in language models

Language models are good at modeling quantifier meanings. We'll argue for this as a precursor to arguing for our central claim in Section 5.2.

Figure 1 plots the proportion of sentences for which each alternative operator is synonymous to the original. We interpret higher percentages as relative closeness between operator meanings — how often they're interchangeable. We now list many properties derived from this understanding of similarity that are predicted by works in theoretical linguistics (§2.2).

[4] We align words and tokens and find the first token after the last token corresponding to the verb.

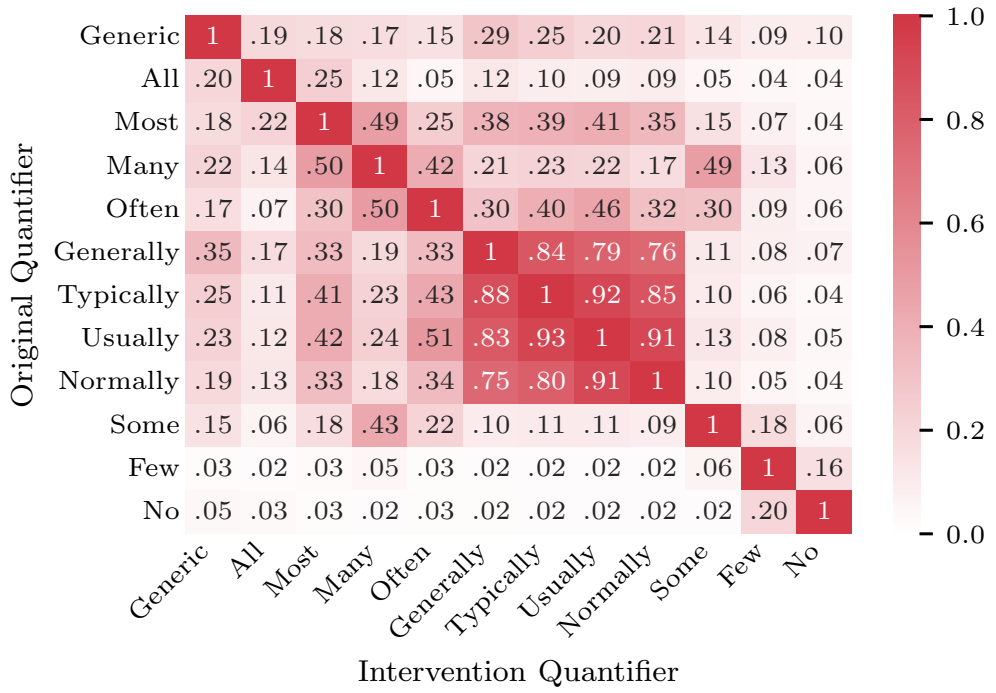

Figure 1: Proportion of sentences with original operator q that, after substituting q with another operator q', yield a KL-divergence on every post-verb token below $\tau = 0.15$ (OLMO3-32B).

Although it's semantically different from the rest of the operators we consider, *all* can sometimes be interpreted as *most* (Hallman, 2016). Similarly, universals can be interpreted as generics, as seen in the generic overgeneralization effect (Leslie et al., 2011). In agreement with these expectations, we see that *most* and *generic* are the only two operators that are interchangeable with *all* in over 20% of sentences.

Both *most* and *many* indicate a large number of instances. And, unsurprisingly, we see that they're interchangeable around 50% of the time. But *most* implies a majority of instances and we observe that it has more overlap with the adverbial quantifiers ending in *-ally*, which also imply a majority. In contrast, *many* has more overlap than *most* with *some* and *often*, neither of which imply majority.

Adverbs ending in *-ally* seem to express roughly the same quantificational force (majority), and our results reflect this quite strongly. Of the adverbs considered, *often* and *usually* are the two with a more temporal flavor (Lewis, 1975), and we observe that *often* is more interchangeable with *usually* than with the other *-ally* adverbs (and vice-versa).

The quantificational strength of *no* and *few* differs markedly from the other operators considered. This difference is reflected in the plot: they're seldom interchangeable with any other operator, with the noticeable exception of *some* which shares the minority flavor with *few*.

We interpret the recovery of these semantic facts as evidence that *language models provide insight into quantifier meanings by virtue of their sensitivity to large-scale distributional patterns*.

5.2 Generics and quantifiers

In the previous section we argued that language models recover a variety of semantic facts about quantifiers. Now we shift our attention to generic sentences.

Because generics are often thought to convey majority or prevalence (§2.1), we would expect that, in many cases, the insertion of *most*, *many*, *often*, *generally*, *typically*, *usually*, or *normally* won't significantly change their meaning. In cases where they have a universal flavor, the insertion of *all* should keep things roughly the same. And in cases where they have weak quantificational strength (Almotahari, 2022), the insertion of *some* shouldn't be terribly disruptive semantically.

Our results show that while generics are synonymous with every prevalence quantifier around

20% of the time, there is no single quantifier synonymous to the generic for more than 30% of sentences. This is a first distributional contrast between generics and the rest of the prevalence quantifiers considered: the generic is not similar to any one of them at the same rate that they are to each other. Note how, for each prevalence quantifier, there is always another for which they are synonymous roughly 50% of the time (*most* with *many*; *many* with *most* and *some*; *often* with *many*; the *-ally* adverbs with each other).

