Does the “most sinfully decadent cake ever” taste good? Answering Yes/No Questions from Figurative Contexts

Geetanjali Rakshit and Jeffrey Flanigan
Computer Science and Engineering Department
UC Santa Cruz
{grakshit,jmflanig}@ucsc.edu

Abstract

Figurative language is commonplace in natural language, and while making communication memorable and creative, can be difficult to understand. In this work, we investigate the robustness of Question Answering (QA) models on figurative text. Yes/no questions, in particular, are a useful probe of figurative language understanding capabilities of large language models. We propose FigurativeQA, a set of 1000 yes/no questions with figurative and non-figurative contexts, extracted from the domains of restaurant and product reviews. We show that state-of-the-art BERT-based QA models exhibit an average performance drop of up to 15% points when answering questions from figurative contexts, as compared to non-figurative ones. While models like GPT-3 and ChatGPT are better at handling figurative texts, we show that further performance gains can be achieved by automatically simplifying the figurative contexts into their non-figurative (literal) counterparts. We find that the best overall model is ChatGPT with chain-of-thought prompting to generate non-figurative contexts. Our work provides a promising direction for building more robust QA models with figurative language understanding capabilities.

The cake was described as ***the most sinfully decadent ever*** *.*

Question: Did the cake taste good?
Answer: Yes

Figure 1: To answer the question “Did the cake taste good?” based on the context, a Question Answering (QA) model needs to be able to correctly infer the meaning of the figurative text “the most sinfully decadent ever”

1 Introduction

> *“Questions are never indiscreet. Answers sometimes are.”*
>
> *- Oscar Wilde*

One of the many interesting phenomena occurring in natural language is the presence of figurative language, which, while making communication creative and memorable (Danescu-Niculescu-Mizil et al., 2012), may sometimes also prove difficult to understand (Zayed et al., 2020). This includes (but is not limited to) linguistic constructs such as idioms, similes, metaphors, rhetorical questions, hyperbole, personification, sarcasm, and irony. It may be particularly difficult for non-native speakers to interpret figurative expressions, and phenomena like sarcasm are often missed altogether (Joshi et al., 2016). Given that figurativeness is commonplace in everyday communication (Lakoff and Johnson, 2008), progress in the field of Natural Language Understanding (NLU) would be incomplete without figurativeness understanding. Consequently, figurative text has been studied in various downstream NLP tasks such as machine translation (Dankers et al., 2022), textual entailment (Agerri, 2008), (Chakrabarty et al., 2021), (Liu et al., 2022) and dialog models (Jhamtani et al., 2021), inter-alia. However, to the best of our knowledge, there has not been a systematic study of figurative language understanding capabilities of question answering models.

We focus on yes/no questions for our question answering (QA) task. Yes/no questions are a good test of figurative language understanding because correctly answering them requires the reader to correctly understand the figurative language. Extractive QA, on the other hand, is not a good test for figurative language understanding because it does not require actually understanding the figurative language.

Proceedings of Recent Advances in Natural Language Processing, pages 926–936
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_100

For example, if we were to pose the question "How did the cake taste?" from the context "The cake was described as the most sinfully decadent ever.", an answer such as "sinfully decadent" from an extractive QA model doesn't really tell us that the model understands the meaning of the figurative text "sinfully decadent". It simply copies the figurative text and it's up to the reader to infer what the answer means.

However, in order to answer a yes/no question such as "Did the cake taste good?", a QA model needs to correctly infer that "sinfully decadent" means *rich and delicious*, or in other words, *really good*, and therefore the answer would be *yes*.

Despite the lack of attention of figurative language for QA tasks, figurative language is extremely common in some important domains, such as online reviews. We randomly sampled 100 reviews from the train split of the Yelp Challenge Dataset[1], and observe that at least 60% of these reviews contain figurative expressions. Users often write strongly-worded reviews, to express highly positive or highly negative opinions about products or services (Mohammad et al., 2016), which tend to contain figurative language.

We show that it can be challenging for existing QA models to draw inferences from figurative text. To do this, we present a new dataset, ***FigurativeQA***, consisting of 1000 yes/no questions and accompanying figurative and non-figurative contexts constructed from Amazon product reviews (Niculae and Danescu-Niculescu-Mizil, 2014) and Yelp restaurant reviews (Oraby et al., 2017). In Figure 2, we show examples from FigurativeQA, in two domains: Amazon product reviews and Yelp restaurant reviews, for both figurative and non-figurative contexts. Each context is accompanied by a question-answer pair, and in the case of figurative contexts, manually constructed and automatically obtained non-figurative versions of the context.

