Human-Model Divergence in the Handling of Vagueness

Elias Stengel-Eskin **Jimena Guallar-Blasco** **Benjamin Van Durme**
Johns Hopkins University
{elias, jgualla1, vandurme}@jhu.edu

Abstract

While aggregate performance metrics can generate valuable insights at a large scale, their dominance means more complex and nuanced language phenomena, such as vagueness, may be overlooked. Focusing on vague terms (e.g. *sunny*, *cloudy*, *young*, etc.) we inspect the behavior of visually grounded and text-only models, finding systematic divergences from human judgments even when a model's overall performance is high. To help explain this disparity, we identify two assumptions made by the datasets and models examined and, guided by the philosophy of vagueness, isolate cases where they do not hold.

1 Introduction

Part of the power of language as a medium for communication is rooted in having a reliable mapping between language and the world: we typically expect language to be used in a consistent fashion, i.e. the word "dog" refers to a relatively invariant group of animals, and not to a different set of items each time we use it. This view of language dovetails with the supervised learning paradigm, where we assume that an approximation of such a mapping can be learned from labeled examples—often collected via manual annotation by crowdworkers. In natural language processing (NLP), this learning typically takes place by treating tasks as classification problems which optimize for log-likelihood. While this paradigm has been extensively and successfully applied in NLP, it is not without both practical and theoretical shortcomings. Guided by notions from the philosophy of language, we propose that borderline cases of vague terms, where the mapping between inputs and outputs is unclear, represent an edge case for the assumptions made by the supervised paradigm, and result in systematic divergences between human and model behavior.

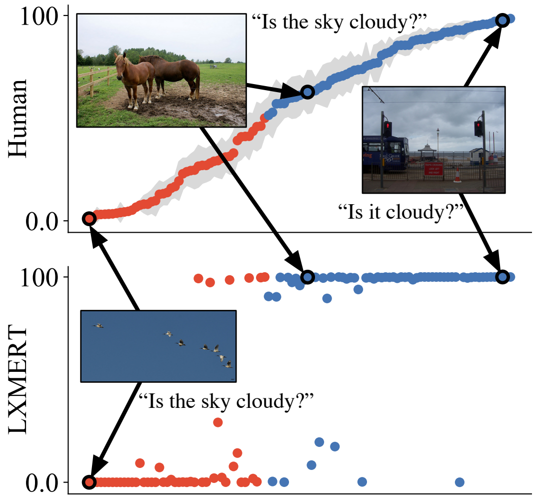


Figure 1: Given a binary question involving a vague term (in this case, *cloudy*) humans hedge between "yes" and "no," following a sigmoid curve with borderline examples falling in the middle. Standard error (grey band) shows that annotator agree even in borderline regions. In contrast, model predictions remain at extreme ends.

To demonstrate this, we begin by identifying a set of canonically vague terms in the binary question subset of the Visual Question Answering (VQA) and GQA datasets (Antol et al., 2015; Goyal et al., 2017; Hudson and Manning, 2019) and isolating a subset of images, questions, and answers from these datasets centered around these terms. Using this subset, we show that while the accuracy of LXMERT (Tan and Bansal, 2019) on non-borderline cases is very high, its performance drops—sometimes dramatically—on borderline cases. We then compare the behavior of the model against that of human annotators, finding that while humans display behavior which aligns with theories of meaning for vague terms, model behavior is less predictable.

We extend our analysis of visually-grounded terms to a text-only case, re-framing the catego-

Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language (UnImplicit 2021), pages 43–57
Bangkok, Thailand (online), August 5, 2021. ©2021 Association for Computational Linguistics

rization of statements into true statements and false ones as a task involving vagueness. Controlling for world knowledge, we find that while probes over contextualized encoders can classify statements significantly better than random, their output distributions are strikingly similar to those observed in the visually-grounded case. When contrasted with scalar annotations collected from crowdworkers, these results support the notion that analytic truth itself admits of borderline cases and poses problems for supervised systems.

In § 2, we provide a more thorough definition of terms used, the motivation for exploring vagueness, and the underlying assumptions of supervised learning that are violated by vague terms.

2 Motivation and Background

Vague terms, broadly speaking, are ones that admit of borderline cases; for example: *cloudy* is vague because, while there are clearly cloudy and not cloudy days, there are also cases where the best response to the question "is it cloudy?" might be "somewhat" rather than a definitive "yes" or "no." Given this definition, we can see that a large portion of the predicates we use in every-day speech are vague. This even encompasses predicates such as *is true* and *is false*, as we might have statements that are true or false to varying degrees.

Vague predicates in particular have been a focus of the philosophy of language, as they represent an interesting edge case for theories of meaning. Take, for example, a canonical example of a vague predicate from philosophy: *is a heap*. There are things that are undeniable heaps, and others that are clearly not. In the extreme case, we can imagine starting with a heap of sand (say, N grains) and removing a single grain of sand from it. Clearly, the resulting mass would still be a heap. This is, however, a dangerous precedent; we can now remove $N - 2$ grains on sand until we have a single grain remaining, whose heap-ness is hard to justify, but which, by induction, is still a heap. This raises important questions: how is it that speakers avoid this paradox and are able to use and understand vague terms, even in borderline cases? Is there a definitive point at which a heap becomes a non-heap? The answers to these questions should influence how we annotate the data from which we aim to learn meaning representations of vague terms.

While the unequivocal instances of heaps fit well into the current paradigm of supervised learning with categorical labels, borderline heaps do present a problem. Recall that the first assumption by supervised learning which we have pointed out is that the ideal mapping between the input (in this case, questions and images) and the the label set (answers) is largely fixed. For example, given the question "Is this a dog?" we assume that the set of things in the world which we call "dog", also known as the *extension* of "dog", remains constant. In that case, the annotator's response to the question corresponds to whether what the image depicts could be plausibly considered as part of the extension of "dog." While we might easily be able to determine the set membership of poodles and terriers, we may have a harder time with Jack London's White Fang: half wolf, half dog. Thus it is clear that the borderline cases of vague terms demand a more nuanced account than merely a forced choice between two extremes. The range of such accounts fall broadly into three classes:

Contextualist theories (Kamp, 1981; Raffman, 1994; Graff, 2000; Shapiro, 2006, i.a.) broadly hold that the interpretation of vague predicates depend on contextual and pragmatic information such as on the speaker's previous commitments, their perceived goals, and the psychological state of the interpreter. This view could in most cases be reconciled with the supervised learning paradigm, provided that the data upon which the interpretation of the vague predicate hinges (i.e. speaker commitments, etc.) is available as input. Past work in modeling the meaning of vague terms has often focused on these accounts (c.f. § 6).

