FLUTE: Figurative Language Understanding through Textual Explanations

Tuhin Chakrabarty[1] **Arkady Saakyan**[1] **Debanjan Ghosh**[2] **Smaranda Muresan** [1]
[1]Department of Computer Science, Columbia University
[2]Educational Testing Service
tuhin.chakr@cs.columbia.edu, a.saakyan@columbia.edu, dghosh@ets.org, smara@cs.columbia.edu

Abstract

Figurative language understanding has been recently framed as a recognizing textual entailment (RTE) task (a.k.a. natural language inference, or NLI). However, similar to classical RTE/NLI datasets, the current benchmarks suffer from spurious correlations and annotation artifacts. To tackle this problem, work on NLI has built explanation-based datasets such as e-SNLI, allowing us to probe whether language models are right for the right reasons. Yet no such data exists for figurative language, making it harder to assess genuine understanding of such expressions. To address this issue, we release FLUTE, a dataset of 9,000 figurative NLI instances with explanations, spanning four categories: Sarcasm, Simile, Metaphor, and Idioms. We collect the data through a model-in-the-loop framework based on GPT-3, crowd workers, and expert annotators. We show how utilizing GPT-3 in conjunction with human annotators (novices and experts) can aid in scaling up the creation of datasets even for such complex linguistic phenomena as figurative language. The baseline performance of the T5 model fine-tuned on FLUTE shows that our dataset can bring us a step closer to developing models that understand figurative language through textual explanations.

1 Introduction

Figurative language such as metaphors, similes or sarcasm plays an important role in enriching human communication, allowing us to express complex ideas and emotions in an implicit way (Roberts and Kreuz, 1994; Fussell and Moss, 1998). However, understanding figurative language still remains a bottleneck for natural language processing (Shutova, 2011). Recently Jhamtani et al. (2021) show that when faced with dialog contexts consisting of figurative language,some models show very large drops in performance compared to contexts without figurative language. Despite the fact that Transformer-based pre-trained language models (LMs) get even larger (Brown et al., 2020; Raffel et al., 2020), they are still unable to comprehend the physical world, cultural knowledge, or social context in which figurative language is embedded (Bisk et al., 2020).

In recent years, there have been several benchmarks dedicated to figurative language understanding, which generally frame "understanding" as a recognizing textual entailment (a.k.a natural language inference (NLI)) task — deciding whether one sentence (premise) entails/contradicts another (hypothesis) (Chakrabarty et al., 2021; Stowe et al., 2022; Srivastava et al., 2022). However, similar to general NLI datasets, these benchmarks suffer from spurious correlations and annotation artifacts (McCoy et al., 2019; Poliak et al., 2018b). These can allow large language models (LLMs) to achieve near human-level performance on in-domain test sets, yet turn brittle when evaluated against out-of-domain or adversarial examples (Glockner et al., 2018; Ribeiro et al., 2016, 2020). To tackle these problems, research in NLI has argued that it is not enough to correctly predict the entail/contradict labels, but also to explain the decision using natural language explanations that are comprehensible to an end-user assessing model's reliability (Camburu et al., 2018; Majumder et al., 2021; Wiegreffe et al., 2021a), leading to novel datasets such as e-SNLI (Camburu et al., 2018). Yet, there is no such dataset for figurative language, hindering our ability to assess LLMs' genuine understanding of figurative language.

In this paper, we make several contributions towards the goal of building models and assessing their ability to understand figurative language:

- **FLUTE: a new benchmark for figurative language understanding through textual explanations**. FLUTE contains 9,000 high-quality <literal, figurative> sentence pairs with entail/contradict labels and the associ-

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7139–7159
December 7-11, 2022 ©2022 Association for Computational Linguistics

Type	Premise (literal)	Hypothesis (figurative*)	Label	Explanation
Paraphrase + Sarcasm	My next door neighbors are *always arguing* in our shared hallway.	It's *so annoying* to have to hear my next door neighbors *argue all the time* in our shared hallway.	E	The sound of arguing neighbors can often be very disruptive and if it happens all the time in a common space like a shared hallway it is natural to find it annoying.
		It's *so pleasant* to have to hear my next door neighbors *argue all the time* in our shared hallway.	C	The sound of arguing neighbors can often be very disruptive and so someone considering it to be pleasant is not really accurate.
Simile	The assembly hall was now *hot and moist*, more so than usual.	In fact, the assembly hall was now *like a steam sauna*.	E	A sauna is a hot and moist environment, so the simile is saying that the hall is even hotter and more moist than usual.
	The assembly hall was now *cold and dry*, more so than usual.		C	A steam sauna is a small room or hut where people go to sweat in steam, so it would be hot and humid, not cold and dry.
Metaphor	He *mentally assimilated* the knowledge or beliefs of his tribe.	He *absorbed the knowledge* or beliefs of his tribe.	E	To absorb something is to take it in and make it part of yourself.
	He *utterly decimated* his tribe's most deeply held beliefs.		C	Absorbed typically means to take in or take up something, while "utterly decimated" means to destroy completely.
Idiom	Lady Southridge was wringing her hands, *trying hard and desperately to salvage* the bleak and miserable situation so that it somehow looks positive.	Lady southridge was wringing her hands, trying *to grasp at straws.*	E	To grasp at straws means to make a desperate attempt to salvage a bad situation, which is exactly what Lady Southridge is trying to do.
	Lady Southridge was wringing her hands, *doing absolutely nothing to overturn* the bleak and miserable situation so that it somehow looks positive.		C	To grasp at straws means to make a desperate attempt to salvage a bad situation, but the sentence describes not doing anything to change the situation

Table 1: FLUTE examples of figurative text (hypothesis) and their respective literal entailment(E) and contradiction (C) premises, along with the associated explanations. * For simile, metaphor, and idiom, figurative examples are the hypothesis whereas for sarcasm, we have both figurative and literal hypotheses (see Section 2).

ated explanations. The benchmark spans four types of figurative language: sarcasm, simile, metaphor, and idiom. Table 1 shows examples from our dataset. A noteworthy property of FLUTE is that both the entailment/contradiction labels and the explanations are w.r.t the figurative language expression (i.e., metaphor, simile, idiom) rather than other parts of the sentence.

- **A scalable model-in-the-loop approach for building FLUTE.** Model-in-the-loop approaches (i.e., GPT-3 (Brown et al., 2020) and crowdsourcing) have been recently proposed to generate NLI datasets, as well as free-form textual explanations (a.k.a natural language explanations (Camburu et al., 2018)) for model decisions (Liu et al., 2022a; Wiegreffe et al., 2021a). For figurative language, Ghosh et al. (2020) has shown that crowdworkers are mostly good at performing minimum edits to generate a literal sentence from a sarcastic one (e.g., using negation or antonyms), which can lead to trivial examples easily classified by LLMs (Chakrabarty et al., 2021). Thus, for building FLUTE, we leverage the power of GPT-3 to generate diverse and high quality literal text (paraphrases/contradictions and/or explanations) using few-shot prompting, coupled with minimal human involvement (e.g., crowdworkers to minimally edit a literal sentence to make it sarcastic and experts for judging and minimally editing GPT-3 output to ensure quality control) (Section 2).

- **Comprehensive set of experiments to assess FLUTE's usefulness towards building models that understand figurative language.** We propose a setup inspired by instruction-based learning (Mishra et al., 2021; Sanh et al., 2021; Wei et al., 2021) and train a T5 (Raffel et al., 2019) model to jointly predict the label (entail/contradict) and explanation. We train two variants: T5 trained on e-SNLI dataset (Camburu et al., 2018) and T5 trained on FLUTE. We evaluate our model on the FLUTE test set (Section 3.2). We propose extensive automatic and human evaluation experiments to assess model understanding through explanations (Section 3.2). We show that the model trained on FLUTE produce higher quality explanations compared to model trained on e-SNLI (Section 4).

Our code and experimental setup is available at [1].Our data can be accessed at [2]

2 Model-in-the-loop for building FLUTE

FLUTE consists of pairs of premises (literal sentences) and hypotheses (figurative sentences)[3], with the corresponding entailment or contradiction labels (NLI instances), along with explanations for each instance (Table 1). We describe the model-in-the-loop methods for creating premise-hypothesis pairs for each type of figurative language (Section 2.1) and the associated explanations (Section 2.2).