We develop a variety of methods for improving QA performance for figurative text. We prompt powerful LLMs like GPT-3 and ChatGPT to convert figurative contexts to literal as an intermediate step to question answering. We then provide these literal contexts as input to state-of-the-art QA models, resulting in considerable gains in performance. The best performance is achieved by the chain-of-thought prompting method from ChatGPT in a few-shot setting, where the model generates a simplified version of the context and then generates the yes/no answer. We also use these LLMs to generate domain-specific training data to fine-tune models specifically for this task.

The outline of the paper is as follows: after reviewing related work (§2), we introduce our new QA dataset for figurative language, FigurativeQA, in (§3). We report baseline QA performance on FigurativeQA and introduce a method for simplifying figurative language to non-figurative by prompting GPT-3 and ChatGPT, which we use to improve our baseline QA models (§4, 5, 6). We report our experiments with chain-of-thought prompting in §7. We prompt ChatGPT to generate in-domain training data for figurative question answering (§8). We finally conclude in (§10). The FigurativeQA dataset can be accessed at `https://github.com/geetanjali-rakshit/figurativeQA`.

2 Related Work

Figurative language has been a difficult problem for many natural language processing (NLP) applications. A number of computational approaches have been proposed to study their occurrence in text (Veale et al., 2016; Qadir et al., 2016; Kordoni, 2018; Mao et al., 2018; Zhang et al., 2017; Troiano et al., 2018), including generation of figurative language (Chakrabarty et al., 2020; Zhou et al., 2021).

The idea of converting metaphors to their literal counterparts has been previously explored for machine translation by Mao et al. (2018), where metaphors in English text are first identified and then converted to a literal version by using word embeddings and WordNet, before doing machine translation into Chinese. In dialog systems, a similar approach was employed by Jhamtani et al. (2021), where idioms and metaphors in utterances are converted to literal versions using a dictionary lookup-based method. Our work is closest to Jhamtani et al. (2021), except that we explore the robustness of QA systems in a machine comprehension setup, instead of dialog models, to figurative language, which, to the best of our knowledge, is a first. Our automatic approach to creating rephrased non-figurative versions of figurative text is done using pre-trained language models, rather than rule-based methods which have been shown to be error-prone (Jhamtani et al., 2021). In a concurrent work,

[1] We use the version in Huggingface Datasets (`https://huggingface.co/datasets/yelp_review_full`), from the paper (Zhang et al., 2015)

Split	Source	Example
Figurative	Amazon	**Context**: *The album , like almost everything Krush has released ,* ***slays*** *.* **Question**: *Is the album good?* **Answer**: *Yes*
		Non-fig. version of the context (manual): *The album is really good, like most of Krush's work.* **Non-fig. version of the context (from GPT-3)**: *The album is really good, like almost everything Krush has released.*
Figurative	Yelp	**Context**: *Although, the menu items doesnt* ***SCREAM*** *French cuisine. Most foods looks like you can get at any American place.* **Question**: *Is the menu authentic french?* **Answer**: *No*
		Non-fig. context (manual): *The menu items aren't typical of French cuisine. Rather, they are common at most American eateries.* **Non-fig. context (from GPT-3)**: *Although, the menu items doesn't look very French. Most foods look like you can get at any American place.*
Non-figurative	Amazon	**Context**: *Nice ring, but the color is paler than the picture .* **Question**: *Is the ring brightly colored?* **Answer**: *No*
Non-figurative	Yelp	**Context**: *the chicken is delicious and so are the ribs* **Question**: *Did the food taste good?* **Answer**: *Yes*

Figure 2: Examples from the figurative and non-figurative splits of FigurativeQA, from Amazon product reviews and Yelp restaurant reviews. The figurative text fragments within the contexts are shown in bold and italics.

Chakrabarty et al. (2022) have also done prompting on GPT-3 to create their figurative NLI dataset, FLUTE, as well as obtain an explanation of the NLI labels in this dataset.

