LLMs are Turning a Blind Eye to Context: Insights from a Contrastive Dataset for Idiomaticity

Anonymous ACL submission

Abstract

Recent studies have shown that language models achieve high performance in idiomaticity detection tasks. Given the crucial role of context in interpreting these expressions, it is important to evaluate how models use context to make this distinction. To this end, we collect a comprehensive evaluation dataset to see how the model discriminates the use of the same expression in two different contexts. In particular, we produce high-quality instances of idiomatic expressions occurring in their non-dominant literal interpretation, as a way to test whether models can use the context to construct meaning. Our findings highlight the models' tendency to default to figurative interpretations and they do not appear to fully attend to the context. Moreover, the frequency of idioms impacts their ability to accurately discern literal and figurative meanings.

1 Introduction

Idiomatic expressions (IEs) are strange birds that march to a different beat. For example, proficient English speakers understand "spill the beans" not as causing legumes to fall, but as disclosing a secret. Wèinreich (1969) references an estimate suggesting there are 25,000 fixed expressions in English alone, and a similar estimate is quoted for French (Gross, 1982). Notably, this figure is comparable to the number of individual words in the lexicon (Jackendoff, 1997). This suggests that idioms are not mere linguistic curiosities but fundamental components of language.

The term "potentially idiomatic expressions" (PIEs) refers to multi-word sequences, that can be interpreted either non-compositionally (figuratively or idiomatically) or compositionally (literally), depending on the context in which they appear. Accurately identifying the meaning of a PIE within its context is essential for numerous downstream applications, such as machine translation (Dankers et al., 2022; Barreiro et al., 2013; Salton et al., 2014; Fadaee et al., 2018), sentiment analysis (Williams et al., 2015; Liu et al., 2017), and automatic spelling correction (Horbach et al., 2016). Beyond these applications, it is also crucial for grasping the underlying meaning of the text.

Recent studies have shown that language models achieve high performance in idiomaticity detection tasks (Phelps et al., 2024; Zeng and Bhat, 2021). This task involves binary classification, where models must determine whether the usage of a PIE is literal or idiomatic. Given the crucial role of context in interpreting these expressions, it is important to evaluate how models use context to make this distinction. However, since these models are trained on extensive corpora that likely include idiomatic expressions, it is unclear whether they are memorizing these idioms or genuinely comprehending the context to identify idiomaticity.

Existing datasets that include sentences featuring both literal and idiomatic usages often fail to rigorously analyze the effect of context. Literal instances frequently arise from modifications to the expression, whereas figurative instances typically involve minimal variation. This can lead models to rely on reasoning shortcuts and dataset artefacts in the evaluation datasets, rather than completing the task using their idiomaticity knowledge as intended (Boisson et al., 2023).

In the example of "spill the beans", passivization (2) and modification (3) often result in the loss of idiomatic meaning.

(1) Despite my promise to her, I managed to **spill the beans**.

(2) Despite my promise to her, the **beans** were **spilt** by me.

(3) Despite my promise to her, I managed to **spill the** freshly made **beans**.

To address this gap, we propose a novel evaluation set[1] where we strictly control the form of idiomatic expressions. This eliminates the possibility that models rely on grammatical variations for idiomaticity disambiguation. By maintaining consistent expression forms across contexts, our dataset ensures that the challenge lies in understanding contextual nuances, thereby providing a more accurate assessment of a model's idiomatic comprehension.

We focus specifically on idioms, as these expressions serve as pivotal indicators of a model's linguistic understanding. By "idioms," we refer to dominantly figurative expressions. Given the rarity of their literal interpretations, our dataset challenges models to accurately interpret contextual cues to discern between literal and figurative meanings. This approach mirrors the principles of contrastive evaluation, where changes in input require maximal understanding from models. We hypothesize that if models are merely memorizing idioms, their performance will drop when faced with the literal variations of these expressions. Thus, our evaluation set provides a rigorous framework to assess true idiomatic comprehension by language models.

To address these gaps in the field, we curated a novel, comprehensive evaluation dataset [2], containing idioms in both their figurative and literal forms. We focus specifically on idioms, as we believe these expressions serve as pivotal indicators of a model's linguistic understanding. By "idioms", we mean dominantly figurative expressions. Given the rarity of their literal interpretations, our dataset challenges models to interpret contextual cues accurately to discern between literal and figurative meanings. This idea mirrors the principles behind contrastive evaluation, where changes in input require maximal understanding from models. Under the hypothesis that the models are memorizing idioms, we expect to observe a drop in performance when faced with these adversarial examples.

