SIGA: A Naturalistic NLI Dataset of English Scalar Implicatures with Gradable Adjectives

Rashid Nizamani,[◦] **Sebastian Schuster,**[◦,⋄] **Vera Demberg**[◦]
[◦]Saarland University [⋄] University College London
rashid_ahmed.nizamani@uni-saarland.de, s.schuster@ucl.ac.uk, vera@lst.uni-saarland.de

Abstract

Many utterances convey meanings that go beyond the literal meaning of a sentence. One class of such meanings is scalar implicatures, a phenomenon by which a speaker conveys the negation of a more informative utterance by producing a less informative utterance. This paper introduces a Natural Language Inference (NLI) dataset designed to investigate the ability of language models to interpret utterances with scalar implicatures. Our dataset is comprised of text extracted from the C4 English text corpus and annotated with both crowd-sourced and expert annotations. We evaluate NLI models based on DeBERTa to investigate 1) whether NLI models can learn to predict pragmatic inferences involving gradable adjectives and 2) whether models generalize to utterances involving unseen adjectives. We find that fine-tuning NLI models on our dataset significantly improves their performance to derive scalar implicatures, both for in-domain and for out-of domain examples. At the same time, we find that the investigated models still perform considerably worse on examples with scalar implicatures than on other types of NLI examples, highlighting that pragmatic inferences still pose challenges for current models.

Keywords: scalar implicature, pragmatics, natural language inference, dataset, English

1. Introduction

A hallmark property of human language understanding is deriving interpretations of utterances that go beyond the literal meaning (Grice, 1975). For example, in many contexts, the determiner *some*, which has the literal meaning of 'at least one,' is intended to convey *some, but not all*, as exemplified in (1).

(1) a. Alex ate **some** of the cookies.
b. ⇝ Alex did not eat **all** of the cookies.

This pragmatic phenomenon, known as scalar implicature (Horn, 1984; Hirschberg, 1985), generally is assumed to arise from listeners reasoning about alternative utterances that the speaker could have used but did not. In this context, assuming that a speaker knows how many cookies Alex ate, a listener reasons that if Alex had eaten all cookies, it would have been more informative to say so, and consequently a listener draws the inference that Alex did not eat all of the cookies. The phenomenon is called scalar implicature because it involves pairs (or sets) of words that can be ordered on a scale, often referred to as *scalar items*.

While for toy examples such as (1), the interpretation of utterances involving scalar implicatures can be straightforwardly computed using models of pragmatic reasoning (e.g., Franke, 2009; Goodman and Frank, 2016), inferring the intended meaning is a lot more challenging for naturalistic utterances for two reasons: First, scalar implicatures are *context-sensitive*. For example, unlike (1), many utterances with *some*, such as "*You sound like you have some small ones in the background*," do not give rise to scalar implicatures (Degen, 2015). Second, scalar items that give rise to implicatures are an *open class*. This class includes determiners like *some* and *all* but also scalar adjectives, such as *good* and *great* (Kennedy and McNally, 2005), or verbs, such as *start* and *finish* (see e.g., van Tiel et al., 2016).

To what extent large language models (LLMs) such as BERT (Devlin et al., 2018) or GPT-3 (Brown et al., 2020) that form the foundation of most of today's natural language understanding systems can draw such pragmatic inferences is still an open question, especially if one moves beyond the implicature from *some* to *some but not all* and if one considers the context sensitivity of these utterances. One piece of evidence for some models struggling with scalar implicature comes from Jeretic et al. (2020) who generated the template-based dataset IMPPRES for evaluating natural language inference (Dagan et al., 2013; Bowman et al., 2015, NLI) models on their ability to derive scalar implicatures. They found that a BERT-based NLI model rarely predicted that *some* entails *not all* or derived any other scalar implicatures. However, these automatically generated utterances were not verified and do not provide a lot of context, so it remains unclear whether humans would actually derive the purported implicatures in all these cases. On the other hand, for the *some-not all* scale, Schuster et al. (2020) showed that a BERT-based model can learn to predict context-sensitive scalar implicatures with relatively high accuracy, if it is trained on similar examples. However, considering that this work is limited to one scale, it remains unclear whether models trained on examples including one

LREC-COLING 2024, pages 14784–14795
20-25 May, 2024.

scale can generalize to other scales.

In this work, we present SIGA (Scalar Implicatures with Gradable Adjectives), a novel English NLI dataset targeting scalar implicatures. Importantly, the examples are extracted from naturalistic corpora together with their context and we annotate all examples using crowd-sourcing and expert annotations. Further, we focus on an open class of scalar items, namely gradable adjectives (Kennedy and McNally, 2005).

