Probing Quantifier Comprehension in Large Language Models: Another Example of Inverse Scaling

Akshat Gupta
AI Research, JPMorgan Chase*
University of California at Berkeley
akshat.gupta@berkeley.edu

Abstract

With their increasing size, large language models (LLMs) are becoming increasingly good at language understanding tasks. But even with high performance on specific downstream task, LLMs fail at simple linguistic tests for negation or quantifier understanding. Previous work on quantifier understanding in LLMs show inverse scaling in understanding *few*-type quantifiers. In this paper, we question the claims of of previous work and show that it is a result of inappropriate testing methodology. We also present alternate methods to measure quantifier comprehension in LLMs and show that LLMs are able to better understand the difference between the meaning of *few*-type and *most*-type quantifiers as their size increases, although they are not particularly good at it. We also observe inverse scaling for *most*-type quantifier understanding, which is contrary to human psycho-linguistic experiments and previous work, where the model's understanding of *most*-type quantifier gets worse as the model size increases. We do this evaluation on models ranging from 125M-175B parameters, which suggests that LLMs do not do as well as expected with quantifiers. We also discuss the possible reasons for this and the relevance of quantifier understanding in evaluating language understanding in LLMs.

1 Introduction

Large Language Models (LLMs) are getting increasingly better at understanding language (Devlin et al., 2018; Radford et al., 2019; Raffel et al., 2020; Zhang et al., 2022; Ouyang et al., 2022; Touvron et al., 2023) which can be seen by their improving performance on various language understanding benchmarks (Wang et al., 2018, 2019). Auto-regressive LLMs including encoder-decoder models like BART (Lewis et al., 2019) and T5 (Raffel et al., 2020) and decoder-only models like GPT (Radford et al., 2018, 2019; Brown et al., 2020; Zhang et al., 2022; Touvron et al., 2023) have been scaled to billions of parameters to improve their language understanding capabilities. With increasing model sizes, the models also gets increasingly better at learning from context and can just be prompted with few examples rather than fine-tuning to do downstream task (Brown et al., 2020; Liu et al., 2023).

Even with this unprecedented yet implicit evidence of increasing language understanding capability of LLMs, these models still fail simple linguistic tests on understanding negation and quantifiers (Jang et al., 2023; Kalouli et al., 2022; Michaelov and Bergen, 2022). Understanding negation and quantifiers is challenging for language models because the presence of a single negating or quantifying word can drastically change the meaning of the sentence. Also, such sentences are infrequently used in pre-training text corpora (Jiménez-Zafra et al., 2020; Michaelov and Bergen, 2022), which makes it hard for the models to account for such situations. Due to this, actual comprehension of *negation* or *quantifier* words is overpowered by the larger context of the sentence, which makes it challenging for language models to deal with these situations.

We focus on one specific linguistic phenomenon, which is the use of *quantifiers*. Quantifiers are words that usually occur before a noun to express the quantity of an object (Kalouli et al., 2022). The presence of different quantifiers can make statements semantically very different from each other. It can be seen from the following example:

$$\begin{aligned} \text{(Ex:1) All Ps are Qs} &\implies P \subseteq Q \\ \text{No Ps are Qs} &\implies P \cap Q = \emptyset \end{aligned}$$

In the above example, two different quantifiers *all* and *no* when applied to the sets P and Q end up

[*] This work was done while at AI Research, JPMorgan.

Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 56–64
December 7, 2023. ©2023 Association for Computational Linguistics

Backbone Phrase	Quantifier	Typicality
postmen carry	M : *Most* postmen carry	(M, T) : *Most* postmen carry mail
		(M, A) : *Most* postmen carry oil
	F : *Few* postmen carry	(F, T) : *Few* postmen carry mail
		(F, A) : *Few* postmen carry oil
	M : *Almost all* postmen carry	(M, T) : *Almost all* postmen carry mail
		(M, A) : *Almost all* postmen carry oil
	F : *Almost no* postmen carry	(F, T) : *Almost no* postmen carry mail
		(F, A) : *Almost no* postmen carry oil

Table 1: An example from the dataset used in this paper where a backbone phrase is modified by quantifiers and followed by typical or atypical critical words.

in polar opposite meanings as can be seen on the right side of respective equations. *All Ps are Qs* means that all objects in the set P belong to the set Q, whereas *No Ps are Qs* means that P and Q are mutually exclusive sets. This minor distinction in the sentence has a drastic effect on the relationship between P and Q.