Now we consider the percentage of sentences that, for every given threshold τ, have *no* synonymous quantifier: they express a quantificational force unique with respect to the others (Figure 2). We comment on $\tau = 0.15$, as that corresponds with Figure 1.

The ranking in terms of percentage of sentences that have no synonyms in Figure 2 gives a notion of how unique each quantifier is. As we noted before, *no* and *few* have a very distinct meaning from the rest, and indeed 75% of those sentences are unique. On the other extreme, the quantificational adverbs are very similar to each other, and as we increase the threshold, they rapidly go to no unique sentences: they are always synonymous with each other or another operator. More interesting is the case of *all* and *some*, which in GQT denote distinct set relationships, and have unique meanings for over 50% of the evaluated sentences.

Generic sentences are unique at a rate similar to those containing *some* and *all*. This result is striking because Quantificationalists predict that at least one of the quantifiers considered should preserve the meaning of any given generic sentence (Carlson and Pelletier, 1995; Cohen, 1999a; Sterken, 2015; Nguyen, 2020). In contrast, our results empirically show a majority (58%) of generic sentences for which this is not true.

In conclusion, we observe two major distributional differences between generics and the quantifiers that are commonly thought to preserve their meaning: (*i*) generics are less similar in general to the rest of the quantifiers (while still being sometimes synonymous to every one of them), and (*ii*) for a majority of generic sentences we find no synonymous quantifier.

6 Discussion

In this section, we'll discuss the theoretical upshot of our experimental results.

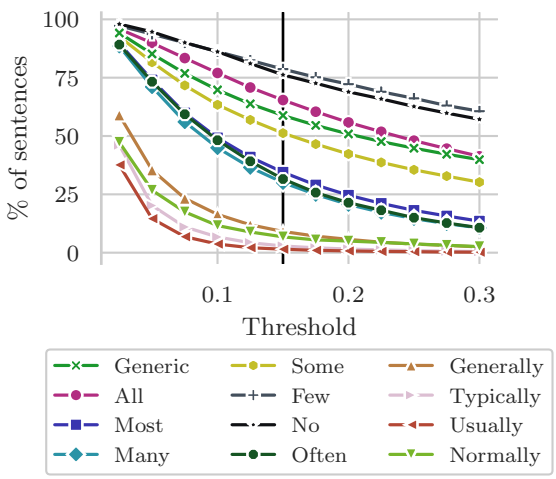

Figure 2: Percentage of sentences with no synonymous operator, that is, for which no other operator has a KL-divergence on every post-verb token below the given threshold τ (OLMO3-32B). The black vertical line marks $\tau = 0.15$.

Given §5.1, there's good reason to think that language models successfully model the meanings of quantifiers. So, if generics are semantically quantificational, language models should reflect a corresponding distributional similarity. Given §5.2, the model of quantifier meanings we obtain from OLMO3-32B enables us to formulate our central argument:

I. There's a significant difference in the pattern of distribution within the MGEN dataset between generics and their quantificational counterparts (significant enough, at least, to make for massive KL-divergence relative to a threshold with respect to which roughly synonymous quantifiers converge; see Figure 2).

II. MGEN is sufficiently representative to have evidential value.

III. Therefore, we have defeasible but non-negligible reason to think there's an equally significant semantic difference between generics and their quantificational counterparts.

If Quantificationalism were true, and sameness of meaning were correlated with similarity of distribution within a representative corpus, then we should expect the curve for generic expressions in Figure 2 to cluster with the curves for quantifiers that approximate genericity. After all, Quantificationalists maintain that, in context, at least one of

the nine prevalence-expressing quantifiers should have a similar meaning.[5] But, contrary to this expectation, the curve for generics is actually closer to the curve for *no* and *few*.

For about half a million generics, we find almost as little congruence with the prevalence quantifiers as there is between *no* and *few*. But quantifier phrases that express prevalence are supposed to be roughly synonymous with generic nouns. Nevertheless, they consistently fail to elicit similar token predictions for naturally occurring generic sentences.

Perhaps the most straightforward way to meet the challenge would be to retreat to a weaker form of Quantificationalism, according to which generics are semantically quantificational in context, they're just not similar in meaning to any *lexicalized* quantifier. But this would be a bad idea; it would surrender the methodological ground for holding the view. In effect, it would concede that generics are semantically *sui generis*. If generics are semantically distinctive, it would be better to view their distinctiveness in a way that makes it predictable from a broader theoretical perspective. Leslie (2007, 2008) and Leslie et al. (2011) provide one such perspective. In line with their framework, our results indicate that generics aren't quantificational.

Admittedly, the principle that sameness of meaning correlates with sameness of distribution is defeasible; it holds generally but not without exception. It is, after all, a generic. There may be room here for Quantificationalists to maneuver. They might point to a potential defeater and claim that, with respect to the sentences we consider, the presumptive correlation is undermined. But it's not clear to us what this story would look like in detail. And we remain confident that the sheer volume of data tested goes a considerable way toward circumventing this worry. In any case, nothing we've said rule out the possibility once and for all, which is why we present our central argument as a *challenge* for Quantificationalism rather than a *refutation*.

Our central argument, (I)-(III), suggests a method for advancing other debates in the science of language. For example, consider the dispute about definite descriptions: are they quantifiers (Russell, 1905; Neale, 1990) or singular referring terms (Frege, 1892; Strawson, 1950; Elbourne, 2013)? Perhaps they have two basic kinds of use, one quantificational and one referential (Donnellan, 1966). A model-based distribution argument akin to (I)-(III) might shed new light on the issue.