We then use this dataset to evaluate a state-of-the-art NLI model based on DeBERTa (He et al., 2020). Specifically, we consider the following two questions:

1. Can NLI models learn to predict scalar implicatures for a broad class of gradable adjectives?

2. Do the predictions of scalar implicatures generalize to examples that involve pairs of gradable adjectives that the model was not finetuned on?

To answer these two questions, we finetune DeBERTa models on both the Multi-Genre Natural Language Inference corpus (MNLI; Williams et al., 2018) and a subset of our novel dataset. We then evaluate the model both on an in-domain evaluation set that is comprised of examples that contain ajdective pairs that appear in the finetuning dataset, and on an out-of-domain evaluation set (Linzen, 2020) that is comprised of examples that contain adjective pairs that do not appear in the finetuning dataset.

We find that performance of NLI models on utterances with implicatures increases significantly when finetuned on our dataset of implicatures compared to a baseline model that was only finetuned on MNLI. Furthermore, we find that they exhibit some level of generalization since we observe similarly large gains on examples involving adjective pairs outside of the finetuning distribution. At the same time however, we find considerably lower accuracy on examples involving implicatures than on other NLI examples, highlighting that such pragmatic inferences still pose challenges for current models. We make our dataset and the experiment code publicly available at https://github.com/Rashid-Ahmed/SIGA-nli.

2. Related work

The work most closely related to ours is Jeretic et al. (2020), which also constructed a dataset for evaluating whether NLI models can derive different types of implicatures or whether they predict labels that are more in line with the literal meaning of sentences. However, unlike our work, they relied on automatically generated examples instead of naturally occurring sentences, and they did not verify whether humans derive implicatures for their examples. Zheng et al. (2021) also used templates to construct the GRICE dataset that contains short dialogs for testing the understanding of various pragmatic inferences, including sclar implicatures. They used this dataset to evaluate different (L)LMs on their ability to learn to derive implicatures. Ruis et al. (2022) evaluated the zero-shot and few-shot abilities of several recent LLMs for their ability to answer indirect questions, which sometimes also require deriving scalar implicatures, Hu et al. (2023a) evaluated the ability of LLMs to draw a wide range of pragmatic inferences using an expert-created set of examples, and (Hu et al., 2023b) showed that language models' predictions of alternatives predict when humans derive scalar implicatures. However, none of these works used naturalistic examples with a broad range of gradable adjectives.

A second line of work investigated how to automatically extract sets of adjectives that map onto the same scales (e.g., *warm* and *hot*). de Marneffe et al. (2010); de Melo and Bansal (2013) and Shivade et al. (2015) used corpus statistics for this purpose. Kim and de Marneffe (2013) investigated to what extent static word embeddings encode adjectival scales, and Kim et al. (2016) retrofit static word embeddings to better encode adjectival scales. In the context of pretrained language models, Garí Soler and Apidianaki (2020) showed that adjectival scales can be reconstructed to some extent from the representations of BERT.

Finally, Liu et al. (2023) automatically generated an NLI dataset for evaluating different properties about the interpretation of gradable adjectives. They found that a model finetuned only on MNLI fails to predict relationships between sentences with different uses of gradable adjectives but additional finetuning on a subset of their dataset led to considerable improvements in model predictions, and that these improvements also generalized to novel adjectives. Lorge and Pierrehumbert (2023) devised a similar task for scalar adverbs, such as *sometimes* and *often*.

These works trying to extract scalar items or probing language models for relationships between scalar items concern an important prerequisite for deriving scalar implicatures, namely identifying the members of the scale which are required to compute the implicature (see also Section 3). However, just because it is possible to extract this information from a language model does not necessarily mean the model uses this information in interpreting utterances, which is the focus of our work.

3. Background

The phenomenon we aim to target with our dataset is scalar implicatures with gradable adjectives.

Gradable adjectives (Kennedy and McNally, 2005) are adjectives that map to some scale. For example, the adjectives *warm* and *hot* are both gradable adjectives that map to a temperature scale. Most semantic theories assume that the meaning of such adjectives is based on a contextually-defined threshold θ (Kennedy and McNally, 2005; Lassiter and Goodman, 2017). A sentence, such as (2a), is then semantically true if the temperature of the coffee exceeds a threshold θ_{warm}, whereas (2b) is true if the temperature of the coffee exceeds a threshold $\theta_{hot} > \theta_{warm}$.

(2) a. The coffee is warm.

b. The coffee is hot.

Importantly, since there is no upper threshold, according to such a meaning, (2a) is also true if the temperature exceeds θ_{hot}, which goes against the intuition that a warm coffee is not hot. This intuition comes from the human ability to draw additional inferences as in this case the derivation of a **scalar implicature**. As mentioned above, theories of scalar implicature (Horn, 1984; Hirschberg, 1985) assume that listeners reason about alternative utterances that a speaker could have said but did not. In this case, when someone utters (2a), a listener reasons that if the coffee was actually hot, it would have been more informative to use the stronger scalar item *hot* and therefore it must be that the coffee has a temperature between θ_{warm} and θ_{hot}.

