Linking language model predictions to human behaviour on scalar implicatures

Yulia Zinova and David Arps and Katharina Spalek and Jacopo Romoli
Heinrich Heine University Düsseldorf
{zinova,david.arps,katharina.spalek,jacopo.romoli}@hhu.de

Abstract

We explore the behaviour of language models on adjectival scales in connection with negation when prompted with material used in human experiments. We propose several metrics extracted from the language model predictions and analyze those metrics in relation to human data. We then use these metrics to propose new items to be tested in both human and model-based experiments.

1 Outline

In this paper, we describe various experiments that explore the relationship between scalar implicatures and language modeling Scalar implicatures are inferences such as the one in (1).

(1) a. The project is difficult.
b. ⇝ The project is not impossible.

Here, *difficult* and *impossible* form a scale. The inference is such that using the weaker item on that scale (1-a) leads to the negation of the stronger item (1-b).

In the first part of the paper, we aim to elicit an implicature (1-b) from the language model as the next token prediction. To do this, we prompt the model with the original sentence (containing a weak scalar item) followed by the repetition of the initial portion of the same sentence and a negation (2). We base this experiment on the material presented in Sun et al. 2024 and explore the output as well as the underlying processing of prompts that include negation.

(2) The project is difficult. This means that the project is not (PROMPT)

Next, we introduce metrics that are based on the model behaviour. These metrics prioritize lexical items that are likely to co-occur in the top predictions of the language model. We use these metrics to automatically extract new pairs of adjectives. From the obtained list of pairs we then select such pairs where the adjectives are on a scale and repeat the negation experiment using corpus data and both established and new adjective pairs. We analyze the model behaviour for both sets of pairs, focusing on the desirable *crossing* pattern of adjectival activation when the model encounters the negation.

In the last part of the paper, we use the proposed model-related metrics in connection with human experiments. On one hand, with the help of one of the metric variants we can explain some part of the variability on the human ratings for scalar implicatures and negative strengthening. On the other hand, we discover that almost all the scales used in human experiments receive low values according to our metrics. We then extract new adjective pairs from the model and propose a set of scalar pairs to use in future human experiments that would be more evenly distributed from the perspective of the language model.[1]

2 Language models and negation

2.1 Introduction

Recent papers have demonstrated time and again that negation poses a challenge for language models that is not resolved by increasing the model and the dataset size (Kassner and Schütze, 2020; Lipkin et al., 2023; Zhang et al., 2023; Sullivan, 2024). This becomes especially relevant in connection with the natural language inference task: the performance on datasets that focus on negation even after fine-tuning is significantly lower than on general datasets where negation does not play a special role (Hossain et al., 2020; Truong et al., 2023).

Another challenge for language models is related to pragmatic inferences that are not tradition-

[1] Our implementation is made available at https://github.com/davidarps/lm-scales

Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning @ COLING 2025, pages 97–106
January 20th, 2025.

ally included in the NLI datasets but are relevant for human daily conversations (Hu et al., 2023). These include presuppositions, scalar implicatures and other related inference types, such as negative strengthening, extensively studied in the theoretical literature (Horn 1984; Hirschberg 1985; Degen 2015; Gotzner and Romoli 2022, among others) but severely underrepresented in NLI datasets (Jeretic et al., 2020). Negative strengthening refers to a type of implicature whereby the meaning of a scalar expression containing a negation (3-a) is enriched using its non-negated antonym (example (3-b), (27) in Gotzner and Romoli 2022).

(3) a. The room is not large.
b. ⇝ The room is (rather) small.

A recent dataset that aims to address the problem of underrepresented inference types provides premise-hypothesis pairs that include scalar items, such as *some/not all* and *warm/hot* (SIGA, Nizamani et al. 2024). It contains premise-hypothesis pairs preceded by a context (4-a) and labeled as *contradiction*, *entailment* or *neutral*. In case of example (4), the label for the pair (4-b)-(4-c) is *contradiction*.

(4) a. Five weeks later, I had my first batch of polished stones in nearly 40 years. I was also disappointed.
b. The polished stone looked good
c. The polished stone looked great

The challenge in creating such datasets, apart from extracting or generating the data, is data annotation, especially given the fact that the rate with which humans predict scalar implicatures in experimental studies varies significantly between items (Van Tiel et al., 2016; Sun et al., 2018; Gotzner et al., 2018b; Ronai and Xiang, 2022). Multiple experimental studies aimed to explain this variation with the help of various linguistic properties as well as the relation to priming (Ronai and Xiang, 2023; Lacina and Gotzner, 2024) but achieved only partial success: no combination of the proposed factors could explain the full range of human rating variation.