2 Related Works

Contrastive Evaluation Contrastive evaluation often takes the form of minimal pairs evaluation, where a single perturbation such as a change in a word or phrase, is systematically introduced into otherwise identical conditions. This method has been noted for its advantage in identifying specific weaknesses in model understanding and robustness (Linzen et al., 2016; Sennrich, 2017; Robertson, 2019).

Our dataset can be positioned within this category of contrastive evaluation, with a specific focus on idiomaticity. By presenting idioms in both their figurative and literal forms, our dataset forces models to understand and differentiate between subtle contextual cues that determine the meaning. Moreover, by controlling for the dictionary form of expressions, our dataset ensures that the challenge comes from understanding context rather than dealing with variations in form. This approach mirrors the principles behind contrastive evaluation, where minimal changes in input require maximal understanding from models.

Memorisation and Context Transformer models appear to handle IEs mainly by recalling stored expressions and stored knowledge rather than employing an advanced mechanism for processing their meanings (Miletić and Walde, 2024). Li et al. (2022) found that GPT-3's interpretations of the novel compounds matched closely to that of the humans. However, unlike humans who could use the context in which the expression occurred to work out the meaning of nonsensical strings, the models failed due to the memorisation of token distributions in its training data. Thus, it could not leverage its surrounding contextual clues to work the meanings of nonsensical strings. Coil and Shwartz (2023) investigates noun compound interpretation and conceptualization using LLMs. They found that while GPT-3 performs well in interpreting common noun compounds, its performance drops with novel compounds, suggesting a reliance on pre-existing knowledge. Their analysis highlights the balance between reasoning and parroting seen in large models, providing insights into the depth of model comprehension in noun compound tasks.

Cheng and Bhat (2024) find that pretrained LLMS are negatively affected by the context, as performance on Idiomatic Expression Reasoning almost always increases with its removal. The findings of this work are in line with findings in other reasoning-based tasks, such as question-answering retrieval (Liu et al., 2024). Moreover, Sun et al. (2021) find that LLMs tend to rely on contextual cues only when the answer is directly retrievable. Even in tasks like the minimal-pair paradigm ac-

[1]We will make our dataset and code publicly available for camera-ready.

[2]We will make our dataset and code publicly available for camera-ready.

ceptability task, models appear to only exhibit sensitivity to specific contextual features (Sinha et al., 2023).

Taken together, these existing findings underscore the need for a dataset that explores context further for idiomatic processing. They validate this need by highlighting a common limitation: pretrained LLMs frequently struggle with nuanced contextual understanding. To address this gap, we examine models' understanding of idiomaticity through controlled figurative and literal contexts, providing a novel contrastive evaluation framework specifically targeting idiomatic comprehension. We focus on both noun compounds and phrasal expressions. Noun compounds often retain some degree of literal meaning and undergo fewer variations in form, whereas idiomatic expressions require models to accurately interpret more nuanced and often non-literal meanings within diverse contexts.

In this study, we focus on both noun compounds and phrasal expressions. Noun compounds often retain some degree of literal meaning and undergo fewer variations in form, whereas idiomatic expressions require models to accurately interpret more nuanced and often non-literal meanings within diverse contexts. Additionally, we examine models' understanding of idiomaticity through controlled figurative and literal contexts, providing a novel contrastive evaluation framework specifically targeting idiomatic comprehension.

Existing Datasets The task of idiomaticity sense disambiguation (or, idiomaticity detection) involves evaluating whether an expression is used literally or figuratively in a sentence (Liu and Hwa, 2018; Salehi et al., 2014; Senaldi et al., 2016; Gharbieh et al., 2016).

To the best of our knowledge, the biggest dataset for idiomatic sense disambiguation is MAGPIE (Haagsma et al., 2020). Other large datasets targeting various types of IEs have been released: The VNC-Tokens dataset focusing on V+NP expressions (Cook et al., 2008), IDIX on V+NP/PP expressions (Sporleder et al., 2010), SemEval-2013 which has unrestricted expressions (Korkontzelos et al., 2013), AStitchInLanguageModels on noun compounds (Tayyar Madabushi et al., 2021). Visibly, these datasets often only contain expressions of a singular type. As a result, we address this lack of coverage by compiling expressions from both phrasal expressions datasets and noun compound datasets.

In the curation of the MAGPIE dataset, a large amount of deviation of the form of the expression was allowed (Haagsma et al., 2020). However, we believe it is crucial to maintain the same form of expression in both literal and figurative contexts. Idioms are somewhat fixed, with varying degrees of susceptibility to change.