While such a mechanistic derivation leads to the intended interpretation in many cases, scalar implicatures tend to highly depend on the context and the specific scalar item (see, e.g., Gotzner and Romoli, 2022). For example, consider the pair of sentences in (3):

(3) a. The student is intelligent.

b. The student is brilliant.

This pair of sentences also contains a weaker (*intelligent*) and a stronger (*brilliant*) scalar item but experimental studies have shown that humans rarely derive the implicature that the student is not brilliant after hearing or reading (3a) (van Tiel et al., 2016). This fact, together with the observation that the same scalar item sometimes gives rise to an implicature and sometimes does not (see Section 1), limits the explanatory power of purely mechanistic accounts. A language model thus needs to be able to consider many linguistic subtleties to derive scalar implicatures in a human-like manner.

4. SIGA: A New Dataset for Pragmatic NLI

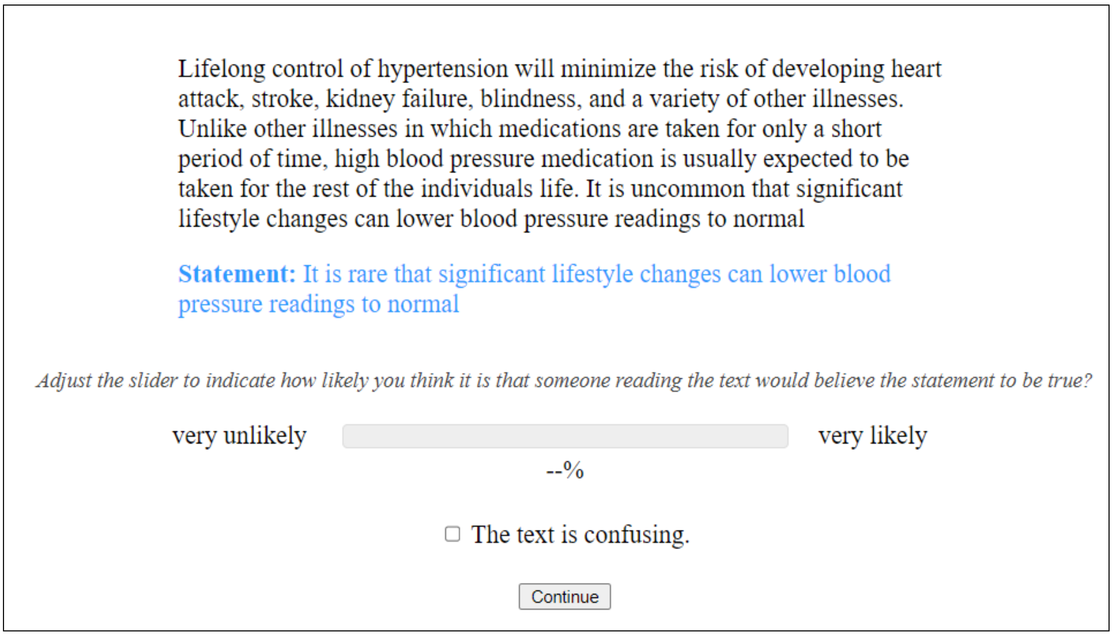
4.1. Data Collection

To construct our dataset we extracted examples from the C4 corpus (Raffel et al., 2020), which is a multilingual corpus of more than 10 languages that contains over 750GB of English language text (Dodge et al., 2021). To find suitable examples, we extracted all sentences that contained two related scalar adjectives separated by "*but not*", e.g., sentences that contained the phrase *good but not great*. Here we are exploiting the observation that implicatures can be explicitly reinforced (Hirschberg, 1985) and one linguistic device for such reinforcements is the frame “WEAK but not STRONG” (Hearst, 1992; de Melo and Bansal, 2013; van Miltenburg, 2015; Pankratz and van Tiel, 2021), where WEAK and STRONG are scalar items on the same scale of varying intensity.

Using this frame has several advantages over extracting sentences with only a weak adjective. First, some adjectives can be part of multiple scales. E.g., depending on the context “The movie was alright” can mean that the plot was not amazing or that it was not particularly funny (Hu et al., 2023b). By extracting utterances that contain both a weak and a strong scalar item, we can be certain which scale a speaker/writer was considering. Second, considering that the speaker/writer explictly reinforced the implicature means that the context does not rule out the implicature. And lastly, in some cases the speaker/writer may have explicitly reinforced the implicature because it is surprising to appear in that context and we hypothesized that this would give us a set of contexts with varying levels of support for the implicature.