In this work, we aim to test and quantify the ability of LLMs to understand quantifiers and how this understanding changes as the models scale. We build upon the work of (Michaelov and Bergen, 2022), who test understanding and sensitivity of LLMs for *most*-type and *few*-type quantifiers. They do these tests on a dataset of 960 sentences created using a previously published study on human response (measured using N400 amplitude) to different quantifiers (Urbach and Kutas, 2010). They find that while LLMs do increasingly well on understanding *most*-type quantifiers, while their understanding of *few*-type quantifiers diminishes as the size of these language models increase. This is an example of an inverse-scaling law (McKenzie et al., 2022; Wei et al., 2022), where the model gets worse at doing a task as the model size increases. Inverse scaling laws are rare in natural language processing and important to identify, yet they must be cautiously evaluated (Wei et al., 2022).

In this paper, we first show that conclusions about the inverse-scaling of *few*-type quantifier comprehension in LLMs (Michaelov and Bergen, 2022) need to be revisited because of a possibly faulty methodology, thus leading to a wrong conclusion about inverse-scaling. We discuss the reasons for this in detail later in the paper. We then propose our own method of measuring *quantifier comprehension* in LLMs. We find that LLMs are able to differentiate between sentences that contain *most*-type versus *few*-type quantifiers quite well and this understanding improves as the model size increases. We measure this by quantitatively evaluating if the models react differently for different types of quantifiers. Although, when we evaluate if the model takes into account the meaning of a quantifier, we find that LLMs comprehend *few*-type quantifiers much better than *most*-type quantifiers. We also find that contrary to the results of (Michaelov and Bergen, 2022), *most*-type quantifier comprehension gets worse with increasing model size, thus showing an inverse-scaling law in *most*-type quantifier comprehension. In this study, we evaluate a number of different language model families, with models ranging from a size of 125 million parameters to 175 billion parameters, and find that the results are consistent for all LLMs.

2 Dataset and Models

The models and dataset used in this paper are identical to the ones used in (Michaelov and Bergen, 2022). This work uses the log probabilities produced by different language models to calculate a quantity called surprisal, which is introduced later in the paper. We do not make additional API calls or query models. We simply use the log probabilities released by (Michaelov and Bergen, 2022), thus mitigating differences due to experimental conditions. This paper aims to provide an alternative way of interpreting the output logits produced by different LLMs compared to (Michaelov and Bergen, 2022).

2.1 Dataset

We use the same dataset as used by (Michaelov and Bergen, 2022) which originates from a set of psycholinguistic experiments done on humans (Urbach and Kutas, 2010). The dataset consists of 120 different backbone phrases, which are modified by two sets of quantifier and completed by a typical and an atypical continuing word. An example can be seen in Table 1.

The backbone phrase shown in the example is ‘postmen carry’, which is modified by a *most*-type and a *few*-type quantifier. Following (Michaelov and Bergen, 2022), in this paper we study the effects of these two quantifiers and how LLMs interpret them. Each backbone phrase is modified by two *most*-type and two *few*-type modifiers. After the quantifiers are used to modify the backbone phrases, if the language model takes into account the meaning of the word, it should be more likely to produce a word with appropriate typicality. Words that are more typically associated with the backbone phrase are labelled *typical (T)*. For examples, the phrase *"postmen carry"* is typically followed by the word *mail* and not by the *atypical (A)* word *oil*. **We expect the language model to take into account the quantifier when assigning probabilities to the word following the quantifier-modified phrase**. Each backbone phrase modified by a quantifier is tested to be followed by a *typical* and an *atypical* word. The *typical/atypical* words are also together referred to as **critical words** in this paper.

The dataset contains a total of 960 sentences, with 120 unique backbone phrases, with 8 modifications to each sentence as shown in Table 1. We have 2 different quantifier types and two quantifiers per quantifier type, thus making four versions of each backbone phrase. Each quantifier-modified backbone phrase is followed by a typical and atypical word, thus making 8 sentences per backbone phrase.