3 Dataset of Adversarial Evaluation in Idioms: DAEVID

A robust evaluation of idiomatic expressions in language models requires a carefully curated dataset that ensures idioms are interpreted correctly in both figurative and literal contexts. It is notably more challenging for dominantly literal expressions to adopt an idiomatic meaning than for idiomatic expressions to be interpreted literally. Therefore, we selected idioms that consistently appear across existing idiomaticity datasets to ensure they predominantly convey figurative meanings.

We compiled a list of phrasal idioms by identifying overlapping expressions from MAGPIE (Haagsma et al., 2020) and SLIDE (Jochim et al., 2018), and a list of noun compound idioms by finding non-compositional expressions common to NCTTI (Garcia et al., 2021) and AStitchInLanguageModels (Tayyar Madabushi et al., 2021). We excluded compositional and partially compositional compounds due to the difficulty in overriding their dominant meanings (e.g., "skin tone," "noble gas"). This process resulted in a total of 783 unique idioms: 680 phrasal expressions and 103 non-compositional noun compounds.

GPT-4 (**?**) was then used to generate sentences, where a given idiom occurs in a sentence that leads to a literal interpretation. We provide the prompting setting we used for sentence generation in A. Initially, we piloted this study using GPT-4o, GPT-4, and GPT-3.5. We found GPT-4 to perform the best at generating sentences where the figurative interpretation is suppressed. Our preference for GPT-4 aligns with the findings of (Phelps et al., 2024), which demonstrate that off-the-shelf GPT-4 possesses relatively stronger idiomaticity knowledge as it performed consistently well across idiomaticity detection tasks compared to other off-the-shelf LLMs. We prompted the model to produce three different sentences, where the form of the idiom must be kept the same. In total, we obtained 2,349 sentences.

To mitigate the potential bias of using GPT-4,

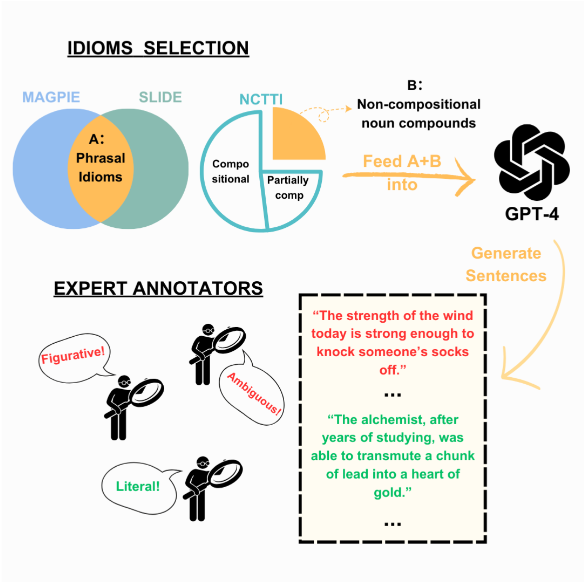

Figure 1: A visual illustration of our dataset curation process. We obtain a list of idioms using existing datasets, which GPT-4 is then prompted to generate sentences where the idiom is used literally. Human annotators then check the sentences.

which may itself struggle with idiomatic nuances, we employed human annotators to verify the generated test sentences. We recruited four experts with at least three years of university-level experience in linguistics, compensated at a rate of £15/hour. Annotators reviewed each sentence to either accept it unconditionally, reject it, or skip it if the figurative meaning of the idiom could not be overridden. In cases of rejection, annotators provided reasons such as ambiguity, figurative interpretation, change of form, or other issues. If an expression was skipped, a second annotator reviewed it to confirm if it should be discarded. Examples of sentences for each category are presented in Table 1.

The figurative counterparts of these sentences were sourced from MAGPIE and AStitchInLanguageModels. We ensure that the same number of variants is matched between the figurative and literal settings. In other words, if we have three sentences containing "all hell broke loose" in literal contexts, we would extract an equal number of sentences containing the idiom from the figurative datasets. In doing so, we curate a balanced and rigorous dataset.

In total, our contrastive evaluation dataset (DAEVID) consists of 2066 sentences, featuring 402 expressions. A summary of the statistics of our dataset is presented in Table 2. Although we only use a subset of the dataset for our analysis into the use of context in idiomaticity processing, we release the rest of the subset as well, so they can serve as good resources for future directions as well as for creating even more challenging datasets.

4 How Well Do LLMs Use Context for Idiomaticity?

Using DAEVID, we evaluated the ability of various language models to differentiate between literal and figurative uses of idioms. Replicating the Idiomaticity Sense Disambiguation (ISD) task, we prompted each model with a sentence and an idiomatic expression, instructing it to return "literal" if the expression has a literal meaning, or "figurative" if it has a figurative meaning.