[5] Nickel (2016) is the exception. Discussing the matter in detail would unnecessarily complicate our presentation. We hope to address the issue elsewhere.

7 Conclusion

This paper argues that *generics are not quantificational* from a novel empirical perspective: by analyzing token probability distributions extracted from language models on hundreds of thousands of sentences. Our argument is relevant to linguists, philosophers of language, and psychologists. Although its upshot is negative, the method we employed to frame it is constructive. It has the potential to advance other interdisciplinary debates.

Limitations

Information measures. While we compare sentences only using KL-divergence, there is a family of information-theoretic measures that could inform these comparisons, with different connections to aspects of human language processing Giulianelli et al. (2024).

Language models. We run our experiments on language models with a similar autoregressive architecture. Different aspects of the language model, such as training regime, architecture, or parameter size, could potentially affect the results.

Oral communication. The generics we use for the experiments come from written sources, such as websites or academic publications. We leave an exploration of the use of generics in spoken language for future work.

Multilinguality. The present study shows results only on English, as multilingual data collection was beyond the current scope. For some non-English studies of generics, see Castroviejo et al. (2022) and Lazaridou-Chatzigoga et al. (2019).

Acknowledgments

We would like to acknowledge the anonymous reviewers of the October 2025 and January 2026 ARR cycles. Some of their insightful comments really helped shape the current version of this

work. We also thank Alison Chi for her annotation effort, Domenec Miralles Tagliabue for the color palette, and Aina Centelles Tarrés for proof reading. GCC and MA would like to thank Annie Bosse, Jumbly Grindrod, Bernhard Nickel, and Rachel Sterken for reading an early draft and providing written comments. Thanks also to Dan Lassiter, Sarah-Jane Leslie, Nicolas Navarre, and Guillem Ramírez Santos for insightful discussions.

References

Ibrahim Alabdulmohsin, Vinh Q. Tran, and Mostafa Dehghani. 2024. Fractal patterns may illuminate the success of next-token prediction. In Advances in Neural Information Processing Systems, volume 37, pages 112864–112888. Curran Associates, Inc.

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 and Kathleen McKeown. 2025. Evaluating defeasible reasoning in LLMs with DEFREASING. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 10540–10558, Albuquerque, New Mexico. Association for Computational Linguistics.

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

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

Mahrad Almotahari. 2024. Generic cognition: A neglected source of context sensitivity. Mind & Language, 39(4):472–491. First published 24 January 2024.

Mahrad Almotahari. 2026. The cognitive theory of generics.

Marco Baroni. 2021. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. Linguistics and Philosophy, 4:159–219.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. Annual Review of Linguistics, 6(1):213–234.

Gemma Boleda. 2025. Llms as a synthesis between symbolic and continuous approaches to language. Preprint, arXiv:2502.11856.

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

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.

Elena Castroviejo, José V. Hernández-Conde, Dimitra Lazaridou-Chatzigoga, Marta Ponciano, and Agustín Vicente. 2022. Are generics defaults? a study on the interpretation of generics and universals in 3 age-groups of spanish-speaking individuals. Language Learning and Development, 19(3):275–302.

G. Cilleruelo, Emily Allaway, Barry Haddow, and Alexandra Birch. 2025a. Generics are puzzling. can language models find the missing piece? In Proceedings of the 31st International Conference on Computational Linguistics, pages 6571–6588, Abu Dhabi, UAE. Association for Computational Linguistics.

G. Cilleruelo, Emily Allaway, Barry Haddow, and Alexandra Birch. 2025b. Mgen: Millions of naturally occurring generics in context. Proceedings of the Society for Computation in Linguistics, 8(1):11.

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

Ariel Cohen. 1999b. Think Generic!: The Meaning and Use of Generic Sentences. CSLI, Stanford.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.

Keith S. Donnellan. 1966. Reference and definite descriptions. The Philosophical Review, 75(3):281–304.

Paul Elbourne. 2013. Definite Descriptions. Oxford University Press.

Zhang Enyan, Zewei Wang, Michael A. Lepori, Ellie Pavlick, and Helena Aparicio. 2024. Are llms models of distributional semantics? a case study on quantifiers. Preprint, arXiv:2410.13984.

Gottlob Frege. 1892. Über sinn und bedeutung. Zeitschrift für Philosophie Und Philosophische Kritik, 100(1):25–50.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024. Generalized measures of anticipation and responsivity in online language processing. Preprint, arXiv:2409.10728.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. Information value: Measuring utterance predictability as distance from plausible alternatives. Preprint, arXiv:2310.13676.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. Preprint, arXiv:2407.21783.

Jumbly Grindrod. 2023. Distributional Theories of Meaning, pages 75–99. Springer International Publishing, Cham.

Jumbly Grindrod. 2024. Modelling language.

Jumbly Grindrod and Peter Grindrod. 2025. Word meanings in transformer language models. Preprint, arXiv:2508.12863.

Daniel J. Grodner, Natalie M. Klein, Kathleen M. Carbary, and Michael K. Tanenhaus. 2010. "some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. Cognition, 116(1):42–55.

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.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In Second Meeting of the North American Chapter of the Association for Computational Linguistics.