These sentences were used to measure human brain response to critical words in association with the quantifier used (Urbach and Kutas, 2010). It was found that humans brain signals produce a spike when an atypical critical word is used with the *most*-type quantifier. This spike in brain activation (called N400 signals) are associated with unexpected events. Hence, these N400 spikes show that the atypical critical words when following a *most*-type quantifier were unexpected/incorrect. A lower activation is seen when the *most*-type quantifier is followed by a typical critical word. This spike in the N400 signal can be explained by a quantity called *surprisal*, which is the negative log-probability of the occurence of a word in that context. This means the less likely the word, the higher the surprisal. It was shown in (Michaelov and Bergen, 2020) that surprisal as measured in language models explain these N400 spikes very well, and that GPT-3 is the best single predictor of these N400 spikes in humans (Michaelov et al., 2023).

2.2 Models

To evaluate quantifier comprehension in LLMs, we use five family of models. We use the GPT2 model family (125M-1.5B parameters) (Radford et al., 2019), ElutherAI’s GPT models (GPT-Neo 125M, GPT-Neo 1.3B, GPT-Neo 2.7B and GPT-J 6B) (Black et al., 2022), the OPT model family (125M - 13B parameters), the GPT-3 model family (2B-175B parameters) and the InstructGPT model family (Ouyang et al., 2022) called GPT3.5 in the rest of the paper (2B-175B parameters).

3 Quantifier Comprehension in LLMs

In this section, we first present how (Michaelov and Bergen, 2022) measure quantifier comprehension in LLMs. Specifically, we present two ideas of *surprisal* and *quantifier accuracy* and ways to measure both properties as proposed by (Michaelov and Bergen, 2022). Alongside, we also highlight shortcomings of these quantifier comprehension evaluation methods.

3.1 Surprisal

As defined in section 2, *surprisal* is the negative log-probability of occurrence of a word given a context, as show below:

$$\mathrm{S}_p(w_i) = -\log p(w_i|w_1, \ldots, w_{i-1}) \qquad (1)$$

where w_i is the critical word under observation and $w_1, \ldots, w_{i-1}$ are the words preceding the critical word in a sentence. The underscore p in the surprisal represents that this is the definition of surprisal in prior work. (Michaelov and Bergen, 2022) acknowledge that words in language models are usually split into subwords. For scenarios when this happens for a critical word, (Michaelov and Bergen, 2022) suggest to sum up the suprisals of each individual subwords. This essentially means multiplying the probabilities of each subword that makes up the critical word. The use of this definition of surprisal is suboptimal as it does not take into account the effects of subword tokenization.

Previous work has shown that just summing up subword probability results in skewing of probability values towards words with shorter length, which is why these quantities are normalized by length (Brown et al., 2020). In our setting, this means the critical words split into larger number of subwords is likely to be assigned lower probability and thus higher suprisal than critical words that are split into fewer or no subwords. To normalize the effect of subword length, we propose normalizing the surprisal values by the subword length of the critical word, depicted by N, following previous works (Brown et al., 2020). Thus, we define surprisal as shown below:

$$\text{S}(w_i) = -\frac{1}{N} \sum_{\forall v_i \in \{w_i\}} \log p(v_i|w_1, \dots, w_{i-1}) \tag{2}$$

where w_i is the critical word split into a set of N-subwords represented by the set $\{w_i\}$ and v_i is a subword that belongs to that set. Surprisal can be understood as a term representing the inverse-probability of occuring of a word in a context. If a word has high probablity of occuring in a context, it will have low surprisal, whereas if a word has a low probablity of occuring in a context, it will have high surprisal. In this work, we will use our definition of surprisal.

3.2 Quantifier Accuracy

(Michaelov and Bergen, 2022) define quantifier accuracy based on the surprisal values for the critical word following a quantifier type. The quantifier accuracy test was motivated by the human brain response experiments done in (Urbach and Kutas, 2010). The aim of defining quantifier accuracy was to measure if language models take into account the meaning of quantifier words when creating the probability distribution over for the critical word. (Michaelov and Bergen, 2022) proposes that if LLMs take into account the meaning of quantifiers in a sentence, then the typical critical words will be predicted with larger probability and thus lower surprisal values following a *most*-type quantifier, and the atypical critical word will be predicted with larger probability and thus lower surprisal value with a *few*-type quantifier .