2.1 FLUTE: Premise-Hypothesis Pair Creation

2.1.1 Sarcasm

When asked to generate literal equivalents of sarcastic sentences crowdworkers on Amazon Mechanical Turk (MTurk) usually perform trivial rephrasings at word/phrase level (Ghosh et al., 2020), which can lead to NLI datasets for sarcasm understanding where LLMs can achieve near-human performance (95%) due to simple lexical cues, such as negation or antonyms (Chakrabarty et al., 2021). Additionally, in many cases sarcasm data is collected from Twitter using hashtags, e.g., #sarcasm, which can be noisy and not diverse.

To address these issues we take a model-in-the-loop approach: given a literal sentence we first use GPT-3 with few-shot prompting to generate a literal paraphrase, and then use crowdworkers to minimally edit this new literal sentence to form a sarcastic one (Figure 1b).[4] Then we pair the original literal sentence with generated literal paraphrase as entailment pair, and with the sarcastic one as a contradiction pair. Below we describe these two steps.

Entailment Pairs. Jointly modeling emotion and sarcasm had been shown beneficial for sarcasm detection (Chauhan et al., 2020). Thus, we select the literal sentences from the Empathetic Dialogue dataset (Rashkin et al., 2019). Each conversation in the dataset is grounded in a situation with emotion label provided. We select literal sentences labeled with negative emotions such as *angry*, *afraid*, *embarrassed*.[5] Including the emotion in the GPT-3 prompts serves two purposes: (a) the generated paraphrases are complex and often more creative than the original literal input, and (b) it is easier for crowdworkers to transform paraphrases with emotional content into sarcastic counterparts with minimal edits. To generate the literal paraphrases, we provide the literal sentence and the associated emotion in the prompt and ask GPT-3 to paraphrase the input (top part of Figure 1b, see prompt in Appendix A.1.2). Every paraphrase generated from GPT-3 is verified by 3 experts. If the quality of the generated paraphrase is deemed insufficient, it is resampled from the model. Any individual example undergoes at most three rounds of sampling, with 15% of them judged as appropriate upon the first round.

Contradiction pairs. We recruit crowd workers on MTurk to convert the manually checked GPT-3-generated literal paraphrases into sarcastic sentences. The workers were provided with the paraphrases and instructed to make minimal edits (e.g., through negations, antonyms) to generate sarcastic sentences. We conducted a qualification test, and recruited 29 distinct workers from the original set of 85 workers.[6] We recruit two independent workers for every paraphrase input. The resulting sarcastic outputs are verified by three experts. 25% of instances were deemed insufficient quality and edited by the experts. Consider the sarcasm example in Table 1. The literal input "My next door ..." is the premise. The first step generated the paraphrase hypothesis by adding the implicitly stated emotion "annoyed" and paraphrasing. Next, Turkers modified the paraphrase to its sarcastic counterpart - by replacing "annoying" (the emotion word) to it antonym "pleasant".

The final FLUTE sarcasm benchmark consists of 2,678 sarcastic sentences contradicted by the

[1] https://github.com/tuhinjubcse/model-in-the-loop-fig-lang

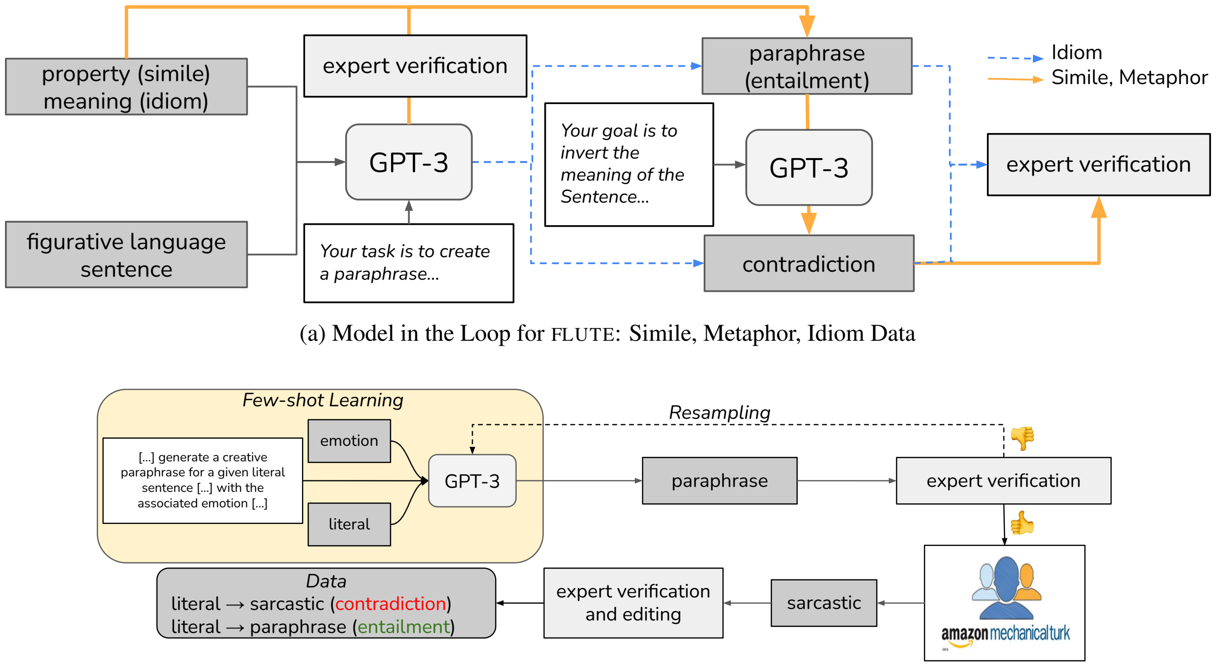
[2] https://huggingface.co/datasets/ColumbiaNLP/FLUTE

[3] Given sarcasm is the opposite of the literal meaning, we would only have contradictions in the dataset, thus we also generate a literal hypothesis that entails the literal premise.

[4] Using GPT-3 to directly generate sarcastic sentences led to low quality output (no semantic consistency, social biases, stereotypes, or toxic content.

[5] Although positive emotions can be used to generate sarcasm, we leave it for the future work.

[6] with native English proficiency, location in the U.S. and a hit success rate of 96%.


property (simile)
meaning (idiom)
figurative language sentence
expert verification
GPT-3
Your task is to create a paraphrase...
paraphrase (entailment)
Your goal is to invert the meaning of the Sentence...
GPT-3
contradiction
expert verification
Idiom
Simile, Metaphor
Few-shot Learning
[...] generate a creative paraphrase for a given literal sentence [...] with the associated emotion [...]
emotion
literal
GPT-3
Resampling
paraphrase
expert verification
amazon mechanical turk
sarcastic
expert verification and editing
Data
literal → sarcastic (contradiction)
literal → paraphrase (entailment)

(a) Model in the Loop for FLUTE: Simile, Metaphor, Idiom Data

(b) Model in the Loop for FLUTE: Sarcasm Data

1,339 seed literal sentences, as well as the 1,339 paraphrased sentences (entailment pairs), as seen in Table 2.

2.1.2 Simile

Entailment Pairs. We start by extracting sentences containing similes from (Chakrabarty et al., 2022, 2021). To generate an entailment pair, we perform two steps (see Fig. 1a): 1) given the sentence that contains a simile, create an auxiliary literal sentence by simply replacing the simile with the simile's property (e.g., replace *like a steam sauna* with *hot and moist*); 2) given the auxiliary literal sentence and the property, prompt GPT-3 to generate a *literal paraphrase* consistent with the property (see prompt in Appendix A.2.3). Experts deemed 720 generated instances satisfactory. To illustrate the pipeline, see the example for simile in Table 1: "In fact the assembly hall was now like a steam sauna". The object *steam sauna* is replaced with its property *hot and moist*, and the resulting sentence is fed into GPT-3 to generate the paraphrase "The assembly hall was now hot and moist, more so than usual".

Contradiction pairs. To generate the contradictions, we prompt GPT-3 to invert the meaning of the *literal paraphrase* (see above) w.r.t the property (see prompt in Appendix A.2.4). Out of 720 generated contradictions, experts deemed 642 satisfactory. We further selected 108 challenging <simile, entailment, contradiction> instances from Liu et al. (2022b), defined by low RoBERTa (Zhuang et al., 2021) logit of the correct option (see Appendix A.2.2 for details), for a total of 750 entailment and 750 contradiction pairs (Table 2).