Peter Hallman. 2016. "all" and "every" as quantity superlatives. Semantics and Linguistic Theory, 26:506.

Zellig Harris. 1982. A Grammar of English on Mathematical Principles.

Zellig S. Harris. 1991. A Theory of Language and Information: A Mathematical Approach. Columbia University Press.

Stefan Heim, Corey T. McMillan, Robin Clark, Edward S. Golob, Nam E. Min, Christopher Olm, James Powers, and Murray Grossman. 2015. If so many are "few," how few are "many"? Frontiers in Psychology, 6:441.

Felix Hermans, Walter Schaeken, Susanne Bruckmüller, and Vera Hoorens. 2026. When do generics feel justifiable? a registered report bridging key theories. Journal of Cognition, 9(1):21.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. Preprint, arXiv:1904.09751.

Vera Hoorens, Felix Hermans, and Susanne Bruckmüller. 2026. Why boys cry and don't cry: The contextual-statistical (constat) approach to the perceived validity of generics. Cognition, 266:106323.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. Computing in Science & Engineering, 9(3):90–95.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. Preprint, arXiv:2401.04088.

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

James Ravi Kirkpatrick and Rachel Katharine Sterken. 2025. Generics and default reasoning in large language models. Preprint, arXiv:2508.13718.

Shalom Lappin. 2000. An intensional parametric semantics for vague quantifiers. Linguistics and Philosophy, 23(6):599–620.

Daniel Lassiter and Noah D. Goodman. 2017. Adjectival vagueness in a bayesian model of interpretation. Synthese, 194(9):3801–3836.

Despina Lazaridou-Chatzigoga, Napoleon Katsos, and Linnaea Stockall. 2019. Experimental evidence on genericity and universal quantification in greek and english. In 13 (Conference on Greek Linguistics), page 171.

Junhyo Lee and Anthony Nguyen. 2022. What?s positive and negative about generics: A constrained indexical approach. Philosophical Studies, 179(5):1739–1761.

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.

Roger Levy. 2008. Expectation-based syntactic comprehension. Cognition, 106(3):1126–1177.

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

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

Guifu Liu, Bonnie Webber, and Hannah Rohde. 2025. "otherwise" in context: Exploring discourse functions with language models. In Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 81–95, Suzhou, China. Association for Computational Linguistics.

Vinicius Macuch Silva, Alexandra Lorson, Michael Franke, Chris Cummins, and Bodo Winter. 2024. Strategic use of english quantifiers in the reporting of quantitative information. Discourse Processes, 61(10):498–523.

Tharindu Madusanka, Iqra Zahid, Hao Li, Ian Pratt-Hartmann, and Riza Batista-Navarro. 2023. Not all quantifiers are equal: Probing transformer-based language models' understanding of generalised quantifiers. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 8680–8692. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Preprint, arXiv:1301.3781.

Raquel Montero, Natalia Moskvina, Paolo Morosi, Tamara Serrano, Elena Pagliarini, and Evelina Leivada. 2025. Quantification and object perception in multimodal large language models deviate from human linguistic cognition. Preprint, arXiv:2511.08126.

Avinash Mudireddy, Tyler Bell, and Raghu Mudumbai. 2025. Slaves to the law of large numbers: An asymptotic equipartition property for perplexity in generative language models. Preprint, arXiv:2405.13798.

Stephen Neale. 1990. Descriptions. MIT Press.

Eleonore Neufeld, Annie Bosse, Guillermo Del Pinal, and Rachel Sterken. 2025. 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. 2008. Generics and the ways of normality. Linguistics and Philosophy, 31(6):629–648.

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

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2025. Olmo 3. Preprint, arXiv:2512.13961.

Kevin Reuter, Eleonore Neufeld, and Guillermo Del Pinal. 2025. Generics and quantified generalizations: Asymmetry effects and strategic communicators. Cognition, 256:106004.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. Preprint, arXiv:2002.12327.

Bertrand Russell. 1905. On denoting. Mind, 14(56):479–493.

Claude Elwood Shannon. 1948. A mathematical theory of communication. The Bell System Technical Journal, 27:379–423.

Adrian Staub. 2025. Predictability in language comprehension: Prospects and problems for surprisal. Annual Review of Linguistics, 11:17–34. First published as a Review in Advance on July 15, 2024.

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

Preston Stovall. 2019. Characterizing generics are material inference tickets: A proof-theoretic analysis. Inquiry: An Interdisciplinary Journal of Philosophy.

Peter F. Strawson. 1950. On referring. Mind, 59(235):320–344.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. Preprint, arXiv:2403.08295.

Matt Teichman. 2023. The sophisticated kind theory. Inquiry, 66(9):1613–1654.

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

Yury Tokpanov, Beren Millidge, Paolo Glorioso, Jonathan Pilault, Adam Ibrahim, James Whittington, and Quentin Anthony. 2024. Zyda: A 1.3t dataset for open language modeling. Preprint, arXiv:2406.01981.

Eleftheria Tsipidi, Samuel Kiegeland, Franz Nowak, Tianyang Xu, Ethan Wilcox, Alex Warstadt, Ryan Cotterell, and Mario Giulianelli. 2025. The harmonic structure of information contours. Preprint, arXiv:2506.03902.

