NPIs Aren't Exactly Easy: Variation in Licensing across Large Language Models

Deanna DeCarlo[*1] **William Palmer**[*1] **Michael Wilson**[2] **Bob Frank**[1]
[1]Department of Linguistics, Yale University
[2]Department of Linguistics & Cognitive Science, University of Delaware
{deanna.decarlo, w.palmer, bob.frank}@yale.edu; mawilson@udel.edu

Abstract

We examine the licensing of negative polarity items (NPIs) in large language models (LLMs) to enrich the picture of how models acquire NPIs as linguistic phenomena at the syntax-semantics interface. NPIs are a class of words which have a restricted distribution, appearing only in certain licensing contexts, prototypically negation. Unlike much of previous work which assumes NPIs and their licensing environments constitute unified classes, we consider NPI distribution in its full complexity: different NPIs are possible in different licensing environments. By studying this phenomenon across a broad range of models, we are able to explore which features of the model architecture, properties of the training data, and linguistic characteristics of the NPI phenomenon itself drive performance.[1]

1 Introduction

Negative polarity items (NPIs) are words or phrases that must be licensed by another element, often negation, that occurs in a syntactically appropriate context (Klima, 1964). Determining the contexts in which such elements are possible has proven to be a difficult problem for language models. Marvin and Linzen (2018) report that LSTM language models fail to systematically distinguish grammatical from ungrammatical occurrences of NPIs across a range of difficult cases. Warstadt et al. (2019) study BERT's ability to determine the possibility of an NPI occuring in a masked position and find improved performance, though success depends upon the mode of evaluation. Hu et al. (2020) report even better results with GPT-2 and GPT-2-XL. Finally, Zhang et al. (2021) show that with sufficiently many parameters and enough training data, as found in RoBERTa-Large (Zhuang et al., 2021), a transformer language model can achieve near-human levels of performance on NPI licensing.

Though this might be thought of as a success story for LLMs, we may still wonder why this particular grammatical regularity has posed such difficulty, as compared to subject-verb agreement, where smaller LSTMs trained on less data achieved quite good performance (Marvin and Linzen, 2018). We suspect that the reasons for this are threefold. First, unlike grammatical subjects that condition agreement on their corresponding verbs in reasonably predictable ways, the set of contexts that license NPIs and the range of NPIs are both rather diverse. Further, while every instance of a finite clause will include a subject and agreeing verb, many contexts that could license NPIs do not include one. Finally, NPI licensing and subject-verb agreement are both dependencies that are unbounded by linear distance; however, the structural distance between a licensing context and an NPI can grow without bound (cf. *I* ***don't*** *see* ***anyone*** and *I* ***don't*** *want to try to see* ***anyone***), unlike the dependency that determines subject-verb agreement.

Previous studies of LLM performance on NPIs, including Warstadt et al. (2019) and Jumelet et al. (2021), have examined the first of these factors: the variability of the licensing environment. Specifically, these studies explored the degree to which the licensing properties of distinct environments are encoded uniformly, with what look like reasonably promising results. Such work assumes implicitly that different environments should be treated identically (though see Bylinina and Tikhonov (2022) for work that does not make this assumption). Similarly, LLM evaluations on NPIs have assumed that different NPIs are licensed in identical environments. However, as we will discuss in Section 2, these assumptions are false: different environments license different NPIs. Learning the distribution of NPIs is thus more complex than previous LLM evaluations have assumed. We aim to develop an

[*] These authors contributed equally to this work

[1] Our code and dataset are available at github.com/clay-lab/condgen-evaluation.

Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 332–341
December 7, 2023. ©2023 Association for Computational Linguistics

approach to evaluate LLMs' knowledge of NPIs in a way that is sensitive to their unique distributional patterns, and to uncover what factors lead to greater success in a model's ability to correctly determine the possibility of an NPI in a given context.

2 Variability in NPI Licensing

As already noted, NPIs are expressions that are only grammatical in a restricted set of contexts, prototypically understood to be negative. Canonical examples of such contexts include the negative quantifiers *no* or *none of the*, or sentential negation *not*. As seen in (1), the English NPI *ever* is possible when it is in the scope of such an element, and ungrammatical otherwise.

(1) a. **No/None of the** packages had *ever* arrived at the yellow house.

b. Packages had **not** *ever* arrived at the yellow house.

c. * Packages had *ever* arrived at the yellow house.

As seen in (2), the NPI *ever* is also licensed by other contexts, including (indirect) yes/no questions, the restrictor of superlatives, and under the scope of *only*, among many others.

(2) a. I wonder **whether** the packages had *ever* arrived at the yellow house.

b. These are the **greatest** packages that had *ever* arrived at the yellow house.

c. **Only** packages had *ever* arrived at the yellow house.