Mining such examples requires a list of adjective pairs that map onto the same scale for which we used a list of 88 pairs compiled by de Melo and Bansal (2013). This procedure resulted in about 10,000 examples, of which we sampled 1,600 examples such that we included all examples with rare adjective pairs and subset of examples with the most common adjective pairs, resulting in a maximally balanced distribution of adjective pairs. Following de Marneffe et al. (2019) and Parrish et al. (2021), we also extracted the two previous sentences (if they exist) along with the sentence containing the scalar items.

In order to generate premise-hypothesis pairs for the NLI task, we then removed "but not STRONG" from the original utterance to form the premise. To form the hypothesis, we copied the last sentence of the premise and replaced the weak scalar item with the strong scalar item. To illustrate this, consider the following example, which shows the context (the preceding two sentences), the original utterance,



Lifelong control of hypertension will minimize the risk of developing heart attack, stroke, kidney failure, blindness, and a variety of other illnesses. Unlike other illnesses in which medications are taken for only a short period of time, high blood pressure medication is usually expected to be taken for the rest of the individuals life. It is uncommon that significant lifestyle changes can lower blood pressure readings to normal

Statement: It is rare that significant lifestyle changes can lower blood pressure readings to normal

Adjust the slider to indicate how likely you think it is that someone reading the text would believe the statement to be true?

Figure 1: An example item in our annotation task.

the modified target utterance, and the hypothesis.

- **Context:** Of those, I watched Criminal Minds-Suspect Behavior, Memphis Beat, and The Protector. See, I told you I watch a lot of cop shows.
- **Original Utterance:** These shows were good, but not great.
- **Target utterance:**[1] These shows were good.
- **Hypothesis:** These shows were great.

4.2. Crowdsourced Judgments

As we mentioned before, a sentence with a scalar item does not necessarily give rise to a scalar implicature. To estimate whether the modified utterances give rise to implicatures, we used a crowdsourcing task that we conducted on Prolific.

Task description For each example, we presented annotators with the context including the premise and a "statement" (the hypothesis). We then asked participants to adjust a slider from 0 to 100 to indicate how likely they thought it was that someone reading the text would believe the statement to be true. Following Parrish et al. (2021), we used a non-linear slider that allowed greater precision at the slider's endpoints assuming that a difference between 99% and 100% is more meaningful than a difference between 50% and 51% (Tversky and Kahneman, 1981). Slider endpoints were labeled "very unlikely" and "very likely." See Figure 1 for an example annotation task.

Using a continuous slider deviates from the classic NLI data collection paradigm that directly asks annotators to provide categorical entailment labels (e.g., Bowman et al., 2015; Williams et al., 2018). Further, we framed the annotation task as asking what someone else would believe rather than asking annotators about their personal beliefs. The reason for these deviations is that pragmatic inferences are much less robust than logical entailment, and importantly they can be explicitly canceled (Hirschberg, 1985). Therefore, we assumed that some annotators would rarely provide contradiction or entailment labels if they interpret the task as providing logical entailment relations, as some annotators have been shown to do (Pavlick and Kwiatkowski, 2019). Asking annotators to provide continuous ratings about a generic listener's beliefs alleviates these issues to some extent as this allows them to express typical interpretations with some uncertainty.

Annotator pool We initially recruited 100 annotators who completed a pre-screener similar to the one used by Parrish et al. (2021). Annotators were paid USD 2.00 which resulted in a hourly rate of approximately USD 15.00/hr. The pre-screener contained 10 questions of which we used 8 for selecting participants whose ratings fell into predefined ranges for at least 75% of the questions. This resulted in a pool of 83 annotators of which 62 completed at least one annotation task.

For the actual annotation tasks, we created batches of 20 examples together with 5 examples

[1] This is after the context, the final sentence of the premise.

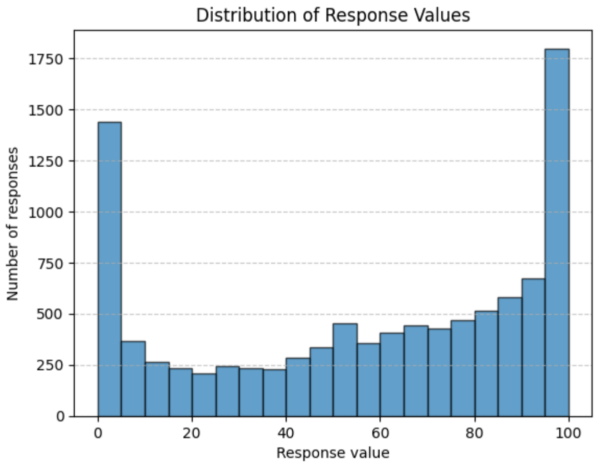


Figure 2: Distribution of raw responses from the annotation task.