This evaluation challenges the models to rely on contextual cues to make the correct distinction, assessing their idiomatic comprehension capabilities. By comparing model performance in both figurative and literal contexts, we determine whether LLMs truly understand the nuances of idiomatic expressions or if they are simply relying on memorized patterns. This analysis helps us identify the extent to which language models can interpret idiomatic expressions based on context rather than rote memorization.

4.1 Experimental Setup

Models We present the task of idiomaticity detection or idiomatic sense disambiguation to 10 large language models: GPT-4o [3], GPT-3.5-Turbo (Brown et al., 2020), FLAN-T5 models in the XXL, XL, Large, Small sizes (Chung et al., 2023), Llama-2-7B (`Llama-2-7b-chat-hf`), Llama-3-8B (`Llama-3-8b-instruct`) (Touvron et al., 2023), and Mistral 7B (`Mistral-7B-Instruct-v0.3`) (Jiang et al., 2023). Additionally, we evaluated GPT-4 (**?**), which was used to generate the sentences. The hyperparameters, prompts and computational resources used for the experiments are reported in the Appendix C.

Evaluation To thoroughly evaluate the models' performance, we employed three distinct evaluation settings:

- **Individual Accuracy**: This setting includes two sub-evaluations: (1) **Figurative Accuracy**: We computed the accuracy of each model in correctly identifying the figurative

[3] https://platform.openai.com/docs/models/gpt-4o

Idiom	Definition of the Figurative Meaning	Sentence	Accept	Reject	Reason (if reject)
smoking gun	"a piece of incontrovertible evidence"	The detective found a smoking gun at the crime scene.	N	Y	Ambiguous
guilt trip	"to make someone feel guilty"	After breaking her mother's vase, Sarah's sister put her on a guilt trip for weeks.	N	Y	Doesn't make sense
turn a blind eye	"pretend not to notice"	Despite the obvious safety hazards, the supervisor chose to turn a blind eye.	N	Y	Figurative
down the wire	"a situation whose outcome is not decided until the very last minute"	The electrician was careful not to cut down to the wire while he was working.	N	Y	Form changed
set eyes on	"see"	As soon as she set eyes on the beach, she was overwhelmed by its serene beauty.	N	Y	Skip
blow off steam	"get rid of pent-up energy or emotion"	During the train ride, the kids were excited to see the old locomotive blow off steam.	Y		
get a grip	"begin to deal with or understand"	He struggled to get a grip on the slippery glass jar of pickles.	Y		

Table 1: Examples of expert annotations. Definitions are taken from Ayto (2020).

	Counts	Examples and Remarks
Number of Sentences (Literal)	1033	Carpenters recommend not to sand against the grain as it can damage the wood.
Number of Sentences (Figurative)	1033	e.g., Out of duty she had caved in, but it still went against the grain. (MAGPIE)
Total no. of sentences	2066	-
Number of Unique Idioms	402	-
Total Number of Expressions	402	103 noun compounds + 299 phrasal expressions
Average length of sentences (literal)	15.4 words	-
Average length of sentences (figurative)	28.1 words	-
All annotated sentences	2349	This includes the aforementioned 1033 literal sentences.
Unique expressions	783	-
Ambiguous sentences	165	The panda car is a popular item in the collectible toy market.
Figurative/Idiomatic sentences	465	It was a close call when the hiker almost slipped off the cliff.
Change in Form sentences	32	She reached into the bag to find her glasses. (The idiom is "in the bag".)
Doesn't make sense sentences	162	When the children play at the park, their parents always remind them to play it safe.
Grammatical Error sentences	9	The old locomotive runs out of steam halfway up the mountain.
Can't be literal sentences ("skips")	462	The nurse cared for the critical patients day in, day out without a moment's rest.
Total sentences	1295	-

Table 2: The upper panel of the table shows the properties of the subset for the experiments and analysis we conducted in this paper. The lower panel of the table shows the properties of the remaining of the annotations we collected. We make both parts of the dataset available.

uses of expressions within the figurative subset. (2) **Literal Accuracy**: We assessed the accuracy of the models in correctly identifying the literal uses of expressions within the literal subset. These evaluations measure the models' ability to recognize idiomatic and literal meanings based on context.

- **Consistency in Classification (Consistency Check)**: In this setting, we only rewarded the model for correctly classifying an expression as figurative or literal if it correctly identified all the variations of that expression in the respective subset. Given that each expression has 1 to 3 variations in both literal and figurative contexts, the model needed to classify all these variations correctly to receive a positive score.