To our knowledge, there are no QA datasets specifically designed for figurative language understanding, but some existing QA datasets do contain figurative language. The FriendsQA dataset (Yang and Choi, 2019) is a dialog-based QA dataset constructed from dialogs from the TV series Friends. While it does contain metaphors and sarcasm, the focus of the dataset is not figurative language, and it is not ideal for testing figurative language understanding as it is unclear how much of the dataset is figurative. The dialog nature of the dataset further contributes to making it challenging and complicates studying the effect of figurativeness. Another dataset that requires figurative language understanding is the RiddleSense dataset (Lin et al., 2021), which comprises of riddles, but unlike ours, it's modeled as an open-domain QA task rather than a machine comprehension task. Parde and Nielsen (2018) show that questions about novel metaphors from literature are judged to be deeper than non-metaphorical or non-conventional metaphors by humans, but their focus is on generating deep questions rather than testing the robustness of QA models. Dankin et al. construct yes/no questions using templates to detect the presence of metaphors in a few-shot setting.

3 FigurativeQA Dataset

The contexts in FigurativeQA comes from two sources: Amazon product reviews (Niculae and Danescu-Niculescu-Mizil, 2014), and Yelp restaurant reviews (Oraby et al., 2017). We extract both figurative and non-figurative contexts from each source. We manually construct yes/no questions and answers on top of these contexts. Figure 2 shows examples from FigurativeQA. The data statistics from each source (Amazon and Yelp) and each split (figurative and non-figurative) are summarized in Table 1.

For the Amazon part of FigurativeQA, we use Niculae and Danescu-Niculescu-Mizil (2014)'s dataset of figurative comparisons. Of the 1260 comparisons in this dataset, we extract instances where all 3 annotators are in agreement about figurativeness (i.e., average figurativeness score of greater than 3). We then randomly pick 150 exam-

Split	Context	fig. construct
Amazon	*The books are* ***like potato chips*** *- you* ***can't eat just one*** *.*	*simile, idiom*
	So when my laptop battery puffed up ***like a balloon*** *, I dreaded paying the cost of replacement .*	*simile, hyperbole*
	Really , this novel feels more ***like a footnote*** *to the series whereas The Gunslinger was a novel that* ***stood extremely well on its own*** *.*	*simile, idiom*
	These horrible recordings ***contain treasure more precious than gold****.*	*simile, sarcasm*
Yelp	*i had the chicken fajitas , which came with a giant flour tortilla that was* ***as hot as hades*** *.*	*simile, hyperbole*
	the cheese was scarce as was the meat , and the taste was nothing to ***write home about*** *.*	*idiom*
	i ate as much as i could because truly , underneath the ***salt mine*** *on my plate , was some damn fine corned beef hash !*	*metaphor, hyperbole*

Figure 3: Examples of figurative constructs observed in the Amazon and Yelp datasets. The figurative text fragments within the contexts are shown in bold and italics. In case of multiple labels occurring in the same context, the first bold fragment corresponds to the first label, and so on. In some cases, the same text fragment may have multiple labels (as in row 2)

	Amazon		**Yelp**	
	Fig.	**Non-fig.**	**Fig.**	**Non-fig.**
Yes	77	76	174	175
No	73	74	176	175
Total	150	150	350	350

Table 1: Distribution of yes/no questions from Amazon product reviews and Yelp restaurant reviews for figurative and non-figurative contexts

Figurative Construct	**Amazon**	**Yelp**
Simile	91	70
Metaphor	20	35
Hyperbole	18	44
Idiom	15	2
Sarcasm	2	20

Table 2: Distribution of occurrences of various kinds of figurative constructs in a random sample of 100 contexts from Amazon and Yelp each. It is common for a context to contain multiple figurative expressions, so these do not add up to 100% (refer to Figure 3 for examples).

ples to form the set of figurative contexts. From the examples with a low average figurativess score, we select 150 examples to form the set of non-figurative contexts.

For the Yelp part of the dataset, the contexts are sourced from (Oraby et al., 2017)'s NLG dataset for the restaurant domain. Since highly positive or highly negative reviews are more likely to contain figurative language, we extract these first, and then, similar to (Niculae and Danescu-Niculescu-Mizil, 2014), use comparator expressions to get a set of reviews likely to be rich in figurative content. We then manually examine these reviews to annotate 350 examples of figurative contexts and non-figurative contexts, each.

The figurative contexts from FigurativeQA tend to contain more *similes*, since comparator patterns (*"like"*, *"as"*, or *"than"*) were used to extract the text. However, we observe that many of these examples also contain other kinds of figurative constructs such as metaphor, idiom, hyperbole, sarcasm, etc. Table 2 shows the number of occurrences of various kinds of figurative constructs that we observe in a random set of 100 figurative contexts, each from Amazon and Yelp in FigurativeQA. (Oraby et al., 2017) note that one of the most prominent characteristics of restaurant reviews in the Yelp corpus is the prevalence of hyperbole, which we also observe in this sample. A context may contain multiple figurative elements, coming from different text fragments within the context. Also, in some cases, the same text fragment may denote multiple kinds of figurative constructs. In Figure 3, we show some examples of various kinds of figurative constructs occurring in FigurativeQA.