2.1.3 Metaphors

Entailment Pairs. We follow a similar model-in-the-loop approach as the one for similes (Figure 1a). We manually select a total of 750 metaphors from the following datasets: (Chakrabarty et al., 2021; Srivastava et al., 2022; Stowe et al., 2022). Next, we prompt GPT-3 to generate paraphrases given the metaphoric sentences (see prompt in Appendix A.4.2). Although the original datasets contain the literal equivalents of the metaphoric sentence, they are not fully adequate for our purposes because: (a) not all metaphor examples have the literal counterpart, and (b) often, the literal counterpart is a minimal modification (one-word edit) of the metaphor (Srivastava et al., 2022; Chakrabarty et al., 2021), which can lead to trivial examples for LLMs.

Contradiction Pairs. To generate contradictions pairs, we start with the GPT-3 generated literal sentence that entails the metaphoric sentence. Consider the metaphor in the Table 1, "He *absorbed* the knowledge or beliefs of his tribe" (taken from (Stowe et al., 2022)). In the original dataset, the non-entailment counterpart is "He *absorbed* the beverages of his tribe", which is created by using a verb in a different sense (literal sense of "absorb")

	Entails	Contradicts	Total
Paraphrase	1339	-	1339
+ Sarcasm	-	2678	2678
Simile	750	750	1500
Metaphor	750	750	1500
Idiom	1000	1000	2000

Table 2: Dataset statistics showing distribution of Figurative Language across FLUTE.

to fit the context of beverage drinking. On the contrary, since we are interested in generating instances that contradict the *metaphor itself*, a more appropriate modification would be "He *utterly decimated* his tribe's most deeply held beliefs" (more examples are in Table 6 in the Appendix). We follow the same method to generate contradiction examples (using GPT-3) as for similes (see prompt in Appendix A.4.3).

Both paraphrases and contradictions are verified by three experts and edited when required. Our FLUTE benchmark contains 750 entailment and 750 contradictions pairs for metaphors (Table 2).

2.1.4 Idioms

Observing the successful generations of paraphrases and contradictions by GPT-3 in the case of simile and metaphors, we jointly generate paraphrases along with contradictions using GPT-3 (See Figure 1a blue dotted lines). We provide the idiom and its meaning in the prompt (see prompt in Appendix A.6). Three experts manually verified all the generated sentences and edited a total of 23% of total generations. We found that jointly generating paraphrases and contradictions greatly eased the data creation process and resulted in relatively high quality of generations.

FLUTE benchmark consists of 1000 entailment and 1000 contradiction pairs for idioms (Table 2).

2.2 FLUTE: Generating Textual Explanations

Our task prediction requires that the model not only correctly infer the label, but also explain *why* a given premise entails or contradicts the hypothesis. Towards this goal, we generate textual explanations for every <premise, hypothesis> pair.

For simile, metaphor, and sarcasm we provide the premise, hypothesis, and label (entailment or contradiction) and prompt GPT-3 to generate an explanation. We provide a natural language instruction followed by several examples. We generate entailment and contradiction explanations separately.

For idioms, the idiom meaning in the seed dataset already makes up for a great explanation. Thus, we utilize the provided idiom meaning to jointly generate the explanation for the entailing premise, as well as for the contradicting premise using GPT-3. Hence, for idiom data in addition to premise, hypothesis, and labels, we also provide the idiom itself and its meaning in the prompt.

LLMs such as GPT-3 have been scrutinized heavily because they can mimic or amplify societal bias (Sheng et al., 2021), religious stereotypes (Abid et al., 2021), and gender stereotypes (Borchers et al., 2022).[7] For example, applications such as story generation often emulate societal bias by including more masculine characters and following social stereotypes based on their training data (Lucy and Bamman, 2021). However, this bias is not evident in the FLUTE dataset probably because the model is told to explain specific figurative instances and not to write creatively. In addition, we are reusing some of the standard datasets (Chakrabarty et al., 2021; Stowe et al., 2022; Srivastava et al., 2022) for FLUTE, which have probably already been removed of any provocative context. Finally, experts manually verified all explanations to ensure their correctness and ability to explain the essence of the entailment or contradiction in reasonable detail rather then learning a simple template (see Table 1). In cases where explanations were not accurate, experts edited them to ensure they are coherent, logically consistent, and grammatical. For explanations pertaining to sarcasm+paraphrase, experts edited a total of 21% of the generated explanations, while for simile, metaphor and idiom it was 27%, 40% and 10% respectively, which further demonstrates the potential of using GPT-3 to significantly reduce the human effort that goes into collecting textual explanations datasets. See Appendix A for details on hyperparameters and prompts.

3 Experimental Setup

3.1 Models

Prior works in explainability have trained two types of models. *Pipeline* models map an input to a rationale (I → R), and then a rationale to an output (R → O).[8] *Joint Self Rationalizing* models map an input to an output and rationale (I → OR). Recently Wiegreffe et al. (2021b) have exposed the short-

[7] For a comprehensive analysis please see Sheng et al. (2021).

[8] Rationales are "textual explanation" in this work, sometime used interchangeably.

comings of free-text pipelines and have empirically shown that joint model rationales are more indicative of labels. Following this, we fine-tune a joint self-rationalizing T5 model. Taking advantage of the text-to-text format of T5 (Raffel et al., 2020) and the recent success of instruction-based models (Sanh et al., 2021; Wei et al., 2021), we design the following instruction for a given literal premise (P) and a figurative hypothesis (H):

*Does the sentence **"P"** entail or contradict the sentence **"H"**? Please answer between **"Entails"** or **"Contradicts"** and explain your decision in a sentence.*

The above instruction is fed to the encoder of T5. The decoder outputs the label followed by the rationale. We fine-tune T5 with the following setups: in the first one, we fine-tune on e-SNLI (Camburu et al., 2018), and in the second, we fine-tune on FLUTE.

T5$_{\text{e-SNLI}}$: e-SNLI (Camburu et al., 2018) dataset comes with supervised ground-truth labels and rationales. We fine-tune the 3B version of T5 on e-SNLI for one epoch with a batch size of 1024, and an AdamW Optimizer with a learning rate of $1e-4$. We remove the *Neutral* examples from e-SNLI because our test data does not have such a category. We take the longest explanation per example in e-SNLI since our data has only one reference explanation. In case the explanations are more than one sentence we join them using 'and' since our data contains single-sentence explanations. This leaves us with 366,603 training and 6,607 validation examples.

T5$_{\text{FLUTE}}$: We fine-tune the 3B version of T5 model for 10 epochs with a batch size of 1024, and an AdamW Optimizer with a learning rate of $1e-4$ in a multitask fashion with data from all the four types of figurative languages combined. Our training data consists of 7,035 samples which is 50X smaller than e-SNLI. For validation we use 500 examples which is used for selecting best checkpoint based on loss.

3.2 Evaluation Setup

To evaluate the above models, we built a test set by randomly selecting 750 instances (i.e., <premise, hypothesis> pairs with associated explanations) from the sarcasm dataset, and 250 examples each from simile, metaphor and idiom datasets, for a total of 1,500 instances.

Below we describe several automatic metrics and human evaluations we consider to assess the models' ability to understand figurative language.

Automatic Metrics To judge the quality of the explanations we compute the average between BERTScore (Zhang et al., 2020) [9] and BLEURT (Sellam et al., 2020), which we refer to as *explanation score* (between 0 and 100). Instead of reporting only label accuracy, we report label accuracy at three thresholds of explanation score (0, 50, and 60). Accuracy@0 is equivalent to simply computing label accuracy, while Accuracy@50 counts as correct only the correctly predicted labels that achieve an explanation score greater than 50.

Rationale Quality Human simulatability (Doshi-Velez and Kim, 2017) has a rich history in machine learning interpretability research as a reliable measure of rationale quality from the lens of utility to an end-user. Simulatability measures the additional predictive ability a rationale R provides over the input I for a given label O, computed as the difference between task performance when a rationale is given as input vs. when it is not (IR → O minus I → O).In prior work on Explanability using Natural Language Wiegreffe et al. (2021b) pointed out that model predictions are often unable to be simulated because they degenerate under high values of noise. Following Wiegreffe et al. (2021b) we thus use a variant of this metric that relies on predicting the gold labels as our measure of rationale quality: (IR → Ô minus I → Ô).

To compute rationale quality, we first train IR → O and I → O for both FLUTE and e-SNLI data. We then compute the test accuracy for FLUTE for both the IR → O and I → O models trained on e-SNLI and FLUTE using predicted rationales from the respective I → OR models. We also compute rationale quality with gold rationales (R*).