	Percentage of Examples
Unanimous Agreement	41.99
Majority Agreement (> 0.6)	40.85
No Majority	17.15

Table 1: Agreement levels for NLI label categories **C**ontradiction, **N**eutral and **E**ntailment.

from MNLI, which we used for calibrating an annotator's use of the scale (see Section 3.3). Each task was completed by five annotators, and annotators completed between 1 and 12 annotation tasks (average: 6.6 tasks). No annotator completed the same annotation task more than once.

Annotators completed a task on average in 10 minutes and were paid USD 2.00 (resulting in a rate of USD 12.00-15.00/hr). We excluded ratings from tasks that were completed in less than 4 minutes, resulting in the exclusion of ratings from 54 tasks (13.5%).

Results Figure 2 shows the distribution of responses on examples with scalar items. As this figure shows, annotators used the entire scale but at the same time, many ratings are concentrated at the endpoints.

4.3. Mapping to NLI Labels

To map the continuous ratings from annotators to the NLI labels “contradiction”, “neutral” and “entailment,” we estimated two thresholds (Chen et al., 2020; Parrish et al., 2021) for each annotator:[2] one threshold for contradiction below which we labeled examples as contradiction and one threshold for entailment above which we labeled examples as entailment. For this estimation we used the ratings for the MNLI examples and their corresponding ground truth labels from the original dataset.[3] Concretely, we optimized the thresholds for each annotator such that the resulting decision boundaries maximize the annotator's F1-score on MNLI examples. As a regularization measure and to prevent extreme thresholds due to outliers, we also considered the MNLI examples from all other annotators when computing the per-annotator thresholds. For each annotator, we computed a weighted micro-averaged F1 with a weight of 0.05 for examples that are from other annotators, resulting in an average F1 score of 82.29 on the MNLI examples.

Using the per-annotator thresholds, we then mapped the examples involving scalar items to the three entailment labels. To assess the quality of this mapping, we computed agreement statistics of these labels (Table 1). For examples with unanimous agreement, the mapped label is the same for all annotators who rated that example; for examples with majority agreement the mapping resulted in labels that were the same for more than 60% of the annotators; for examples without a majority, no label was the same for more than 60% of the annotators. As is common practice for NLI datasets (e.g. Bowman et al., 2015), we discarded all examples for which there was no majority label, resulting in the removal of 500 examples.[4]

4.4. Expert annotations

A manual inspection of some of the examples without unanimous agreement revealed that many ratings were close to the threshold between neutral and contradiction or close to the threshold between neutral and entailment. Since these distinctions are theoretically important (e.g., a hypothesis-premise pair labeled contradiction indicates a scalar implicature), we further refined the labels through a two-round expert annotation process.

In the first round, the first author annotated all examples with an NLI label (without access to the crowdsourced majority label). For 730 examples (66.36%), their label agreed with the crowdsourced majority label. For the remaining 370 examples (33.63%) for which their label was different from the majority label, the second author provided a

[2] We estimated individual thresholds for each annotator since different annotators likely used the rating scale differently.

[3] We sampled a balanced number of contradiction, neutral, and entailment examples. In order to avoid examples for which there may be disagreements betweeen different annotators, we exclusively sampled examples where all five annotators provided the same label.

[4] Most NLI datasets (e.g., SNLI and MNLI) exclude data points only if there is no label that more than half of the annotators assigned. We opted for the higher threshold of 60% with the goal of eliminating examples that exhibit systematic disagreements (see e.g., Pavlick and Kwiatkowski, 2019).

	C	N	E	Σ
training dataset	291	77	232	600
id-test dataset	85	17	73	175
ood-test dataset	99	24	111	234
Σ	475	118	416	1,009

Table 2: Number of examples per label for train and in-domain (id) and out-of-domain (ood) test splits.

Scalar items	train dataset	id-test
good/great	336	94
uncomfortable/painful	79	25
small/tiny	86	29
uncommon/rare	72	20
possible/practical	27	7

Table 3: Distribution of scalar items in training and in-domain test set. See Appendix A for the distribution of scalar items in the out-of-domain test set.

third NLI label for adjudication. For 331 examples, the third label succesfully adjudicated between the crowdsourced label and the annotation by the first author, for the remaining 39 examples, all three labels disagreed and the examples were discarded.

Lastly, the expert annotation process revealed 52 examples where the target utterance either consisted only of one word (e.g., just "Great.") or it no longer made sense after removing "but not STRONG" (e.g., "so" is likely interpreted differently in the original utterance "So small but not tiny." than in the modified utterance "So small."). We also removed these examples, resulting in a final dataset size of 1,009 examples.

