Interpretation of Generalization in Masked Language Models: An Investigation Straddling Quantifiers and Generics

Claudia Collacciani[1,*,†], Giulia Rambelli[1,†]

[1]University of Bologna, Italy

Abstract

Generics are statements that express generalizations and are used to communicate generalizable knowledge. While generics convey general truths (e.g., *Birds can fly*), they often allow for exceptions (e.g., penguins do not fly). Nonetheless, generics form the basis of how we communicate our commonsense about the world [1, 2]. We explored the interpretation of generics in Masked Language Models (MLMs), building on psycholinguistic experimental designs. As this interpretation requires a comparison with overtly quantified sentences, we investigated i) the probability of quantifiers, ii) the internal representation of nouns in generic vs. quantified sentences, and iii) whether the presence of a generic sentence as context influences quantifiers' probabilities. The outcomes confirm that MLMs are insensitive to quantification; nevertheless, they appear to encode a meaning associated with the generic form, which leads them to reshape the probability associated with various quantifiers when the generic sentence is provided as context.

Keywords

Generics, Quantifiers, Masked Language Models, Commonsense Knowledge, Pragmatics

1. Introduction

Generic generalizations, or *generics*, are sentences such as *Birds fly* and *Cars have four wheels*, which allow us to convey information about categories, or *kinds*, of individuals. They are used to communicate information that extends beyond the present context and express our knowledge about the world, including beliefs, stereotypes, or prejudices (e.g., *Women are more sensitive than men*, as well as the less harmful *Italians eat spaghetti*). Generics can be considered one of the cornerstones of human cognition since they allow us to conceptualize the properties we attribute to categories and thus organize our experience of the world [3].

The most distinctive feature of generics is that they allow for exceptions [4]. For example, *Birds fly* is judged true even if there are birds that cannot fly (e.g., penguins): in this case, therefore, the corresponding universal statement (*All birds fly*) is false. Different generalizations tolerate exceptions to varying degrees. Thus, some generic statements might be better paraphrased with *all*, others with *most*, and others with *some*, but —unlike quantified statements —they do not explicitly contain information about the prevalence of the property in the category (i.e., how many members of the category possess the property). Similarly, there is no unambiguous relationship between the prevalence of a property among category members and the acceptability of the corresponding generic as true. For example, the generalization *Lions have manes* is accepted even if only male adult lions have manes, but the generalization *Lions are males* is rejected.

Given these properties, the meaning of generalizations can be considered "vague", and their interpretation can be assumed to be derived by people through world knowledge and pragmatic skills [5]. Most of the experimental studies conducted on generics are cognitively driven and based mainly on contrasting generics with overtly quantified sentences [3]; in other words, quantifiers are used to approximate the vague meaning of generics.

In this paper, we investigate the interpretation of generalizations in Large Language Models (LLMs) of the Transformer family, building on psycholinguistic experimental designs (in particular, Leslie et al. [6] and Cimpian et al. [7]). Since comparison with quantification seems to be necessary to decode the meaning of generics, we also used quantifiers.

We present three tasks related to different but complementary research questions:

1. *Are LLMs biased towards some quantifiers more than others?* We computed the probability of several quantifiers appended to a generic statement. This analysis serves as a baseline to understand the probability distribution of quantifiers and if

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 – Dec 02, 2023, Venice, Italy

[*] Corresponding author.

[†] These authors contributed equally.

claudia.collacciani2@unibo.it (C. Collacciani); giulia.rambelli4@unibo.it (G. Rambelli)

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

there is any bias towards some of them (for example, a generic overgeneralization effect such as that found in humans by Leslie et al. [6]).

2. *Are the hidden representations of generics similar to those of quantified phrases?* We extracted the hidden representation of words in generic and quantified sentences and compared their representation pairwise to understand which quantified nominal phrase approximates the meaning of generics better.
3. *Are LLMs showing the same prevalence effect as humans?* We reproduced the experimental design of Cimpian et al. [7] Implied Prevalence task to test whether the presence of the generic as a premise impacts the probability of selected quantifiers.

The data and code that we used for the experiments are publicly available[1].

2. Related Works

2.1. Generics in Human Cognition

Experimental evidence has revealed a generic bias for which people tend to overgeneralize from the truth of a generic to the truth of the corresponding universal statement [6, 8, 9]. For example, people tend to accept the statement *All lions have manes* as true, even though it is not, because they rely on the truth of the corresponding generic *Lions have manes.* This effect is known as the generic overgeneralization (GOG) effect [6]. It is detected only on certain categories of generics, namely those of minority characteristic and majority characteristic generics, i.e., generics that predicate properties that are true for a minority or the majority of category members.

Cimpian et al. [7] conducted a series of studies investigating the relationship between genericity and prevalence and found an inferential asymmetry in the meaning of generics. People tend to judge a generic sentence about a novel category as true even if they have been informed that only a certain percentage of the kind (on average, up to less than 70 percent) possess the property in question (Truth Condition task). However, when asked to estimate how many members of the kind possess the property, given the generic (Implied Prevalence task), they tend to assign very high percentages (on average, very close to 100 percent). This study indicates that generic sentences require little evidence to be judged true but have substantial implications, since the properties they predicate tend to be interpreted as applying to virtually all members of the category.