Since the only available naturalistic dataset for scalar inferences (SIGA, Nizamani et al. 2024) focuses on the implicatures or their absence in a positive context, it does not allow to evaluate the behaviour of the language models with respect to the scalar terms in the context of negation. Such an evaluation is an important missing step, since the underlying process of implicature computation involves reasoning about the alternatives and their negated variant (Van Tiel et al., 2016; Gotzner et al., 2018a). For this reason, in the first experiment we test the behaviour of the (smallest) OPT language model for next word prediction, trying to elicit a completion following a prompt that includes a negation similar to the experimental setup of Van Tiel et al. 2016.

We show that the language model exhibits a significant amount of copying in such a scenario, what on the surface level looks like ignoring the negation (and leads to a contradicting sentence completion). We examine the underlying representations and find evidence for the desired trends in processing the negation that often do not reach the level to become visible in the output.

2.2 Experiment 1

In the first experiment we evaluate negation processing by a language model using both scales and contexts from Sun et al. (2024). To approach this task, we test whether a language model is likely to predict an adjective compatible with a scalar implicature as the next word. We use a setting that is compatible with computing a scalar implicature based on the gradable adjectives.

Model Previous work has shown that models of different sizes show similar performance on token-level predictions related to scalar implicatures (Arps and Zinova, 2024). Therefore, all experiments are conducted with only one model, namely OPT-125m (Zhang et al., 2022). OPT-125m is a decoder-only (causal) language model with twelve layers and an embedding size of 768. It has been trained on next-token prediction on 180B tokens of predominantly English books and web-crawled data from different domains.

Data In this experiment we use the scales and the sentences from Sun et al. (2024). The prompts for the experiment were constructed following the scheme in (5): the first sentence contains a weak adjective (5-a) and is taken from the material of Sun et al. (2024). In our prompt, this sentence is followed by a second sentence that starts with a connector (5-b) and continues with the same prompt as in (5-a) repeated up to the adjective position and followed by a negation (5-c). We then obtained the model predictions over all the vocabulary for the next word following the complete prompt (5) (including negation). The expected item according

to the implicature pattern would be (the negation of) *brilliant*, the stronger alternative of *intelligent*.

(5) a. This student is intelligent.
b. Put differently,[2]
c. this student is not ...

The results of this experiment are in line with the previous predictions concerning language models and negation: in most cases the model predicted the same weak adjective it observed in the first part of the prompt as one of the top predictions. The same weak adjective has rank 0 after negation in 383 out of 1276 cases (30%), rank 1 in 102 cases (8%) and rank between 2 and 4 in another 145 cases (11%). This means that in 30% of all the cases the resulting sentence (in this case *"Put differently, this student is not intelligent"*) contradicts the preceding part of the prompt (5-a). This is not surprising given the difficulty of the task, previous findings and the absence of fine-tuning.

In order to check whether the model ignores the negation, as suggested by Kassner and Schütze (2020) and by the surface evaluation above, we have traced the activation of the weak and the strong gradable adjectives at the position before the weak gradable adjective is introduced in (5-a), at the same point (before the negation) in the last part of the prompt (5-c) and after the negation (end of the prompt (5-c)).

Despite the very high surface copying rate, we can observe that the model does not ignore the negation, which is visible on the cumulative representations of scalar adjective activation for a specific scale. To obtain such a representation, we have collected the logit activation at the following points, accumulating them over various prompts: [0] at the beginning of the first sentence, [1] before the scalar adjective in the first sentence, [2] after the adjective in the first sentence, [3] at the beginning of the second sentence, [4] before the negation in the second sentence, and [5] after the negation in the second sentence. Example (6) shows these points in the exemplar prompt provided above in (5).

(6) [0] This student is [1] intelligent [2]. Put differently, [3] this student is [4] not [5] ...

The case illustrated in Fig. 1 demonstrates the desired behaviour of a language model in the context

[2]We have tested various connectors such as *It means that* and *In other words* as well as an empty connector, but did not observe any significant variation in the model behaviour.