Epistemic accounts (Sorensen, 2001; Williamson, 1994, i.a.) bite the proverbial bullet, allowing for a hard boundary between heaps and non-heaps to exist, but claiming that its location is unknowable. This is in contrast to the supervised paradigm, where the boundary is treated as known.

Logic-based approaches tackle the paradox induced by vagueness, either by claiming that borderline examples do not admit of truth values (supervaluationism), or by adapting logic to permit more granular classifications (many-valued logic; Sorensen, 2018). The latter approach can sometimes accommodate the supervised paradigm.[1]

[1] It may still be incompatible with log-likelihood. Treating *ordinal* many-valued logic as a k-way classification problem requires that all values be equidistant, i.e. predicting a value of 1/5 when the true value is 4/5 is as bad as rating it 3/5.

Ambiguity and Under-specification It is important to distinguish vagueness from under-specification (imprecision in the input making the output difficult to recover) and ambiguity (the presence of multiple valid answers), both alternative explanations for annotator disagreement. Indeed, Bhattacharya et al. (2019) include both in their taxonomy of VQA images-question pairs with high annotator disagreement. While they are major challenges in any language-based task, both are often defeasible in nature: we can provide additional information that would reveal the "correct" answer to an annotator, i.e. we could provide a better, sharper version of the image, or more contextual information. Vagueness is non-defeasible: even if one were to know the exact number of grains of sand, the predicate "*is a heap*" would remain vague.

3 Visually Grounded Vagueness

The interpretation of vague terms as described in § 1 typically occurs in a grounded setting; the question "Is this a dog?" is only meaningful in the context of some state of affairs (or depiction thereof). We focus on binary questions about images, taking examples from VQA and GQA; this ensures that the vague term is the question's focus, excluding open-ended queries like "What is the old man doing?" which only implicitly involve vagueness.

Data collection We begin by isolating a number of vague descriptors (*sunny*, *cloudy*, *adult*, *young*, *new*, *old*) in the VQA and GQA datasets. We then use high-recall regular expressions to match questions from these descriptors in the development sets of both datasets, manually filtering the results to obtain high-precision examples. Here, we make the simplifying assumption that a group of predicates involving these terms, such as "is *x*", "seems *x*" and "looks *x*" are approximately equivalent and used interchangeably.

This process results in a variable number of questions per descriptor, with *sunny* and *cloudy* typically having far more representation. Given the size of the whole development sets, and the fact that the data presented is being used merely for analysis rather than for training models, we annotate between 32 and 264 examples, depending on the data availability for each predicate.[2]

While the VQA development data contains 10 annotations per example, GQA does not, and thus, in order to verify the quality of the VQA annotations and to collect annotations for GQA, we solicited 10-way redundant annotations from Mechanical Turk, presenting annotators with a question and its corresponding image from the vision-and-language dataset (e.g. "Is it sunny?").[3] Rather than providing categorical labels (e.g. "yes", "no") workers were asked to use a slider bar ranging from "no" to "yes", whose values range from 0 to 100, using an interface inspired by Sakaguchi and Van Durme (2018). Examples were provided in groups of 8.[4] The resulting annotations are normalized per annotator by the following formula $x' = (x - x_{\text{min}})/x_{\text{max}}$ where x_{min} and x_{max} are the annotators minimum and maximum scores. This accounts for differences in slider bar usage by different annotators. Inter-annotator agreement is measured via majority voting, where an annotator is said to agree with others when their judgement falls on the same side of the slider bar scale (i.e. > 50, < 50). Using this metric, we exclude annotators with $< 75\%$ agreement. After exclusion, all predicates had $> 90\%$ average agreement.[5].

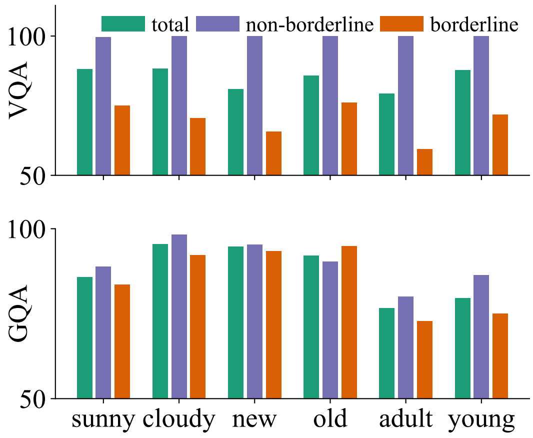


Figure 2: Accuracy of LXMERT on VQA and GQA Yes/No questions per predicate is highest for non-borderline examples, but drops in "borderline" regions.

Vagueness and accuracy We begin by demonstrating that vagueness is not merely a theoretical problem: Fig. 2 shows that while the total accuracy of LXMERT (Tan and Bansal, 2019) is fairly high, it drops on all descriptors (except for "old" for GQA) when looking only at accuracy in the borderline regions. For VQA, we take advantage of the

[2] Note that for some predicates (e.g. *sunny* and *cloudy*, more data was available.

[3] Since we were merely verifying the data quality for VQA, we only ran two descriptors: "sunny" and "cloudy".

[4] c.f. Appendix A for more on the collection protocol.

[5] All data is available at `website.com`

existing 10-way redundant annotations, defining borderline examples as those for which there was any disagreement between annotators, i.e. even if 9 annotators responded "yes" and one responded "no" for a given example, it is considered borderline. This results in 49.24% borderline cases. We find that for GQA, defining borderline examples as having mean normalized scores $\in [15.0, 85.0]$ yields roughly the same percentage (47.20% borderline).

The contrast between borderline and non-borderline regions is especially dramatic for VQA, with the minimum non-borderline accuracy being 99.67% for "sunny," while the accuracy in the borderline region drops to 69.78%. Though the results are less dramatic for GQA, they generally trend in the same direction. We argue that, given that these borderline examples account for roughly half of the data examined, the relatively high aggregate performance obtained by models on binary questions in VQA and GQA may be partially attributed to an absence of vague terms rather than to the strength of the model. Conversely, given a shifted evaluation dataset with more vague terms, the performance would likely drop dramatically.

Vagueness in detail Having demonstrated that model performance is diminished on borderline cases, we seek to further explore the divergence in model and human behavior.

Fig. 1 plots the mean human scores in the top plot, with examples ordered by their mean human rating. The bottom plot shows LXMERT output scores for the same examples. The human scores display a sigmoid shape, while the model scores are saturated at either 0 or 1. For the sake of space, the remaining plots are reported in Appendix B, and we constrain ourselves to a quantitative analysis to demonstrate that a similar trend holds across the remaining descriptors.