2.2. Genericity in NLP

Most of the NLP literature on genericity has focused on the creation and annotation of resources for identifying generic expressions as opposed to non-generic ones and, based on these resources, on the development of automatic annotation systems [10, 11, 12, among others]

To the best of our knowledge, there are no studies investigating the interpretation of generalizations by LLMs, except for the recent work by Ralethe and Buys [13], which addresses the generic overgeneralization effect in BERT and RoBERTa. The authors argue that these models suffer from overgeneralization by assessing how many times one or more of *all, every, most, some, few* and *many* are predicted in a masked sentence like *[MASK] lions have manes*: the higher the rank of the quantifiers, the stronger the LM exhibits the GOG effect. However, the GOG effect refers to the acceptance of *universally quantified* sentences, not just quantified ones; therefore, we can speak of overgeneralization only when the preferred quantifier is the universal one (*all* or *every*). For this reason, we first propose a similar task to evaluate the probability distribution of various quantifiers, distinguishing between them qualitatively.

3. Materials and Methods

Data For this study, we selected the generic sentences from the dataset of Allaway et al. [14]. The authors extracted 653 generics about objects, animals, and plants from Bhagavatula et al. [15] and annotated them into three categories obtained unifying theories from linguistics and philosophy, by condensing the five types of generics proposed by Leslie [16, 17] and Khemlani et al. [18]. In **quasi-definitional** sentences, the property is essential to a concept, thus is considered a defining characteristic of the concept (e.g., *triangles have three sides*). In this type of sentences, the generic is *de facto* equivalent to the corresponding universal quantified statement (e.g., *all triangles have three sides*). In **principled** sentences, the property has a strong association with the concept. This category includes both properties that are viewed as inherent, or connected in a principled way with a concept (e.g., *birds can fly.*), and properties that are uncommon and often dangerous (e.g., *sharks attack swimmers*); this last case is the one that Leslie [16, 17] defines *striking*. Finally, **characterizing** sentences express a non-accidental relationship between property and concept, based only on absolute or relative prevalence among category members. These generics concern properties that are neither deeply connected to the concept nor striking, but occur in the majority (*majority characteristic* generics for [16, 17], e.g., *Cars have radios*) or in the minority of members of the category (*minority characteristic* generics for [16, 17], e.g., *Lions have manes*).

[1] https://github.com/claudiacollacciani/Interpretation-of-Generalization-in-Masked-Language-Models

From the original batch, we restricted our choice to 207 generic sentences, picking only the ones in the bare plural form (e.g., *Tigers are striped*), excluding indefinite and definite singulars (e.g., *A/The tiger is striped*). All these syntactic forms can express generic meanings, but the bare plural is the only surface form in English that gives rise to a generic interpretation unambiguously [19]. For this and other reasons, this is considered as the paradigmatic case, and it is the one that has been used in the psycholinguistic experiments from which we draw inspiration.

Models We experimented with BERT and RoBERTa, two bidirectional Masked Language Models (MLMs) based on the Transformer architecture. **BERT** [20] is trained both on a masked language modeling task and on a next sentence prediction task, as the model receives sentence pairs in input and has to predict whether the second sentence is after the first one in the training data. BERT has been trained on the BookCorpus and the English Wikipedia for around 3300M tokens. We employed the `bert-base-uncased` and `bert-large-uncased` pre-trained versions, which differ in terms of parameters (110M and 340M parameters, respectively). On the other hand, **RoBERTa** [21] has the same architecture as BERT; however, it introduces several parameter optimization choices, such as dynamic masking, a larger batch and vocabulary size, and the removal of the next sentence prediction objective. Another key difference is the larger training corpus: RoBERTa was trained on 160GB of texts. We relied on the Huggingface's Transformers[2] Library to load the models and carry on our experiments.

4. Experiments

4.1. Experiment 1: Probability distribution of Quantifiers in MLMs

In the first place, we needed to assess what was the most expected quantifier for the sentences in our dataset. Therefore, we modified the original generic sentences by placing the special token [MASK] at the beginning of each sentence, as in '[MASK] *strawberries have a sweet flavor.*' Then, we computed the conditional log probability of quantifiers *few, some, many, most*, and *all* in the masked position, following previous works in quantification [22, 13, 23]. The conditional log probability is defined as

$$p(w_i) = logP_{MLM}(w_i|c) \tag{1}$$

where c are the words preceding and following the critical word in the sentence.

This analysis serves as a baseline to understand the probability distribution of quantifiers, if there is any bias towards some of them (i.e., overgeneralization effect), and possibly whether the belonging of sentences to different categories impacts it (as observed for humans by Leslie et al. [6]).

Results Figure 1 reports the quantifiers distributions for the base models, as the larger counterparts show a similar trend (all boxplots are in Appendix B). Overall, all models consider *few* the least likely. This outcome reflects our expectations: as the selected sentences are all generalizations, they are, in most cases, referable to a substantial number of category members, rarely to 'few' members. Apart from that, BERT and RoBERTa show different probability distributions of quantifiers, regarding *some* and *all* in particular. BERT models assign a higher probability to the existential and proportional quantifiers (*some*, *many*, and *most*) than the universal quantifier *all*, and *some* is overall the most expected. The differences among the quantifier scores are statistically significant[3], with few exceptions (see Appendix B). It is