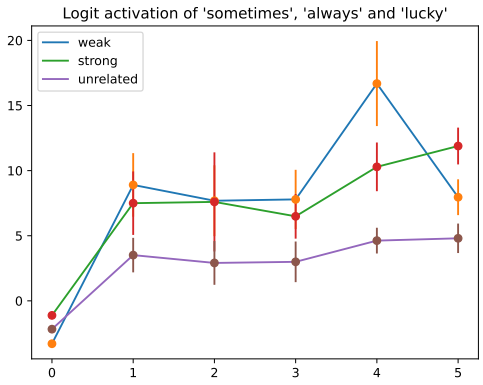

Figure 1: Activation of *sometimes*, *always* and *lucky* (an unrelated adjective) across various prompts at the different points in the prompt.

of implicature computation given a negation: although the activation of the weak item (the one present in the first sentence) is higher at point [4] (before the negation), the insertion of a negation leads to a drop of the activation of the weak adjective and a rise of the activation of the strong adjective at point [5]. The magnitude of these effects is such that the activation of the strong adjective after negation is higher than that of a weak adjective and the model does not copy the adjective that occurred in the prompt. The standard deviation bars on the plot show that in this case the effect can be reliably observed over individual prompts. We will call this behaviour of the model *crossing*. Note that *crossing* does not guarantee that the strong adjective will appear as the most likely token after negation, it only guarantees the non-copying behaviour of the model.

In the other case, illustrated in Fig. 2, both the effect of decreasing the activation of the weak scalar item and the effect of increasing the activation of the respective strong scalar item is observed, so the approaching trend of the two activations does not reach the level at which we could observe a reflection of this trend in the next word prediction behaviour: the weak item remains the most likely continuation and the model exhibits the copying behaviour. We will call this scenario *approaching*.

The last scenario illustrated in Fig. 3 includes the already observed effect of decreasing the activation of the weak scalar item but the activation of the strong scalar item also drops slightly. As a result, the difference in the activations decreases but similar to the *approaching* case there is no surface evidence of this trend. We will call this scenario

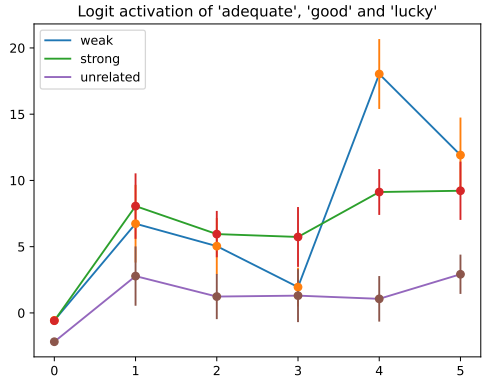

Figure 2: Activation of *adequate*, *good* and *lucky* (an unrelated adjective) across various prompts at the different points in the prompt.

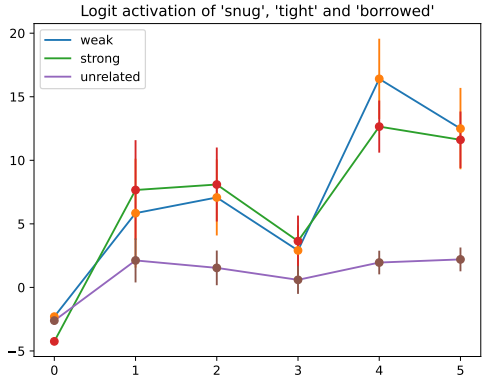

Figure 3: Activation of *snug*, *tight* and *borrowed* (an unrelated adjective) across various prompts at the different points in the prompt.

difference lowering.

These results show that introducing a negation in the second sentence does influence underlying activation of the language model but in most cases does not lead to a visible change of the output. In the next section, we introduce a negation-independent metric for adjective pairs. We hypothesize that the probability of the two adjectives to be simultaneously encountered among the top candidates for the next token in a positive scenario (corresponding to close and high activations of both adjectives at point [1]) correlates with the probability of *crossing* behaviour of the model if it is prompted following the schema (5).

2.3 Extracting new adjective pairs from the language model

We propose a method to evaluate the quality of scales from the literature, using language model behavior.

Corpus Our corpus-based experiments are performed on the training data of the BabyLM 2023 challenge (Warstadt et al., 2023). This data consists of mostly transcribed and child-directed speech from different sources. We preprocess the data using the Boot-BERT pipeline (Samuel, 2023).