Following Item Response Theory (Reise et al., 2005; Lalor et al., 2016) – a modeling paradigm for psychological tests premised on variability among respondents – we posit a 2-parameter sigmoid response function given by $\big(1 + \exp\big(-k * (x - x_0)\big)\big)^{-1}$ where k and x_0 are scale and shift parameters, respectively. This parameterization reflects the intuition that non-borderline examples are found near the spectrum's ends (0 and 100) while borderline examples form a curve in the spectrum's center. In other words, it defines an "ideal" curve in the sigmoid family that fits the data collected from annotators. In some cases, this curve is stretched, nearing a line, while in others it is more pronounced.

We fit three separate logistic regressions: one to the mean of the annotator responses, one to the model response obtained from LXMERT, and a baseline fit against data drawn from a uniform distribution. The quality of the fit, measured by root mean squared error (RMSE) on 10% held-out data, repeated across 10 folds of cross-validation, is given in Fig. 3. For both datasets, sigmoid functions fit to model predictions have an RMSE comparable to those fit to uniformly random data, while the functions fit to human data have errors an order of magnitude lower.

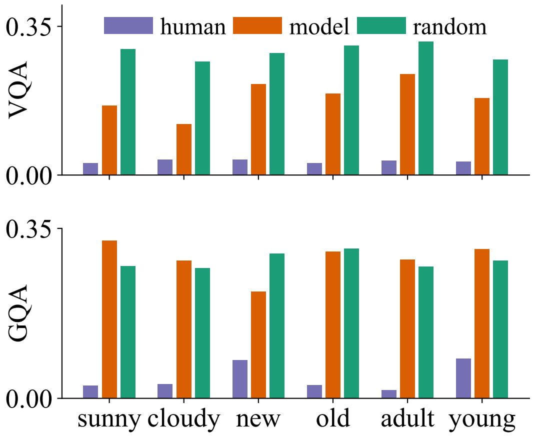


Figure 3: Mean RMSE from sigmoid fit to VQA and GQA data using 10-fold cross-validation. Human predictions result in a far better sigmoid fit, while model predictions have similar fit to data $\sim \mathcal{U}(0, 1)$.

This indicates that the remaining GQA and VQA predicates follow a similar pattern to the one seen in Fig. 1. While model predictions often fall on the correct side of the middle threshold, as examples become borderline, some predictions become erratic while others are confidently misclassified. Note that this is doubly problematic: firstly, the model only makes use of a small region of the label space. While the output vocabulary includes entries such as "partly cloudy" and "overcast," for all examples tested, the model assigns $> 98\%$ of its probability mass to "yes" and "no."

Even within this constrained assignment, the model has the possibility of hedging using the output logits (e.g. $p(\text{yes}|x) = 0.40$ etc.). *Prima facie* we might hope that, given a large categorically-labeled dataset, the model would learn the correct output distribution, as Pavlick and Kwiatkowski (2019) put it, "for free." We do not find this to be the case: the prediction generally heavily favors one label alone, posing problems for any downstream task as well as active learning setups using uncertainty sampling (Lewis and Catlett, 1994).

In contrast, annotators display hedging between the labels, reliably using the slider-bar interface to equivocate between extremes in borderline cases. These results suggest that the first assumption described in § 2, namely that images can be identified as being in the extension of a descriptor or not (e.g. in the set of scenes described as "cloudy"), holds only at the ends of the example range, and is not warranted in the borderline region. In contrast, the training data which LXMERT sees makes the assumption that the descriptor either applies (examples with a "yes" label) or does not apply (examples labelled "no") in all regions; we see that this is perhaps too strong of an assumption when trying to capture the nuances of vague terms.

Note also that the annotators' standard error (grey band) is generally fairly low even in the central region, where we would expect greater disagreement. This trend holds across descriptors, and perhaps implies that the second assumption, that annotators can reliably recover the mapping between inputs and outputs, does to hold as long as the annotators are provided the proper interface for expressing their intuitions.

4 Text-only Vagueness

§ 3 explored predicates grounded in another representation of the world, namely images. However, much of NLP deals with text in isolation, without grounding to some external modality. In an ungrounded setting, it is unproductive to evaluate models on external knowledge that they would not have access to—thus, we cannot evaluate a text-only model's performance on vague predicates the same way as a grounded model's performance. In other words, we need to develop a paradigm which does not rely on knowledge about a state of the world, but rather on linguistic knowledge. This is precisely the analytic-synthetic distinction, with analytic truths being truths *by virtue of meaning alone* (e.g. "a bachelor is an unmarried man") and synthetic truths being those which require verification against a state of affairs (e.g. "Garfield is a bachelor"). To avoid evaluating our text-only models on their ability to reason against a world which they are not privy to, we restrict our analysis to analytic truths and falsehoods, which we construct by pairing words either with their true definition or with a distractor definition, creating statements that are analytically true and false. Recall from § 2 that even the predicates *is true* and *is false* may be seen as vague; there are statements which are only partially true or false, and we can speak meaningfully of some statements being truer than others.

Sentence	T/F	Mark
journalism is newspapers and magazines collectively	T	◇
T-shirt is an archaic term for clothing	F	△
T-shirt is a close-fitting pullover shirt	T	●
a teammate is someone who is under suspicion	F	■

Table 1: Example sentences, with their label in the created dataset and corresponding color in Fig. 4.

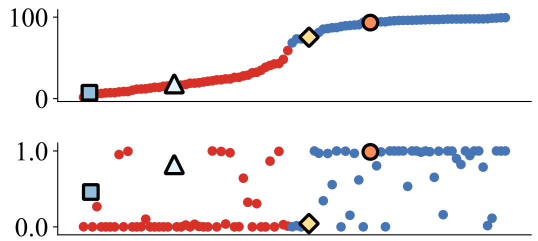


Figure 4: *Top:* mean truth score given by humans on 96 statements. False statements colored red, true blue; statements from Table 1 overlaid. *Bottom:* P(true) assigned by the best probing classifier (XLNet + `[CLS]`).