A major step forward in our understanding of the distribution of NPIs came from attempts to characterize these licensing environments in a uniform fashion (Ladusaw, 1979). However, it was quickly observed that not all NPIs are licensed by the same contexts. For example, the English NPI (adverbial) *any* is licensed by negation and indirect yes/no questions but not by superlatives.

(3) a. **No** masons build cathedrals *any* better than that.

b. * These are the **greatest** masons that build cathedrals *any* better than that.

c. I wonder **whether** the masons have built a cathedral *any* better than that.

NPIs like *exactly* are even more restrictive, occurring only with negation:

(4) a. **None of the** students have *exactly* been getting good grades.

b. * These are the **smartest** students that have *exactly* been getting good grades.

c. * I wonder **whether** the students have *exactly* been getting good grades.

Recent research in formal semantics has aimed to understand this variation. Under the proposal of Zwarts (1998), which was further refined in Giannakidou (1998), licensing contexts are characterized according to their semantic properties. Zwarts and Giannakidou provide four increasingly demanding semantic criteria for characterizing contexts, each of which entails the previous one. Such a semantic classification allows us to characterize the distribution of different NPIs. Each NPI is associated with a certain minimal requirement on its licensing context, and will therefore be allowed in all more restrictive contexts. This creates a hierarchy of NPIs, ranging from superweak NPIs, which require environments satisfying only the weakest condition, to superstrong NPIs, which are require environments satisfying the strongest condition.

Elegant as this approach is, Hoeksema (2012) shows that this classification is not completely adequate, as it does not capture the full complexity and diversity in the distribution of different NPIs. Table 1 reports Hoeksema's characterization of licensing contexts for a number of NPIs.[2,3] Among other things, this table demonstrates that the relationship between the set of licensing contexts for different NPIs does not follow the subset-superset relationship that would be expected from the proposal just outlined: the set of contexts that license *yet* is neither a subset nor a superset of those licens-

[2] The data in this table reflects Hoeksema's reports from the NPI literature and his own corpus analysis. The authors, all native English speakers, have checked and agree with these judgments. Following Bylinina and Tikhonov (2022), we believe it would be useful to compare model performance to experimental measures of NPI acceptability. While there is a rich body of experimental work on this topic, including Chemla et al. (2011), Geurts (2003), and Denić et al. (2021), none of these studies consider the range of contexts and NPIs explored in the current work, so detailed comparison with human judgments and behavior will need to wait for future research.

[3] We exclude from consideration expressions that are not uniquely identifiable as NPIs from their position in the sentence, as opposed to their interpretation (e.g., *either*, *can help*). Furthermore, we also exclude NPIs that consist of multiple words (e.g., *at all* or *in years*), since, as an anonymous reviewer pointed out, it is a non-trivial matter to assess whether a language model "accepts" them.

	1	2	3	4	5	6	7	8	9
any	+	+	+	+	+	+	+	+	+
ever	+	+	+	+	+	+	+	+	+
remotely	+	+	+	+	+	+	+	-	+
adv. any	+	+	+	+	+	+	+	-	+
yet	+	+	+	-	+	-	+	+	+
anymore	+	-	-	-	-	-	+	-	-
squat	+	-	-	-	-	-	+	-	-
exactly	+	-	-	-	-	-	-	-	-

Table 1: Licensing contexts for English negative polarity items (modified from Hoeksema 2012). Contexts: 1 = negation, 2 = indirect y/n questions, 3 = matrix y/n questions, 4 = *wh*-questions, 5 = conditional clauses, 6 = universal restrictors, 7 = *the only* restrictor, 8 = superlative restrictors, 9 = scope of *only*

ing adverbial *any*. Nonetheless, the distribution of NPIs instantiated in this table is something that a language model should master. Further, we may expect that differences in restrictiveness, both of NPIs (in terms of the number of contexts in which they are licensed) and contexts (in terms of the number of NPIs they license) have an impact on the feasibility of learning the distributions. We turn now to exploring these questions.

3 Experiment

3.1 Models

In recent years, LLMs have been developed with a variety of architectures, model sizes and training datasets. While smaller models with smaller datasets are easier to train and work with, larger models with larger datasets typically perform better on linguistic tasks. In the current work, we consider as broad a range of LLMs as was feasible, with the limitation that many state-of-the-art models are proprietary and do not provide access to the detailed information our experiments require. Specifically, we consider three broad classes of transformer architectures: Encoder-only Masked Language Models (MLMs), Decoder-only Language Models (LMs), and Encoder-Decoder Sequence to Sequence Models (Seq2Seqs). Parameter counts in these models ranged from 11 million to 175 billion, and training data ranged from from 6 GB to 4.7 TB. Details of the models studied are in Table 2.

3.2 Materials and Methods

We construct a test dataset that includes each of the 8 NPIs and 9 contexts listed in Table 1. To these contexts, we add an additional Null context that does not license any NPIs. For each NPI, we create 6 distinct sentence templates each of which can be prefixed by a carrier of the licensing context. Some contexts support more than one carrier prefix, which yielded at total of 12 distinct instantiations per sentence template. An example of how this works is shown in Table 3. In total, there are 576 sentences in the test dataset.