Identifying Matches The method starts with an unlabeled tokenized text corpus $\mathcal{C}$, a next word prediction language model $\mathcal{M}$ and a collection $\mathcal{A} = \{a_1 \dots a_m\}$ of m (adjective) terms. $\mathcal{M}$ provides, for each token $s_{i,j}$ in each sentence $s_i \in \mathcal{C}$, a probability distribution $p_{\mathcal{M}}(s_{i,j+1}|s_{i,:j})$ over all possible next tokens $s_{i,j+1}$ given a prefix $s_{i,:j}$. Specifically, we collect all $k = 10$ most likely continuations for every prefix in the corpus, and filter the corpus for prefixes $s_{:i}$ where a scalar term is among the k most likely next tokens. We call these situations *matches*. Matches are independent from whether that scalar term is actually present in $\mathcal{C}$ as a continuation.

Co-occurrence counts from matches Assume that count(a) is the number of times that the adjective a is matched across the corpus. Further assume that cc(a_r, a_s) is the number of times that two adjectives a_r and a_s are matched at the same prefix $s_{i,j}$ in the corpus. To account for the fact that the adjectives occur with different frequencies, we compute the following scores:

The scaled cooccurence score, conditioned on one of the terms:

$$\text{cc}_{log}(a_r, a_s) = \frac{\log \text{cc}(a_r, a_s)}{\log \text{count}(a_r)}$$

By this we obtain two coocurence scores, conditioned either on the weak or on the strong scalar item. We call the following score *scale by strong*:

$$\text{cc}_{log}(a_{weak}, a_{strong}) = \frac{\log \text{cc}(a_{weak}, a_{strong})}{\log \text{count}(a_{weak})}$$

And the following score *scale by weak*:

$$\text{cc}_{log}(a_{strong}, a_{weak}) = \frac{\log \text{cc}(a_{strong}, a_{weak})}{\log \text{count}(a_{strong})}$$

Thus we collect two scores that can be used either separately or combined. One way to combine them and make the resulting metric symmetric is calculating the harmonic mean of these scaled cooccurence scores:

$$\text{cc-hm}(a_r, a_s) = 2\frac{\text{cc}_{log}(a_r, a_s) * \text{cc}_{log}(a_s, a_r)}{\text{cc}_{log}(a_r, a_s) + \text{cc}_{log}(a_s, a_r)}$$

The harmonic mean prioritizes pairs of adjectives such that each of them is likely to be found in the top predictions of the model when the other one is in the top predictions. We can now sort all cooccuring adjective pairs a_r, a_s by their cc-hm(a_r, a_s), and put special focus on the pairs with very high cooccurence scores.

Finding new scalar adjective pairs From the pairs obtained on the previous step we have manually selected those that are scalar alternatives and identified the weaker and the stronger scale mates between them (see Table 5 in the Appendix). We have also attempted an automatic filtering of non-scalar pairs and automatic strength evaluation following the proposal by de Melo and Bansal (2013).

Among the versions we attempted were (1) rank extraction from the language model for the patterns suggested by de Melo and Bansal (2013) when the first adjective of the pattern is in the ground truth as well as (2) corpus search in the corpus used for the other experiments as well as (3) Google n-gramm inquiry.

Neither method brought results reliable enough to justify automatic scale and strength extraction from the proposed list. Since for human experiments items have to be often evaluated according to additional criteria, our proposal at the moment is to supply a list of pairs with their model scores and leave it to the linguists to select the suitable pairs.

2.4 Experiment 2

In this experiment we have tested the behaviour of the language model on the adjective pairs from Sun et al. 2024 together with some extra scales from Lacina and Gotzner 2024 and on the scales extracted in the previous section using the harmonic mean score (all the scales are provided in the Appendix). For a fair evaluation, we have used the same corpus (BabyLM challenge, Warstadt et al. 2023) for all the scales. We have identified all the sentences where the weak adjective from either list occurs in the corpus and applied the pattern shown in (5) to all such sentences.

For each example we have computed the logit of the weak and the strong adjectives before and after negation (analogous to the points [4] and [5] in the first experiment). We have then computed the proportion of cases when the activation of the weak item decreases after the negation is introduced (matched term lowering), the proportion of cases where the activation of the strong item after negation exceeds the activation of the weak item (crossing) and the proportion of cases where the two activations approach each other but the activation of the weak item remains higher than that of the strong item (approaching). The results are presented in Table 1.

	Sun et al	New pairs
Matched term lowering	1.00	0.99
Crossing	0.31	0.33
Difference lowering	0.99	0.99
Approaching	0.22	0.15

Table 1: Comparison of model behaviour for scales from human experiments and scales extracted on the basis of the language model data.