Human Evaluation. Finally, we measure the quality of the generated textual explanations from T5$_{\text{e-SNLI}}$ and T5$_{\text{FLUTE}}$ models via the MTurk platform. We recruit 79 crowd workers with at least 98% HIT approval rate. We compute human judgement scores (H$_{\text{score}}$), identical to the e-ViL score in Kayser et al. (2021). For each NLI instance (a total of 200 random samples, 50 per figurative language type), we present two textual explanations generated by the two models (T5$_{\text{e-SNLI}}$ and T5$_{\text{FLUTE}}$) and ask three workers the following question: *Given*

[9] We use the deberta-mnli version that has shown to have highest correlation with human judges.

	$\text{T5}_{\text{e-SNLI}}$						T5_{FLUTE}					
	Acc @0	Acc @50	Acc @60	H_{score}	Yes%	No%	Acc @0	Acc @50	Acc @60	H_{score}	Yes%	No%
Sarcasm	60.6	15.7	2.4	34.2	14.7	52.0	**91.6**	**86.2**	**56.2**	**85.3**	**75.3**	**8.7**
Simile	61.2	22.8	3.6	43.6	22.0	40.7	**62.8**	**57.2**	**30.4**	**84.9**	**74.7**	**8.0**
Metaphor	**81.8**	31.8	11.6	55.3	36.0	28.0	73.3	**55.6**	**23.7**	**80.2**	**64.0**	**6.0**
Idiom	**84.8**	46.4	7.6	60.9	37.3	24.7	79.2	**77.2**	**66.8**	**83.1**	**69.3**	**8.7**

Table 3: Accuracy scores across four figurative language types by varying thresholds of explanation score, along with human evaluation scores H_{score}, Yes% (higher is better), and No% (lower is better) for explanations generated by T5 fine-tuned on e-SNLI ($\text{T5}_{\text{e-SNLI}}$) and T5 fine-tuned on FLUTE (T5_{FLUTE}). $p < 0.001$ via Wilcoxon signed-rank test for all bolded results.

the two sentences, does the explanation justify the answer above? We provide four options: *Yes* (1), *Weak Yes* ($\frac{2}{3}$), *Weak No* ($\frac{1}{3}$), and *No* (0). For each explanation, we average the scores by the three annotators and report the sample average in Table 3 as H_{score}. If the answer is anything other than *Yes*, we ask to categorize the shortcomings of the explanation: *Insufficient Justification*, *Too Trivial*, *To Verbose*, *Untrue to Input*, *Violates Common Sense* (Majumder et al., 2021).

Three workers were recruited for each instance and the IAA using Krippendorff's α (Krippendorff, 2011) between the workers is 0.45, indicating moderate agreement. See Appendix B for details on the H_{score} computation and Figure 4 for a screenshot of the MTurk task interface.

	Ac (IR → O)	Ac (I → O)	RQ.
e-SNLI	68.4	74.5	-6.1
FLUTE	89.3	90.5	**-1.2**
FLUTE (R*)	95.6	90.5	**5.1**

Table 4: Rationale Quality ($p < 0.001$ via Wilcoxon signed-rank test) (higher is better) on FLUTE test set using accuracy of IR → O and I → O models trained on e-SNLI and FLUTE respectively. We use the predicted rationale R obtained from respective I → OR models $\text{T5}_{\text{e-SNLI}}$ and T5_{FLUTE} as well as gold rationale (R*). Rationale Quality and accuracy are abbreviated as RQ. and Ac.

4 Results and Discussion

Table 3 shows accuracy at varying *explanation score* thresholds. A threshold of 0% does not account for the quality of the textual explanation and is equivalent to simply reporting label accuracy. With an increase in threshold to greater than 50% we see accuracy scores dropping almost by half for $\text{T5}_{\text{e-SNLI}}$, showing most explanations generated from the model trained on e-SNLI are of poor quality. By increasing the threshold to greater than 60%, the accuracy scores further decrease, demonstrating that models like T5 fine-tuned on e-SNLI still struggle generating correct explanations even when the label predictions are correct. On the contrary, Table 3 shows that the accuracy scores for T5_{FLUTE} are significantly higher for each type of figurative language, indicating higher quality of explanations achieved by fine-tuning the model on our dataset.

In terms of the Rational Quality (Table 4), using predicted rationales from I → OR we observe that model fine-tuned on e-SNLI generates explanations that are less useful for gold label prediction compared to a model trained on FLUTE. Simulatability scores obtained using predicted rationales are negative compared to gold rationales since rationales generated by the model are still imperfect and may introduce noise that negatively affects the prediction. The positive score of gold rationales indicates that explanations provided in the FLUTE dataset are indeed useful for the model to make predictions.

Table 3 also presents the H_{score} scores for explanations collected from human evaluation (see Appendix B and prior work (Kayser et al., 2021; Majumder et al., 2021) for details on computation). We notice the scores for T5_{FLUTE} are 51.1, 41.3, 24.9, 22.2 points better than for $\text{T5}_{\text{e-SNLI}}$ for Sarcasm, Simile, Metaphor, Idiom (in that order). Crowd workers answered with solid 'Yes' (that the explanation justifies the label) in 43.4% more cases on average when presented explanations from T5_{FLUTE} compared to $\text{T5}_{\text{e-SNLI}}$. Likewise, they answered 'No' in 28.5% less cases on average when presented explanations from T5_{FLUTE} compared to $\text{T5}_{\text{e-SNLI}}$. These results demonstrate a notably higher quality of explanations generated

Premise	The camp life really had energized him.
Hypothesis	Life in the camp wilted him.
Label	Contradiction.
Gold_Exp	To wilt means to lose vitality or strength, while to energize means to give strength or energy to.
T5$_{\text{e-SNLI}}$	If the camp life energized him, then he is not wilted.
T5$_{\text{FLUTE}}$	To wilt something means to lose its energy or vitality, while to have energized something mean to feel a sense of excitement or life.
Premise	No one came to my wedding.
Hypothesis	No big deal that there were no guests at my wedding reception considering it's the only one I'll ever have.
Label	Contradiction.
Gold_Exp	A wedding reception is supposed to be a joyous event with many guests, so not having any guests at it would make it a very uneventful and pointless one.
T5$_{\text{e-SNLI}}$	No one is same as No big deal.
T5$_{\text{FLUTE}}$	A wedding reception is a celebration of the couple's union and is usually attended by family and friends so not having guests there would be seen as snub.

Table 5: Examples of T5$_{\text{e-SNLI}}$ and T5$_{\text{FLUTE}}$ model generated explanations vs. gold explanations for NLI involving metaphor (top) and sarcasm (bottom). More examples in Table 7 in Appendix.

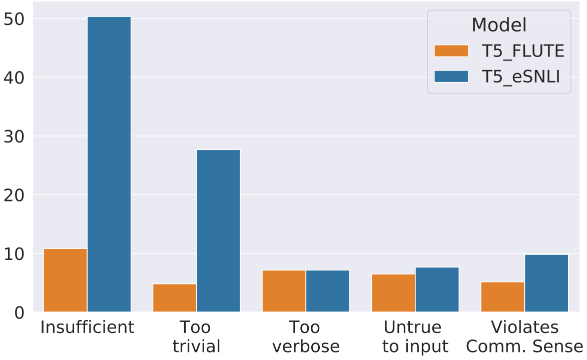


Figure 2: Bar plot of the number of crowd worker-identified shortcomings of explanations generated by T5$_{\text{e-SNLI}}$ and T5$_{\text{FLUTE}}$ by shortcoming and model type as percent of the sample (lower means fewer shortcomings). An extended plot by figurative language type is available in Appendix, Figure 5.

by the model fine-tuned on FLUTE compared to the model fine-tuned on e-SNLI, despite the significantly larger size of the latter dataset.

Figure 2 shows that for cases where crowd workers did not choose an absolute 'Yes', the most common error category to be found is *Insufficient Justification*. The percentage of explanations for which shortcomings were identified is higher or the same for across all categories for T5$_{\text{e-SNLI}}$. The second example ("No one is same …") in Table 5 shows an insufficient explanation generated by T5$_{\text{e-SNLI}}$. T5$_{\text{e-SNLI}}$ generated explanations were also more frequently marked as *Too Trivial*. Often, they do not explain the reasoning but rather follow a standard template *if A then not B* or *if A then B*, such as the first example in Table 5 for T5$_{\text{e-SNLI}}$. We share more such examples of erroneous explanations in the Appendix, Table 8.