4.5. Data splits and statistics

We divided the examples into three splits: a training split, an in-domain test split, and an out-of-domain test split. For the training and in-domain splits, we considered all 775 examples that include one of the 5 most frequent scalar items (see Table 3) and then randomly split this dataset into a training and an in-domain test set. The out-of-domain test set includes all utterances with the remaining 23 pairs of scalar items. See Appendix A for the distribution of scalar items in the out-of-domain test set.

Table 2 shows the distribution of labels in the three data splits. Most examples were either labeled as contradiction or entailment. Recall that in all examples, the target utterance in the premise and the hypothesis differ only by the adjective: the premise contains a weaker scalar item, and the hypothesis contains a stronger scalar item. Therefore, an NLI pair labeled contradiction indicates that the premise gives rise to a scalar implicature, as in the following pair:

(4) a. [...] There are very mixed opinions, but some came back several times. The quality was mentioned as **good**.
b. The quality was mentioned as **great**.
(Label: Contradiction)

Entailment examples, on the other hand, indicate that the use of the weaker and stronger scalar item is seen as synonymous, as in the following NLI pair:

(5) a. I felt pretty good with it. I did that 17 miler at 7:50 pace. My quad felt **good**.
b. My quad felt great.
(Label: Entailment)

Finally, for most examples that were labeled as neutral, the context was not informative enough to indicate how the utterance should be interpreted and annotators considered it possible that the stronger statement in the hypothesis could be either true or false.

(6) a. but after cramping at the Yuengling Shamrock Marathon in March I decided to transfer my registration from the 50-mile to the 50k (31 miles). My training leading up to the race was **good**.
b. My training leading up to the race was **great**.
(Label: Neutral)

In summary, as both the individual examples as well as the aggregate statistics show, our dataset covers the heterogeneity of utterances with gradable adjectives. In some contexts, they give rise to implicatures and in others, they do not. Together with the division of examples into in-domain and out-of-domain test sets, this provides us with an optimal dataset for investigating whether NLI models can (learn to) draw scalar implicatures, as described in the following section.

5. Modeling Experiments

As mentioned in the introduction, we used our novel NLI benchmark to assess the ability of pre-trained language models to derive scalar implicatures on a diverse set of utterances.

5.1. Implementation details

Base model. We conducted all our experiments using DeBERTa (He et al., 2020), a masked language model similar to BERT (Devlin et al., 2019) with an improved attention mechanism. We used the weights of the "large" variant of the model[5]

[5] Accessed through the HuggingFace Tansformers library (Wolf et al., 2020) at `https://huggingface.co/microsoft/deberta-large-mnli`.

Test data	Model	Accuracy	F1 Score		
			C	N	E
SIGA in-domain	DeBERTa-large MNLI	37.71	8.88	3.71	59.22
	DeBERTa-large MNLI+SIGA	57.14	62.76	33.33	56.45
	Majority baseline	48.67	65.38	0.00	0.00
SIGA out-of-domain	DeBERTa-large MNLI	42.30	5.71	11.90	65.23
	DeBERTa-large MNLI+SIGA	52.56	54.27	19.60	58.71
	Majority baseline	42.31	59.46	0.00	0.00
MNLI	DeBERTa-large MNLI	91.29	93.82	87.71	92.23
	DeBERTa-large MNLI+SIGA	87.07	88.72	80.49	90.55

Table 4: Accuracy and F1 scores of models for in-domain and out-of-domain datasets. “MNLI” models were finetuned only on MNLI (base model); “MNLI+SIGA” were additionally finetuned on our training examples (finetuned model). C: Contradiction, N: Neutral, E: Entailment.

which has 350M parameters that were finetuned on the MNLI (Williams et al., 2018) dataset. We chose this model because it achieves above 90% accuracy on the MNLI evaluation splits and it has been shown to perform well on other pragmatic tasks such as predicting the content of presuppositions (Parrish et al., 2021).

Furthermore, a manual analysis of a subset of MNLI revealed that the dataset contains very few instances of pragmatic inferences (Jeretic et al., 2020). We confirm this finding through a more targeted anlaysis of all examples in MNLI for which the premise and the hypothesis contain different adjectives modifying the same noun phrase, which may be an indication for a pragmatic inference (e.g., as in the premise-hypothesis pair “I saw a good movie” and “I saw a great movie”). While there exist 1,771 examples with different adjectives across the two statements, in almost all cases the adjectives are intended to be synonyms (e.g., *small* and *little*) or they are ordinals (e.g., *first* and *second*) and thus do not constitute examples involving scalar implicatures. This lack of implicature examples in MNLI allows us to evaluate how much pretraining by itself equips models with the ability to draw pragmatic inferences.