As can be seen in Table 1, exchanging the scalar pairs in the experiment led to an increase of the crossing instances, but this increase remained small. Another interesting observation is related to the approaching scenario: The number of approaching instances reduces when the scale selection is performed according to the model cooccurence scores. This can be interpreted as that, loosely said, stronger related items tend to increase and decrease their activation together, which is not a desirable trend in the current setup. At the same time we observe an approaching behaviour in almost all (99%) of the cases with either selection of the items.

This means that if the trend behind the negation processing could be magnified, in principle it would be possible to achieve a desired (non-copying) behaviour under negation in almost all the cases. It is left for future research to explore such possibilities.

3 Model metrics and human behaviour

In order to evaluate the obtained metrics on human data, we extract all the values as described above for the scalar adjectival pairs that were used in human experiments (Gotzner et al., 2018a,b; Lacina and Gotzner, 2024). We then compute the correlations between our metrics and human data. This reveals that the most helpful metric is *scale by strong*: that of the high ranking of the weak item in those sentences where the strong item is ranked

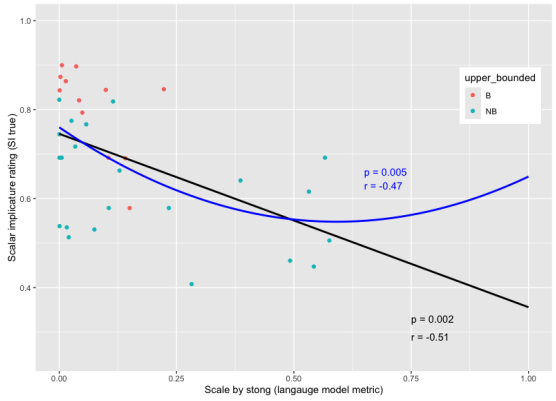

Figure 4: Correlation of *scale by strong* metric from the OPT language model and human scalar implicature ratings.

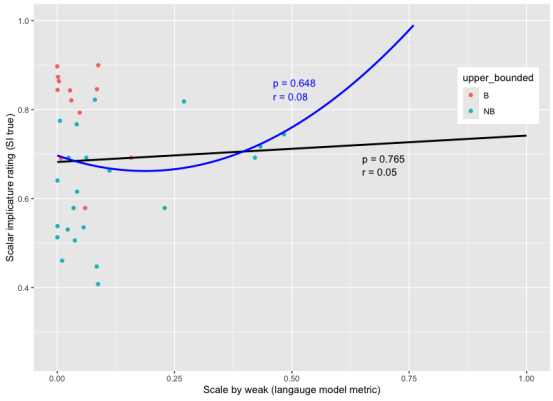

Figure 5: Correlation of *scale by weak* metric from the OPT language model and human scalar implicature ratings.

high (e.g. how often is *attractive* present in the top ten prediction of those sentences that have *stunning* in top ten predictions). This metric strongly negatively correlates with human scalar implicature rating $(-0,51)$ and positively correlates with human rating for negative strengthening $(0,34)$. It must be noted, however, that our results are limited by the range of scales for which human data is available: as described above, for the metrics we have calculated how often the two adjectives of the pair co-occur within the top ten predictions of the language model. We have discovered that within this metric, although values between zero and one are possible (see Appendix), the actual values for the original experimental items (see Figure 4) are very low (less than 0.6).

Figure 4 illustrates that both linear and quadratic correlations are plausible ($r = -0.51$ for linear and $r = -0.47$ for quadratic correlation) given the data from the previous experiments due to the limited range of values of the *scale by strong* metric. The extension of the two correlation curves demonstrates that obtaining the experimental results for items that lie on the right spectrum of the *scale by strong* metric is essential for making a decision about the validity of either correlation.

One parameter to take into account is the upper boundedness of the scale. It has been shown to be the main predictor for the human scalar implicature and negative strengthening ratings from a collection of linguistic features (Van Tiel et al. 2016 as well as Sun et al. 2018; another relevant but less strong predictor is semantic similarity). Although statistical evaluation of non-upper-bounded scales is not possible due to the insufficient amount of data, the two categories are shown on Figure 5. It can be observed that all the human ratings of bounded scales are higher than these of non-bounded scales and at the same time almost all the language model scores for those scales are very low. The question whether this accidental or systematic can be studied by exploring the linguistic properties of the scales that receive high scores by the language model metrics.