To illustrate this, we refer to the examples shown in Table 1. For the backbone prompt modified by a *most*-type quantifier - "*Most* postmen carry", an LLM is consider accurate if surprisal for the word *oil* is more than surprisal of the word *mail*, or in other words, $p(\text{mail} \mid \text{Most postmen carry}) > p(\text{oil} \mid \text{Most postmen carry})$. To succinctly express this, a sentence in the dataset is considered to be *most*-type accurate if for a ***m****ost*-type quantifier modified **b**ackbone **p**hrase (MBP),

$$\text{S}(typ|MBP) < \text{S}(atyp|MBP) \tag{3}$$

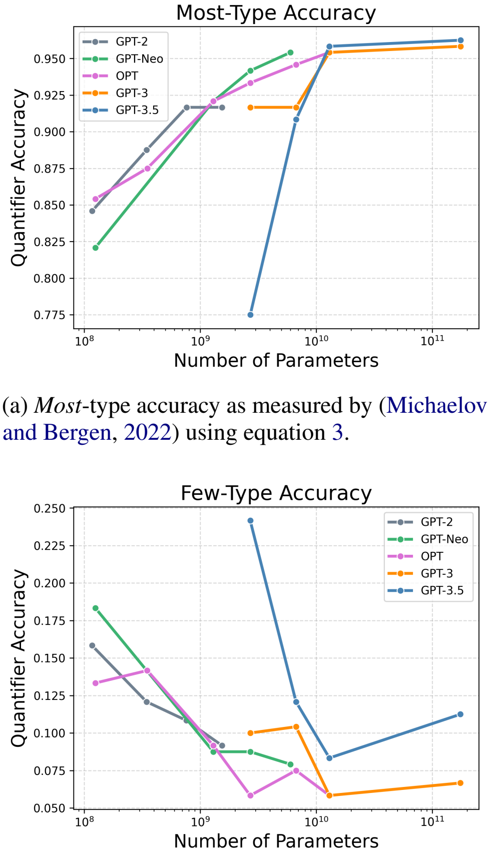
Similarly, for a backbone prompt modified by a *few*-type quantifier - "*Few* postmen carry", an LLM is considered accurate if $p(\text{oil} \mid \text{Few postmen carry}) > p(\text{mail} \mid \text{Few postmen carry})$. This means that the atypical word is more likely to occur with the *few*-type quantifier. Thus, a sentence is considered to be *few*-type accurate for a ***f****ew*-type quantifier modified **b**ackbone **p**hrase (FBP) if for that phrase,

$$\text{S}(atyp|FBP) > \text{S}(typ|FBP) \tag{4}$$

As proposed by (Michaelov and Bergen, 2022), the *most*-type and *few*-type quantifier accuracy is then calculated as the ratio of sentences following the above equations for the respective quantifiers. Figure 1 shows *most*-type and *few*-type accuracy for different LLMs as a function of the number of parameters in the model. We also see the inverse-scaling of *few*-type quantifier understanding very clearly. As shown by the plot, as the number of parameters increase, the *few*-type quantifier comprehension gets worse. Figure 1 is created using our normalized definition of surprisal taking into account the subword tokenization, and is thus slightly different from the original paper.