Finetuning. To test whether models can learn to derive context-sensitive scalar implicatures, we also finetuned DeBERTa on our training split. We finetuned the model for 3 epochs with a learning rate of 5e-06 using the AdamW optimizer (Loshchilov and Hutter, 2019).

Evaluation. We evaluate all models separately on the in-domain test split and the out-of-domain test split. Since the predicted labels are theoretically meaningful in our dataset, we report both the overall accuracy as well as F1 scores for each prediction category.

5.2. Results and Discussion

Table 4 shows the results of the original model as well as the finetuned models on both test splits as well as the MNLI test split. The model that was exclusively finetuned on MNLI, i.e., on virtually no examples that require scalar implicature computations, performs very poorly on both the in-domain and the out-of-domain test split. This low performance is particularly pronounced for the contradiction and neutral examples, which indicates that the model incorrectly considers most examples involving scalar implicatures as entailment. This is not suprising considering that most examples in MNLI whose premise and hypothesis contain different adjectives on the same scale are contexts in which annotators treated the different adjectives as synonyms. Therefore, models trained on MNLI may have learned the heuristic that a premise and hypothesis involving two different adjectives on the same scale indicate an entailment relation.

If we additionally finetune the model on the SIGA training examples, we observe large increases in accuracy (significant at a level of $\alpha = 0.05$ according to a non-parametric permutation test) on both the in-domain and the out-of-domain test sets and for both datasets, and the models outperform a majority baseline that always predicts contradiction. This suggests that the model can to some extent learn when humans would draw scalar implicatures if it is trained on such examples and this behavior also transfers to examples involving unseen scalar items. At the same time, however, even with these additional training examples, the performance on the SIGA test sets is still a lot lower than the perfomance on MNLI, which highlights that examples with scalar implicatures still pose challenges for NLI models, even with explicit supervision.

Table 4 also shows that while both models achieve very high accuracy on MNLI, additional finetuning on SIGA examples led to a non-negligible drop in performance. In combination with the considerably improved performance on contradiction examples in the SIGA test sets, this raises the question whether the differences on all test sets primarily stem from the model predicting contradiction more often. An analysis of the differences in predicted labels between the base model and the finetuned model ruled out that explanation: While indeed many labels changed from entailment or neutral to contradiction on examples in SIGA, the finetuned model predicted a lot more neutral and entailment labels than the base model. However, for the MNLI examples the predictions changes were also very heterogeneous, so it was not the case that finetuning on SIGA examples only affected predictions on examples with scalar adjectives.

In summary, the comparisons between the base model and the finetuned models suggest that on the one hand, finetuning on a small number of examples involving scalar implicatures improves model performance on examples with both seen and unseen scalar items. On the other hand, however, as evidenced by the change in predictions on MNLI, the additional finetuning affects model behavior on a broad class of examples, which suggests that the additional finetuning targets not only the weights that are involved in deriving scalar implicatures.

6. General Discussion and Conclusion

In this work we set out to investigate to what extent NLI models based on pretrained language models can derive and can learn to derive scalar implicatures. For this purpose, we created SIGA, a high quality dataset consisting of naturalistic utterances involving gradable adjectives and annotated this dataset through a combination of crowdsourcing and expert annotations. In experiments with a state-of-the-art NLI model based on DeBERTa, we found that without phenomenon-specific finetuning, the model fails to derive scalar implicatures in almost all cases. With additional finetuning, on the other hand, we found that the model performs considerably better on examples involving this phenomenon and these improvements also transferred to examples with unseen scalar items.

Our findings corroborate recent findings that pretrained language model representations encode some information about gradable adjectives and their degrees of intensity on the corresponding scale (Garí Soler and Apidianaki, 2020; Liu et al., 2023). Further, our results echo the finding that additional finetuning is necessary to guide the model to make use of this information encoded in the representation: Like Liu et al. (2023), we found that the desired behavior started to surface only through finetuning on task-specific examples.

Despite the positive findings, our results and especially the comparison to other NLI datasets also highlights that current NLI models still struggle with drawing pragmatic inferences and there is still a considerable gap between human and model behavior.

7. Limitations

Despite the obvious limitation of constructing an English dataset and evaluating only models in English, there are some limitations of our work that should be considered when drawing conclusions from our findings.

First, while DeBERTa is one of the best performing masked language models for NLI, we did not evaluate more recent autoregressive models such as GPT-3/GPT-4 (Brown et al., 2020) or Llama 2 (Touvron et al., 2023) and therefore we cannot make definitive claims about larger models trained with additional objectives such as reinforcement learning from human feedback (Ouyang et al., 2022). However, the results by Ruis et al. (2022) and Hu et al. (2023a) suggest that many pragmatic inferences still pose challenges for even the latest large language models.