In relation to this it is also worth exploring Fig. 5 that depicts the possible correlations of human scalar ratings with the model scores *scale by weak*. Although the statistical analysis produces very low correlation values ($r = 0.05$ for linear and $r = 0.08$ for quadratic correlation), the visual inspection reveals that most of the values for the model scores are below 0.12, so the correlation analysis can not be reliably performed on the basis of this data.

Since *scale by weak* score is very low for most of the scales, the harmonic mean score that takes into account both *scale by strong* and *scale by weak* also does not provide a significant correlation for the available set of data. Similarly to the case depicted on Fig. 5 the set of data does not exclude the possibility of discovering such a correlation given a different set of data.

Since the value of the proposed metrics are very low for most of the items found in the experimental literature, we suggest that more experiments should be performed with different scales that are better distributed according to those metrics. As described above, we propose a list of pairs that satisfies these criteria from the model perspective and leave if to the linguists to pick the best experimental items from it. This list is provided in the Appendix and the suggestions are marked in bold.

4 Discussion

In this paper we described several experiments related to scalar adjectives. First of all, we could establish that despite the overt copying behaviour of the analyzed language model the underlying activation exhibits desirable trends. Second, we have proposed model-based metrics to evaluate the scales and experimented with scales that receive high ratings according to these metrics. We could achieve a slight increase in desirable behaviour (crossing pattern between the activations of the weak and the strong items), although these results provided a lesser increase than we had expected.

We believe that the observed underlying behaviour of the language model while processing the negation opens new perspectives in adjusting the model predictions by magnifying the desired trends.

Finally, we have explored the connection between the human experimental results and proposed metrics. We could observe that one of the metrics (*scale by strong*) can be used to partially explain the variability in human ratings of the scalar implicatures associated with different scales. At the same time time we could see that the scales used in human experiments have very low scores on all the proposed language model metrics. In this light we suggest new material for further experiments.

This type of work opens a field of automatic generation of possible experimental material as well as running the experiments using the language model before transferring them to the lab. This can lead to a significant decrease in time needed for experimental design as well as to lowering the cost of running various versions of the same experiment since some relevant design problems can be already observed and corrected on the level of the language model experiments.

References

David Arps and Yulia Zinova. 2024. It is difficult, but not impossible: Measuring scalar activation in language models. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Trento, Italy. SEMDIAL.

Gerard de Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Judith Degen. 2015. Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8:11–1.

Nicole Gotzner and Jacopo Romoli. 2022. Meaning and alternatives. *Annual Review of Linguistics*, 8(Volume 8, 2022):213–234.

Nicole Gotzner, Stephanie Solt, and Anton Benz. 2018a. Adjectival scales and three types of implicature. In *Semantics and Linguistic Theory*, pages 409–432.

Nicole Gotzner, Stephanie Solt, and Anton Benz. 2018b. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in psychology*, 9:1659.

Julia Bell Hirschberg. 1985. *A theory of scalar implicature (natural languages, pragmatics, inference)*. Ph.D. thesis, University of Pennsylvania.

Laurence Horn. 1984. Towards a new taxonomy for pragmatic inference: Q-and r-based implicature. *Meaning, form and use in context*.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Radim Lacina and Nicole Gotzner. 2024. Exploring scalar diversity through priming: A lexical decision study with adjectives. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. 2023. Evaluating statistical language models as pragmatic reasoners. *Preprint*, arXiv:2305.01020.

Rashid Nizamani, Sebastian Schuster, and Vera Demberg. 2024. SIGA: A naturalistic NLI dataset of English scalar implicatures with gradable adjectives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14784–14795, Torino, Italia. ELRA and ICCL.

Eszter Ronai and Ming Xiang. 2022. Three factors in explaining scalar diversity. In *Proceedings of sinn und bedeutung*, volume 26, pages 716–733.

Eszter Ronai and Ming Xiang. 2023. Tracking the activation of scalar alternatives with semantic priming. *Experiments in Linguistic Meaning*, 2:229–240.

David Samuel. 2023. Mean BERTs make erratic language teachers: the effectiveness of latent bootstrapping in low-resource settings. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 221–237, Singapore. Association for Computational Linguistics.

Michael Sullivan. 2024. It is not true that transformers are inductive learners: Probing NLI models with external negation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1924–1945, St. Julian's, Malta. Association for Computational Linguistics.

Chao Sun, Ye Tian, and Richard Breheny. 2018. A link between local enrichment and scalar diversity. *Frontiers in Psychology*, 9:2092.

Chao Sun, Ye Tian, and Richard Breheny. 2024. A corpus-based examination of scalar diversity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 50(5):808.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: An analysis of language models on negation benchmarks. *Preprint*, arXiv:2306.08189.