Task	Architecture	# Models	# Params. (M)	Dataset Size (GB)
MLM	AlBERT	8	11 - 206	6
MLM	BERT	6	66 - 335	49
MLM	MultiBERTs	25	110	49
MLM	RoBERTa	4	82 - 355	16
MLM	Electra	3	14 - 51	14
LM	LLaMA	4	6738 - 65286	4700
LM	OPT	9	125 - 174604	800
LM	GPT2	4	124 - 1558	55
Seq2Seq	T5 Efficient	26	16 - 11307	305

Table 2: LLMs used in current experiment.

3.3 Testing Procedure

Testing of the different model types proceeds in slightly differently fashions. For MLMs, we replace the NPIs in the test examples with mask tokens, feeding the resulting string to the model. We then extract the log probability of the NPI at the position of the mask token. For the Seq2Seq models, which were trained on span-mask denoising, we used a similar procedure, giving the appropriate masked sequence to the model, and then extracting the log probability of the NPI given by the decoder as the filler for the mask. Autoregressive LMs are tested by truncating the sentence to the position immediately before the NPI. We then feed this sequence to the model, using teacher, and then obtain the log probability for the NPI at the following token position. In all cases, we ensure that the NPIs under study constitute single tokens in the model vocabulary.[4]

It is immediately clear that there is an asymmetry between the MLMs and Seq2Seq models on the one hand and the LMs on the other: the former models see both the left and right context in assessing the likelihood of the NPI, while the LMs only see left context. We have done our best to construct stimuli in which the right context provides no information about the possible presence of an NPI (the right context is fully acceptable in the Null context in the absence of an NPI), as this would penalize the

[4] Cases where NPIs were more than one token and thus excluded were the following: DistilBERT Base Cased, BERT Base Cased, and BERT Large Cased lacked *squat*; all LLaMA models lacked *squat* and *remotely*; and all Seq2Seq models (which were all T5 models) lacked *squat*.

Licensing Context	Carrier prefix(es)	*any* Example
*Null	0, The	*Laws have done *any* harm.
Negation	No, None of the	**No** laws have done *any* harm.
Indirect y/n question	I wonder whether the	**I wonder whether the** laws have done *any* harm.
Matrix y/n question	Is it likely that	**Is it likely that** laws have done *any* harm.
Indirect *wh*-question	I wonder which	**I wonder which** laws have done *any* harm.
Conditional clauses	They will notify everyone if the	**They will notify everyone if the** laws have done *any* harm.
Universal restrictor	These are all of the <that>	**These are all of the** laws **that** have done *any* harm.
The only restrictor	These are the only <that>	**These are the only** laws **that** have done *any* harm.
Superlative restrictor	These are the greatest <that>	**These are the greatest** laws **that** have done *any* harm.
Scope of *only*	Only	**Only** laws have done *any* harm.

Table 3: Test examples created from the template *laws have done any harm* with the italicized NPI *any*. For each licensing context, this template is prefixed by one or more of the carriers in bold to produce the test.

LMs.

3.4 Analysis via Point-Biserial Correlations

There is no absolute probability that can tell us whether a model licenses an NPI in a particular context. Instead, we must compare relative probabilities: how much more (or less) likely is an NPI in a particular licensing context compared to a context that does not license any NPIs? For this reason, we "adjust" the probability by subtracting from it the probability of the Null context, which we know is not a licensing context for any NPI.[5] The resulting value tells us whether an NPI in a particular context is more or less probable compared to a minimally different context that does not license any NPIs: a positive value indicates that an NPI is predicted to be more likely than in the baseline context (i.e., the model "licenses" it in that context to some degree), while a 0 or negative value indicates the opposite (i.e., the model does not license it in that context).

To explore to what degree LLMs are sensitive to the contours of NPI distributions in the same way that humans are, we compute the point-biserial correlation between the (adjusted) model log probabilities extracted during testing and the dichotomous human judgments given in Table 1. The analysis from here on is bifurcated into evaluation by NPI and by licensing context. For each analysis, the log probabilities obtained in Section 3.3 are grouped by model instance and NPI or context. The point-biserial correlation is performed separately for each probability group with a binary encoding of the human judgments from Table 1 using the following equation:

$$r = \frac{\bar{x}_1 - \bar{x}_0}{s_x} \cdot \sqrt{\frac{n_1 \cdot n_0}{n^2}}$$

For both analyses (NPI and context), n and s_x have the same interpretation:

n = # of test items
s_x = s.d. of human binary judgments

For analysis by an NPI N, the interpretation of these variables is as follows:

$\bar{x}_1$ = mean adj. log prob of N in items that license it
$\bar{x}_0$ = mean adj. log prob of N in items that do not license it
n_1 = # of items in which N is licensed
n_0 = # of items in which N is not licensed

We can take n_1 to be a rough proxy for strength of the NPI: a weak NPI will have a higher value, while a strong NPI will have a lower one. [6]

For analysis by a licensing context C, the interpretation is:

$\bar{x}_1$ = mean adj. log prob across NPIs licensed by C
$\bar{x}_0$ = mean adj. log prob across NPIs not licensed by C
n_1 = # of C items with licensed NPI
n_0 = # of C items with unlicensed NPI

Note that in the analysis by licensing context, the negation context is excluded as there is no variance in its associated human licensing judgments; it licenses all of the NPIs.