3.2.1 What's wrong with this way of defining quantifier accuracy?

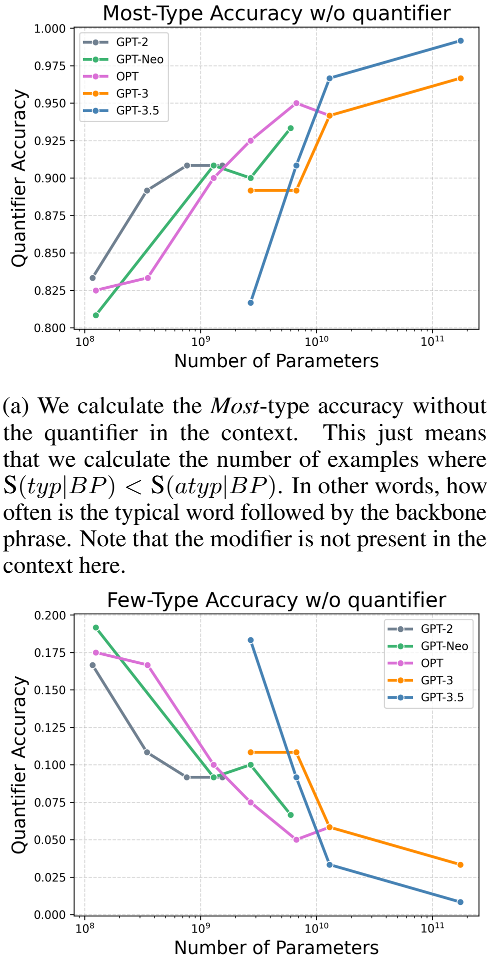
Quantifier accuracy as defined in equations 3 and 4 have a few drawbacks. The first is the assumption that *typicality* of a word for humans is the same as that for language models. A word deemed "typical" for a backbone phrase would indeed be in the top few words used by a human, but the same might not be true for language models. To experimentally confirm this, we analyse the output distribution of generated words following a backbone phrase. We find that the "typical" word in the dataset does not even fall into the top-100 most likely words following a backbone phrase for gpt-2 large. This



(a) *Most*-type accuracy as measured by (Michaelov and Bergen, 2022) using equation 3.

(b) *Few*-type accuracy as measured by (Michaelov and Bergen, 2022) using equation 3.

Figure 1: Quantifier accuracy as a function of model parameters for different models as defined in (Michaelov and Bergen, 2022).



(a) We calculate the *Most*-type accuracy without the quantifier in the context. This just means that we calculate the number of examples where $\mathrm{S}(typ|BP) < \mathrm{S}(atyp|BP)$. In other words, how often is the typical word followed by the backbone phrase. Note that the modifier is not present in the context here.

(b) We calculate the *Few*-type accuracy without the quantifier in the context. This just means that we calculate the number of examples where the atypical word is **not** followed by the backbone phrase, or $\mathrm{S}(atyp|BP) > \mathrm{S}(typ|BP)$.

Figure 2: Here we calculate the percentage of times the typical words occurs with larger probability than the atypical word in Figure 2a and vice versa in Figure 2b. These are similar to the quantities calculated in Figure 1 without the quantifier present in the context.

is true for ALL of the sentences in the dataset. **This shows that the typical token for humans is not necessarily typical for language models**.

The second assumption is that the chosen atypical word in the dataset is **the only** complementary word corresponding to the typical word. While the "typical" word is the most common follow up word for a given backbone phrase, we can have many alternative "atypical" words to follow the backbone phrase. For example, if we consider the phrase - "Most postmen carry ", the atypical word *oil* is just as atypical as the word *fish*. In fact, for GPT2-large, *fish* has a larger surprisal value compared to *oil*, which means according to GPT2-large, *fish* is more atypical than *oil* and is thus a more ideal candidate as an "atypical" word for comparison in equations 3 and 4. Just like the critical word *fish*, we can find many atypical words that are just as atypical if not more, than the chosen words in the dataset. **This means that if the given atypical word does not satisfy the equations 3 and 4, there might still exist an unknown number of other atypical words that might be able to satisfy this criteria**. These reasons renders the accuracy metric as defined by (Michaelov and Bergen, 2022) incorrect.