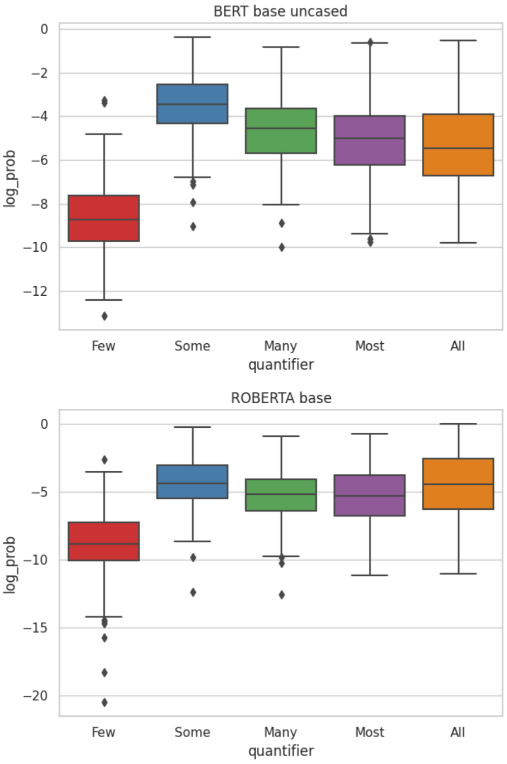

Figure 1: Probability distributions per quantifier for MLM-base variants in Experiment 1.

[2] https://huggingface.co/docs/transformers/index

[3] By relying on Wilcoxon Signed-Rank Test statistical test.

worth noticing that the reported distributions remain the same even when we separate the analysis by sentence categories. In other words, BERT models are not sensitive to the pragmatic differences of the selected sentences.

Conversely, both *all* and *some* are the most expected quantifiers for RoBERTa. Accordingly, the universal quantifier *all* seems to be more expected by RoBERTa than by BERT, but this distribution is not constant for all three sentence conditions. In quasi-definitional and characterizing sentences, *some* and *all* have the same probability; alternatively, *some* is more expected in principled sentences. We could draw that, for principled sentences, the model prefers not to overgeneralize the property to all members of the category. This behavior seems to approximate Leslie et al. [6] results: people tend to overgeneralize (i.e., to accept as true the universal sentence corresponding to the generic) in the case of characterizing sentences, while they do not overgeneralize (correctly) in the case of striking sentences (included in the principled category).

Regardless, the observed trends could be determined by the overall frequency of the quantifiers. As a sanity check, we extracted their frequency from a large corpus of English, enTenTen21 [24, 25]. We found that $freq(some) > freq(all) > freq(many) > freq(few) > freq(most)$ (frequencies are reported in Appendix A). This pattern confirms that *few* is not the less probable because of a frequency effect but for the properties of the sentence. Conversely, *most* is the less frequent but has a probability score similar to the more frequent *many* and *all*. Finally, *some* is overall the most frequent quantifier. This observation could partially reflect the probability outputs of BERT; however, it is not the case for RoBERTa scores.

4.2. Experiment 2: Representation of words in Generics and Quantified Sentences

The architecture of MLMs allows us to follow the transformations of each token throughout the neural network. Previous works in BERTology have reported that internal representations, also known as contextualized embeddings, encode syntactic and semantic properties in different hidden layers [26]. However, it is complex to localize semantic phenomena, as they spread across the entire model [27]. For our purposes, we decided to compare the contextualized embedding of a target token (*strawberries*) in the generic sentence (*<u>strawberries</u> have a sweet flavor*) with the embedding of the target token in each of the corresponding quantified sentences (e.g., *all <u>strawberries</u> have..*). Following Timkey and van Schijndel [28], for each layer, we computed the similarity of the two contextualized embeddings by relying on Spearman's ρ correlation[4].

This study has a twofold aim: i) Identify the quantifier that shifts the noun representation closer to that of the generic statement (if possible), and ii) Localize the layers where quantification emerges. To the best of our knowledge, no previous work has explored the internal representations of quantified expressions in relation to genericity.

Results Figure 2 illustrates how the Spearman's ρ between the noun and the quantified version changes with respect to each hidden layer in BERT and RoBERTa base (see Appendix C for the plots of larger models and the ones reporting correlations by sentence categories). The first layers do not show a difference among correlations, meaning that representations characterized by the different quantifiers are practically identical; this is expected, as the context is limitedly attended by the model in these layers. The following layers show a gradual change in correlation values, but BERT and RoBERTa show different patterns. For the first, we observe a slight decrease in scores from layers 3 to 9 (but the correlation values are still above 0.9). Conversely, the peak of the curve is at layer 5 for RoBERTa (ρ 0.76), while the other internal

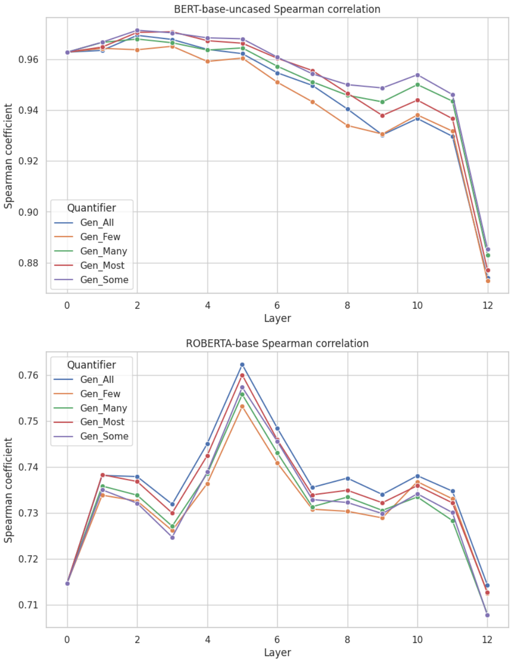

Figure 2: Spearman's ρ against model layers.