5 Related Work

In recent years, evaluating how well RTE models can capture specific linguistic phenomena such as figurative language has attracted many NLP researchers. In earlier work, (Agerri, 2008) analyze metaphor examples in the RTE-1 (Dagan et al., 2006) and RTE-2 (Bar-Haim et al., 2006) datasets, whereas (Poliak et al., 2018a)'s diverse collection of RTE examples contains one type of figurative language - puns. Our research is closer to (Chakrabarty et al., 2021), however, FLUTE has better diversity and also contains explanations for each NLI instance, as well as multiple expert checks to ensure higher quality (see Section 2). A portion of FLUTE's metaphors are based on the Big-bench corpora (Srivastava et al., 2022) as well as the IMPLI dataset (Stowe et al., 2022), which is inspired by (Zhou et al., 2021)'s prior work on the paired idiomatic and literal dataset. However, there are several distinctions between these datasets and FLUTE. First, not all the metaphors in Big-bench have literal paraphrase and contradictions. Second, in case of both the Big-bench and IMPLI dataset, the non-entailment examples are created via minimal edits to the original metaphor, often resulting in neutral examples or contradiction to the non-metaphoric part of the sentence. In FLUTE, we ensure that the non-entailment examples are in contradiction to the metaphor (Table 6).

One of the motives for having NLEs with <literal, figurative> sentence pairs like FLUTE does is to evaluate model's ability to explain their decisions. Recent datasets such as CoS-E (Rajani et al., 2019), Movie Reviews (Zaidan and Eisner, 2008), and e-SNLI (Camburu et al., 2018) have been released in a similar vein. Recent work has also leveraged large language models to explain humor in image captions (Hessel et al., 2022) or sarcasm in dialouges (Kumar et al., 2022). The e-SNLI dataset (i.e., NLE of the entailment relations in the SNLI dataset) has been used in related work (Narang et al., 2020; Yordanov et al., 2021; Majumder et al., 2021; Feng et al., 2022) for explanation generation. In contrast to e-SNLI, which was created via crowdsourcing, we rely on a model-in-the-loop framework for FLUTE influenced by (Wiegreffe et al., 2021a). We have utilized the e-SNLI dataset for explanation generation and observed a T5 model trained on FLUTE performs notably better.

6 Conclusion

We release FLUTE, a dataset for figurative language understanding spanning across Sarcasm, Similes, Metaphors, and Idioms collected via a model-in-the-loop framework. To encourage genuine understanding of figurative language, our data also contains free-form textual explanations. Upon conducting baseline experiments with state-of-the-art benchmark models (i.e., models trained on the the e-SNLI dataset), we notice those models perform poorly. In contrast, performance of the T5 model fine-tuned on FLUTE shows that our dataset can bring us a step closer to developing models that understand figurative language through textual explanations. We hope our research on explanation generation for figurative language will be a fruitful future direction, and our dataset will be a challenging testbed for experimentation.

Limitations

While we focused on four types of figurative language and generated a diverse dataset, we believe it is just a first step towards capturing figurative NLI instances and their explanations, since figurative language is able to draw on a wide variety of cultural knowledge and contexts. Although the sarcasm portion captures the most common type of incongruity between sarcastic context and sentiment, sarcasm can manifest in many different forms - situational, underplayed, or dramatic, for which examples and explanations will differ. Finally this study doesn't explicitly focus on faithfulness of model generated Natural Language Explanations, however we hope to evaluate the faithfulness of these using methods described in contemporaneous literature on faithfulness of NLE's (Sia et al., 2022; Chan et al., 2022). While GPT-3 did not generate any examples of societal bias in this study, prior research has investigated the reliability or faithfulness of generations (Wiegreffe et al., 2021a). Likewise, we could also conduct a human study asking specific questions (e.g., whether the explanations mimic any bias, are credible, etc.); we leave this for future study.

Ethics Statement

We use a model-in-the-loop framework for content generation. Although we use language models trained on data collected from the Web, which have been shown to have issues with gender bias and abusive language, we have verified carefully that our FLUTE data does not contain any toxic text and it underwent manual inspection by the authors and experts. We pay Amazon Mechanical Turkers at the rate of 15 $/hr which is compliant with the minimum hourly wage in United States.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments. We are also grateful to Esin Durmus and Faisal Ladakh for the initial brainstorming on model in the loop strategies for dataset creation. Tuhin is funded by Columbia Center of Artificial Intelligence & Technology (CAIT) and the Amazon Science Ph.D. Fellowship.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Rodrigo Agerri. 2008. Metaphor in textual entailment. In *Coling 2008: Companion volume: Posters*, pages 3–6, Manchester, UK. Coling 2008 Organizing Committee.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, and Bernardo Magnini. 2006. The second pascal recognising textual entailment challenge. In *Computer Science*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Conrad Borchers, Dalia Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Kirk. 2022. Looking for a handsome carpenter!

debiasing GPT-3 job advertisements. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 212–224, Seattle, Washington. 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, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not Rocket Science: Interpreting Figurative Language in Narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Aaron Chan, Shaoliang Nie, Liang Tan, Xiaochang Peng, Hamed Firooz, Maziar Sanjabi, and Xiang Ren. 2022. Frame: Evaluating simulatability metrics for free-text rationales. *arXiv preprint arXiv:2207.00779*.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4351–4360, Online. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Yufei Feng, Xiaoyu Yang, Xiaodan Zhu, and Michael Greenspan. 2022. Neuro-symbolic natural logic with introspective revision for natural language inference. *Transactions of the Association for Computational Linguistics*, 10:240–256.

Susan R Fussell and Mallie M Moss. 1998. Figurative language in emotional communication. *Social and cognitive approaches to interpersonal communication*, pages 113–141.

Debanjan Ghosh, Elena Musi, and Smaranda Muresan. 2020. Interpreting verbal irony: Linguistic strategies and the connection to theType of semantic incongruity. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 82–93, New York, New York. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Jack Hessel, Ana Marasović, Jena D Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2022. Do androids laugh at electric sheep? humor" understanding" benchmarks from the new yorker caption contest. *arXiv preprint arXiv:2209.06293*.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. E-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1244–1254.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. *University of Pennsylvania ScholarlyCommons*.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5956–5968, Dublin, Ireland. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022a. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955*.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022b. Testing the ability of language models to interpret figurative language.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2021. Rationale-inspired natural language explanations with commonsense. *arXiv preprint arXiv:2106.13876*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018a. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 337–340, Brussels, Belgium. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018b. Hypothesis only baselines in natural language inference. *arXiv preprint arXiv:1805.01042*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

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 of Machine Learning Research*, 21:1–67.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Richard M Roberts and Roger J Kreuz. 1994. Why do people use figurative language? *Psychological science*, 5(3):159–163.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. 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 4275–4293, Online. Association for Computational Linguistics.

Ekaterina V Shutova. 2011. Computational approaches to figurative language. Technical report, University of Cambridge, Computer Laboratory.

Suzanna Sia, Anton Belyy, Amjad Almahairi, Madian Khabsa, Luke Zettlemoyer, and Lambert Mathias. 2022. Logical satisfiability of counterfactuals for faithful explanations in nli. *arXiv preprint arXiv:2205.12469*.

Aarohi Srivastava, Abhinav Rastogi, and Abhishek Rao. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. In *In preparation*.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *arXiv*.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2021a. Reframing human-ai collaboration for generating free-text explanations. *arXiv preprint arXiv:2112.08674*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021b. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2021. Few-shot out-of-domain transfer learning of natural language explanations. *arXiv preprint arXiv:2112.06204*.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, 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.

A Appendix

In this section we report the details of the experiments (e.g., hyperparameters used for GPT-3 based generations, examples, as well as the prompts for the figurative language.

A.1 Sarcasm dataset

A.1.1 Hyperparameters for the sarcasm dataset

We use GPT-3-Davinci-001 model for auxillary paraphrase generations from which crowd workers create sarcasm. To generate paraphrase, we use the following hyperparameters: `temperature=1, max tokens=100, top p=0.9, best of=1, frequency penalty=0.5, presence penalty=0.5`.

To generate explanations, we use the following hyperparameters: `temperature=1.0, max tokens=100, top p=0.9, frequency penalty=0.5, presence penalty=0.5, stop=["."]`.

A.1.2 Prompts for generation of Paraphrase from which Sarcasm is created

You will be presented with examples of some literal input sentences and their creative paraphrases. For each example, we also provide the associated emotion. Your task is then to generate a creative paraphrase for a given literal sentence where the creative paraphrase should reflect the associated emotion without changing its meaning. Make sure to use some sort of humor and commonsense about everyday events and concepts

> 1) **Literal**: A lot of people have got engaged recently.
>
> **Emotion**: surprised
>
> **Creative Paraphrase**: The way all the couples are pairing off lately and naming the big day, I think Cupid's really busy.
>
> 2) **Literal**: We have enough candles mom
>
> **Emotion** annoyed
>
> **Creative Paraphrase**: I think the Catholic church is going to have to canonize a whole new generation of saints to justify our candle use mom
>
> . . .