3.5 Beta Regressions

In order to understand what factors are responsible for the variation in correlations across different NPIs and contexts, as well as across models, we perform a number of beta regressions. Because beta regressions require a dependent variable in the range $[0, 1]$ and our values are correlations with a possible range of $[-1, 1]$, we scale them to the appropriate range by adding 1 and dividing by 2.

[5] We then exclude the adjusted probability of the NPI in the baseline context from further analysis, as it is always 0 following this procedure.

[6] We recognize that this quantification of NPI strength flattens distinctions among NPIs that are not characterized in terms of a subset-superset relationship among licensing environments, as seen in "bagel" environments, though such elements do not exist in English.

Context	Occurrences
Indirect y/n question	266
Matrix y/n question	358
Indirect *wh*-question	866
Conditional clauses	1960
Universal restrictor	755
The only restrictor	187
Superlative restrictor	807
Scope of *only*	142

Table 4: Frequency of licensing contexts predictions in the parsed Penn Treebank datasets Brown and WSJ.

For each of the correlation data sets (by NPI and by context), we run two types of regressions. In the first, we regress the scaled correlations on the log of the number of parameters in a given model and a linguistic quantity we call *licensing number*. We define the *licensing number of an NPI* as the number of distinct environments that license it according to Table 1. For instance, the licensing number of *any* is 9, while the licensing number of *exactly* is 1. Similarly, we define the *licensing number of a context* as the number of NPIs it licenses; for example, the licensing number of negation is 8, while the licensing number of superlative restrictors in 3. Because some of the LLMs we evaluate do not include particular NPIs as single tokens, we convert the licensing number of contexts to a ratio by dividing it by the number of NPIs that occur in the model's vocabulary as a single token. This gives us the proportion of the available NPIs that a model licenses in a context. As a second type of regression, we use as predictor variables the individual NPIs or contexts for the analysis by NPI or context, respectively. NPIs and contexts are converted to one-hot encodings, and the resulting vectors are used as predictors.

3.6 Context Frequencies

We also considered as an additional predictor of model correlations the frequencies of the different licensing contexts. To do this, we used the Brown and WSJ parsed datasets from the Penn Treebank (Marcus et al., 1999) to estimate the frequency of our licensing contexts in natural text, which we expect to be indicative of the frequency of the licensing contexts in the models' training corpora. We searched the datasets using Tregex (Levy and Andrew, 2006). Due to inconsistencies in assigned structures in the corpus, the frequencies reported in Table 4 are imprecise, but we believe that they are reasonably representative.

Ideally, one would determine the frequencies of NPIs in natural text as well. However, such a pursuit is difficult, since many NPIs have non-NPI uses that may occur in non-licensing environments. For example, *any* lives a double life as an NPI and as a word indicating "free-choice":

(5) a. Nobody had any questions. (NPI)
b. Pick a card, any card! (free-choice)

What's more, a possible NPI appearing in the scope of a licensor is insufficient to ensure it is interpreted as an NPI:

(6) John isn't remotely working.

Here, *remotely* can be read as an NPI, with the resulting interpretation that John isn't even close to doing anything that could be considered working. However, it could also be interpreted literally, as saying that John is working in-person. To our knowledge, all (English) NPIs suffer from one type of ambiguity or another in a similar way. Searching for NPIs in corpora is thus not as straightforward as one might hope because it involves not only the relatively simple task of finding specific words, but also the more complicated task of determining how those words are meant to be interpreted in a particular context.

4 Results

4.1 By NPI

Figure 1 shows the result of the correlations by NPI, with parameter count as the independent variable. Here we see considerable variation in performance across NPIs, particularly in the LMs and Seq2Seq models, where *ever* has a relatively high correlation across model sizes, while *yet* has a much lower correlation.

A beta regression on the licensing number and parameter count, shows a significant positive effect of licensing number ($\beta = 0.196, p < 0.001$), as well as a significant positive effect of the number of parameters ($\beta = 0.230, p < 0.001$). When we investigate our results by model type (LM, MLM, Seq2Seq) separately, we find that the relationships hold only for Seq2Seq models, but not for MLM and LM models, with a positive relationship for licensing number ($\beta = 0.190, p < 0.05$) and a positive relationship for number of parameters ($\beta = 0.226, p < 0.05$).