For each context in FigurativeQA, we construct a yes/no question. For the figurative contexts, we make sure to pose a question such that answering it would require an understanding of the figurative text present in the context. For the non-figurative contexts, we construct questions similar to the ones for the figurative contexts. Additionally, for the figurative contexts extracted from Amazon and Yelp, we manually create non-figurative counterparts that preserve the meaning and overall content.

3.1 Inter-annotator Agreement

Annotations for all the data in FigurativeQA (figurativeness scores for the examples from Yelp, construction of question-answer pairs, manual conversion of figurative contexts to non-figurative) were done by an in-house-trained graduate-student annotator. To assess the quality of figurative and non-figurative contexts for the Yelp contexts, we perform a second round of annotations with another trained annotator on a random sample of 50 contexts. This resulted in an inter-annotator agreement of 0.72 on figurativeness, calculated by Cohen's κ.

Similarly, to assess the overall quality of FigurativeQA, we randomly sample 50 figurative contexts for double annotation, which gives an additional set of annotations for the answers to the questions. The inter-annotator agreement on the answers was found to be 0.96, calculated by Cohen's κ. To validate the effectiveness of the questions for figurativeness comprehension, we also asked the annotators to indicate if answering the question required them to understand figurative text fragments present in the context. In the random sample of 50, in 49 cases the annotators were in agreement that this was indeed the case.

4 Do QA models find answering questions from figurative contexts harder?

Using FigurativeQA as a test set, we show that current models struggle to do well on figurative text compared to literal ones. We use a RoBERTa-based (Liu et al., 2019) QA model fine-tuned on BoolQ to show this. The BoolQ dataset (Clark et al., 2019) consists of yes/no questions from the Natural Questions dataset. We use the training split of BoolQ containing 9,427 examples to fine-tune RoBERTa-base and report its performance on FigurativeQA in Table 3. We find that the RoBERTa QA model performs poorly on the figurative contexts compared to the non-figurative contexts, with a drop in performance of ~8.5% points for Amazon, and ~23% points for Yelp. We observe that switching the figurative contexts for their manually created non-figurative counterparts shoots these numbers up in both cases, by ~10% points and ~23% points, for Amazon and Yelp, respectively. More powerful models like ChatGPT (in a few-shot setting) perform significantly better on figurative contexts, but still don't match the results on non-figurative versions of the contexts. This indicates that the conversion of figurative language to non-figurative language may help improve QA performance.

	Amazon	Yelp
RoBERTa-BoolQ		
Fig (Original)	83.4 ± 0.7	66.8 ± 1.4
Fig (manual non-fig)	**93.5 ± 1.1***	**90.0** ± **1.4***
Non-fig (Original)	92.0 ± 1.4	89.8 ± 1.7
ChatGPT(few-shot)		
Fig (Original)	92.6±1.1	80.6±0.7
Fig (manual non-fig)	**93.8 ±0.3***	**83.3±1.6***
Non-fig (Original)	93.5 ± 0.3*	88.7 ± 1.8*

Table 3: Accuracy of RoBERTa-base fine-tuned on BoolQ, and ChatGPT (few-shot), on the figurative split, manually created non-figurative version of the figurative split, and non-figurative split of FigurativeQA. (We reran experiments 1000 times with bootstrap resampling. The numbers reported are the mean and std-dev. * denotes statistically significant results, with $p < 0.05$ calculated using the Wilcoxon signed-rank test. The numbers in **bold** are the best results.)

5 Can prompting or finetuning LLMs help simplify figurative contexts?

We posit that answering questions from figurative contexts is harder, and that simplifying the figurative context into its literal/non-figurative version improves QA performance. However, since the task of manually converting figurative text to non-figurative is expensive and time-consuming, we propose to do this automatically by prompting GPT-3 (Brown et al., 2020) in two ways. First, we use GPT-3 (da-vinci-003) and ChatGPT in a few-shot setting to generate non-figurative/literal versions of all the figurative contexts in FigurativeQA.[2] We also used a similar approach to prompt ChatGPT. Please refer to Appendix A for model details and the prompts used. Second, we use a trained version of GPT-3 (da-vinci-002) fine-tuned specifically for the task of converting figurative to literal text.