,

A.1.3 Prompts for generation of Explanation for Paraphrase from which Sarcasm is created (Entailment)

You will be presented with examples of two sentences typically a premise along with an entailing paraphrase of the premise called the hypothesis. Your task is to generate natural language explanations to justify the Entailment between the premise and the hypothesis.

> 1) **Premise**: Awful seeing a naked man run through my neighborhood.
>
> **Hypothesis**: The sight of a man running through my neighborhood sans clothes was pretty disgusting.
>
> **Explanation**: It is socially unacceptable to not wear clothes and step out of one's house so seeing a man who is running naked in the neighborhood is pretty shameful and disgusting.
>
> 2) **Premise**: My mother didn't cook her chicken all the way through at dinner the other night.
>
> **Hypothesis**: The fact that my mother didn't cook her chicken all the way through at dinner makes me feel like I'm going to vomit.
>
> **Explanation**: Eating undercooked chicken can cause food poisoning and so finding out that the chicken at dinner wasn't cooked all the way through often makes people throw up
>
> . . .

A.1.4 Prompts for generation of Explanation for Sarcasm (Contradiction)

You will be presented with examples of some literal and sarcastic sentences. Your task is then to write explanations to justify why it is sarcastic w.r.t the literal

> 1) **Literal**: When I moved into my apartment it was full of bugs
>
> **Sarcasm**: I absolutely loved when I moved into my apartment and found it crawling with bugs.
>
> **Explanation**: Bugs are usually disgusting and most people are terrified of them therefore it is unlikely to love seeing someone's

apartment infested by them.

2) **Literal**: I've been hearing some strange noises around the house at night.

Sarcasm: I am completely comforted by the weird noises I keep hearing around the house at night.

Explanation: Hearing weird noises around the house at night could invoke a potential danger such as a robbery or someone breaking in with malicious intent which makes someone scared rather than comforted. . . .

A.2 Simile dataset

A.2.1 Hyperparameters for the simile dataset

We use GPT-3-Davinci-002 model for simile data generations. To generate paraphrase, we use the following hyperparameters: `temperature=1`, `max tokens=256`, `top p=0.5`, `best of=1`, `frequency penalty=0.5`, `presence penalty=0.1`.

To generate contradictions, we use the following hyperparameters: `temperature=0`, `max tokens=100`, `top p=1`, `frequency penalty=0`, `presence penalty=0`.

To generate explanations, we use the following hyperparameters: `temperature=0.7`, `max tokens=86`, `top p=1`, `frequency penalty=0`, `presence penalty=0`, `stop=["."]`.

A.2.2 Challenging instances for the simile dataset

To select challenging instances from FigQA dataset Liu et al. (2022b), we use RoBERTa fine-tuned on SNLI, MNLI, FEVER-NLI, and ANLI (R1, R2, R3). We choose <simile, literal, contradiction> instances by taking the average between RoBERTa logit for entailment given <simile, literal> as input and the logit for contradiction given the <simile, contradiction> input. We then choose instances with the lowest such score. In this way, we are able to select instances for which RoBERTa is essentially confusing entailment and contradiction. From our observation, these are usually ironic similes, e.g. for the simile *I'm sharp as a pillow* and the literal sentence *I'm not sharp*, RoBERTa would have a low logit for entailment, while for the sentence *I'm sharp* it would have a low logit for contradiction.

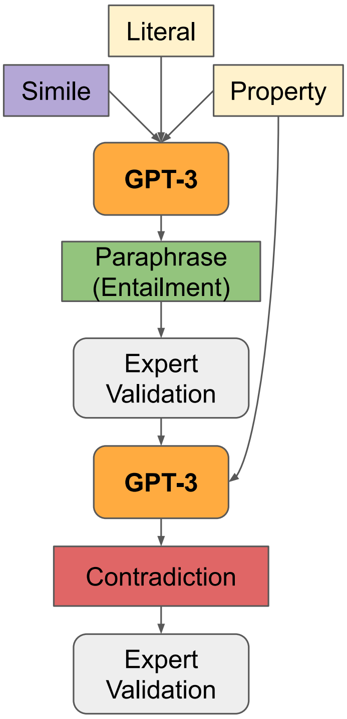


Figure 3: Model-in-the-loop to generate a Simile NLI dataset.

A.2.3 Prompts for Simile Paraphrase Generation

You will be presented with examples of some literal input sentences and their creative paraphrases. You will also be presented words that need to be preserved. Your task is to generate a creative paraphrase for a given literal sentence consistent in meaning. DO NOT CHANGE words after "Preserve:" keyword.

1. **Sentence:** overwhelmingly , it began to draw him in.
Preserve: overwhelmingly
Creative Paraphrase: He was overwhelmingly obsessed with it.
. . .

A.2.4 Prompts for Simile Contradiction Generation

You will be presented with a Sentence and a Property. Your goal is to invert the meaning of the Sentence with respect to the Property via a minimal edit.

1. **Sentence:** The place looked impenetrable and inescapable.
Property: impenetrable and inescapable
Inversion: This place looked easy to walk

into and exit from.

. . .

A.2.5 Prompts for Simile Contradiction Explanation Generation

You will be provided with a Simile and a contradictory sentence after the word "Contradiction". Your task is to explain why the contradictory sentence contradicts the Simile.

1. **Simile:** like a psychic whirlpool , it began to draw him in.
Contradiction: Mildly, it began to draw him in
Explanation: A whirlpool is a strong current, so a psychic whirlpool drawing in indicates that it was drawing him in intensely, rather than mildly.
. . .

A.3 Prompts for Simile Entailment Explanation Generation

You will be presented with a sentence containing a simile (Simile Sentence) and an entailing sentence (Entail Sentence). Please provide an explanation for why Simile Sentence is implied by the Entail Sentence.

1) **Simile Sentence:** The place looked like a fortress
Entail Sentence: The place looked impenetrable and inescapable
Explanation: A fortress is a military stronghold, hence it would be very hard to walk into, or in other words impenetrable.
. . .

A.4 Metaphor dataset

A.4.1 Hyperparameters for the metaphor dataset

We use GPT-3-Davinci-002 model for metaphor data generations. To generate paraphrase, we use the following hyperparameters: `temperature=1, max tokens=100, top p=0.8, best of=1, frequency penalty=0.5, presence penalty=0.1`.

To generate contradictions, we use the following hyperparameters: `temperature=0, max tokens=100, top p=0.8, frequency penalty=0.5, presence penalty=0.1`.

To generate explanations, we use the following hyperparameters: `temperature=0.8, max tokens=100, top p=0.9, frequency penalty=0, presence penalty=0, stop=["."]`.

A.4.2 Prompts for Metaphor Paraphrase Generation

You will be presented with examples of some metaphor input sentences and their creative paraphrases. Your task is to generate a creative paraphrase for a given literal sentence consistent in meaning.

1. **Sentence:** A golden sun shines high in the sky.
Creative Paraphrase: a very bright sun shines high in the sky.
. . .

A.4.3 Prompts for Metaphor Contradiction Generation

You will be presented with examples of some literal input sentences and their contradictions. Your task is then to generate contradiction of the new sentences via a minimal edit.

1. **Sentence:** The company released him after many years of service.
Contradiction: The company hired him after many years of service.
. . .

A.4.4 Prompts for Metaphor Entailment Explanation Generation

You will be presented with examples of sentences containing a metaphor along with an entailing paraphrase of the sentence. Your task is to generate natural language explanations to justify the Entailment with the input.

1. **Metaphor:** Krishna is an early bird.
Contradiction: Krishna wakes up early everyday.
Explanation: Early bird means the person who wakes up early in the morning.
. . .

A.4.5 Prompts for Metaphor Contradiction Explanation Generation

You will be provided with a sentence containing a Metaphor and a contradictory sentence after the word “Contradiction". Your task is to explain why the contradictory sentence contradicts the Metaphor.

1. **Metaphor:** Joseph has the heart of a lion.
Contradiction: Joseph has the calm demeanor of a lamb.
Explanation: A lamb is typically seen as a gentle and timid creature while a lion is seen as a brave and fierce creature.
. . .

Metaphor	A weather vane *crowns* the building.
Original Entailment	The king crowned the prince.
GPT-3 Contradiction	A weather vane mars the building.
Metaphor	A *golden* sun shines high in the sky.
Original Entailment	A very expensive sun shines high in the sky A sunset is setting in the west.
GPT-3 Contradiction	A sunset is setting in the west.
Metaphor	Fear had changed him to a *shaken jelly*.
Original Entailment	He was afraid of shaking jelly.
GPT-3 Contradiction	He had conquered his fear and now he is a strong and capable person.