Following Ettinger et al. (2018), these statements are created artificially, mitigating annotator bias. Definitions of the 2542 most frequent English nouns[6] are then obtained from WordNet (Miller, 1995; Fellbaum, 1998) using the NLTK interface (Bird, 2006). By pairing a "trigger" word with its definition, we create an analytically true statement (c.f. row 3 in Table 1). In order to create analytically false statements, we pair the same word with a definition for a related but distinct term. A set of candidate terms is created recursively taking the hypernym of the trigger word's top wordsense[7] for three levels (i.e. the hyper-hyper-hypernym) and adding all its hyponyms, excluding the trigger's siblings. The best distractor candidate is chosen using lexical overlap, where the candidate with the lowest overlap with the true definition is chosen. Note that as a simplifying assumption we ignore polysemy here; it is possible that via polysemy the

[6] `https://www.wordfrequency.info`

[7] Based on pilot evaluations, we exclude chemistry-related wordsenses, as their definitions often contain low-frequency technical terms.

chosen distractor definition is not strictly analytically false. However, this result is unlikely given that human annotators reliably recognized distractor definitions. We expect that, while the examples are categorically labeled *true* and *false*, annotators will determine that certain statements fall into a borderline region between these extremes, corresponding to notions like "partially true" or "mostly false."[8] Crucially, where in § 3 the vagueness was present in the question itself (i.e. the task was to determine whether the object in question, e.g. the sky, in the image fell into the extension of the vague term e.g. things that are cloudy) here it is in the label set; the task becomes determining whether the statement as a whole falls into the set of true statements. The data is split into 4000 train, 500 development, and 536 test sentences. For all triggers, both statements are found in the same split.

96 sentences were sampled from the development set and annotated with 10-way redundancy by vetted crowdworkers on Mechanical Turk. Using a similar interface as in § 3, annotators were presented with sentences and asked to rate the sentence's truth using a sliding bar (ranging from 0 to 100) from false to true. In addition, an "I don't know" checkbox was provided to avoid forcing a choice. Sentences were presented in groups of 8. Additional details on the annotation interface can be found in Appendix A.

4.1 Encoders and Models

While the text-only experiments also focus on examining vagueness, several important contrasts to § 3 must be drawn. In the visual setting, the entire LXMERT model was separately finetuned on the whole GQA and VQA train splits, and analysis examples were sourced from the development data. In the text-only case, we do not have a pre-made dataset and construct our own. Due to the smaller size of our dataset, we have opted to only fine-tune the classification layer, freezing the weights of the contextualized encoders, unlike in the visual setting where we trained the entire model. This is far less computationally expensive, and allows us to expand our text-only analysis to a range of encoder types and model architectures. We examine three different contextualized encoders:

BERT BERT (Devlin et al., 2019) is a transformer-based model which uses a word's context to predict its identity; during training, words in the input are randomly replaced with a `[MASK]` token; the model then predicts masked words based on their contexts—a cloze-style task known as masked language modeling (MLM). BERT also uses a next-sentence prediction objective.

RoBERTa RoBERTa (Liu et al., 2019) uses roughly the same methodology as BERT, but trains the model for more epochs with larger batch sizes while removing the next-sentence prediction task.

XLNet While traditional language models only consider one factorization (in the forwards or the backwards direction), Yang et al. (2019) maximize the expected log-likelihood with respect to all factorizations input's joint probability.

Drawing on the observations of Warstadt et al. (2019) that probing results can change dramatically depending on how an encoder is probed, we introduce three probing classifiers:

Mean-pool The mean-pool classifier takes the average across all dimensions of the encoder output at each input token, yielding one vector for the whole sentence. This vector is then passed to a 2-layer multi-layer perceptron (MLP) with ReLU activations, which produces a classification over the 2D output space.

Sequence The sequence classifier uses the encoder representation at the index of the `[CLS]` token, which it then passes to a 2-layer MLP with twice as many hidden units as input units.

Bilinear This classifier splits the probing prompt into a trigger word (e.g. "bachelor") and a definition (e.g. "an unmarried man"); it encodes both into vectors, mean-pooling the definition to produce two vectors, which are projected through two linear layers. The projected representations x_{trig} and x_{def} are then passed through a bilinear layer, given by $f(x_{\text{trig}}, x_{\text{def}}) = x_{\text{trig}}^T \mathbf{A}\, x_{\text{def}}$, where $\mathbf{A}$ is a 3-dimensional learned parameter.

Control Tasks Following Hewitt and Liang (2019), we construct control tasks for all of our models and encoders. A control task is one where labels and inputs are paired randomly; the purpose of such a task is to disentangle what portion of the probing classifier's performance can be attributed to the strength of the classifier, and what portion is present in the representation.[9]

[8] Note that this conceptualization of truth diverges from that of classical logic, but may be more faithful to actual usage.

[9] All models are trained for 100 epochs with the Adam optimizer using a learning rate of 0.0001. The best model was chosen by validation performance.

5 Results and Analysis

We find that our control classifiers perform randomly, indicating our task has very low sensitivity. Fig. 5 shows the test accuracies of all (non-control) models in all settings. We see that all models fall well below human performance, but well above the random baseline of 50%. Among the probing methods, `[CLS]` pooling slightly outperforms mean-pooling. The bilinear method consistently underperforms the pooling methods, suggesting that the gap between human and model performance is not due to malformed prompts (e.g. incorrect articles in the definition or trigger phrase). Appendix C gives some examples and model predictions.

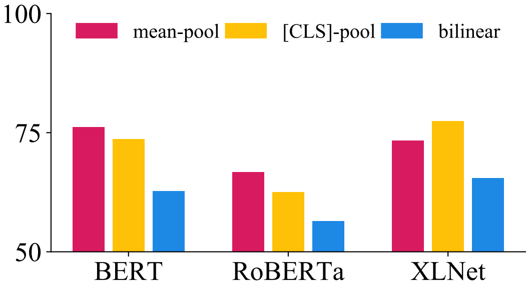


Figure 5: Test accuracy across encoders and probing methods; all models perform well above chance.

Human annotators are able to perform the task with high reliability, achieving an accuracy of 88.54 with majority voting. Fig. 4 shows that certain sentences are easily classified as either true or false, while a smaller number of sentences are considered borderline. A qualitative analysis of these sentences reveals that they typically fall into two categories: sentences where the trigger described is very abstract (e.g. "a separation is the state of lacking unity") and those where the distractor definition is very closely related to the trigger (e.g. "a baby is a person's brother or sister"). Intuitively, both of these phenomena can make a sentence only partially true or false.

While Fig. 5 suggests the models are performing reasonably well in the aggregate, Fig. 4 demonstrates a similar trend to those seen in § 3, showing that the classification patterns of humans differ drastically from those of the best model, as illustrated by the overlaid examples. We also see the same overconfidence in the output distribution of the model, with predictions saturating at either end of the simplex. Fig. 6 further reinforces this; here, we perform the same analysis as in § 3, fitting a 2-parameter logistic regression to the aggregate human scores, the model predictions, and samples of a uniformly-distributed random variable, computing the RMSE between the best-fit sigmoid and the data. Across all models and all encoder types, we see that the RMSE of a sigmoid fit to the model predictions is close to or higher than the RMSE of a sigmoid fit to uniformly random data ($\text{RMSE}_{\text{random}} = 0.326$), as evidenced by the overlaid red horizontal line, while the sigmoid fit to human performance has a far lower RSME ($\text{RMSE}_{\text{human}} = 0.051$). This quantitatively reinforces the qualitative difference seen in Fig. 4.