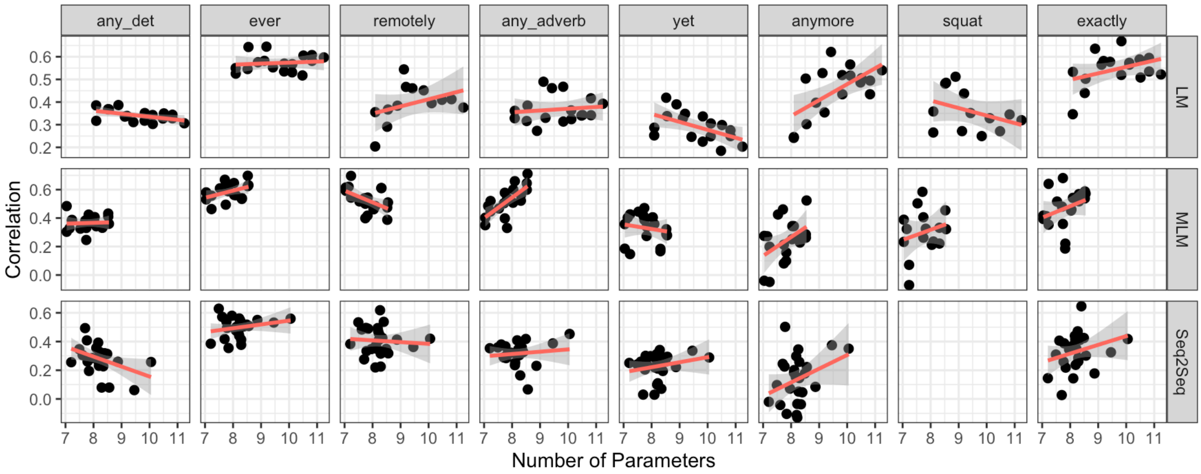

Figure 1: Relationship between the number of model parameters and the correlation between model predictions and human judgments for NPIs. NPIs are presented from most to least licensed, from left to right.

We also run a series of beta regressions to evaluate the relative performance of different NPIs. To do this, we code the NPIs as separate one-hot predictors, and regress on all but one of the one-hot vectors in turn. The omitted vector can be interpreted as the regression model's "baseline," with effects associated with the other predictors revealing how model performance on the associated NPIs compare to this baseline. By leaving out each NPI, we obtain a partial ordering that describes the relative degree to which the models' predictions accord with human judgments.

Figure 2 shows the partial orderings obtained for each model type, represented as Hasse diagrams. The sequence from left to right represents MLMs, LMs, and Seq2Seqs. The visual ranking of one NPI over another indicates that the models' judgments for the higher ranked NPI more closely match human judgments than the lower ranked NPI. Our regression results for the whole dataset are reflected in the MLM Hasse diagram as NPIs with higher licensing numbers (*ever*, *remotely*, and *any (adv.)*) generally appear nearer to the top, while NPIs with lower licensing numbers (*squat* and *anymore*) appear toward the bottom. An interesting exception is that *exactly* has the lowest licensing number, being licensed by only 1 context, yet it is one of the highest ranked NPIs.

For LMs, we similarly see many NPIs in a position in the Hasse diagram consistent with the regression results. For example, *ever* and *exactly* appear again at the top. Additionally, we see that *any (adv.)* is in the middle of the LM Hasse diagram, just as it is in the other two diagrams.

For Seq2Seqs, we see *ever* and *remotely* at the top, in a similar position as in the MLM and LM Hasse diagrams. We also see that *yet*, *anymore*, *exactly*, and *any (det.)* have many of the same or similar relative orderings as in the MLM Hasse diagram.

4.2 By Context

The correlation results broken down by model type and context are illustrated in Figure 3. We see that there is considerable variability across contexts, particularly in the LM results, with *Indirect Wh-Q*s having a correlation near 0, and *Indirect Y/N-Q*s and *Conditional*s having a correlation generally above 0.5.

A beta regression on the results by context with the licensing number, parameter count, and their interaction as independent variables found significant effects of licensing number ($\beta = -3.286$, $p < 0.001$) and number of parameters ($\beta = -0.175$, $p < 0.05$). These effects were qualified by a significant interaction of licensing number and number of parameters ($\beta = 0.302$, $p < 0.01$). The directionality of these effects indicates that while smaller models tend to display behavior less correlated with human judgments for contexts that license more NPIs, this penalty decreases for larger models [7]. Additionally, a regression on context frequency

[7] It is worth noting that the licensing contexts in our study exhibit limited variability, typically licensing 4 to 5 NPIs. Notable exceptions in our study are *The Only Restrictor* and *Superlative Restrictor*, which license 7 and 3 out of 8 NPIs, respectively. As such, our results are sensitive to the choice of the set of NPIs to a good degree.