Second, compared to broader datasets such as MNLI or automatically generated datasets such as IMPPRES, our dataset is relatively small. Considering the substantial performance differences across models, we are nevertheless confident that our evaluation reveals systematic differences. Furthermore, our dataset is comparable in size to other specialized datasets such as the CommitmentBank (de Marneffe et al., 2019), which has been an invaluable resource to study human and model behavior with regards to speaker commitment.

Third, while we extracted all our examples from a corpus, the modification due to the removal of "but not STRONG" did lead to slightly unnatural sentences in some cases. For example, without "but not great", the sentence "It's good (but not great) and I'm not sure I'd pay $25 for it." becomes a bit odd due to the combination of the positive adjective *good* and the negative second conjunct of the sentence.

Finally, in this work, we made the simplifying assumption that all English language users would derive the same implicatures and we did not consider individual differences. However, both work in experimental pragmatics (e.g., Mayn and Demberg, 2022) and in NLP (Pavlick and Kwiatkowski, 2019; Jiang and de Marneffe, 2022; Plank, 2022) has recently highlighted that for many linguistic phenomena, there exist systematic disagreements. We consider investigating such disagreements both in our raw human ratings as well as in similar datasets an important direction for future work.

Despite these limitations, we consider SIGA a highly valuable resource for developing and evaluating natural language understanding models on their ability to draw human-like pragmatic inferences.

Acknowledgments

We thank the anonymous reviewers for their thoughtful feedback. This research was supported by the the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (Grant Agreement #948878).

8. Bibliographical References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. *Recognizing textual entailment: Models and applications*, volume 6 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2010. "was it good? it was provocative." learning the meaning of scalar adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176, Uppsala, Sweden. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

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):1–55.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Franke. 2009. *Signal to Act: Game Theory in Pragmatics*. Ph.D. thesis, Universiteit van Amsterdam.

Aina Garí Soler and Marianna Apidianaki. 2020. BERT knows Punta Cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online. Association for Computational Linguistics.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11):818–829.

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

Herbert P. Grice. 1975. Logic and Conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Speech Acts*, volume 3, pages 41–58. Academic Press.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Julia Bell Hirschberg. 1985. *A Theory of Scalar Implicature (Natural Languages, Pragmatics, Inference)*. PhD 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*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023a. 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.

Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. 2023b. Expectations over Unspoken Alternatives Predict Pragmatic Inferences. *Transactions of the Association for Computational Linguistics*, 11:885–901.

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.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Christopher Kennedy and Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, pages 345–381.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630, Seattle, Washington, USA. Association for Computational Linguistics.

Joo-Kyung Kim, Marie-Catherine de Marneffe, and Eric Fosler-Lussier. 2016. Adjusting word embeddings with semantic intensity orders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 62–69, Berlin, Germany. Association for Computational Linguistics.

Daniel Lassiter and Noah D. Goodman. 2017. Adjectival vagueness in a Bayesian model of interpretation. *Synthese*, 194(10):3801–3836.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Wei Liu, Ming Xiang, and Nai Ding. 2023. Adjective scale probe: Can language models encode formal semantics information? In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Isabelle Lorge and Janet Pierrehumbert. 2023. Not wacky vs. definitely wacky: A study of scalar adverbs in pretrained language models. *arXiv preprint arXiv:2305.16426*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Alexandra Mayn and Vera Demberg. 2022. Individual differences in a pragmatic reference game. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Elizabeth Pankratz and Bob van Tiel. 2021. The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition*, 13(4):562–594. Edition: 2021/08/16 Publisher: Cambridge University Press.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. *Journal Machine Learning Research*, 21(1).

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics.

Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M. Lai. 2015. Corpus-based discovery of semantic intensity scales. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–493, Denver, Colorado. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Emiel van Miltenburg. 2015. Detecting and ordering adjectival scalemates. In *Proceedings of MAPLEX*, Yamagata, Japan.

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

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online. Association for Computational Linguistics.

A. Scalar Items in Out-of-distribution Test Set

Scalar items	out of distribution dataset
big/huge	43
good/best	28
thin/skinny	21
uncommon/unusual	17
dim/dark	17
interesting/fascinating	16
interesting/exciting	16
clean/spotless	15
neglected/forgotten	13
unfortunate/fatal	9
known/famous	7
big/large	6
further/far	5
thick/impenetrable	4
unusual/strange	4
unfortunate/disastrous	3
bleak/hopeless	3
great/best	2
small/little	2
uncommon/extraordinary	1
sufficient/ample	1
strange/weird	1

Table 5: Distribution of scalar items in out of distribution test set.