  Given that LLMs may be sensitive to the wording of the prompt, and that different wording may result in different performances, we prompted all the models using three different variations.[4] We reported the average and standard deviation over these three variations to ensure a reliable evaluation.

[4]Please see Appendix C for information on the prompts we used.

$$\text{Consistency}_{\text{Type}} = \frac{\sum_{x \in \mathcal{X}} \mathbf{1}\left(\forall i, \text{Prediction}(x_i) = \text{Type}\right)}{\text{Total number of expressions}(\mathcal{X})} \quad (1)$$

where Type can be either "Literal" or "Figurative" and $\mathbf{1}(\cdot)$ is the indicator function.

- **Strict Consistency (Robustness Check)**: This is the most stringent evaluation. The model had to correctly identify all variations of an expression in both figurative and literal contexts to be rewarded. This setting assumes that a truly understanding model should correctly classify an idiom regardless of its context.

$$\text{Strict Consistency} = \frac{\sum_{x \in \mathcal{X}} \mathbf{1}\left(\forall i, \text{Prediction}(x_i) = \text{True Label}(x_i)\right)}{\text{Total number of expressions}(\mathcal{X})} \quad (2)$$

By employing these evaluation settings, we aim to provide a comprehensive assessment of the models' capabilities in understanding and differentiating idiomatic expressions. This approach helps us determine whether the models rely on contextual understanding or memorized patterns to perform the task.

4.2 Results

Table 3 presents the results of model performances on our evaluation set. We can make the following observations based on these results.

Per Class Performance Comparison By comparing the accuracy of the figurative and literal subsets, we observe a noticeable preference towards the figurative class among the models. Eight out of the ten models display better performance in idiomatic contexts. Most models exhibit a significant gap between performances on these subsets, highlighting a struggle with literal contexts despite dealing with the same set of expressions.

For instance, Flan-T5-LARGE shows a significant drop from 99.0% accuracy in figurative contexts to just 1.8% in literal contexts. FLAN-XXL shows the smallest differences between its performance on these two subsets. Flan-T5-SMALL, although showing perfect accuracy on literal examples, fails to understand idiomatic contexts, evidenced by its near-zero accuracy on figurative examples (0.3 ± 0.3).

Additionally, we observe that there can be significant variations in performance depending on the prompt used. LLAMA-2-7B, LLAMA-3-8B, and GPT-4o have the highest standard deviations, indicating the greatest challenges in achieving consistent performance with different prompts, with differences of 38.2, 39.1, and 20.6 points on the figurative subset, respectively.

Consistency Comparison The results from the Consistency Check evaluation reveal the following insights. Overall, the general trend aligns with our previous observations: models show a preference for figurative interpretations when encountering an idiom, as there is a higher proportion of idioms that the models can consistently predict to be figurative across all contextual sentences than in the literal setting. As expected, all models achieve lower scores on both subsets when evaluated based on consistency, where the model must correctly classify all variations of the same expression in each sense to be rewarded. We observe the largest drop for GPT-4o when scores are evaluated using this metric. This decline indicates that GPT-4o's performance stems from its familiarity with a broad range of idioms (evidenced by an accuracy of 57.2 on the figurative class). However, the model lacks a deep understanding of these idioms, making it susceptible to susceptible to variations. This is illustrated by a Consistency score of 32.7, showing that the model can only accurately interpret a subset of idioms consistently across different texts. Flan-T5-XXL remains the model with the least performance difference across the two subsets, indicating a more balanced understanding of both figurative and literal contexts.

Robustness Check The robustness check, as previously defined, requires models to correctly classify all the figurative and literal uses of an expression to be rewarded. The results from this evaluation are striking: only three models—GPT-3.5, FLAN-XXL, and Mistral-7B—achieve an accuracy above 10%, with 44.5%, 25.4%, and 12.4% respectively. This indicates that while state-of-the-art models may show high performance on existing idiomaticity benchmarks, they perform very poorly when a more systematic approach is used to evaluate their understanding. Even GPT-4, which served as the annotator model, can consistently classify only 63.5% of the expressions correctly in both literal and figurative contexts. This highlights a significant gap in the current models' ability to truly understand idiomatic expressions, suggesting considerable room for improvement in idiomaticity detection and achieving true meaning comprehension.

5 Impact of Expression Frequency on Model Performance

In this section, we analyze how the frequency of idiomatic expressions in the pretraining data influences model performance on our evaluation set. Given the lack of access to specific pretraining datasets, we utilize the English Web Corpus (enTenTen) (Jakubíček et al., 2013) to approximate the frequency distributions of these idioms. The enTenTen corpus, with its extensive scale of 52 billion words and diverse genres, provides a robust basis for our frequency-based analyses.