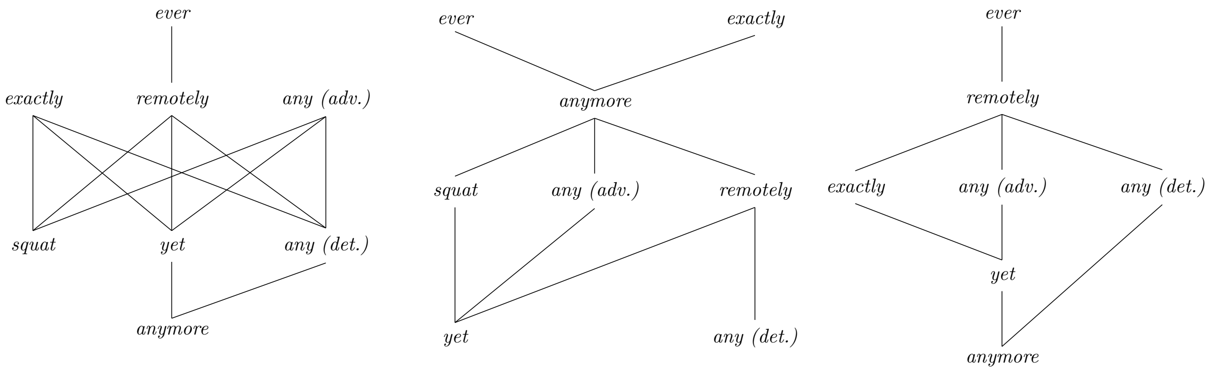

Figure 2: The diagrams are presented in the following sequence, from left to right: MLM, LM, and Seq2Seq. These Hasse diagrams depict the partial ordering of NPIs based on the results of the beta regressions performed on the one-hot encodings. The orderings are derived from correlations between the NPIs and human licensing judgments, with certain NPIs demonstrating notably higher correlations. These diagrams utilize vertical positioning to visually represent the relative ordering: an NPI positioned higher in the diagram indicates a stronger correlation with human judgments compared to any NPI reachable by following a downward line. As an illustration, in the Seq2Seq diagram, *ever* occupies the topmost position. Every other NPI can be reached by tracing a line downward from *ever*, signifying its comparatively greater correlation with the human judgments. NPIs that are not connected in this manner did not exhibit any statistically significant relative relationships in the regression results.

found significant positive effects ($\beta = 0.210$, $p < 0.001$).

As with the NPIs, we perform beta regressions on the licensing contexts encoded as one-hot vectors, excluding each licensing context in turn as a "baseline." We thus obtain a partial ordering of model performance on particular contexts across all NPIs that a context licenses.

Figure 4 shows the partial orderings of the licensing contexts obtained for each model type. The sequence from left to right represents MLMs, LMs, and Seq2Seqs. MLMs' predictions most closely correlate with human judgments for *Conditional*s, with the least similarity found for the licensor *The Only Restrictor*. Figure 4 also demonstrates that LMs' performance most closely accords with human preferences in the *Conditional* and *Indirect Y/N-Q* contexts and least so in the *Indirect Wh-Q* context. For the Seq2Seq model, performance is best for the *Conditional*, *Superlative Restrictor*, and *Indirect Wh-Q* contexts.

Across all models, *Conditional* licensing environments are associated with the highest model performance. *Indirect Y/N-Q* and *Universal Restrictor* contexts tend to be associated with an upper middling performance across the model tasks. Other licensing environments, namely *Superlative Restrictor* and *Indirect Wh-Q*, are associated with all levels of performance, appearing toward the top, middle, and bottom of the different model diagrams.

5 Discussion

Our results paint a complex picture, where both model size and the number of licensing contexts of a given NPI contribute to higher correlations with human judgments. Nevertheless, we observe substantial variation in the correlations between model and human preferences across NPIs and contexts.

For the results by NPI, we see relatively consistent positions of NPIs across different model architectures when considering their relative relationships with the correlations, indicating that interesting structural patterns exist within the class of NPIs. While some aspects of traditional NPI theory are reflected, namely the number of licensing contexts for a given NPI (which is informative about its relationship to other NPIs), there is still much complexity that does not fit in with the specifics of NPI theory. For example, *exactly* is consistently among the NPIs on which the models perform best, despite being licensed in only one of the contexts we consider. This indicates that the behavior of the LLMs we considered does not fully capture the distinctions relevant to the licensing of NPIs proposed in traditional linguistic theories. It is also possible that because *exactly* is not licensed by any contexts other than negation, its licensing conditions are easier to learn. In other words, *exactly* is such a strong NPI in a prototypical sense that models may find it easier to distinguish the context that licenses it from the contexts that do not in comparison to