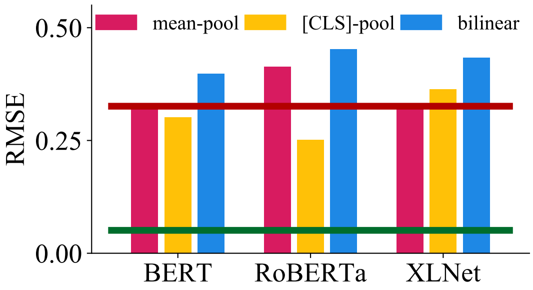


Figure 6: 10-fold cross-validated RMSE against model of 2-parameter sigmoid against model predictions from each encoder and model pairing. RMSE to human performance (green line, bottom) and against random data (red line, top) are overlaid. RMSE to model predictions is close to or worse than to random data.

6 Related Work

Human-model divergence In similar vein to our work, Pavlick and Kwiatkowski (2019) observe that human annotators consistently disagree on natural language inference (NLI) labels, and that the disagreement cannot be attributed to a lack of annotations. They similarly find that models do not implicitly learn to capture human uncertainty from categorical data. In contrast, our work seeks to pinpoint vagueness as a cause for some of the difference in behavior.[10]

Other work has looked at annotating data to accommodate the kinds of disagreements seen in Pavlick and Kwiatkowski. Chen et al. (2020) extends the EASL framework (Sakaguchi and Van Durme, 2018) for efficiently eliciting reliable scalar judgements from crowdworkers to NLI, ob-

[10] We examined high-disagreement examples from the data released by Pavlick and Kwiatkowski, which largely seem not to be caused by vagueness except for some examples from JOCI (Zhang et al., 2017), e.g. *P*: "I loved apple sauce", *H*: "The sauce is a condiment" may have high disagreement due to vagueness in the predicate *isACondiment(x)*".

taining scalar NLI judgements rather than categorical labels. In a similar context, Li et al. (2019) argue that for tasks involving plausibility, the use of cross-entropy loss drives model predictions to the extremes of the simplex, and demonstrate the benefits of shifting to a margin-based loss on the Choice of Plausible Alternatives (Roemmele et al., 2011) task. These results dovetail with our observations regarding various models' output distributions, especially in the text-only setting, where our task is very similar to tasks measuring plausibility.

While Pavlick and Kwiatkowski (2019) focus on NLI data, Bhattacharya et al. (2019) have noted that similar disagreements exist in the visual domain, specifically on the VQA data set, where they find that certain image-question pairs are less reliably answered than others. The ontology they propose to classify these images includes ambiguity and under-specification, but not vagueness.

Vagueness Past work in vagueness has often focused on modeling it as a phenomenon, while our work is concerned with analyzing model performance on vague predicates, rather than capturing the semantics of vague predicates, which has been the focus of previous work such as Meo et al. (2014) and McMahan and Stone (2015). Although color terms provide a particularly rich substrate for modeling the semantics of vague terms, we have chosen to exclude them as we feel they demand a level of psychophysical analysis beyond the scope of this work. This work deals instead with gradable terms, following work such as Fernández and Larsson (2014), who present a type-theory record account of vagueness for learning the semantics of gradable adjectives, DeVault and Stone (2004), who use vagueness to illustrate the need for context in a dialog-driven drawing task, and Lassiter and Goodman (2017), who introduce a Bayesian pragmatic model of gradable adjective usage. These lines of previous work draw on the contextualist account of vagueness, holding that the meaning of vague predicates shifts with respect to the interests of the parties communicating, a notion that naturally expresses itself in rational pragmatic models of dialog. Rather than modeling vagueness, we use it as a tool to examine model behavior, focusing on single interactions instead of a dialog. We refer the reader to Juhl and Loomis (2009) for a full account of the analytic/synthetic distinction.

Text-only semantic probing The challenge of analyzing the semantic content of sentence encodings precedes the contextual encoders studied herein; Ettinger et al. (2016) introduce a suite of simple classification tasks for probing the compositionality of LSTM-based sentence embeddings, while Conneau et al. (2018) present 10 linguistically-motivated probing tasks, including 3 semantic tasks, for LSTM- and CNN-based sentence embeddings. Ettinger et al. (2018) create a set of artificial prompts, as done in this work, to probe the compositionality of InferSent (Conneau et al., 2017), while Dasgupta et al. (2018) use NLI-style prompts for the same purpose.

Similar probing suites have been proposed since the advent of contextual encoders; Tenney et al. (2019b) propose a set of edge-probing tasks that examine semantic content, and Tenney et al. (2019a) find that semantic information is typically encoded at higher transformer layers. Presenting a suite of negative polarity item-based tasks, Warstadt et al. (2019) expand on the observation that different transformer layers account for different phenomena, noting that additionally, the manner in which a probing task is framed often makes a large impact.

Dictionary Embeddings Dictionary embeddings, as described by Hill et al. (2016), use dictionary resources to learn a mapping from phrases to word vectors. Dictionaries have also been used with a view to augmenting the semantic information in word embeddings, as in Tissier et al. (2017) and Bosc and Vincent (2018). In contrast to these approaches, we use definitions to investigate the semantic content of existing mappings.

7 Conclusion

We have identified clashes between the assumptions made under the current NLP paradigm and the realities of language use by focusing on the phenomenon of vagueness. By isolating a subset of examples from VQA and GQA involving vagueness, we were able to pinpoint some key divergences between model and human behavior which result in lower model performance. We then created an artificial text-only dataset, controlling for world knowledge, which we used to contrast multiple models building on multiple contextualized encoders, finding similar human-model contrasts. In closing, we would like to advocate for the broader use of concepts from the philosophy of language, such as vagueness, in challenging current models and providing additional insights beyond aggregate statistics and leaderboards.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. Why does a visual question have different answers? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4271–4280.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Tom Bosc and Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532.

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.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single &!#* vector: Probing sentence embeddings for linguistic properties. In *ACL 2018-56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2126–2136. Association for Computational Linguistics.

Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J Gershman, and Noah D Goodman. 2018. Evaluating compositionality in sentence embeddings. *arXiv preprint arXiv:1802.04302*.