Table 6: Example of Metaphors and their contradictions (from prior work and generated for this paper.). Note, examples from prior work replace the metaphor sentence with adding words that fits into their context whereas in this work we generate examples that *contradicts* the metaphor.

A.5 Idiom dataset

A.5.1 Hyperparameters for the idiom dataset

We use GPT-3-Davinci-002 model for idiom data generations. To jointly generate paraphrase and contradiction, we use the following hyperparameters: `temperature=1, max tokens=200, top p=0.9, best of=1, frequency penalty=0.5, presence penalty=0.5, stop=[".."]`.

To jointly generate explanations for entailment and contradiction, we use the following hyperparameters: `temperature=0.7, max tokens=256, top p=0.9, frequency penalty=0, presence penalty=0, stop=[".."]`.

A.6 Prompts for joint paraphrase and contradiction generation for idioms

You will be presented with examples of some input sentences containing an idiom. You will be provided with the meaning of the idiom. Your task is to first generate a paraphrase that complies with the meaning of the idiom and then generate a negation of the paraphrase that contradicts the meaning of the idiom. Please look at the span within bold tags when performing paraphrase and negation.

1) **Sentence**: He looked great, and he was smiling <b>to beat the band</b>.

Idiom: to beat the band

Meaning: To a huge or the greatest possible extent or degree.

Paraphrase: He looked awesome and was smiling <b>hilariously in an uncontrollable manner</b>

Negation: He looked awesome and was smiling <b> in a very coy and restrained manner</b>.
. . .

A.6.1 Prompts for joint generation of Explanation for idioms

You will be presented with examples of sentences containing an idiom along with an entailing and contradictory paraphrase of the sentence. Your task to generate natural language explanations to justify the Entailment or Contradiction with the input.

1) **Sentence**: Not to share the bank with the table, or to take some minor part of it, but to go the whole hog.

Idiom: go the whole hog

Meaning: To do something as thoroughly as possible or without restraint.

Entailment: Not to share the bank with the table, or to take some minor part of it, but to take it all for themselves without any restraint.

Contradiction: Not to share the bank with the table, or to take some minor part of it, but to show some restraint and not go overboard.

Entail_Explanation: Usually to go the whole hog refers to do something as thoroughly as possible , taking it all for oneself or without any restraint.

Contra_Explanation: Usually to go the whole hog refers to do something as thoroughly as possible, without any sort of restraint and is often characterized by being extreme or overboard..

2) **Sentence**: I told her almost everything, including my reticence about seeing this work of literature go through the mill which was vanity publishing. Idiom: go through the mill

Idiom: go through the mill

Meaning: To be abused or treated very harshly; to suffer intense anguish, stress, or grief.

Entailment: I told her almost everything, including my reticence about seeing this work of literature be abused and treated in an extremely harsh manner which was vanity publishing.

Contradiction: I told her almost everything, including my reticence about seeing this work of literature be celebrated and treated in an extremely proper manner which was vanity publishing.

Entail_Explanation: To go through the mill in the context here refers that vanity publishing will abuse and treat the work of literature being very poorly.

Contra_Explanation: Usually when we say go through the mill it does not mean something being celebrated and treated well but instead being abused and treated poorly which is being said here in the context of vanity publishing doing to the work of literature..

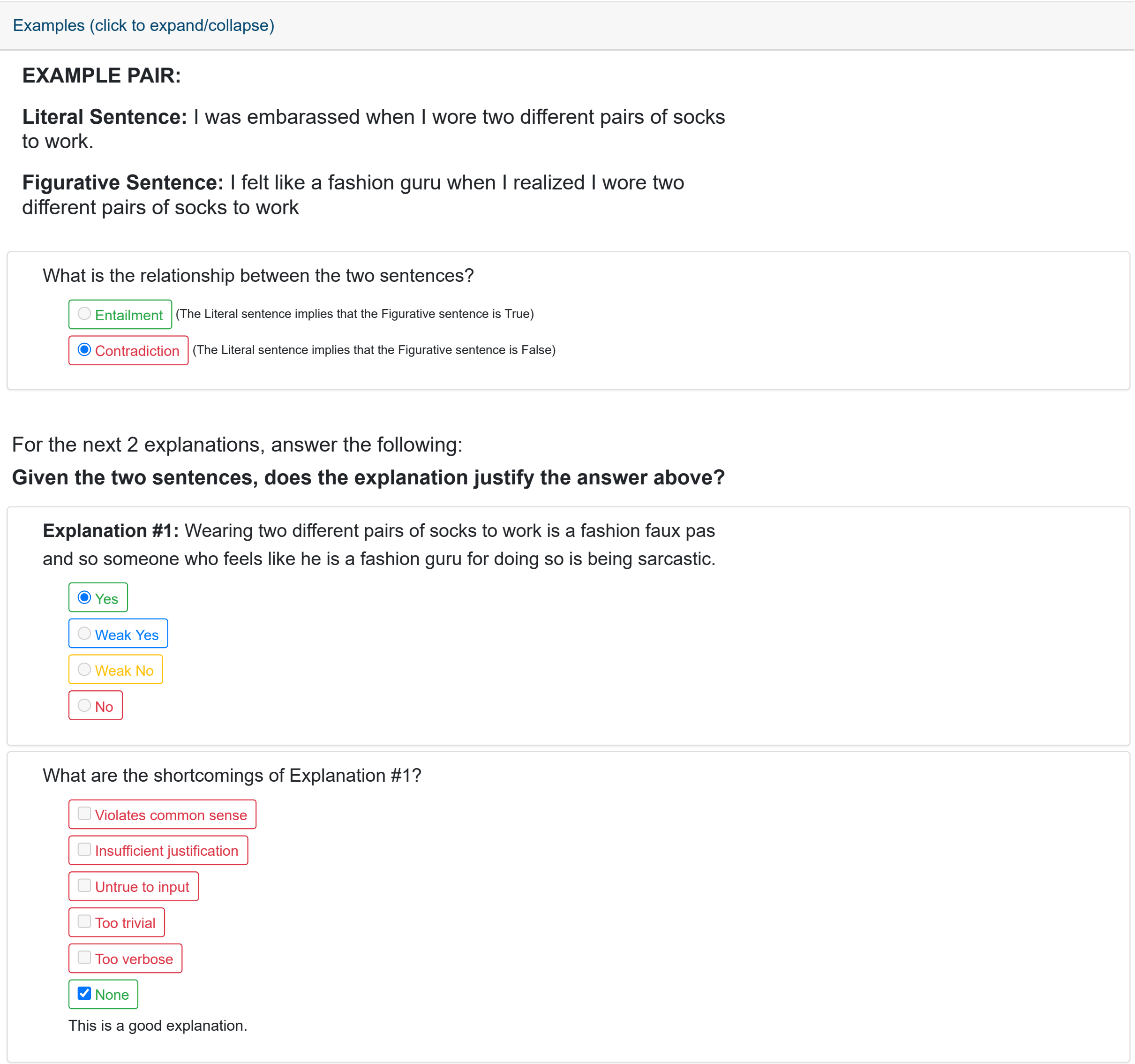
. . .

B Details of Human evaluation

We follow Kayser et al. (2021) in all the below human evaluation procedures. Refer to Figure 4 for the example of the interface for crowdworkers. We first ask the crowdworkers to identify the relationship between the literal and figurative sentence (whether it is a contradiction or entailment). Using in-browser checks, we ensure that the crowdworkers understood the NLI pair by only accepting the submission if the relationship was identified correctly.

Then, we provide 2 explanations: one generated by T5_{FLUTE} and one generated by $\text{T5}_{\text{e-SNLI}}$. The crowdworkers do not know which one is which. For each NLE, we ask: *Given the two sentences, does the explanation justify the answer above?*, and provide four options: Yes, Weak Yes, Weak No, and No. We also ask to provide the shortcomings of the explanations if the worker selected a score lower than Yes. The workers have the following options to choose from, following prior work by (Majumder et al., 2021; Kayser et al., 2021): *Violates Common Sense*, *Insufficient Justification*, *Untrue to Input*, *Too Trivial*, *To Verbose*. Some examples of these shortcoming are provided in Table 8. In addition to the Figure 2, we provide the bar plot of the number of shortcomings as percentage of the sample by figurative language type in Figure 5.