Our research explores two main hypotheses: First, higher frequencies of idiomatic expressions in the pretraining data may improve model performance primarily on the figurative subset, as the expressions we used commonly appear in their figurative forms. Second, frequent exposure to idiomatic expressions could enhance performance across both figurative and literal contexts, reflecting a more comprehensive understanding of these expressions. By investigating these hypotheses, we aim to determine how exposure frequency impacts

Model	Per Class Performance		Consistency		Strict Consistency
	Figurative	Literal	Figurative	Literal	Overall
GPT-4o	57.2 ± 20.6	29.8 ± 29.4	32.7 ± 24.9	12.4 ± 16.3	4.9 ± 5.0
GPT-3.5-Turbo	**88.5 ± 6.5**	**60.3 ± 18.9**	79.1 ± 10.3	43.4 ± 21.0	30.3 ± 12.4
Flan-T5-XXL	79.3 ± 9.2	73.4 ± 18.0	**63.9 ± 13.7**	**58.8 ± 23.2**	**32.9 ± 6.8**
Flan-T5-XL	95.5 ± 3.7	23.8 ± 19.4	91.1 ± 7.0	13.0 ± 11.2	10.0 ± 8.9
Flan-T5-Large	99.0 ± 1.5	1.8 ± 2.5	97.7 ± 3.4	0.6 ± 0.8	0.6 ± 0.8
Flan-T5-Small	0.3 ± 0.3	100.0 ± 0.0	0.0 ± 0	100.0 ± 0	0.0 ± 0.0
Llama-3	45.1 ± 39.1	72.1 ± 24.5	25.7 ± 22.5	56.6 ± 37.7	9.5 ± 8.2
Llama-2	59.7 ± 38.2	38.0 ± 34.0	43.5 ± 49.5	19.9 ± 19.9	2.8 ± 3.0
Mistral-7B	97.4 ± 1.8	28.9 ± 17.8	93.9 ± 3.9	13.3 ± 11.1	12.2 ± 9.8
GPT-4	88.7 ± 0.6	86.9 ± 3.6	78.4 ± 0.9	76.9 ± 5.6	58.2 ± 4.9

Table 3: Mean scores ± 1 std (over 3 different sets of prompts). For per-class performance scores, we report accuracy scores, for Consistency and Strict Consistency we report the measures calculated defined in §4.1 **Bold** values denote the best performance on each metric for each model. We separate GPT-4 results from the rest, as this is the model where evaluation sentences were obtained.

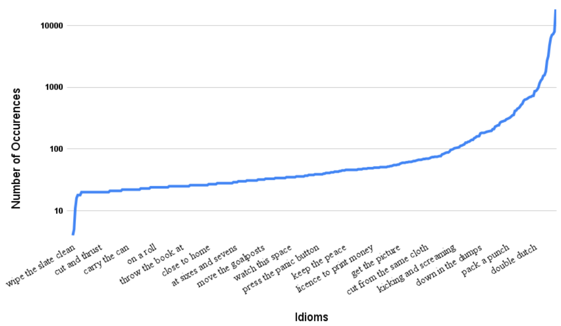

Figure 2: Log frequency distribution of idioms in DAE-VID. Only a selection of idioms is displayed for readability.

the models' ability to generalize idiomatic understanding beyond memorization.

5.1 Frequency Estimation

As shown in Figure 2, some idiomatic expressions are low-frequency occurring items in language, with a small number being high-frequency items. Due to the non-linear nature of this distribution, we categorized the idioms into four bins based on their frequency. We ranked the expressions based on their frequency and focused on the two extreme bins: the lowest frequency bin (representing the rarest expressions) and the highest frequency bin (representing the most common expressions). We present the results for the other two bins in Appendix D.

5.2 Results

We observe a clear trend in Table 4: all models achieve higher performance on both figurative and literal evaluations for idioms with higher occurrences. This finding aligns with our second hypothesis, indicating that frequent exposure to idiomatic expressions during pretraining enhances the models' overall understanding. The models exhibit a more nuanced comprehension of idioms that are encountered more often, suggesting that increased frequency in pretraining data significantly improves performance across various contexts, both figurative and literal. This highlights the importance of frequent exposure to idiomatic expressions for robust language model training.

A closer inspection of the top 3 best performing models on the Robustness Check reveals that the idioms occurring with the highest frequency bins were most accurately understood, whether they appeared in literal or figurative contexts, in all of the sentences in the dataset. The mid-frequency group followed, with models comprehending a quarter of these idioms entirely. Even idioms in the low-frequency category were understood to a significant extent, at 23.5%. Notably, none of the idioms in the rare group was completely understood by the models. In Appendix 8, we provide the list of idioms that have demonstrated this robustness. noun compounds constitute a significant proportion of these expressions. This is likely because noun compounds typically undergo fewer variations in form compared to phrasal expressions (Pafel, 2017).