David DeVault and Matthew Stone. 2004. Interpreting vague utterances in context. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 1247–1253.

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.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database cambridge. *MA: MIT Press*.

Raquel Fernández and Staffan Larsson. 2014. Vagueness and learning: A type-theoretic approach. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (* SEM 2014)*, pages 151–159.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Delia Graff. 2000. Shifting sands: An interest-relative theory of vagueness. *Philosophical topics*, 28(1):45–81.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

Cory Juhl and Eric Loomis. 2009. *Analyticity*. Routledge.

Hans Kamp. 1981. The paradox of the heap. In *Aspects of Philosophical Logic*, pages 225–277. Springer.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

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

David D Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier.

Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. 2019. Learning to rank for plausible plausibility. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4818–4823.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Brian McMahan and Matthew Stone. 2015. A bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Timothy Meo, Brian McMahan, and Matthew Stone. 2014. Generating and resolving vague color references. In *Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*, pages 107–115.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

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

Diana Raffman. 1994. Vagueness without paradox. *The Philosophical Review*, 103(1):41–74.

Steven P. Reise, Andrew T. Ainsworth, and Mark G. Haviland. 2005. Item response theory: Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science*, 14(2):95–101.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218.

Stewart Shapiro. 2006. *Vagueness in context*. Oxford University Press on Demand.

Roy Sorensen. 2001. *Vagueness and contradiction*. Clarendon Press.

Roy Sorensen. 2018. Vagueness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2018 edition. Metaphysics Research Lab, Stanford University.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Julien Tissier, Christopher Gravier, and Amaury Habrard. 2017. Dict2vec: Learning word embeddings using lexical dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 2019. Investigating bert's knowledge of language: Five analysis methods with npis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2870–2880.

Timothy Williamson. 1994. *Vagueness*. Routledge.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

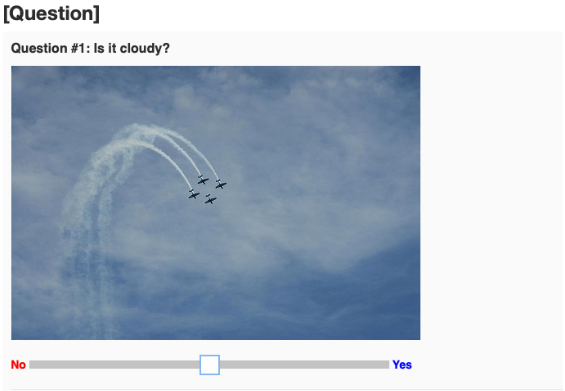
Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

A Data Collection

Figure 12 shows that on certain examples human annotators vary in their truth judgements, with some sentences receiving a high score (i.e. “True”) from certain annotators and a low score (i.e. “False”) from others. Further inspection reveals that many of the highest-variance examples have one annotator who is an extreme outlier.

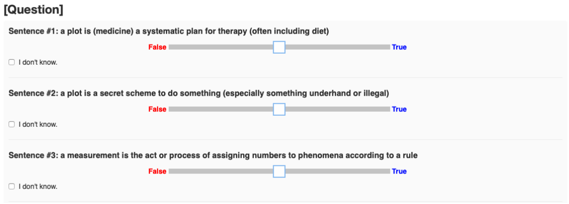
Figure 7 shows the MechanicalTurk annotator interface for collecting VQA and GQA annotations. The task was only available to annotators in the US with an approval rating $> 98\%$ and more than 500 recorded HITs. Instructions asked annotators to respond to the questions by using the sliding bar. They were provided with a comment box to use in case any issues arose.

Similarly, Figure 8 shows the interface for collecting text-only annotations. Here, the task was only shown to annotators from a list of reliable workers. Instructions asked annotators to rate how true a sentence was, and told that sentences may be true or false. They were instructed to use the “I don’t know” checkbox in cases where they did not know a word in the statement.


[Question]

Question #1: Is it cloudy?

Figure 7: Mechanical Turk annotation template for visual annotations.


[Question]

Sentence #1: a plot is (medicine) a systematic plan for therapy (often including diet)

□ I don't know.

Sentence #2: a plot is a secret scheme to do something (especially something underhand or illegal)

□ I don't know.

Sentence #3: a measurement is the act or process of assigning numbers to phenomena according to a rule

□ I don't know.

Figure 8: Mechanical Turk annotation template for text annotations.

B Plots

Figures 9 and 10 show human annotations plotted against model predictions for all of the predicates examined. In all cases, we see major divergences between human and model data, as quantified in Fig. 3. We also see that the standard error between annotators is fairly low. Furthermore, we see similar trends between descriptors across the two datasets, with “new” being skewed towards the higher end for both.

Figure 11 verifies that for the descriptors examined (“sunny” and “cloudy”) the mean score obtained from annotators on Mechanical Turk and the mean score from the VQA development roughly correspond, justifying the use of the VQA development data in § 3. However, we do note some divergence between the two annotation formats, likely due to the forced choice presented to the original VQA annotators.

C Text Examples

Table 2 contains 28 example sentences from the validation set, with human classifications derived by majority voting over the annotators who did not use the “I don’t know” box, as well as classifications obtained by the `[CLS]` model.