As an intrinsic evaluation of the effectiveness of our prompting method, we manually evaluate the correctness of the non-figurative/literal contexts generated by prompting GPT-3 on a random sam-

[2] The experiments for this method to convert figurative text to non-figurative were performed by running API calls to the OpenAI da-vinci model. For each context, this took less than 1 second, for a total of less than 18 min and cost less than 8 USD for the entire dataset.

ple of 100 instances each, from Amazon and Yelp in FigurativeQA. We label each generated literal version as either **"correct"**, where none of the figurative expressions are present but the meaning is preserved, or **"incorrect"** where the generated output is the same/similar to the original context or the meaning has changed. Please note that this is a rather strict evaluation of correctness, as in some cases, some of the figurative text fragments present in the context is converted to literal, while the context may still be left with some amount of figurativeness (possibly arising from multiple figurative text fragments present in the context). Table 4 shows the results from the manual evaluation of the GPT-3 and ChatGPT outputs. We observe that these models are pretty good at converting figurative language in FigurativeQA to literal, with nearly 89% and 81% of the outputs from GPT-3 judged to be correct in Amazon and Yelp, respectively, and 92% and 88% for ChatGPT. In Figure 4, we show examples of non-figurative text generated from GPT-3 and ChatGPT.

	Amazon	**Yelp**
GPT-3	89%	81%
ChatGPT	**92%**	**88%**
Finetuned GPT-3	80%	77%

Table 4: Evaluation of non-figurative outputs from GPT-3 and ChatGPT, showing the percentage of generated outputs that do not contain figurative expressions, but preserve the original meaning of the figurative context.

We next explore using a fine-tuned version of GPT-3 to generate literal versions of figurative texts. Chakrabarty et al. (2022) propose the FLUTE dataset for Natural Language Inference (NLI), which has 9,000 figurative NLI instances, and explanations for the NLI labels. We extract the premise-hypothesis pairs with the label *"entailment"* from the training split of FLUTE to fine-tune GPT-3 (3,182 examples in total). We used the *davinci* model from OpenAI as the base model and fine-tuned for 4 epochs, with all default settings. We didn't perform any hyper-parameter tuning.[3] Table 4 (row 3) shows the results from manual evaluation of the fine-tuned GPT-3 outputs.

6 Can automatically generated non-figurative text improve QA performance?

We observed that ChatGPT has a much stronger performance on FigurativeQA than the baseline model of RoBERTa finetuned on BoolQ (section 4), and both of these models do better on non-figurative texts. We showed that both GPT-3 and ChatGPT can be effectively used to simplify figurative text into their non-figurative counterparts (section 5). We next experiment with simplifying contexts to boost QA performance. As competitive baselines, we also report zero-shot and few-shot QA performance[4] of GPT-3 and ChatGPT in Table 5. Besides the RoBERTa-finetuned-on-BoolQ baseline (previously described in section 4, we also fine-tune GPT-3 on the training split of BoolQ. For fine-tuning GPT-3, we used the *davinci* model from OpenAI as the base model and fine-tuned for 4 epochs, with all default settings. We didn't perform any additional hyper-parameter tuning.

In our experiments, we do not require knowing which contexts are figurative and which are non-figurative. We simply input both figurative and non-figurative contexts to the LLM to simplify any figurative language that is present, regardless if the context actually contains figurative language. In Table 5, we show that this method exhibits significant gains over the baseline RoBERTa model. We also report the performance of using GPT-3-finetuned-FLUTE as input to the RoBERTa baseline.

7 Can we use chain-of-thought prompting for improving QA performance on FigurativeQA?

Wei et al. (2022) have shown chain-of-thought prompting in Large Language Models (LLMs) to be effective for solving tasks requiring complex reasoning. Since understanding figurative language often requires implicit reasoning, we investigate the effect of applying chain-of-thought prompting for FigurativeQA using ChatGPT. (Our few-shot prompt for the chain-of-thought method is described in Appendix C.) This approach gives us the highest overall accuracy on FigurativeQA (Table 5).

[3] To fine-tune GPT-3 on the FLUTE dataset, it cost about 15 USD and took 62 minutes.

[4] Please refer to Appendix B for details about prompting GPT-3 and ChatGPT as a QA system.