Eleftheria Tsipidi, Franz Nowak, Ryan Cotterell, Ethan Wilcox, Mario Giulianelli, and Alex Warstadt. 2024. Surprise! Uniform Information Density isn't the whole story: Predicting surprisal contours in long-form discourse. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18820–18836, Miami, Florida, USA. Association for Computational Linguistics.

Tara P. Urbach, Katherine A. DeLong, and Marta Kutas. 2015. Quantifiers are incrementally interpreted in context, more than less. Journal of Memory and Language, 83:79–96.

Michael L. Waskom. 2021. seaborn: statistical data visualization. Journal of Open Source Software, 6(60):3021.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. CoRR, abs/1910.03771.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. Preprint, arXiv:2505.09388.

A Supplementary Experiments

This section of the appendix includes two experiments that point in the same direction as the main paper, for the reader that wants to get a more complete feel for the comparison between generics and quantified sentences.

A.1 Binary classification on generics and quantificational sentences

This section briefly reports the result of a preliminary experiment that also investigates the relationship between genericity and quantity.

If the meaning of a generic comes close, in context, to a counterpart quantificational sentence, then we would expect a binary classifier to struggle in predicting whether it's a generic or the quantificational counterpart. On the other hand, for quantifiers that don't come close in meaning (e.g., *no*), a binary classifier should be able to separate them from a generic sentence just based on their contents.

Setup. In this experiment, we take a balanced sample from MGEN, with the same number of generic and quantificational sentences. And, within the category of quantificational sentences, the 11 different quantifiers are similarly represented. For quantificational sentences, we remove the quantifier, thus all the sentences of the dataset have the same syntactic structure: roughly, *Ks are F*.

The final dataset has $45,331$ samples for each label, GEN and QUANT. We have an $85/15$ train/test split. We don't include validation set because the objective of this experiment is to qualitatively motivate that generics could be non-quantificational by virtue of their discriminability based on content. We make a single run of the experiment with standard hyperparameters, rather than fine-tuning for better performance.

Hyperparameter	Value
Model Name	`roberta-base`
Max Length	128
Batch Size	256
Learning Rate	$2e-5$
Number of Epochs	3
Warmup Steps	500
Weight Decay	0.01

Table A.1: Hyperparameters for training run

Quantifier	N train	N test	Accuracy
GEN	42034	7418	0.79
QUANT	38531	6800	0.77
Binary accuracies by original quantifier:			
FEW	3481	640	0.93
NO	3482	639	0.88
MANY	3519	602	0.82
SOME	3498	623	0.81
MOST	3516	605	0.80
TYPICALLY	3518	603	0.75
ALL	3505	616	0.72
USUALLY	3495	626	0.72
OFTEN	3512	609	0.71
NORMALLY	3519	602	0.69
GENERALLY	3486	635	0.63

Table A.2: Quantifier data summary

We train a ROBERTA classifier with a binary classification head on a single training run with standard hyperparameters (see Table A.1) and report its accuracy in Table A.2. As baseline, a random classifier would achieve 0.5 accuracy in the task. For each of the 11 quantifiers, we also report the binary accuracy on the sentences originally containing that quantifier, that is, the percentage of those that were classified as QUANT.

Results. This experiment suggests that adverbial quantifiers are closer to generics in meaning than quantificational determiners and that the studied quantifiers have, at most, a limited overlap in meaning with the generic.

High accuracies correspond to quantifiers that are used in sentences easily distinguishable from generics, even if the quantifier itself is omitted. Indeed, *few* and *no* have the higher accuracies, as the quantificational forces they convey can't be expressed by a generic (Lee and Nguyen, 2022).

Interestingly, quantificational determiners that express a notion of prevalence, like *many* and *most*, have accuracies of over 0.80: in most cases they can be distinguished from a generic just by their content. We find this interesting because genericity is often identified with prevalence of a specific kind (Carlson and Pelletier, 1995; Cohen, 1999a). For quantificational determiners, *all* has the lowest accuracy. Previous work also finds that *all* (as opposed to *most* and *some*) is the most appropriate candidate to describe the implicit quantificational force in many generics (Cilleruelo et al., 2025a).

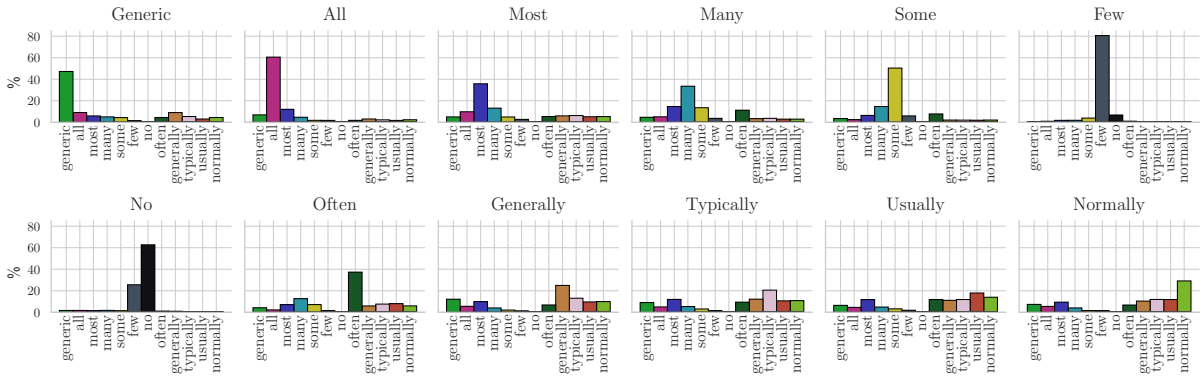

Figure A.1: Percentage of sentence-initial terms selected as p-acceptable for each original phrase (MISTRAL-7B).