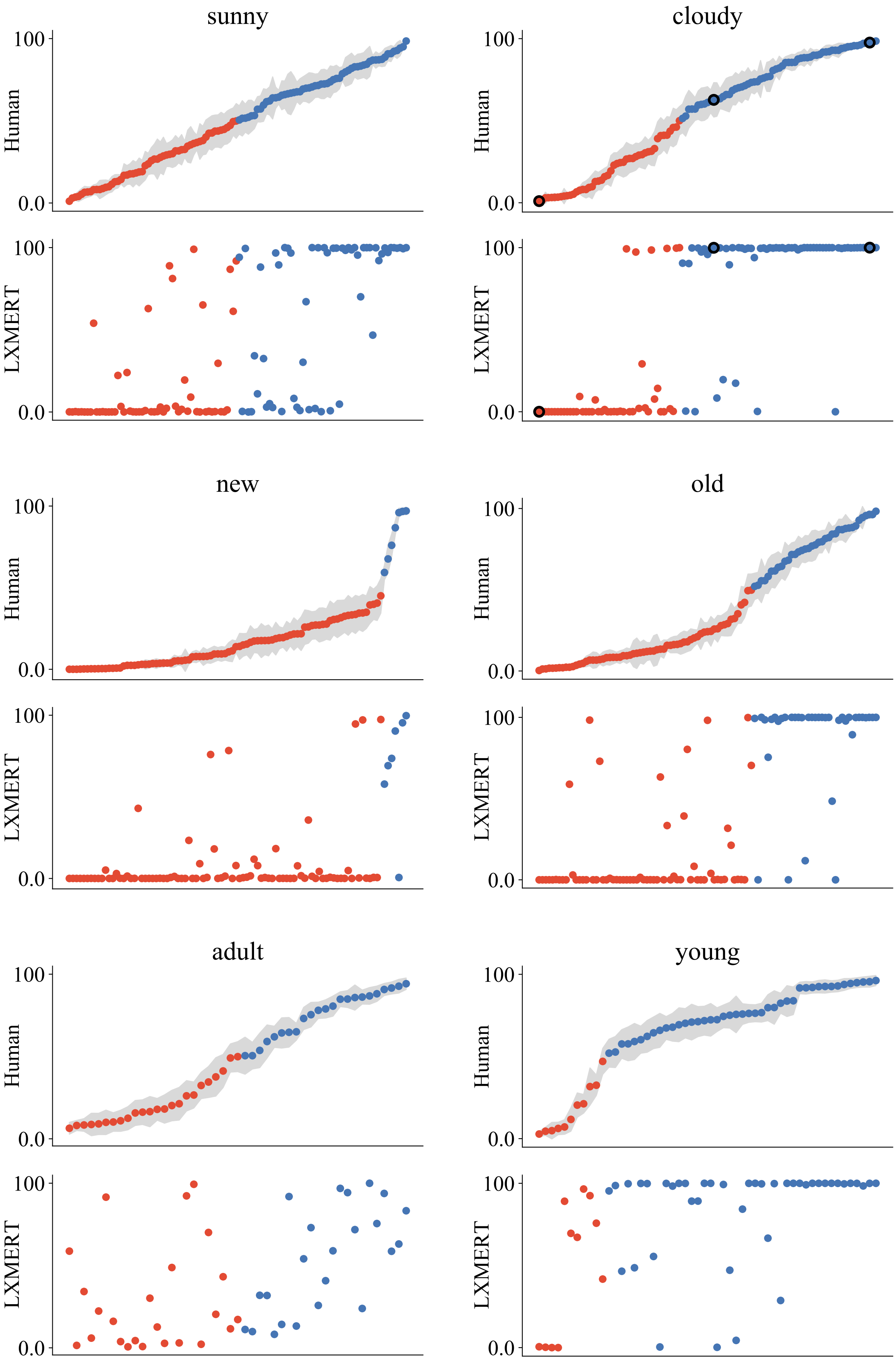


Figure 9: Human and model scores for questions containing vague terms from the GQA dataset.

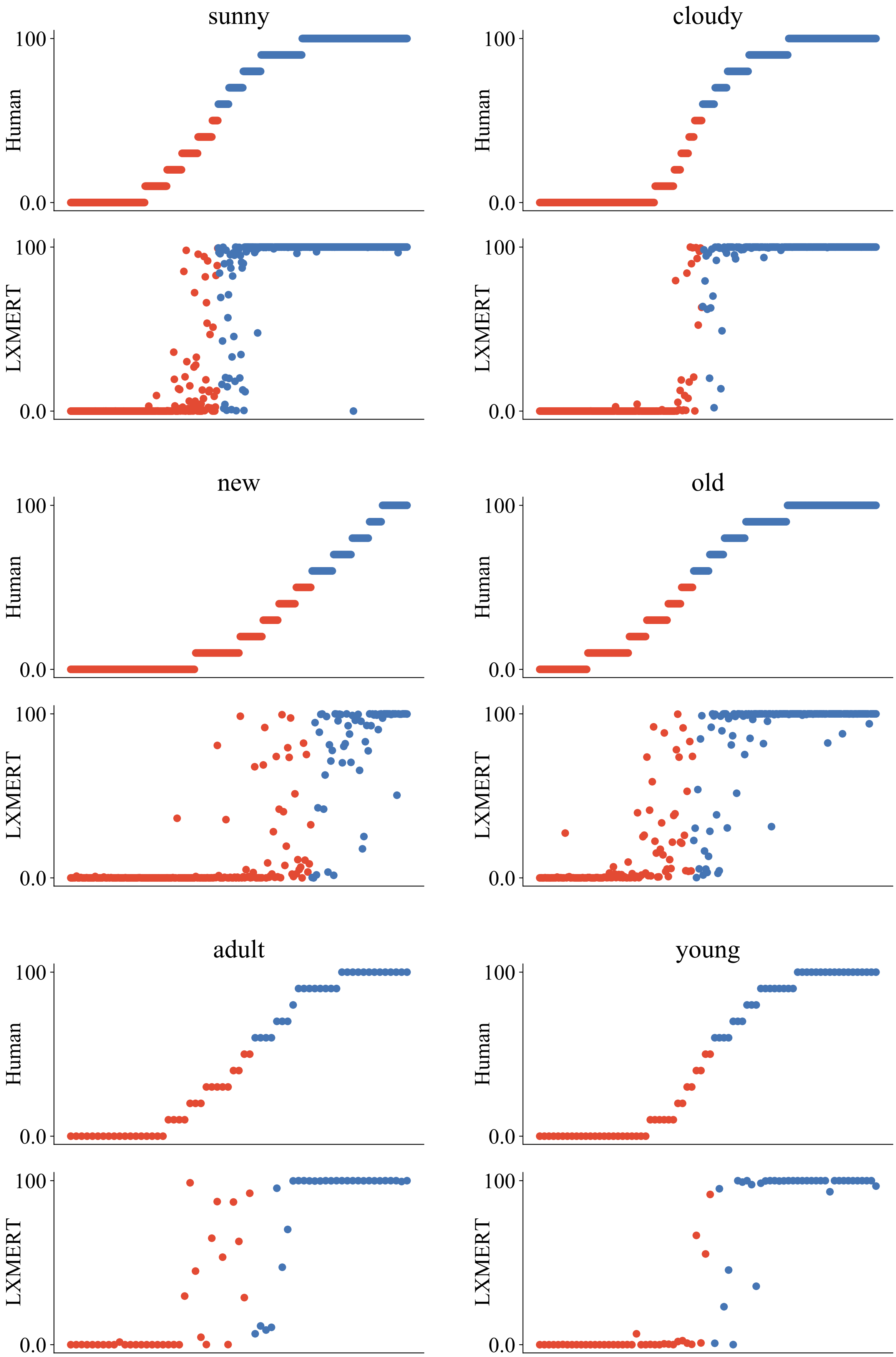


Figure 10: Average annotator scores and model scores for questions containing vague terms on the VQA dataset.