Amazon	**Figurative Context**: *However , the obvious problem with Eragon hits* ***like a brick wall*** *.* **[CORRECT] Non-fig. version from GPT-3**: However, the obvious problem with Eragon is glaringly obvious. **[CORRECT] Non-fig. version from ChatGPT**: However, the obvious problem with Eragon is very clear.
	Figurative Context: *Not a storybook , by any means , this one is more* ***like a visit to the zoo*** *.* **[INCORRECT] Non-fig. version from GPT-3**: *Not a fairytale, by any means, this one is more like a visit to the zoo.* **[INCORRECT] Non-fig. version from ChatGPT**: *Not a fairytale, by any means, this one is more like a visit to the zoo.*
Yelp	**Figurative Context**: *this is as authentic thai* ***as much as imitation crab is authentic crab*** *.* **[INCORRECT] Non-fig. version (from GPT-3)**: *this is as authentic thai as much as imitation crab is genuine crab.* **[CORRECT] Non-fig. version from ChatGPT**: *This is not authentic Thai, just as imitation crab is not authentic crab.*
	Figurative Context: *the same thing with the steak and potatoes , it was almost as if they tried to* ***decorate the plate with salt*** *.* **[CORRECT] Non-fig. version from GPT-3**: *The steak and potatoes were heavily salted, as if they were trying to make the plate look more appealing.* **[CORRECT] Non-fig. version from ChatGPT**: *The steak and potatoes were oversalted and appeared to be more about presentation than taste.*

Figure 4: Examples of non-figurative contexts generated from GPT-3, for Amazon and Yelp. The figurative text fragments within the contexts are shown in **bold** and *italics*.

8 Can we prompt LLMs to generate training data for FigurativeQA?

Due to the lack of training data for question answering with figurative contexts, our supervised models are all finetuned on BoolQ. We hypothesize that adding synthetically generated QA pairs for this task will improve performance of the fine-tuned models. We prompt ChatGPT to generate synthetic training data (we tried a variety of prompts – refer to Appendix D for the prompt used). We use contexts from both Amazon and Yelp domains to generate question answer pairs from ChatGPT. For the Amazon contexts, we randomly sample reviews from 4 categories (Books, Electronics, Jewelry and Digital Music) from Amazon Product reviews from (McAuley and Leskovec, 2013). From these reviews, we extract sentences containing comparator patterns ("like", "as", "than") and use them as contexts, as they are more likely to contain figurative expressions. For the Yelp contexts, we extract sentences from (Oraby et al., 2017)'s NLG dataset also containing the same comparator patterns, but not already included in FigurativeQA. (Refer to Appendix E for statistics of the data generated for training.)

We find that further finetuning RoBERTa-finetuned-on-BoolQ on synthetic QA data generated from ChatGPT yields the best performance on the figurative split of both Amazon and Yelp (Table 5).

9 How much does the prompting method help with handling figurativeness?

Our experiments show that the process of converting figurative text into literal by prompting GPT-3 may effectively be used for improving question answering performance. We also study the effect of our method on the degree of figurativeness present in the text. The Amazon reviews data from (Niculae and Danescu-Niculescu-Mizil, 2014) comes labeled with figurativeness scores of 1-4, with 3 sets of annotations. Using the average figurativeness scores, we bin the Amazon reviews examples in FigurativeQA into 4 splits, and compute the improvement in QA performance when using our method over the baseline. As evident from Figure 5, the more figurative examples show a higher gain in QA performance.

10 Conclusion and Future Work

We demonstrate that current QA models have reduced accuracy when answering questions from

	Fig.		**Non-fig.**		**Overall**	
	Amazon	**Yelp**	**Amazon**	**Yelp**	**Amazon**	**Yelp**
Zero-Shot						
GPT-3 (zero)	71.9±1.2	60.2±3.2	88.7±0.9	86.0±2.2	80.3±1.1	73.1±2.1
ChatGPT (zero)	91.0±0.7	87.4±2.6	93.0±0.3	88.6±2.4	92.0±0.5	88.0±2.3
Few-Shot						
GPT-3 (few)	85.7±1.8	64.1±3.7	90.2±0.8	88.3±1.9	88.0±1.1	76.2±2.7
ChatGPT (few)	92.6±1.1	80.6±0.7	93.5±0.3	88.7 ± 1.8	93.0±0.7	84.7±1.1
Supervised						
RoBERTa	83.2±1.1	66.8±2.6	92.2±1.4	89.6±1.7	87.7±0.9	78.2±1.6
GPT-3-BoolQ	86.3±2.1	69.2±3.8	88.7±0.9	86.5±1.2	87.5±1.4	77.9±2.2
RoBERTa +synthetic	**95.3±0.5**	**92.3±0.7**	95.8±1.2	90.8±1.6	95.5±0.7	**91.5±0.9**
Simplified Contexts						
GPT-3+ RoBERTa	86.5 ± 1.1	73.4 ± 1.7	92.4 ± 1.1	89.4 ± 1.7	89.5 ± 3.2	81.5 ± 1.2
GPT-3-FLUTE +RoBERTa	88±0.7	69.4±2.1	92.0±0.4	89.5±1.2	90.0 ± 1.4*	79.4 ± 2.3*
ChatGPT+ RoBERTa	88.7±1.6	75.3±3.5	92.2±1.1	89.5±2.1	90.5±1.2	82.4±3.2
ChatGPT+ ChatGPT (few)	89.3±0.8	91.0±0.3	95.7±0.7	91.2±0.2	92.5±0.4	91.1±0.3
ChatGPT+CoT	94.7±0.3	91.6±1.2	**96.4±1.1**	**91.4±0.7**	**95.6±0.9**	**91.5±1.1**