3.2.2 What do these scaling graphs actually measure?

Finally, we want to explain what the scaling in Figure 1 and (Michaelov and Bergen, 2022) actually depicts. To see this, we want to refer the reader to Figure 2, which shows the accuracy metric as defined in equations 3 and 4 for a critical word following a backbone phrase without the

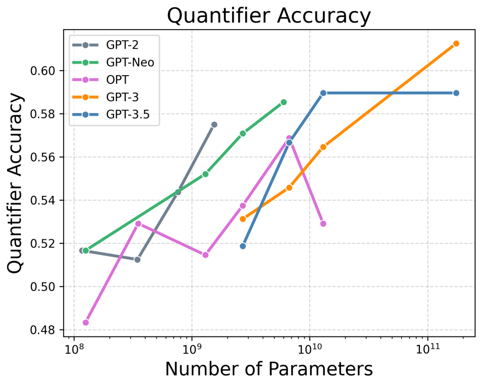


Figure 3: This figure shows that large language models get increasingly better at differentiating between *most*-type quantifiers and *few*-type quantifiers as they scale.

quantifier. This means that Figure 2a measures the count when $\mathrm{S}(typ|BP) < \mathrm{S}(atyp|BP)$, or how often is the typical word followed by the backbone phrase. Similarly, figure 2b measures $\mathrm{S}(atyp|BP) > \mathrm{S}(typ|BP)$, or how often the atypical word is **not** followed by the backbone phrase.

The scaling in Figure 2 looks almost identical to Figure 1. This indicates that the method defined by (Michaelov and Bergen, 2022) to measure the effect of quantifier is not even accounting for the presence of the quantifier, and **we end up just measuring how often the typical word is more probable than the atypical word**. Thus, the method proposed to evaluate quantifier comprehension using equation 3 and 4 in (Michaelov and Bergen, 2022) is not actually measuring quantifier comprehension, **it is measuring typicality**.

In fact, what these scaling plots show is that as the size of the model increase, the typical words in LLMs get more probable and the atypical words get less probable. This essentially means that the model is getting better at understanding language as typically used by humans, and is able to associate the typical word in a given context with larger probability than the atypical words.

4 Proposed Evaluation of Quantifier Comprehension in LLMs

In this section, we present a more robust way of measuring quantifier comprehension in LLMs. Measuring quantifier comprehension in LLMs in the setting defined by (Michaelov and Bergen, 2022) has to be grounded in the principle that the typical and atypical words chosen in the dataset are not unique, and hence to measure the effect of presence of quantifier in a context, we should do measurements on the same critical word. We propose two tests do this.

4.1 EXPERIMENT-1 : Differentiating Between Different Types of Quantifiers

In this section, we check if the models are able to differentiate between the meanings of two types of quantifiers and react appropriately. To check this, we fix a critical word (either typical or atypical), and change the quantifier and see how the surprisal value of the critical word is affected. We expect that when we have a typical critical word, the *few*-type quantifier should lead to a higher surprisal value or make the typical word less probable. For example, for the phrase *"Most/Few postmen carry mail"*, the surprisal for the word mail should be more when accompanied by a *few*-type quantifier than when compared to a *most*-type quantifier. Similarly, for an atypical word, surprisal values for *most*-type quantifiers should be larger than when observed with *few*-type quantifiers. In summary, an LLM is able to differentiate between two types of quantifiers if for a critical word, one of the following is true depending on the type of critical word under observation:

$$\mathrm{S}(typ|MBP) < \mathrm{S}(typ|FBP) \tag{5}$$

$$\mathrm{S}(atyp|MBP) > \mathrm{S}(atyp|FBP) \tag{6}$$

The results of Experiment-1 are shown in Figure 3. We see that LLMs get increasingly better at differentiating between the two types of quantifiers and are able to adapt their output probability distribution at the critical word to reflect this understanding. This improvement of quantifier comprehension scales with increasing model size just like other capabilities of LLMs. Although the absolute value of quantifier accuracy peaks only at about 61% for the 175 billion parameter GPT-3 model, which shows that for a majority of sentences, the meaning of the quantifier is not reflected in the output probability distribution at the critical word. This shows that although LLMs are getting better at understanding quantifiers as they scale, they are far from perfect.