Bob Van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of semantics*, 33(1):137–175.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Yuhui Zhang, Michihiro Yasunaga, Zhengping Zhou, Jeff Z. HaoChen, James Zou, Percy Liang, and Serena Yeung. 2023. Beyond positive scaling: How negation impacts scaling trends of language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7479–7498, Toronto, Canada. Association for Computational Linguistics.

Appendix: established scale used for the experiments and proposed scales

	weak	strong
0	adequate	good
1	allowed	obligatory
2	attractive	stunning
3	big	enormous
4	cheap	free
5	dark	black
6	difficult	impossible
7	few	none
8	funny	hilarious
9	hard	unsolvable
10	hungry	starving
11	intelligent	brilliant
12	low	depleted
13	memorable	unforgettable
14	old	ancient
15	possible	certain
16	rare	extinct
17	scarce	unavailable
18	silly	ridiculous
19	small	tiny
20	snug	tight
21	some	all
22	special	unique
23	tired	exhausted
24	ugly	hideous
25	warm	hot
26	wary	scared

Table 2: Scales from (Sun et al., 2024) that we used in our experiments

	weak	strong
0	angry	annoyed
1	bad	mediocre
2	good	excellent
3	overweight	obese
4	pretty	beautiful
5	warm	hot

Table 3: Additional scales from (Lacina and Gotzner, 2024)

	weak	strong
0	afraid	scared
1	amazing	incredible
2	angry	mad
3	bad	terrible
4	big	huge
5	calm	quiet
6	clear	obvious
7	courageous	fearless
8	damaged	destroyed
9	difficult	hard
10	frightening	terrifying
11	good	great
12	great	awesome
13	great	good
14	honest	frank
15	odd	strange
16	overweight	obese
17	pleased	proud
18	popular	famous
19	pretty	beautiful
20	silly	stupid
21	small	tiny
22	smart	intelligent
23	some	all
24	surprised	shocked
25	tasty	delicious
26	useless	worthless
27	warm	hot
28	wealthy	rich

Table 4: Scales annotated by the authors which have a high cc-hm score.