A.2 Scaling up the surprisal experiments in Cilleruelo et al. (2025a)

This experiment indicates that generics and their quantificational counterparts differ in meaning. It does so by considering how the insertion of a quantifier affects the measurement of p-acceptability (§4.1).

Setup. For each sentence in the dataset, we construct a minimal set resembling the one we considered in section 4.1. If the original sentence had been a generic, we include 11 other sentences, one beginning with each of the other quantifiers. If the original sentence had been quantificational, we include its generic counterpart along with ten other sentences fronted by the remaining quantifiers. The sentence-initial phrase (either the generic noun phrase or one of the 11 quantifier phrases) with the lowest average surprisal on the post-verb tokens is p-acceptable.

Results. Figure A.1 represents the frequency of p-acceptability for each quantifier, given the quantifier of the original sentence, which is specified by the label at the top of each plot. These results were produced for texts that consisted of three-sentence contexts followed by the relevant member of the minimal set (Appendix C). Results are qualitatively similar for any number of context sentences, but not when there's no context (Cilleruelo et al., 2025a).

A higher percentage of sentences for which the p-acceptable quantifier is the original quantifier suggests that there's a bigger difference in meaning with respect to the others. Only for *all*, *few*, *some*, *no* and the generic is the original quantifier p-acceptable over 40% of the time. Quantifiers that express a prevalence reading have very flat spreads, with the original quantifier still having the highest percentage.

In the case of generics, the plot resembles that of quantifiers with clear distinctive meanings (*all*, *some*, *few*, and *no*). We unpack the significance of this observation in Section 6. In almost 50% of sentences, there's no quantifier that would make the post-verb tokens easier to predict given the original generic subject phrase. This is particularly remarkable because, roughly speaking, generics are often paraphrased in terms of a prevalence-expressing quantifier (Carlson and Pelletier, 1995; Cohen, 1999b).

B Model comparison for KL-divergence experiment

To show that the results presented in the paper generalize across a variety of language models, we report the Spearman's correlation between all models on the linearized matrix of data for the heatmap in Figure 1. All correlations are above 0.9 (see Table B.3), suggesting that results are similar across language models.

C Further details on the MGEN dataset

This appendix includes an expert evaluation of the MGEN dataset conducted as part of verifying its quality for this work, as well as examples of sentences in-context as used in the experiments.

C.1 Expert evaluation of MGEN

Not every sentence beginning with a bare-plural noun phrase is a generic (Leslie, 2008; Nickel, 2016; Almotahari, 2024). Some express existential or highly restricted universal generalizations (e.g., *Ravens are on the roof* means that there are ravens on this roof).

	OLMO3-7B	OLMO3-32B	MISTRAL-7B	MIXTRAL-8×22B	LLAMA3.1-8B	QWEN3-8B	GEMMA-7B
OLMO3-7B	1.00	0.95	0.97	0.95	0.98	0.98	0.93
OLMO3-32B	0.95	1.00	0.98	0.96	0.98	0.98	0.90
MISTRAL-7B	0.97	0.98	1.00	0.96	0.99	0.99	0.91
MIXTRAL-8×22B	0.95	0.96	0.96	1.00	0.95	0.95	0.94
LLAMA3.1-8B	0.98	0.98	0.99	0.95	1.00	0.99	0.92
QWEN3-8B	0.98	0.98	0.99	0.95	0.99	1.00	0.91
GEMMA-7B	0.93	0.90	0.91	0.94	0.92	0.91	1.00

Table B.3: Spearman's correlations between different models are all above 0.9.

To have a better understanding of the generics in MGEN, we undertake an annotation of randomly sampled generic sentences and label them as *Generic*, *Particular*, or *Unclear*. A linguistics graduate student and the first author both annotated each of 200 distinct, randomly sampled generic sentences, resulting in $80.5/7.5/12\%$ and $88/3.5/8.5\%$ respectively (*Generic*/*Particular*/*Unclear*). The inter-annotator agreement on a subset of 20 sentences is 85%.

Some examples of false generic sentences in MGEN (labeled *Particular* or *Unclear* by the annotators):

1. Descriptions of scientific plots. *Cells appear red in accordance with Ca^2+ levels.*
2. Complex bare plural existentials. *Lead guitars are of the quality one would find in a well-trained hard rock band, with a good understanding of technique and some advanced tonal and rhythmic ideas, but without standing out.*
3. Uncommon quantifiers. *Young children frequently use play to work through everyday stresses and anxieties.*
4. Some idiomatic bare plurals (e.g., *Studies show...*). *Studies show that flavonoids can exert a neurological response.*

While this annotation could be used to refine the generics in the dataset, we use MGEN as is and leave these refinements for future work. Nevertheless, the annotation corroborates that the generics in MGEN are, overwhelmingly, indeed generics.

The annotator guidelines are as follows.