Table 5: QA accuracy on FigurativeQA. (We reran experiments 1000 times with bootstrap resampling. The numbers reported are the mean and std-dev. * denotes results that are not statistically significant compared to the best results, with $p < 0.05$ calculated using the Wilcoxon signed-rank test. The numbers in **bold** are the best results.) GPT-3 finetuned models use da-vinci-002 as the base model.

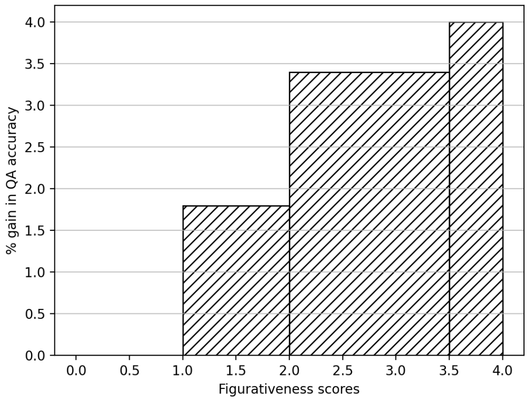

Figure 5: Figurativenss Vs Accuracy for the instances from Amazon reviews

figurative contexts compared to literal ones. This indicates the need for QA models that are robust to figurative language. By manually creating non-figurative versions of these contexts, we observe a significant improvement in performance. To automate this approach, we propose a method of prompting GPT-3 to produce simplified, non-figurative contexts, which yields significant performance gains over the baseline. Chain-of-thought prompting using ChatGPT has the best overall performance on FigurativeQA. We hope that our method and dataset will spur more research into question answering with figurative language.

11 Acknowledgments

This research was supported in part by a research gift from eBay Inc.

Limitations

Our dataset contains the specific domains of Amazon and Yelp reviews, which is English-only, and results and conclusions may not generalize to other domains or languages. The text generated by prompting GPT-3 may sometimes produce text that is not faithful to the original figurative text.

References

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

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.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. *arXiv preprint arXiv:2106.01195*.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. *arXiv preprint arXiv:2009.08942*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding and textual explanations. *arXiv preprint arXiv:2205.12404*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You had me at hello: How phrasing affects memorability. *arXiv preprint arXiv:1203.6360*.

Verna Dankers, Christopher G Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. *arXiv preprint arXiv:2205.15301*.

Lena Dankin, Kfir Bar, and Nachum Dershowitz. Can yes–no question-answering models be useful for few-shot metaphor detection?

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. *arXiv preprint arXiv:2110.00687*.

Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016. How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99.

Valia Kordoni. 2018. Beyond multiword expressions: Processing idioms and metaphors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 15–16.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Bill Yuchen Lin, Ziyi Wu, Yichi Yang, Dong-Ho Lee, and Xiang Ren. 2021. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. *arXiv preprint arXiv:2101.00376*.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. *arXiv preprint arXiv:2204.12632*.

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*.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th annual meeting of the association for computational linguistics*. Association for Computational Linguistics (ACL).

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2008–2018.

Shereen Oraby, Sheideh Homayon, and Marilyn Walker. 2017. Harvesting creative templates for generating stylistically varied restaurant reviews. *arXiv preprint arXiv:1709.05308*.

Natalie Parde and Rodney Nielsen. 2018. Automatically generating questions about novel metaphors in literature. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 264–273.