[4] The authors reported that Spearman's ρ is more robust to rogue dimensions contextual language models than cosine or Euclidean similarity measures.

layers have a constant value of around 0.73. In the last intermediate layers (8-10), we observe some drifting away between the values for the different quantifiers, indicating that the contribution of the quantifiers differs from one to another. For instance, in BERT-base and large the noun quantified by *some* is the most similar to the corresponding non-quantified. At the same time, RoBERTa has generics similar to *all*, which could be interpreted as an overgeneralization effect.

Intriguingly, the most similar representations are the most probable quantifiers of the previous experiment, thus confirming that the choice of the quantifier is not a frequency effect but is related to providing a representation closer to that of the generic statement. However, the values for the various quantifiers always remain close to each other and follow the same trend, so it is hard to disclose how the meaning of the quantifier affects the noun representation. Finally, all models worsened their performance at the last layer - the one producing the most context-specific representations [29], indicating that contextual information weakens the quantification signal.

4.3. Experiment 3: Implied Prevalence effects in MLMs

As observed above, MLMs are not particularly sensitive to quantifiers, and the probability choices are independent of the sentence's meaning. This outcome is mainly due to the fact that these models are agnostic to world knowledge. Therefore, we decided to test the relation between quantification and generalization from a more formal point of view: we examined how models interpret generalizations aside from their content, that is, whether they contain any linguistic information associated with the form of generics.

We reproduce the experimental design of Cimpian et al. [7] Implied Prevalence task, in which people were presented with a generic sentence about a novel animal category and then asked to estimate how many members of the category possess the characteristic predicated by the generic (e.g., Information: *Morseths have silver fur.* Question: *What percentage of morseths do you think have silver fur?*). In this case, world knowledge is not called into play, unlike in Leslie et al. [6]'s experiments: the categories employed are made up, and thus lack associations to properties in the speakers' mind. Since models do not seem to encode the world knowledge necessary to interpret generics on account of their content (with the partial exception of RoBERTa), this experimental design may be suitable for investigating instead the default interpretation they associate with a generic form.

We build the stimuli using the generic sentence as the premise in the following way: *Strawberries have a sweet flavor means that* [MASK] *strawberries have a sweet flavor.* As in Experiment 1, we compute the log probability of quantifiers (*few, some, many, most,* and *all*) in the masked position. This last study should answer the following question: Does the presence of a generic sentence impact the quantifier preference, compared to the a-contextualized version of the first analysis?

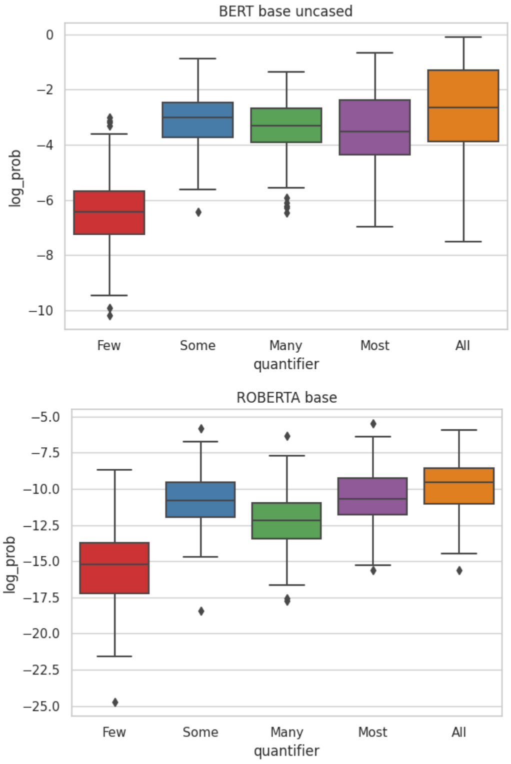

Figure 3: Probability distributions per quantifier for MLM-base variants in Experiment 3.

Results Figure 3 reports the quantifiers distributions for the base models, as the larger counterparts show the same trend (all boxplots are in Appendix D). Surprisingly, the most expected quantifier is now *all* for all models[5]. BERT shows an inversion in the ratio between the probability associated with the universal quantifier *all* and that associated with the other quantifiers (apart from *few*), which in the first experiment were all more expected than *all*, whereas now are all less expected. The pattern exhibited by RoBERTa shows a less striking change than BERT; however, the probability of *all* with respect to the other quantifiers is still higher than in Experiment 1. As an additional check, we performed the same test with a very small dataset of 24 generic sentences about novel

[5] The differences among the quantifier scores are statistically significant; a few exceptions are listed in Appendix D.

categories (non-words) from Cella et al. [30],[6] to be sure that the content of the sentences would not affect the results. As we expected, we obtained the same pattern as with our dataset.

The results of this experiment, when compared with the baseline for the choice of quantifiers in Experiment 1, suggest that the presence of a generic sentence as a premise does indeed have an impact on the preference of the quantifier by the models. When the generalization is provided as context, the preferred quantifier becomes *all*. This behavior mirrors that of people, observed in Cimpian et al. [7] experiments and later replications: people tend to estimate very high percentages (on average very close to 100 percent) in the Implicit Prevalence task.

5. Discussion and Conclusions

In this paper, we analyzed the interpretation of generics in MLMs through psycholinguistic experimental designs that exploited quantified expressions to investigate the understanding of generic ones.