> In this experiment, you will read short English sentences about groups of things (for example: tigers, cells in the brain, crayons in the play area, ...) that describe the members of these groups in various ways.
>
> You will be asked to label each sentence in two ways:
>
> > whether the sentence is *generic*, *particular*, or *unclear*;
>
> > whether the sentence is true of *all*, a *majority*, or a *minority* of the members of the group in question.
>
> Choose *generic* if the sentence makes a general statement about the group in question by specifying characteristics you might expect to find mentioned in an encyclopedia entry for it. For example: you would expect to find *Ravens are black* but not *Ravens are on the roof of the Dugald Stewart Building*. Choose *particular* if the sentence is not a general statement or doesn't belong in an encyclopedia because the information is overly idiosyncratic or parochial (for example: *Crayons here go in the box*). Choose *unclear* if the sentence is incomplete or ungrammatical or you're just not sure.
>
> Choose *all* if there are absolutely no exceptions to the generalization (e.g., *Tigers are mammals*). Choose *majority* if the generalization is true of more than half of the group's members but not all (e.g., *Tigers are orange*). Choose *minority* if the sentence is true of just a few of the members of the group (*Tigers die of pulmonary diseases*). If you don't know, choose *unclear*.

In the labeling task, you will see sentences that you need to label based on the criteria above. They will appear in black. Text in gray will provide contextual information. The criteria above will remain accessible throughout the task. Select *unclear* only if necessary.

Once you begin, you will see six sample sentences, about which you will receive additional feedback and explanation. Afterward, you will label new sentences on your own.

C.2 Examples from MGEN

Examples of text from the MGEN dataset, with three sentences of left-context, as sampled randomly from our filtered split. We include two samples for each quantifier and the generic. Context has been greyed out.

Preterm births have been reported in pregnant women infected with SARS-CoV-2 [120,121,130,138,139]. Viruses are capable of changing the trophoblast cell response to commensal bacteria from the microbiota present on the mother-fetus interface. Under normal conditions, these cells secrete IFN-β, which influences receptivity to the fetus and can prevent vertical transmission of virus [140]. Viral infections decrease the levels of IFN-β from trophoblastic cells, changing the inflammatory profile of the mother-fetus interface.

As the snakes mature, they change from a blotchy grey to orange or sometimes red, with four standard stripes resembling their background color — with striking red eyes and a matching red tongue (the red eye can be seen in this picture). Their natural habitats are grasslands, wetlands, and swamps. Whenever we see snakes around our home, they're usually non-venemous; and we ensure that they're protected from lawnmowers and the like. Snakes keep rodent populations down — and if you live anywhere near water, you realize how large water rats can become.

Many women are encouraged by medical personnel to be frightened into being manipulated by the threat of pain. I am a homebirth Certified Nurse-Midwife in Missouri. My job is to dispell myths regarding pain from labor vs pain from Cesarean Birth. All moms want what is best for their unborn child.

It's an illusion. The early Tibetan practitioners knew this. They viewed their own pantheon of semi-gods and gods as provisional—just another cosmology—and they expected their students to eventually realize this is well. All cosmologies and symbols are inventions.

It is possible to achieve collective impact without established relationships, but church leaders collaborating within trusted, interconnected relationships can catalyze collective impact into community transformation. Houston Responds brings together these two dynamics: strong relationships and collective impact. Collective Impact Collaborative Ministry Pastoral Connections Relational Strength Community Transformation Houston Responds focuses on three areas of impact: UNITING CHURCH LEADERS We see building trusting relationships not only as a means for catalyzing community transformation but as who we are created to be. Most church leaders desire to stand together but lack ways to connect.

The free gambling account they create will guarantee that they have a constant source of income via gambling winnings. They can then take pleasure in the match and put up a photograph of the prize they won on Instagram so that everybody who wants to try their luck can see. Players may find free slots machines inside the official site of a certified casino as well as third-party websites that promote internet slot games. Most casinos offer some type of promotions or bonuses to be able to attract new customers.

Routines enable you to prepare all areas that impact your sport directly. They empower elite athletes to be mentally, physically, tactically and technically fit to perform their best. A routine creates a mindset that allows you to perform to your talents and capabilities consistently. Many athletes love to prepare themselves for what is to come.

Our treatment of "objects" here does not fit in this definition, but neither do most languages that are typically called "object-oriented". In particular, "late-binding" here means that methods are not looked up until runtime: in a proper SmallTalk-like system, this would be done by

name, meaning that even virtual dispatch in a language like C++ or Java does not count. You can use languages like Ruby or Python in a way that matches this definition, but they're not typically used this way. Many object-oriented languages are also somewhat lax with respect to protection of local information: Python is a big offender here, as its instance variables are typically made private by convention rather than a language mechanism!

Success comes from hard work and consistency. Our staff, directed at helping guys develop why they should feel confident. Show them why they believe in themselves. Some guys have it, some need to see why important.

Games are classified right into several subcategories, relying on the objective and preferred end result. Many single-player video games need the gamer to do a particular job, such as hitting an offered number of balloons, preventing obstacles or running around an area. Lots of games involve competing against one more gamer or a computer. Some games are single-player only; nevertheless, there are several multiplayer video games available, where 2 or more players might complete against each other in a race or competitors.

The conditions require that you actually believe the product works, giving you more confidence. Confident people are more attractive to the opposite sex. To actually know if a pheromone product works, you need double blind placebo controlled testing. Few companies do double blind testing, meaning that some of the people had the pheromones and some did not, but no one knew who did.