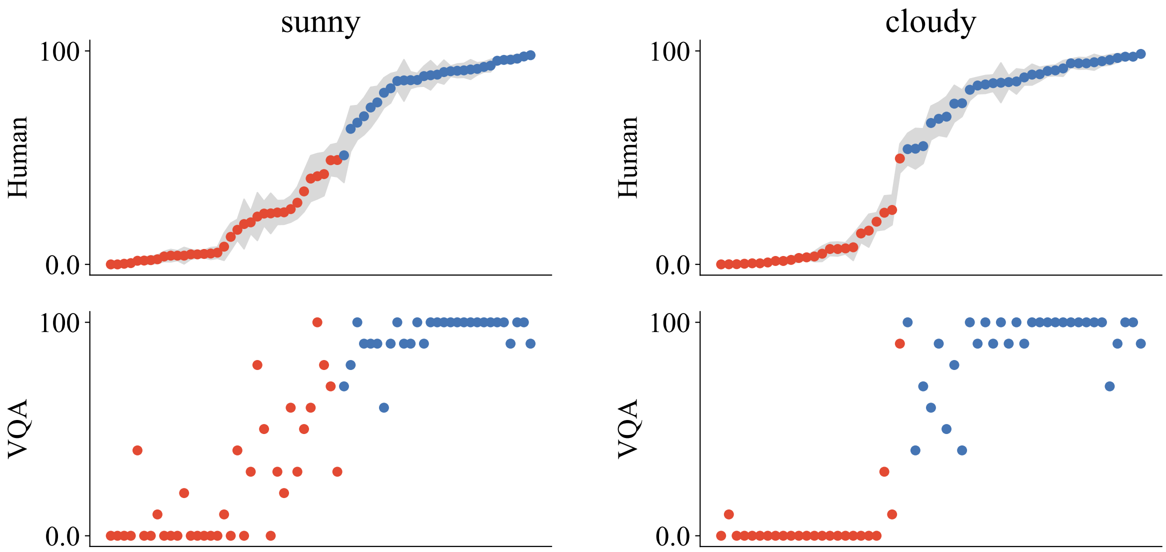


Figure 11: Manual verification of VQA plots shows that Mechanical Turker's judgments largely correspond to those present in the development set, with some divergence.

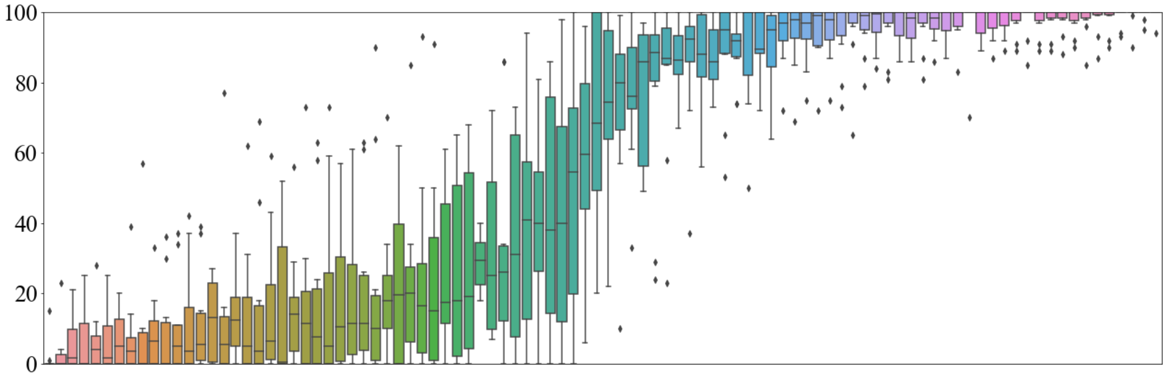


Figure 12: Human means and quartiles for examples ranked by average score

Sentence	Label	Human	XLNet	BERT	RoBERTa
a plot is (medicine) a systematic plan for therapy (often including diet)	F	21.90	1.00	0.22	0.49
a plot is a secret scheme to do something (especially something underhand or illegal)	T	95.60	1.00	0.38	0.31
a measurement is the act or process of assigning numbers to phenomena according to a rule	T	73.30	0.02	0.88	0.39
a measurement is a sudden event that imparts energy or excitement, usually with a dramatic impact	F	8.70	0.00	0.47	0.44
one is the product of two equal terms	F	21.33	0.04	0.72	0.49
one is the smallest whole number or a numeral representing this number	T	94.22	0.15	0.97	0.69
an exit is an opening that permits escape or release	T	97.90	0.94	0.93	0.79
an exit is a man-made object taken as a whole	F	7.30	0.00	0.01	0.09
a label is a brief description given for purposes of identification	T	95.20	1.00	0.62	0.33
a label is the act of having on your person as a covering or adornment	F	20.40	0.00	0.22	0.41
a ritual is the act of prolonging something	F	25.22	0.64	0.26	0.92
a ritual is any customary observance or practice	T	97.90	1.00	1.00	0.80
distance is faulty position	F	5.90	0.27	0.00	0.71
distance is the property created by the space between two objects or points	T	97.90	1.00	0.98	0.40
a shock is a lack of gratitude	F	7.67	0.00	0.29	0.27
a shock is the feeling of distress and disbelief that you have when something bad happens accidentally	T	96.10	0.53	0.03	0.74
a route is the frozen part of a body of water	F	7.30	0.00	0.88	0.73
a route is an established line of travel or access	T	97.90	0.79	0.89	1.00
a ban is a decree that prohibits something	T	97.70	1.00	0.75	0.83
a ban is a legal instrument authorizing someone to act as the grantor's agent	F	5.70	0.00	0.19	0.88
citizenship is the status of a citizen with rights and duties	T	96.20	1.00	0.91	1.00
citizenship is the state of having been made ready or prepared for use or action (especially military action)	F	12.56	0.00	0.07	1.00
an accent is distinctive manner of oral expression	T	90.30	0.97	0.58	0.53
an accent is (language) communication by word of mouth	F	47.56	0.03	0.08	0.22
journalism is newspapers and magazines collectively	T	81.89	0.02	0.96	0.32
journalism is an artifact made of hard brittle material produced from nonmetallic minerals by firing at high temperatures	F	1.60	0.00	0.00	0.88
atmosphere is a particular environment or surrounding influence	T	87.20	1.00	0.52	0.72
atmosphere is any attribute or immaterial possession that is inherited from ancestors	F	12.56	0.00	0.00	0.31

Table 2: Sentences, labels, human means and model logits for 28 sample validation examples.