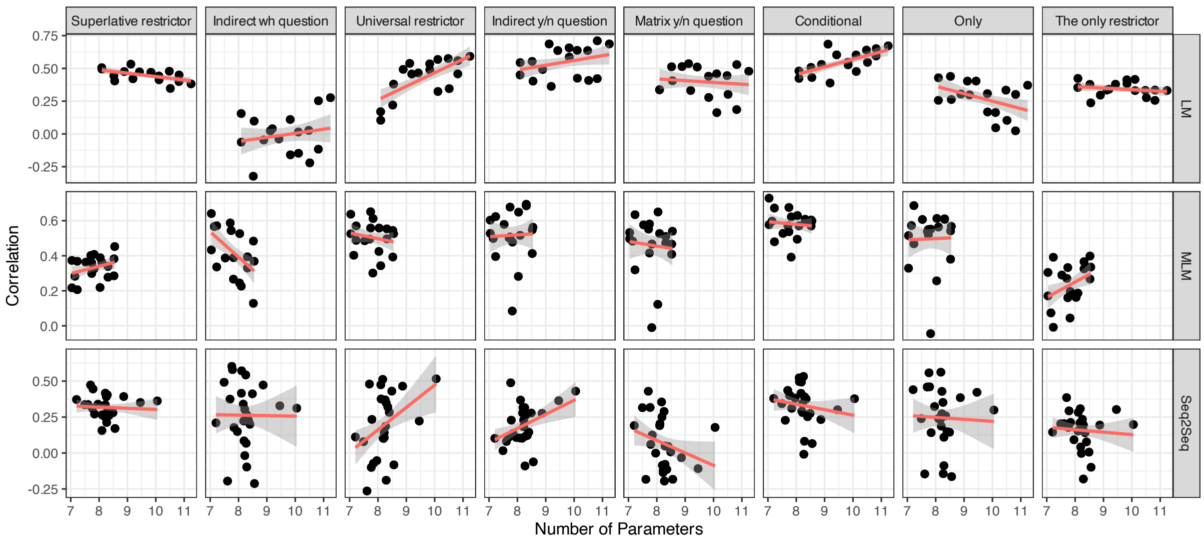

Figure 3: Relationship between the number of model parameters and the correlation between model predictions and human judgments for the licensing contexts. Contexts are presented from least to most licensing, from left to right.

NPIs that are subject to greater complexity.

Although much contemporary research has shown that larger LLMs trained on larger datasets tend to exhibit better performance compared to smaller models trained on smaller datasets, our findings make us less sanguine about the prospect of big models and big data leading to fully human-like linguistic behavior. While the models do acquire a degree of knowledge pertaining to NPI licensing contexts, many of the subtleties are lost. It is plausible that within this domain, the notion that larger LLMs are inherently superior may not hold true at the level of detail we investigate. Additionally, it is worth noting that the context frequency in natural text seems to be related to model performance in some way, though a more extensive investigation may better distinguish its effects from the effects of licensing number and model features.

The elevated performance observed in *Conditional* contexts across all three model types may be plausibly attributed to the syntactic characteristics of this licensing environment. Specifically, the use of the word *if* serves as a distinguishing marker for a *Conditional*, while other contexts may be identified only by more abstract structural properties. This easy-to-identify distinguishing feature may render proficiency in this licensing construction relatively more obtainable. In future research, more robust NPI theory could provide additional explanatory power for understanding the relationship that LLMs learn about NPIs and their licensing environments.

6 Conclusion

We investigated NPI licensing in LLMs by analyzing the similarities between model and human judgments and their relationship with certain linguistic and model features. Analysis by NPI reveals a significant positive relationship between both model size and model performance, as well as between licensing number and model performance. However, analysis by licensing context reveals that larger LLMs may not be inherently better than smaller LLMs at particular levels of granularity and that model performance may not be influenced by all of the anticipated factors in the most intuitive way. Additionally, we have determined hierarchies among NPIs and licensing contexts, which provide a broader perspective on NPI licensing across model tasks. Several patterns emerged: while traditional semantic classifications of NPIs were not reflected, a key feature, namely the number of contexts that license a given NPI, does appear to have an impact on the hierarchies, though with some clear exceptions. Similarly, hierarchies among licensing contexts may be influenced by the syntactic characteristics of the environments. This complex situation seems to reflect the complexity of NPIs, which are linguistically heterogeneous.