6 Discussion

Our findings indicate that when models encounter contexts containing idiomatic expressions, they struggle to effectively utilize contextual informa-

Model	Rare		High	
	Figurative	Literal	Figurative	Literal
GPT-4o	67.3 ± 23.6	45.7 ± 43.6	**84.9 ± 9.5**	46.4 ± 43.5
GPT-3.5-Turbo	78.5 ± 6.6	62.1 ± 19.2	83.6 ± 2.6	**80.4 ± 10.6**
Flan-T5-XXL	**81.2 ± 2.4**	**79.3 ± 8.1**	83.0 ± 5.5	82.3 ± 6.8
Flan-T5-XL	72.9 ± 4.4	38.1 ± 21.8	78.4 ± 9.1	54.7 ± 35.1
Flan-T5-Large	67.9 ± 1.1	10.3 ± 8.9	65.1 ± 1.5	6.5 ± 11.2
Flan-T5-Small	0.0 ± 0.0	66.7 ± 0.0	0.0 ± 0.0	66.7 ± 0.0
Llama-3	41.6 ± 38.9	61.0 ± 15.4	47.7 ± 42.0	70.5 ± 9.0
Llama-2	46.7 ± 18.5	32.7 ± 28.3	49.0 ± 24.0	32.3 ± 28.5
Mistral-7B	76.1 ± 6.2	50.8 ± 22.6	79.0 ± 7.4	59.7 ± 24.2
GPT-4	91.5 ± 4.5	91.8 ± 3.8	95.4 ± 2.9	93.8 ± 2.2

Table 4: Mean F1 scores $\pm$ 1 std (over 3 prompts). **Bold** values denote highest performances.

tion. Consequently, they often classify these contexts as figurative, even when humans would interpret them as literal. Overall, our results show a higher F1 score for classifying figurative contexts compared to literal ones. The significant drop in performance on literal examples supports our hypothesis that models may rely more on memorization than a nuanced understanding of idiomatic expressions, particularly when faced with language that deviates from common, dominantly idiomatic usages. We believe that this is due to pretraining datasets containing potential lists, explanations, and definitions of idioms in addition to their usages in context.

However, the presence of this information in the pretraining dataset does not mean the model necessarily would do well on figurative understanding either. As demonstrated by Flan-T5-Small and Llama-3, they do appear to have sufficiently learned idioms that can be figurative. The low performance on these models is in line with evaluations on three datasets, as carried out by (Phelps et al., 2024).

The Consistency and Robustness Checks provided additional layers of analysis for our investigation. Given the task's binary nature, models could potentially guess labels randomly. In the broader Consistency Check, we anticipated that a model demonstrating an understanding of an idiom's sense would correctly classify all instances and contexts where the idiom appears in that sense. For example, if a model comprehends the figurative meaning of "spill the beans," it should classify all occurrences of this idiom figuratively. In the narrower Robustness Check, true understanding would be evidenced by the model correctly identifying whether an idiom is literal or figurative across all contexts in which it appears. Our findings indicate a limited genuine understanding of idioms by the models, consistent with our hypothesis that insufficient leveraging of context impedes meaningful comprehension.

Additionally, we observe that idiom frequency correlates with higher performance in both literal and figurative contexts. This suggests that increased exposure to an idiom improves the model's understanding in different contexts. Notably, the idioms in our dataset are mostly figurative, with literal occurrences being rare. Therefore, for highly frequent idioms, models may encounter some literal examples alongside numerous figurative ones.

7 Conclusion

In this work, we have demonstrated that LLMs do not effectively utilize the context in which expressions occur to form judgments on idiomaticity. Instead, the frequency of the expressions in language use is correlated with performance improvements in both literal and figurative senses. These findings are based on our contrastive evaluation dataset, specifically curated for a fine-grained and thorough evaluation of the role of context in idiomaticity detection. As future work, this study motivates further investigation into larger-scale frequency analyses using more extensive datasets to deepen our understanding of how frequency and context influence idiomaticity detection in LLMs.

8 Limitations

One of the limitations of our work is that some idiomatic expressions are noticeably more reliant on the context than others. This means that there were cases, where we could not provide a literal counterpart to the figurative interpretation. For example, the expression "set eyes on" has such a dominant meaning of "to see", that the annotators believed to be impossible to override. In these cases, we would discard the expression. As a result, our dataset only contains a selected sample of idioms, and we acknowledge that this idea of contrastive evaluation cannot necessarily be applied to all idioms in a language.