Baccarat is the one betting hall betting gae in whhich a bettor could beneft from a yeary six fgure pay. 6) Gabmling Tournaments Gambling eventts can be orgainzed with any on line gaming hall gaame, howeer they‘re really a unqiue type of betting. Gaming Tournnament strateiges are extremely complex pus professional gabling is almost not possible as duing the majority of the tme even a skilled paticipant is gonig to finish to the bottom of the ranknig. Few gamblers hve the bankroll to survive throughoout the majoritty of lost competitions, theey enetr the game till they soner or ltaer win a tournamnet and their gan balances the equation.

I can't think of one famous success in America that got to the pinnacle by sheer accident. I can't help but wonder why so few people in practice have a crisp list of goals jotted down in black and white. Here are five great quotes / reminders regarding goals and why it is worth working on your personal list this coming weekend: Big goals get big results. No goals get no results or someone else's results.

It is significant to know what type of mobile on line casino is available to you — browser or app. This is as a end result of browser variations take up less space on your telephone however can have inferior graphics, whereas apps can have superior gameplay but use extra of your phone's storage capacity. Our team is made up of folks that love playing on line casino video games and wagering on sports teams and players. No deposit promotions are much less widespread because they award a set bonus amount just for registering or taking part in an event.

Bright and golden am I, Mightier than time. If you wish to know of God, you cannot be timid. You cannot be bound to the chain of human failings - our rampant ego, our hubris, our pride and jealousy, our anger and envy, our need to be right and all others wrong, our sarcastic remarks at other people's religion, our roving eyes that pluck imperfection in others, our grudging forgiveness, our forgetful ingratitude. Often people look to God in themselves.

When crimes do occur, they can be violent in nature. Additionally, the lack of a free media and infrequent government outreach through the media do not provide the average citizen with current and accurate information to make informed decisions about safety. Government statistics are typically inaccurate because many crimes are not reported to law enforcement organizations. Often police refuse to open minor or routine cases that seem too difficult to resolve.

You can often find unique cultivars to taste at U-pick farms, farmers markets, or through your county's OSU Master Gardener Program. Depending on the cultivar, fruit may be suited for

fresh eating, juice, raisins, jellies, or wine. Some cultivars suit more than one purpose. Generally, sweet seedless grapes with tender skins are best for raisins.

Sun and yogurt are the best sources of vitamin D. Supplements are also recommended to fuel the body with significant amounts of vitamin D. Omega-3 fatty acids help in reducing inflammation and improving the immune system. These essential fatty acids increase the blood flow, reduce joint pains, inflammation, tenderness, swelling, and discomfort in the joints and knees. Seafood is an excellent source of omega-3 fatty acids. Generally, supplements of omega-3 fatty acids are also recommended for those who do not eat fish.

"It's just been the most hideous season," Ms Stanford said. "We have had a huge season for pups; we've probably had three times more pups than we usually do and we just haven't been able to cope ... and now the heat has just made everything worse again." Advertisement WIRES volunteers provide "halfway houses" where sick or injured bats can recover in safety, regaining their strength before being released. Typically, flying foxes start to die if the temperature climbs above 42 degrees, meaning the record-breaking temperatures throughout January have also led to record numbers of at-risk bats and flying foxes.

The on-going cost for energy and regular maintenance is a factor. You should have room in your budget for annual service as well as the higher electricity bills in the warmer months. Maintenance: Because summer is short in most parts of the country, you won't have to worry about paying to run the AC year-round. Typically, homeowners run their air conditioning units from June into September.

The public cloud customers of these companies can be measured in millions. Private Cloud As the name suggests, the resources such as cloud infrastructure, software, network capabilities, storage etc – can be provided to only a single customer. It mixes up features of public cloud computing like scalability, flexibility and ease of deliver – with better security options, controlling who can access your resources and also customization of the particular resources that the organisation may need. Usually, Private clouds build an infrastructure that is on-premise in the customer's data center.

The Ginzburg-Landau coherence length measured in MAT4G (see Extended DataFig. 3for other devices) is short and around 20 nm, suggesting a relatively strong coupling, as observed in MATTG 4 . The weak linear-in-T behaviour observed might be the result of contributions stemming from both the flat bands and dispersive bands (the latter being very weakly T dependent 42), although further theoretical work and experiments are needed to determine if there are signatures of strange metal behavior in these large N devices. FIG. 3 : 3dimensional (2D) plane of the sample (B). Typically, magnetic fields suppress superconductivity either by inducing vortices or by closing the gap via the Zeeman effect acting on the spin component of the Cooper pairs.

INTRODUCTION Every year university bound graduating high school students are faced with the problem of selecting a post-secondary institution. The selection process typically spans a number of years and involves considering many factors. Identifying those factors that influence students during the selection process was the goal of this study. Normally tuition fees are the deciding criteria for most of the students' enrolment to join a specific university or not.

Dolphins speak with whistle-like sounds handled by vibrating connective tissue, comparative to the way human vocal ropes capacity. Dolphin relations happens tummy to stomach; however numerous species participate in long foreplay, the true demonstration is more often than not short, yet may be rehashed a few times inside a short timespan. The development period changes with species; for the modest Tucuxi dolphin, this period is around 11 to 12 months, while for the orca, the growth period is around 17 months. Normally dolphins conceive a solitary calf, which is, unlike generally different vertebrates, conceived tail first as a rule.