4.2 EXPERIMENT-2: Measuring Quantifier-Specific Accuracy

Here we want to measure how good LLMs are at understanding a specific quantifier. To measure this, we compare how the surprisal of a critical word is affected as we add a quantifier in the context. When we add *most*-type quantifiers, the surprisal should decrease for a typical word whereas it should increase for an atypical word. In other words, a sentence is accurate for *most*-type quantifier comprehension if:

$$S(typ|MBP) < S(typ|BP) \quad (7)$$

$$S(atyp|MBP) > S(atyp|BP) \quad (8)$$

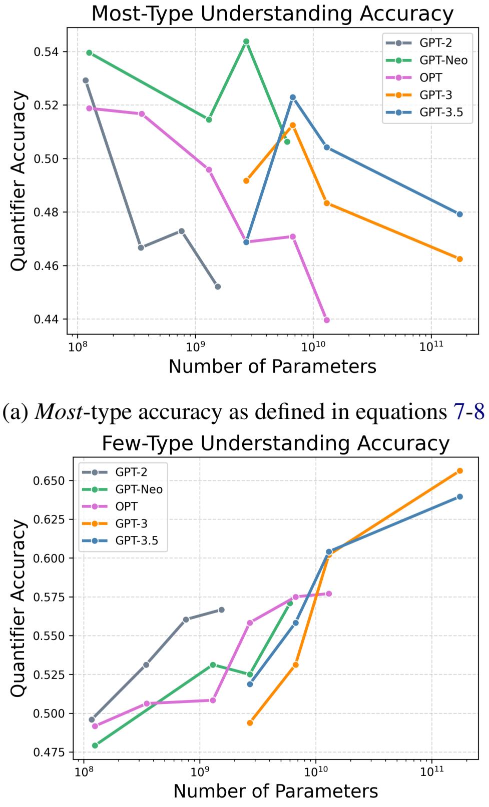
Here, MBP is a ***m**ost*-type quantifier modified **b**ackbone **p**hrase, such as *"Most postmen carry"* and BP is just a **b**ackbone **p**hrase without modifier, such as *"Postmen carry"*. Similarly, for *few*-type quantifiers, the surprisal should decrease for atypical critical words and increase for typical words. Specifically, sentence is considered accurate for a *few*-type quantifier comprehension if:

$$S(typ|FBP) > S(typ|BP) \quad (9)$$

$$S(atyp|FBP) < S(atyp|BP) \quad (10)$$

Figure 4 shows the quantifier-specific comprehension ability of models as defined in equations 7-10. Although section 4.1 showed that models are able to differentiate between *most*-type and *few*-type quantifiers, we see in Figure 4 that they don't necessarily incorporate the meaning of quantifiers when quantifiers are added to a sentence. We see that LLMs become increasingly better at incorporating the meaning *few*-type quantifiers as model size increases by changing the relative probability values of the critical words given the change in context. But this is not observed in the case of *most*-type quantifiers, where we find that the models get increasingly worse at taking into account quantifier meaning, thus showing an **inverse-scaling in *most*-type quantifier comprehension**. This shows that the model gets increasingly worse at understanding *most*-type quantifier as the size of the model increases.

Note that in this work, to calculate suprisal, we never compare two different critical words as can be seen in equations 5-10. This circumvents any affects due to subword tokenization and the non-uniqueness of the chosen critical words in the dataset. All the comparisons are made with respect to a single critical word.



(a) *Most*-type accuracy as defined in equations 7-8

(b) *Few*-type accuracy as defined in equations 9-10

Figure 4: Quantifier specific accuracy as defined in equations 7-10.

5 Discussion

The above two tests for evaluating quantifier comprehension in LLMs show that these models are far from perfect. The underlying premise of the method used in this paper and (Michaelov and Bergen, 2022) is that the presence of a quantifier should increase or decrease the probability of a critical word depending on its typicality (Michaelov and Bergen, 2022). But both tests described in section 4 show that this is not ubiquitously observed. The accuracy numbers for both tests are around 50-60%, which means that the probability distributions do not incorporate quantifier meaning for a large majority of sentences. A test like this makes a fundamental assumption that understanding of meaning can be measured by studying the relative ranking of tokens in the generated word logit. While this is a fair assumption, we think it is necessary to explicitly point this out

Incorporating quantifier meaning in this way is not a necessary condition for models to perform well, as can be seen by their consistent improvement across different benchmark (Wang et al., 2018, 2019; Brown et al., 2020; Touvron et al., 2023). Also, it has been shown in previous studies that humans are not that great at quantifier comprehension as well (Urbach and Kutas, 2010), and continue to have a preference towards the more typical word in a context irrespective of the quantifier. These observations suggest two things. Firstly, that LLMs are not good at quantifier comprehension. Secondly, we also observe this lack of sensitivity to quantifier meaning in humans. This combined with the fact that despite lack of quantifier comprehension, LLMs get increasingly better at language understanding, we can argue that quantifier comprehension is not as necessary of a task in language processing and understanding as we thought it was.