Limitations

Many prominent LLMs today are proprietary and restrict the ability to collect the probability the model gives to an arbitrary token at an arbitrary

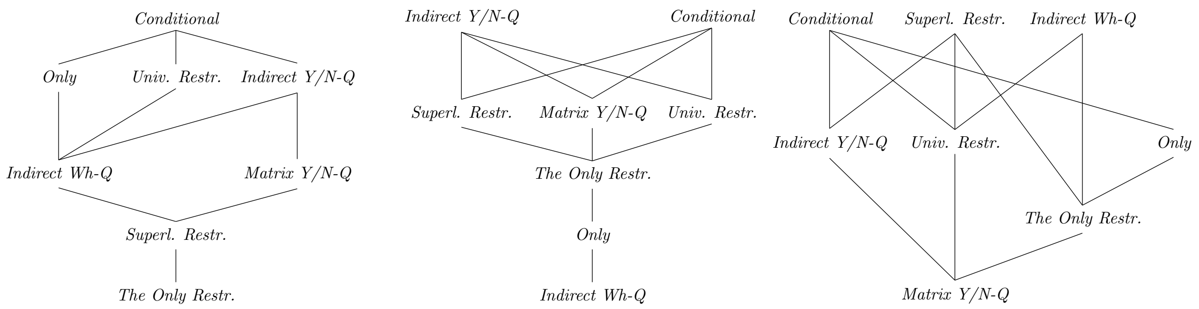

Figure 4: The diagrams are presented in the following sequence, from left to right: MLM, LM, and Seq2Seq. These Hasse diagrams depict the partial ordering of licensing contexts based on the results of the beta regressions performed on the one-hot encodings. The orderings are derived from correlations between the contexts and human licensing judgments, with certain contexts demonstrating notably higher correlations. These diagrams utilize vertical positioning to visually represent the relative ordering: a context positioned higher in the diagram indicates a stronger correlation with human judgments compared to any context reachable by following a downward line. As an illustration, in the LM diagram, *Conditional* occupies one of the topmost positions. Every context other than *Indirect Y/N-Q* can be reached by tracing a line downward from *Conditional*, signifying its comparatively greater correlation with the human judgments. Contexts that are not connected in this manner did not exhibit any statistically significant relative relationships in the regression results.

position. As a result, our approach does not allow us to evaluate such models.

Additionally, the incorporation of training dataset size as predictor of performance is complicated due to a lack of consistent documentation of this potentially crucial part of the pre-training regimen. Many papers that present new LLMs either omit information regarding the size of the training dataset, or else present it in units that are difficult to convert to a standardized measure, including compressed disk size, uncompressed disk size, token count, and word count. While it seems clear that larger datasets should lead to increased performance, it is difficult to determine precisely what the relationship between dataset size and performance on various tasks is for this reason.

Moreover, the nature of progress in terms of available computational resources naturally leads to a confound between model size and dataset size. As more computational resources become more available over time, models and datasets tend to grow in tandem. Furthermore, the fact of when MLMs, Seq2Seqs, or LMs happen to be *en vogue*, and the particular computing resources available at that time, leads to a confound between a model's pretraining task, its size, and the size of the dataset used to (pre-)train it. Training and making available a more systematically varied set of LLMs, where task, model size, and dataset size are intentionally varied independently, could help alleviate our current inability to distinguish the effect of such differences on various tasks. Such an undertaking would be, however, out of reach for all but those with the most computational resources at hand, given the current size of state-of-the-art LLMs and the datasets they are pre-trained on.

Finally, the available data on the assessment of NPI licensing is not entirely comprehensive, as we find the subtleties of NPI/context combinations fit for our purposes represented by binary judgments. A more detailed empirical investigation could well reveal more gradient human judgments, which may alter future analysis of LLM knowledge of NPI licensing.

References

Lisa Bylinina and Alexey Tikhonov. 2022. Transformers in the loop: Polarity in neural models of language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6601–6610, Dublin, Ireland. Association for Computational Linguistics.

Emmanuel Chemla, Vincent Homer, and Daniel Rothschild. 2011. Modularity and intuitions in formal semantics: The case of polarity items. *Linguistics and Philosophy*, 34:537–570.

Milica Denić, Vincent Homer, Daniel Rothschild, and Emmanuel Chemla. 2021. The influence of polarity items on inferential judgments. *Cognition*, 215:104791.

Bart Geurts. 2003. Reasoning with quantifiers. *Cognition*, 86(3):223—251.

Anastasia Giannakidou. 1998. *Polarity Sensitivity as (non)Veridical Dependency*. John Benjamins, Amsterdam.

Jack Hoeksema. 2012. On the natural history of negative polarity items. *Linguistic Analysis*, 44:3–33.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jaap Jumelet, Milica Denić, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess npi licensing.

Edward Klima. 1964. Negation in english. In Jerry A. Fodor and Jerrold Katz, editors, *The Structure of Language*, pages 246–323. Prentice Hall, Englewood Cliffs, NJ.

William A. Ladusaw. 1979. *Polarity Sensitivity as Inherent Scope Relations*. Ph.D. thesis, University of Texas, Austin.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1999. Treebank-3 ldc99t42. Philadelphia: Linguistic Data Consortium.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Volume: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Frans Zwarts. 1998. Three types of polarity. In F. Hamm and E. W. Hinrichs, editors, *Plurality and Quantification*, pages 177–238. Kluwer.