The first two experiments raise questions about the codification of quantifiers, as it seems that the models do not substantially exhibit a strong sensitivity to quantifiers and do not encode a semantic difference in the representation of quantification. Altogether, our results suggest that the models do not appear to contain the commonsense knowledge required to interpret generics that differ in content through quantifiers. However, they seem to have encoded a meaning associated with the generic form, which leads them to reshape the probability associated with various quantifiers when the generic sentence is provided as context. In the last experiment, we observed that the models prefer the universal quantifier unanimously if preceded by a generic utterance. People behave similarly when tested on novel categories, that is, non-word categories for which subjects have no prior understanding. However, people can modulate their interpretations of generalizations in a real language setting through their world knowledge of real categories. Regardless, MLMs tend to treat real and invented categories equally, being agnostic to world knowledge. For this reason, this could be a potentially harmful bias.

The presented analysis has theoretical and methodological implications. First, we observed that the language of generalization is a complex phenomenon that is hard to investigate in human processing and even more in LLMs, mostly because the investigation of generics' interpretation makes use of quantifiers, and language models often fail in tasks related to quantification. Another problem lies in the fact that it is difficult to test autoregressive models (e.g., GPT family) on tasks such as the one used in Experiment 1 because, as they do not have access to the right context, they do not have sufficient information to modulate the probabilities associated with the various quantifiers accordingly. Finding ways to test autoregressive models in addition to MLMs would be desirable.

In this paper, we have not directly investigated the aspect of the attention in MLMs. Future research could address this aspect. Furthermore, future work could involve the definition of alternative tasks for investigating generalizations to make comparing models and human interpretations easier. Psycholinguistic tests on this phenomenon often rely on truth judgments, and we should be cautious about comparing human truth judgments with model outputs since they lack commonsense knowledge comparable to that of humans. Overall, further investigations are needed to clarify the interpretation of generics in language models.

Acknowledgments

This research is funded by the European Union (GRANT AGREEMENT: ERC-2021-STG-101039777). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. We thank the anonymous reviewers for their useful suggestions.

References

[1] J. A. Hampton, Generics as reflecting conceptual knowledge, Recherches linguistiques de Vincennes (2012) 9–24.

[2] S.-J. Leslie, Carving up the social world with generics, Oxford studies in experimental philosophy 1 (2014).

[3] D. L. Chatzigoga, Genericity, in: The Oxford Handbook of Experimental Semantics and Pragmatics, Oxford University Press, 2019, pp. 156–177.

[4] M. Krifka, F. J. Pelletier, G. Carlson, A. ter Meulen, G. Chierchia, G. Link, Genericity: An introduction, in: G. N. Carlson, F. J. Pelletier (Eds.), The Generic Book, University of Chicago Press, 1995, pp. 1–124.

[5] M. H. Tessler, N. D. Goodman, The language of generalization., Psychological review 126 (2019) 395.

[6] S.-J. Leslie, S. Khemlani, S. Glucksberg, Do all ducks lay eggs? the generic overgeneralization ef-

[6] The authors reproduced the experiment of Cimpian et al. [7], obtaining the same results.

fect, Journal of Memory and Language 65 (2011) 15–31.

[7] A. Cimpian, A. C. Brandone, S. A. Gelman, Generic statements require little evidence for acceptance but have powerful implications, Cognitive science 34 (2010) 1452–1482.

[8] S. Khemlani, S.-J. Leslie, S. Glucksberg, Inferences about members of kinds: The generics hypothesis, Language and Cognitive Processes 27 (2012) 887–900.

[9] S.-J. Leslie, S. A. Gelman, Quantified statements are recalled as generics: Evidence from preschool children and adults, Cognitive psychology 64 (2012) 186–214.

[10] N. Reiter, A. Frank, Identifying generic noun phrases, in: Proceedings of the 48th annual meeting of the association for computational linguistics, 2010, pp. 40–49.

[11] A. Friedrich, A. Palmer, M. P. Sørensen, M. Pinkal, Annotating genericity: a survey, a scheme, and a corpus, in: Proceedings of the 9th Linguistic Annotation Workshop, 2015, pp. 21–30.

[12] V. Govindarajan, B. V. Durme, A. S. White, Decomposing generalization: Models of generic, habitual, and episodic statements, Transactions of the Association for Computational Linguistics 7 (2019) 501–517.

[13] S. Ralethe, J. Buys, Generic overgeneralization in pre-trained language models, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3187–3196. URL: https://aclanthology.org/2022.coling-1.282.

[14] E. Allaway, J. D. Hwang, C. Bhagavatula, K. McKeown, D. Downey, Y. Choi, Penguins don't fly: Reasoning about generics through instantiations and exceptions, arXiv preprint arXiv:2205.11658 (2022).

[15] C. Bhagavatula, J. D. Hwang, D. Downey, R. Le Bras, X. Lu, L. Qin, K. Sakaguchi, S. Swayamdipta, P. West, Y. Choi, I2D2: Inductive knowledge distillation with NeuroLogic and self-imitation, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 9614–9630. URL: https://aclanthology.org/2023.acl-long.535.

[16] S.-J. Leslie, Generics and the structure of the mind, Philosophical perspectives 21 (2007) 375–403.

[17] S.-J. Leslie, Generics: Cognition and acquisition, Philosophical Review 117 (2008) 1–47.

[18] S. Khemlani, S.-J. Leslie, S. Glucksberg, Generics, prevalence, and default inferences, in: Proceedings of the 31st annual conference of the cognitive science society. Austin, TX: Cognitive Science Society, 2009.