We map the answers to $1, \frac{2}{3}, \frac{1}{3}, 0$ respectively. Then, we compute the average score across 3 workers per entry, and the sample average per figurative language type for the corresponding model.



Examples (click to expand/collapse)

EXAMPLE PAIR:

Literal Sentence: I was embarassed when I wore two different pairs of socks to work.

Figurative Sentence: I felt like a fashion guru when I realized I wore two different pairs of socks to work

What is the relationship between the two sentences?

○ Entailment (The Literal sentence implies that the Figurative sentence is True)

● Contradiction (The Literal sentence implies that the Figurative sentence is False)

For the next 2 explanations, answer the following:

Given the two sentences, does the explanation justify the answer above?

Explanation #1: Wearing two different pairs of socks to work is a fashion faux pas and so someone who feels like he is a fashion guru for doing so is being sarcastic.

● Yes

○ Weak Yes

○ Weak No

○ No

What are the shortcomings of Explanation #1?

□ Violates common sense

□ Insufficient justification

□ Untrue to input

□ Too trivial

□ Too verbose

☑ None

This is a good explanation.

Figure 4: Amazon Mechanical Turk interface, borrowed from (Kayser et al., 2021), asking to first select the relationship between the sentences, and then evaluate the explanation.

Premise	My father continuously gambles all his earnings away even though he has two minors to support.
Hypothesis	My father is the best dad in the world who gambles all his earnings away and leaves me and my sister struggling.
Gold	A great parent is someone who provides for their children and protects them but gambling away all the money earned leaves the children in a difficult position and hence the father cannot be considered the best dad.
FLUTE	Gambling away one's earnings is not a good thing and it can lead to the father not being able to support his children who are dependent on him.
e-SNLI	Gambling all his earnings away is not the best dad in the world.

Premise	He ran slowly.
Hypothesis	He ran like an olympic sprinter.
Gold	An Olympic sprinter is someone who runs extremely fast, so saying someone ran like one would imply they ran quickly, not slowly
FLUTE	Olympic sprinters are known to run very fast, so saying someone ran like an olympic sprinter would mean they ran very quickly
e-SNLI	Slowly and sprinter are not the same

Premise	He's totally trustworthy as an executive secretary, but he's relatively new and inexperienced and does not know what to do on occasions like this
Hypothesis	He's totally trustworthy as an executive secretary, and has been around long enough to know the ropes on occasions like this.
Gold	To know the ropes means to understand or be familiar with the details, but the person is relatively new and inexperienced
FLUTE	To know the ropes means to have experience in the job, but in this context the person is relatively new and inexperienced
e-SNLI	He's not been around long enough to know the ropes and he is relatively new and inexperienced.

Table 7: Example of Sarcasm, Simile and Idiom NLI with respective gold and generated explanations from T5 by finetuning on FLUTE and e-SNLI

Shortcoming	Premise	Hypothesis	Label	Explanation	
				T5$_{\text{e-SNLI}}$	T5$_{\text{FLUTE}}$
Violates Common Sense	The gesture was awkward.	The gesture was like a ballet move	C	A gesture is not a ballet move.	A ballet move is graceful and fluid, while an awkward gesture is not graceful or fluid.
Insufficient Justification	I stayed in a house with a rat problem for a while, it was not fun.	I simply cannot wait to stay in a house again that had a rat problem because it was not only a lot of fun but also super enjoyable	C	It was either not fun or not fun.	Rats are often associated with being filthy and unhygienic and so staying in a house with rat problems would not be fun at all.
Untrue to Input	Even though he was strong in his magic, it felt weak compared to the inferno I'd faced when I killed Mab.	He was strong in his magic but, compared to the blazing inferno that i'd faced when i'd killed mab, his power felt like a candle flame.	E	The candle flame is the inferno.	A candle flame is a very small flame, which is weak compared to blazing inferno.
Too Trivial	He ran fast.	He ran like an olympic sprinter.	E	He ran fast like an olympic sprinter.	An olympic sprinter is a very fast person, so if he ran like an sprinter, it would mean that he ran fast.
Too Verbose	She asked the witnesses for information.	She petitioned the witnesses for information.	E	She asked the witnesses for information and petitioned the witness.	Petitioning someone means asking them for information, which is what she did in this sentence.

Table 8: Examples of shortcomings of T5$_{\text{e-SNLI}}$ explanations. For this table, T5$_{\text{e-SNLI}}$ explanations were sorted by most crowd workers votes for a respective shortcoming.

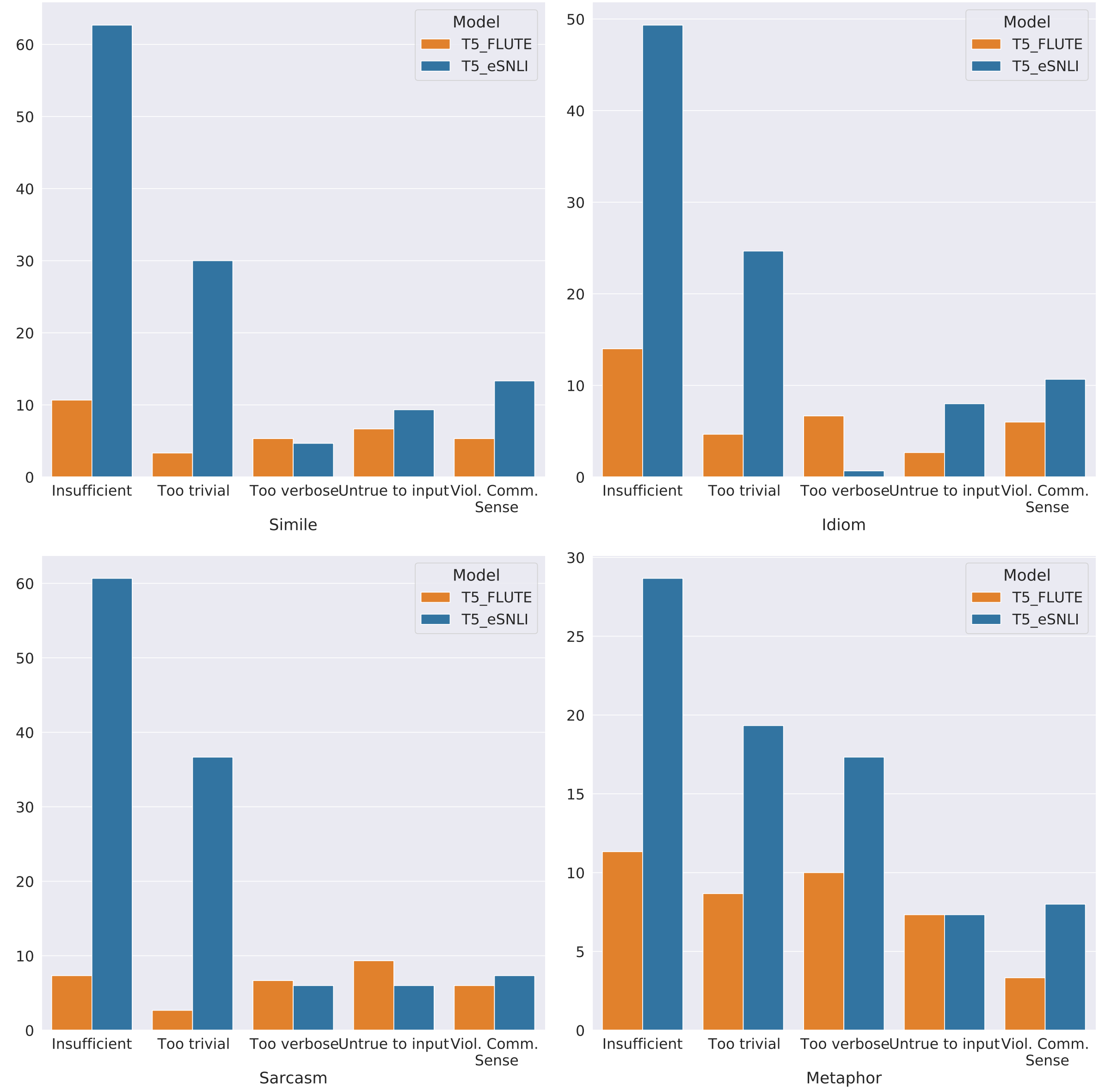


Figure 5: Bar plot of the number of crowd worker-identified shortcomings of explanations generated by T5$_{\text{e-SNLI}}$ and T5$_{\text{FLUTE}}$ by type of shortcoming, figurative language type, and by type of model as percent of the sample (lower means fewer shortcomings).