Ashequl Qadir, Ellen Riloff, and Marilyn A Walker. 2016. *Automatically inferring implicit properties in similes*.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies*, 9(1):1–160.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Zhengzhe Yang and Jinho D Choi. 2019. Friendsqa: Open-domain question answering on tv show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197.

Omnia Zayed, John Philip McCrae, and Paul Buitelaar. 2020. Figure me out: a gold standard dataset for metaphor interpretation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5810–5819.

Justine Zhang, Arthur Spirling, and Cristian Danescu-Niculescu-Mizil. 2017. Asking too much? the rhetorical role of questions in political discourse. *arXiv preprint arXiv:1708.02254*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Jianing Zhou, Hongyu Gong, Srihari Nanniyur, and Suma Bhat. 2021. From solving a problem boldly to cutting the gordian knot: Idiomatic text generation. *arXiv preprint arXiv:2104.06541*.

A Appendix A: Prompts for GPT-3 and ChatGPT for simplifying figurative text

For GPT-3, we use the da-vinci-003 model with temperature set to 0 and max-length set to 100. For ChatGPT, we use gpt-3.5-turbo. In each case, we use a prompt with 5 examples, as shown in Figure 6.

B Appendix B: Prompts for GPT-3 and ChatGPT for QA

For GPT-3, we use the da-vinci-003 model with temperature set to 0 and max-length set to 1. For ChatGPT, we use gpt-3.5-turbo. In each case, we use a prompt with 2 examples, as shown in Figure 7.

C Appendix C: Chain of Thought Prompting ChatGPT for QA

We use the gpt-3.5-turbo model. We used a prompt with 2 examples, as shown in Figure 8.

For the following inputs, if the text contains figurative language, convert it to a literal version. Otherwise, output the same text as the input.

Input: It's inevitable. Their love was built on sand and this is why their marriage has landed on the rocks.
Output: It's inevitable. Their love was unstable and this is why their marriage has failed.

Input: The weather forecast predicted a heatwave this week across most of the country.
Output: The weather forecast predicted a heatwave this week across most of the country.

Input: During the heatwave, the entire house was like a furnace.
Output: During the heatwave, the entire house was uncomfortably hot.

Input: The brisket is nothing to write home about.
Output: There is nothing particularly remarkable about the brisket.

Input: The fries were served cold.
Output: The fries were served cold.

Input: The lamb had a melt in the mouth texture.
Output: The lamb was soft and well-cooked.

Input: The adapter worked like a charm.
Output: The adapter worked perfectly.

Figure 6: Prompt to generate non-figurative versions of the figurative contexts from GPT-3 and ChatGPT.

Answer the following question with a yes or no based on the passage.

Passage: The chocolate cake was sinfully decadent.
Question: Did the cake taste good?
Answer: Yes

Passage: The camera in the phone freezes every few minutes
Question: Does the camera work well?
Answer: No

Figure 7: Prompt to get yes/no answers from GPT-3 and ChatGPT.

Generate a simplified version of the passage and then answer the following question with a yes or no based on the meaning of the passage.

Passage: The chocolate cake was sinfully decadent.
Question: Did the cake taste good?
Simplified Passage: The chocolate cake was rich and delicious.
Answer: Yes

Passage: The camera in the phone freezes every few minutes.
Question: Does the camera work well?
Simplified Passage: The camera stopped working every few minutes.
Answer: No

Figure 8: Chain-of-thought prompting with ChatGPT

D Appendix D: Prompting ChatGPT to generate Synthetic Question Answer pairs from figurative and non-figurative contexts

We use the gpt-3.5-turbo model. We used a prompt with 4 examples, as shown in Figure 9.

E Appendix E: Data Statistics for Synthetic Training Data

Table 6 shows the distribution of synthetic training data generated from ChatGPT for the task of question answering from figurative and non-figurative contexts.

Domain	**Yes**	**No**	**Total**
Yelp	1270	484	1754
Amazon	3320	2102	5422

Table 6: Distribution of yes/no questions generated by prompting ChatGPT

From the following text, generate a yes/no question that requires understanding the literal meaning of the text, and an answer. Refer to the examples provided.

Text: She was a peacock in everything but looks.
Question: Was she pretty?
Answer: No

Text: They seemed to have spared no chilli peppers in the sauce.
Question: Was the sauce hot?
Answer: Yes

Text: The chicken was well-cooked and flavorful.
Question: Did the chicken taste good?
Answer: Yes

Text: The pearls in the studs sparkled like the moon.
Question: Were the earrings dull? the?
Answer: No

Figure 9: Prompt to generate question answer pairs from ChatGPT