[19] G. N. Carlson, Reference to kinds in English., University of Massachusetts Amherst, 1977.

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of NAACL, 2019.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[22] M. Apidianaki, A. Garí Soler, ALL dolphins are intelligent and SOME are friendly: Probing BERT for nouns' semantic properties and their prototypicality, in: Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 79–94. URL: https://aclanthology.org/2021.blackboxnlp-1.7. doi:10.18653/v1/2021.blackboxnlp-1.7.

[23] A. Gupta, Probing quantifier comprehension in large language models, arXiv preprint arXiv:2306.07384 (2023).

[24] A. Kilgarriff, Kovář; v.; rychlý, p.; suchomel, v. the tenten corpus family, in: 7th International Corpus Linguistics Conference CL, 2013.

[25] V. Suchomel, Better web corpora for corpus linguistics and nlp, Doctoral Theses. Brno: Masaryk University, Faculty of Informatics (2020).

[26] A. Rogers, O. Kovaleva, A. Rumshisky, A primer in BERTology: What we know about how BERT works, Transactions of the Association for Computational Linguistics 8 (2020) 842–866. URL: https://aclanthology.org/2020.tacl-1.54. doi:10.1162/tacl_a_00349.

[27] I. Tenney, D. Das, E. Pavlick, BERT rediscovers the classical NLP pipeline, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4593–4601. URL: https://aclanthology.org/P19-1452. doi:10.18653/v1/P19-1452.

[28] W. Timkey, M. van Schijndel, All bark and no bite: Rogue dimensions in transformer language models obscure representational quality, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 4527–4546. URL: https://aclanthology.org/2021.emnlp-main.372. doi:10.18653/v1/2021.emnlp-main.372.

[29] K. Ethayarajh, How contextual are contextualized word representations? Comparing the geometry

of BERT, ELMo, and GPT-2 embeddings, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 55–65. URL: https://aclanthology.org/D19-1006. doi:10.18653/v1/D19-1006.

[30] F. Cella, K. A. Marchak, C. Bianchi, S. A. Gelman, Generic language for social and animal kinds: An examination of the asymmetry between acceptance and inferences, Cognitive Science 46 (2022) e13209.

A. Quantifiers Frequencies in enTenTen21

We extracted the frequencies from enTenTen21. The corpus is made up of texts collected from the Internet consisting of more than 60 billion tokens. The texts were downloaded in October–December 2021 and January 2022. We relied on the concordance tool provided by SketchEngine to extract the frequencies in the form '[quantifier][noun]'.

	N. hits	N. hits per million tokens	% of whole corpus
some	43,436,065	705.29	0.07053%
all	39,201,717	636.54	0.06365%
many	30,529,532	495.72	0.04957%
few	19,067,122	309.6	0.03096%
most	11,187,850	181.66	0.01817%

Table 1
Distribution of quantifiers in enTenTen21 corpus.

B. Experiment 1: Boxplots and Wilcoxon statistical analysis

We report the boxplots for the base and large versions of BERT and RoBERTa for Experiment 1.

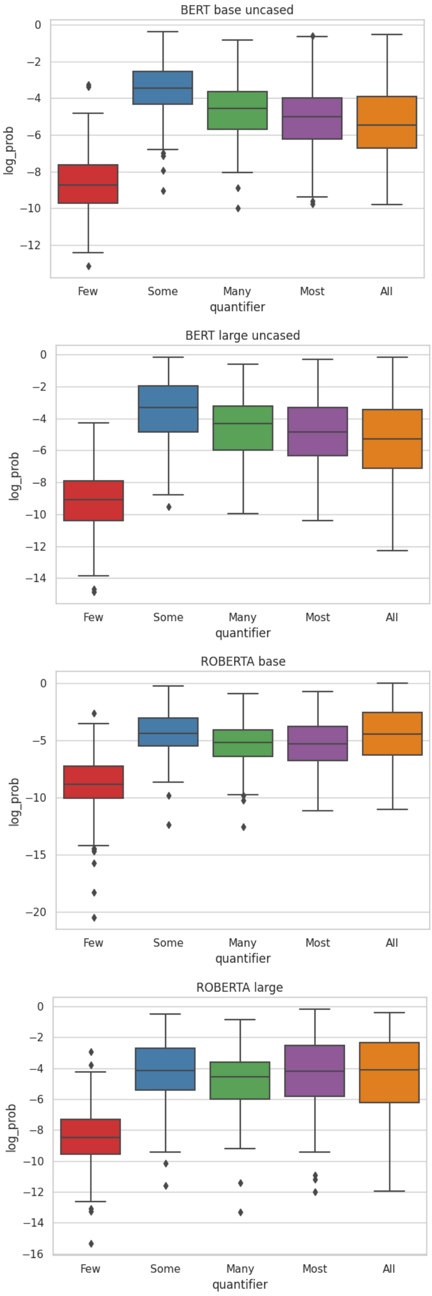

Figure 4: Probability distributions per quantifier for BERT and RoBERTa variants in Experiment 1.

In the tables below, we also report the results of the statistical test performed to verify if the difference in probabilities of quantifiers is statistically significant or not.