6 Related Work

Inverse scaling laws were introduced as a competition (McKenzie et al., 2022) to incentivize research towards finding scenarios where language models get worse as their size increases. As the field of NLP moves towards scaling models to larger and larger sizes, it is important to know the scenarios where this scaling becomes detrimental (Wei et al., 2022; McKenzie et al., 2023).

As language models get increasingly better, some common linguistic tests that they are put through revolve around understanding negation and quantifiers. Studying the affects of negation has been the subject of focus for many studies (Kassner and Schütze, 2019; Kalouli et al., 2022; Ettinger, 2020) for different encoder-based masked language models. These studies find that these language models are not sensitive to negations. Studies on quantifiers (Kalouli et al., 2022) also seem to show similar results for masked language models. (Michaelov and Bergen, 2022) was the first work to study the quantifier understanding in decoder-based LLMs.

7 Conclusion

In this paper, we conduct a study to evaluate how well large language models understand quantifiers. Specifically, we study two types of quantifiers - *most*-type and *few*-type quantifiers. We present a set of experiments to evaluate quantifier comprehension of large language models and show that these models are able to differentiate between *most*-type and *few*-type quantifiers as they scale. We also show that LLMs struggle incorporate the meaning of *most*-type quantifier comprehension when compared to *few*-type quantifiers. We also show that *most*-type quantifier comprehension demonstrates an inverse-scaling law and their understanding of *most*-type quantifiers get worse as the model size increases. This study indicates that LLMs do not take into account the meaning of quantifiers that strongly, as shown by low accuracy scores in Figures 3 and 4. Even so, these models get increasingly better at language understanding tasks, thus indicating that quantifier understanding might not be the best test to evaluate language understanding in LLMs.

Acknowledgements

This paper was prepared for informational purposes in part by the Artificial Intelligence Research Group of JPMorgan Chase & Co and its affiliates ("J.P. Morgan") and is not a product of the Research Department of J.P. Morgan. J.P. Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy, or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer, or solicitation for the purchase or sale of any security, financial instrument, financial product, or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person if such solicitation under such jurisdiction or to such person would be unlawful.

References

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

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

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can large language models truly understand prompts? a case study with negated prompts. In *Transfer Learning for Natural Language Processing Workshop*, pages 52–62. PMLR.

Salud María Jiménez-Zafra, Roser Morante, M Teresa Martín-Valdivia, and L Alfonso Urena Lopez. 2020. Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1):1–52.

Aikaterini-Lida Kalouli, Rita Sevastjanova, Christin Beck, and Maribel Romero. 2022. Negation, coordination, and quantifiers in contextualized language models. *arXiv preprint arXiv:2209.07836*.

Nora Kassner and Hinrich Schütze. 2019. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Ian McKenzie, Alexander Lyzhov, Alicia Parrish, Ameya Prabhu, Aaron Mueller, Najoung Kim, Sam Bowman, and Ethan Perez. 2022. The inverse scaling prize.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. 2023. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*.

James A Michaelov, Megan D Bardolph, Cyma K Van Petten, Benjamin K Bergen, and Seana Coulson. 2023. Strong prediction: Language model surprisal explains multiple n400 effects. *Neurobiology of Language*, pages 1–71.

James A Michaelov and Benjamin K Bergen. 2020. How well does surprisal explain n400 amplitude under different experimental conditions? *arXiv preprint arXiv:2010.04844*.

James A Michaelov and Benjamin K Bergen. 2022. 'rarely'a problem? language models exhibit inverse scaling in their predictions following'few'-type quantifiers. *arXiv preprint arXiv:2212.08700*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Jason Wei, Yi Tay, and Quoc V Le. 2022. Inverse scaling can become u-shaped. *arXiv preprint arXiv:2211.02011*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.