	weak	strong	*scale by weak*	*scale by strong*		weak	strong	*scale by weak*	*scale by strong*		weak	strong	*scale by weak*	*scale by strong*
0	red	purple	0.78	0.98	50	new	recent	0.68	0.85	100	interesting	remarkable	0.53	0.77
1	second	third	0.90	0.98	51	**unconscious**	**dead**	0.85	0.57	101	full	crowded	0.49	0.77
2	better	worse	0.82	0.97	52	pale	white	0.85	0.63	102	lost	all	0.76	0.59
3	yellow	white	0.97	0.81	53	difficult	dangerous	0.81	0.85	103	necessary	vital	0.57	0.76
4	pink	red	0.96	0.79	54	full	empty	0.72	0.85	104	best	better	0.70	0.76
5	purple	black	0.96	0.73	55	**difficult**	**impossible**	0.77	0.85	105	long	difficult	0.64	0.76
6	**good**	**great**	0.92	0.96	56	**several**	**all**	0.85	0.67	106	understandable	acceptable	0.75	0.54
7	pink	white	0.95	0.75	57	important	critical	0.65	0.85	107	violent	cruel	0.65	0.75
8	orange	red	0.95	0.77	58	good	complete	0.61	0.85	108	full	all	0.75	0.60
9	interesting	amusing	0.61	0.95	59	**damaged**	**broken**	0.84	0.73	109	good	superior	0.46	0.75
10	**important**	**essential**	0.70	0.95	60	**similar**	**identical**	0.67	0.84	110	authorized	required	0.75	0.50
11	yellow	black	0.94	0.79	61	good	high	0.74	0.84	111	nervous	afraid	0.75	0.66
12	bulky	heavy	0.94	0.49	62	dark	black	0.84	0.81	112	damaged	lost	0.75	0.59
13	**different**	**separate**	0.74	0.94	63	oval	round	0.84	0.52	113	light	white	0.74	0.75
14	interesting	exciting	0.78	0.94	64	good	large	0.71	0.84	114	**comfortable**	**luxurious**	0.44	0.75
15	gray	white	0.94	0.64	65	**polite**	**friendly**	0.83	0.71	115	deep	loud	0.69	0.74
16	brown	black	0.94	0.78	66	sad	angry	0.82	0.83	116	emotional	moral	0.71	0.74
17	brown	white	0.94	0.78	67	**concerned**	**alarmed**	0.54	0.83	117	**unusual**	**unique**	0.74	0.62
18	gray	black	0.94	0.63	68	**possible**	**probable**	0.47	0.83	118	serious	dangerous	0.71	0.74
19	happy	proud	0.87	0.93	69	interesting	unusual	0.62	0.83	119	**sick**	**dead**	0.74	0.61
20	brown	blue	0.93	0.83	70	accurate	true	0.83	0.67	120	**cool**	**cold**	0.72	0.74
21	red	black	0.93	0.89	71	useful	valuable	0.79	0.83	121	certain	all	0.74	0.56
22	blue	black	0.92	0.86	72	steep	high	0.83	0.55	122	tender	soft	0.74	0.68
23	concerned	worried	0.90	0.92	73	good	perfect	0.65	0.83	123	acceptable	necessary	0.73	0.60
24	green	white	0.92	0.82	74	violent	dangerous	0.82	0.73	124	**cheap**	**free**	0.73	0.57
25	**tired**	**exhausted**	0.69	0.92	75	new	better	0.74	0.82	125	personal	private	0.72	0.73
26	**kind**	**generous**	0.60	0.92	76	**mediocre**	**poor**	0.82	0.43	126	smooth	shiny	0.64	0.73
27	**overweight**	**obese**	0.85	0.92	77	**allowed**	**required**	0.75	0.82	127	**worried**	**frightened**	0.55	0.73
28	**tired**	**sleepy**	0.57	0.91	78	serious	fatal	0.60	0.82	128	**possible**	**likely**	0.69	0.73
29	bruised	broken	0.91	0.59	79	interesting	attractive	0.69	0.82	129	knowing	caring	0.61	0.73
30	interesting	important	0.91	0.85	80	amused	pleased	0.81	0.64	130	dangerous	fatal	0.57	0.73
31	tedious	difficult	0.91	0.51	81	**special**	**unique**	0.74	0.81	131	**good**	**excellent**	0.53	0.73
32	**damaged**	**destroyed**	0.90	0.79	82	useless	impossible	0.81	0.69	132	third	half	0.73	0.72
33	good	better	0.83	0.90	83	interesting	beautiful	0.81	0.75	133	**bent**	**broken**	0.72	0.50
34	green	black	0.90	0.80	84	**uncommon**	**rare**	0.80	0.58	134	reasonable	high	0.72	0.52
35	attractive	beautiful	0.89	0.70	85	great	perfect	0.66	0.80	135	different	inferior	0.42	0.72
36	**moist**	**wet**	0.89	0.69	86	pleased	surprised	0.80	0.76	136	black	all	0.72	0.57
37	important	vital	0.58	0.89	87	severe	fatal	0.67	0.80	137	interesting	true	0.71	0.69
38	**some**	**all**	0.89	0.86	88	equal	identical	0.71	0.80	138	unpleasant	dangerous	0.71	0.47
39	relieved	happy	0.89	0.54	89	bad	evil	0.67	0.80	139	active	aggressive	0.62	0.71
40	harmful	dangerous	0.89	0.69	90	simple	obvious	0.76	0.80	140	surprised	frightened	0.52	0.71
41	good	cheap	0.58	0.88	91	near	close	0.76	0.80	141	general	any	0.71	0.57
42	**hot**	**boiling**	0.63	0.88	92	important	necessary	0.68	0.79	142	white	all	0.71	0.56
43	good	new	0.87	0.86	93	free	all	0.79	0.62	143	bright	warm	0.70	0.71
44	**warm**	**hot**	0.87	0.80	94	ready	willing	0.70	0.79	144	artistic	scientific	0.70	0.59
45	important	obvious	0.73	0.87	95	interesting	valuable	0.64	0.78	145	smooth	glossy	0.38	0.70
46	**old**	**ancient**	0.73	0.86	96	great	certain	0.67	0.78	146	some	black	0.58	0.70
47	sunny	warm	0.86	0.69	97	**willing**	**eager**	0.56	0.78	147	possible	any	0.70	0.61
48	distinct	separate	0.86	0.77	98	**aggressive**	**violent**	0.77	0.73	148	easy	pleasant	0.51	0.70
49	impractical	impossible	0.86	0.45	99	**neglected**	**abandoned**	0.77	0.60					

Table 5: All adjective pairs obtained from the ngram-based filtering in Sec. 2.3. Candidates for scale are marked in boldface.