Model	Group1	Group2	p-value	Significance
bert-base	few	some	1.023e-35	significant
bert-base	few	many	1.023e-35	significant
bert-base	few	most	7.204e-35	significant
bert-base	few	all	2.538e-34	significant
bert-base	some	many	4.439e-30	significant
bert-base	some	most	2.390e-26	significant
bert-base	some	all	6.709e-21	significant
bert-base	many	most	8.841e-04	significant
bert-base	many	all	8.854e-06	significant
bert-base	most	all	3.054e-03	significant
bert-large	few	some	1.023e-35	significant
bert-large	few	many	1.084e-35	significant
bert-large	few	most	1.536e-35	significant
bert-large	few	all	7.630e-35	significant
bert-large	some	many	4.051e-24	significant
bert-large	some	most	1.296e-19	significant
bert-large	some	all	1.958e-17	significant
bert-large	many	most	5.408e-02	not significant
bert-large	many	all	3.407e-05	significant
bert-large	most	all	6.600e-05	significant

Table 2
Wilcoxon Signed-Rank Test on BERT variants for Experiment 1.

Model	Group1	Group2	p-value	Significance
RoBERTa-base	few	some	1.450e-35	significant
RoBERTa-base	few	many	1.291e-35	significant
RoBERTa-base	few	most	6.319e-33	significant
RoBERTa-base	few	all	9.364e-34	significant
RoBERTa-base	some	many	2.023e-19	significant
RoBERTa-base	some	most	2.724e-08	significant
RoBERTa-base	some	all	4.231e-02	significant
RoBERTa-base	many	most	8.949e-01	not significant
RoBERTa-base	many	all	8.148e-03	significant
RoBERTa-base	most	all	6.013e-05	significant
RoBERTa-large	few	some	1.023e-35	significant
RoBERTa-large	few	many	1.272e-35	significant
RoBERTa-large	few	most	7.414e-35	significant
RoBERTa-large	few	all	3.472e-34	significant
RoBERTa-large	some	many	4.334e-14	significant
RoBERTa-large	some	most	1.612e-01	not significant
RoBERTa-large	some	all	5.842e-02	not significant
RoBERTa-large	many	most	6.901e-06	significant
RoBERTa-large	many	all	1.639e-02	significant
RoBERTa-large	most	all	5.972e-01	not significant

Table 3
Wilcoxon Signed-Rank Test on RoBERTa variants for Experiment 1.

C. Experiment 2: A layer-wise analysis of MLMs representations

We report the plots for the base and large variants of BERT and RoBERTa, with respect to each hidden layer.

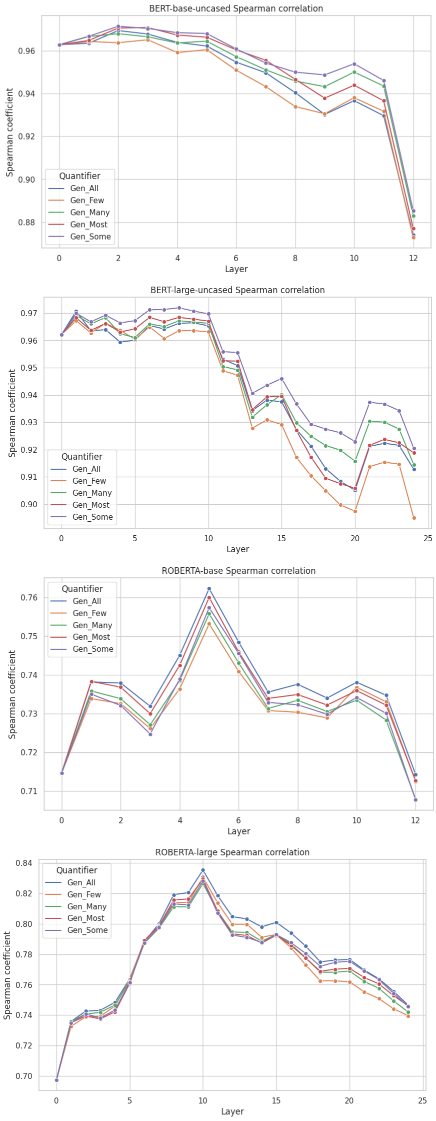

Figure 5: Spearman's ρ between the noun in generic sentence and its quantified variant across models layers.

We also report the same plots with the correlation scores for each sentence category. While the trends are the same for the three conditions, the values have a slight difference in the means, with quasi-definitional sentences having a higher correlation than the other two types.

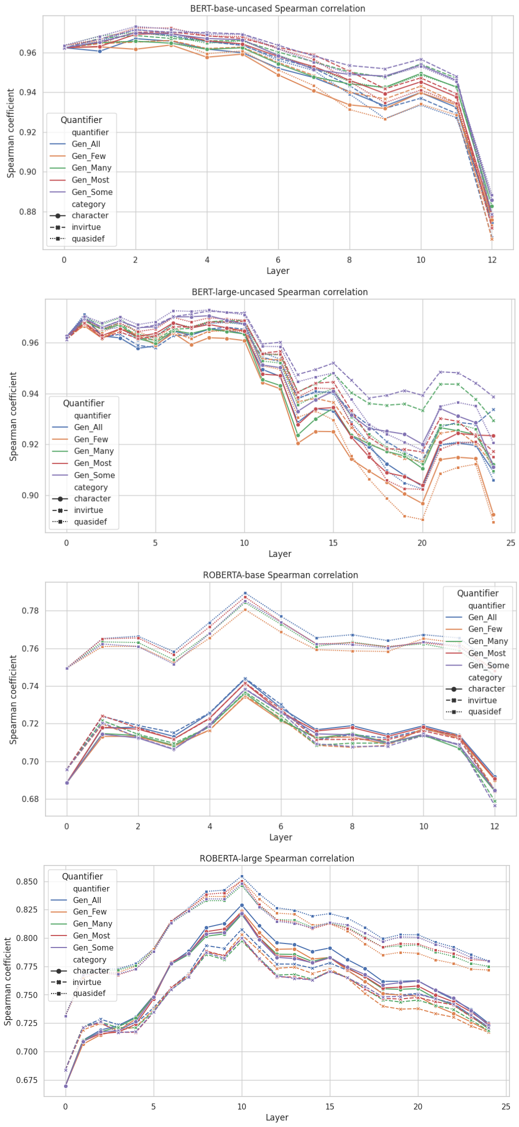

Figure 6: Spearman's ρ between the noun in generic sentence and its quantified variant across models layers, by sentence categories.

D. Experiment 3: Boxplots and Wilcoxon statistical analysis

We report the boxplots for the base and large versions of BERT and RoBERTa for experiment 3.

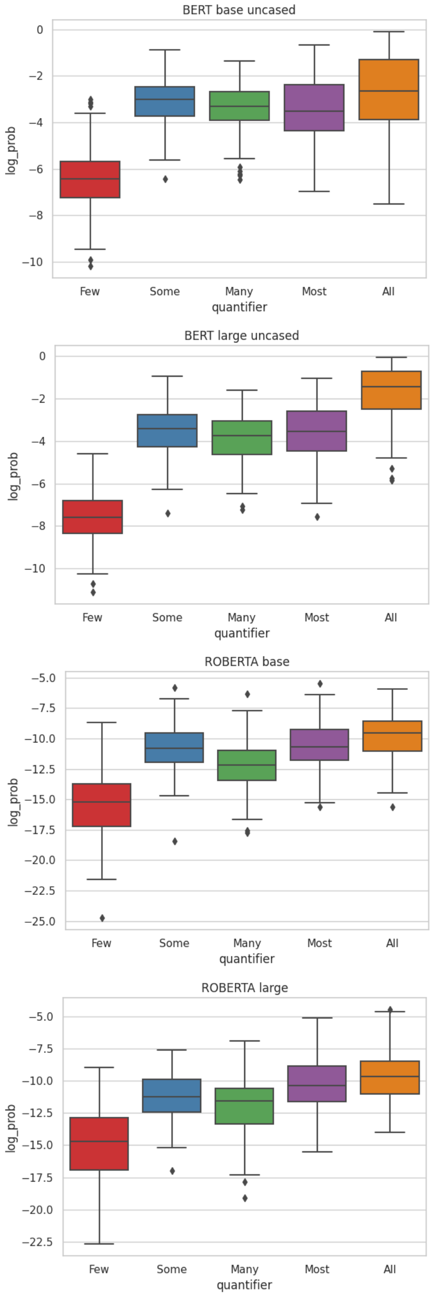

Figure 7: Probability distributions per quantifier for BERT and RoBERTa variants in Experiment 3.

In the tables below, we also report the results of the statistical test performed to verify if the difference in probabilities of quantifiers is statistically significant or not.

Model	Group1	Group2	p-value	Significance
bert-base	few	some	1.701e-35	significant
bert-base	few	many	1.116e-35	significant
bert-base	few	most	1.140e-34	significant
bert-base	few	all	3.357e-35	significant
bert-base	some	many	2.399e-07	significant
bert-base	some	most	1.335e-04	significant
bert-base	some	all	2.599e-02	significant
bert-base	many	most	7.970e-01	not significant
bert-base	many	all	7.696e-07	significant
bert-base	most	all	3.920e-10	significant
bert-large	few	some	1.023e-35	significant
bert-large	few	many	1.084e-35	significant
bert-large	few	most	1.536e-35	significant
bert-large	few	all	7.630e-35	significant
bert-large	some	many	4.051e-24	significant
bert-large	some	most	1.296e-19	significant
bert-large	some	all	1.958e-17	significant
bert-large	many	most	5.408e-02	not significant
bert-large	many	all	3.407e-05	significant
bert-large	most	all	6.600e-05	significant

Table 4
Wilcoxon Signed-Rank Test on BERT variants for Experiment 3.

Model	Group1	Group2	p-value	Significance
RoBERTa-base	few	some	1.084e-35	significant
RoBERTa-base	few	many	5.788e-34	significant
RoBERTa-base	few	most	3.992e-35	significant
RoBERTa-base	few	all	1.652e-35	significant
RoBERTa-base	some	many	4.769e-21	significant
RoBERTa-base	some	most	2.381e-01	not significant
RoBERTa-base	some	all	5.573e-08	significant
RoBERTa-base	many	most	1.141e-19	significant
RoBERTa-base	many	all	2.409e-27	significant
RoBERTa-base	most	all	4.718e-10	significant
RoBERTa-large	few	some	3.169e-35	significant
RoBERTa-large	few	many	3.326e-34	significant
RoBERTa-large	few	most	7.521e-35	significant
RoBERTa-large	few	all	2.515e-35	significant
RoBERTa-large	some	many	3.979e-10	significant
RoBERTa-large	some	most	5.351e-11	significant
RoBERTa-large	some	all	3.152e-20	significant
RoBERTa-large	many	most	1.723e-23	significant
RoBERTa-large	many	all	7.571e-28	significant
RoBERTa-large	most	all	3.265e-07	significant

Table 5
Wilcoxon Signed-Rank Test on RoBERTa variants for Experiment 3.