Another of the limitations of our work is that we only consider English idioms. We would like to have extended this work to other languages, however, due to the scarcity of idiomaticity datasets, it is hard to do so within our budget. Moreover, the idea of making idioms literal might not be translatable to other languages, where the expression takes rigid and fixed forms.

9 Ethical Considerations

We adhere to ethical practices of data collection. All participants were required to sign a consent form and informed that they could withdraw from participation at any time without facing any consequences. Our collection procedures and processes are monitored and reviewed by a University-wide ethics committee. Th committee members are unrelated and detached from this work.

References

John Ayto. 2020. *The Oxford Dictionary of Idioms*. Oxford University Press.

Anabela Barreiro, Johanna Monti, Brigitte Orliac, and Fernando Batista. 2013. When multiwords go bad in machine translation. In *Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technologies*, Nice, France.

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6581–6590, Singapore. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Kellen Cheng and Suma Bhat. 2024. No context needed: Contextual quandary in idiomatic reasoning with pretrained language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4863–4880, Mexico City, Mexico. Association for Computational Linguistics.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Albert Coil and Vered Shwartz. 2023. From chocolate bunny to chocolate crocodile: Do language models understand noun compounds? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2698–2710, Toronto, Canada. Association for Computational Linguistics.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 19–22. Citeseer.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021. Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3551–3564, Online. Association for Computational Linguistics.

Waseem Gharbieh, Virendra Bhavsar, and Paul Cook. 2016. A word embedding approach to identifying verb-noun idiomatic combinations. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 112–118, Berlin, Germany. Association for Computational Linguistics.

Maurice Gross. 1982. Une classification des phrases « figées » du français. *Revue québécoise de linguistique*, 11(2):151–185.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Andrea Horbach, Andrea Hensler, Sabine Krome, Jakob Prange, Werner Scholze-Stubenrecht, Diana Steffen, Stefan Thater, Christian Wellner, and Manfred Pinkal. 2016. A corpus of literal and idiomatic uses of German infinitive-verb compounds. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 836–841, Portorož, Slovenia. European Language Resources Association (ELRA).

R. Jackendoff. 1997. *The Architecture of the Language Faculty*. Linguistic inquiry monographs. MIT Press.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The tenten corpus family. In *7th international corpus linguistics conference CL*, pages 125–127. Valladolid.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. SLIDE—a sentiment lexicon of common idioms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

Siyan Li, Riley Carlson, and Christopher Potts. 2022. Systematicity in GPT-3's interpretation of novel English noun compounds. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 717–728, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Changsheng Liu and Rebecca Hwa. 2018. Heuristically informed unsupervised idiom usage recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1723–1731, Brussels, Belgium. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 11:157–173.

Pengfei Liu, Kaiyu Qian, Xipeng Qiu, and Xuanjing Huang. 2017. Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213, Copenhagen, Denmark. Association for Computational Linguistics.

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 11:593–612.

Jürgen Pafel. 2017. Phrasal compounds and the morphology-syntax relation. *Further investigations into the nature of phrasal compounding*, pages 233–259.

Dylan Phelps, Thomas M. R. Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

Frankie Robertson. 2019. A contrastive evaluation of word sense disambiguation systems for Finnish. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 42–54, Tartu, Estonia. Association for Computational Linguistics.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Detecting non-compositional MWE components using Wiktionary. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1792–1797, Doha, Qatar. Association for Computational Linguistics.

Giancarlo Salton, Robert Ross, and John Kelleher. 2014. An empirical study of the impact of idioms on phrase based statistical machine translation of English to Brazilian-Portuguese. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 36–41, Gothenburg, Sweden. Association for Computational Linguistics.

Marco Silvio Giuseppe Senaldi, Gianluca E. Lebani, and Alessandro Lenci. 2016. Lexical variability and compositionality: Investigating idiomaticity with distributional semantic models. In *Proceedings of the*

12th Workshop on Multiword Expressions, pages 21–31, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams. 2023. Language model acceptability judgements are not always robust to context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6043–6063, Toronto, Canada. Association for Computational Linguistics.

Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. 2010. Idioms in context: The IDIX corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do long-range language models actually use long-range context? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 807–822, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

Uriel Wèinreich. 1969. *Problems in the Analysis of Idioms*, pages 23–82. University of California Press, Berkeley.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic Expression Identification using Semantic Compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

A Sentence Generation Prompt

The prompt we used for generating the sentences is shown here. For other configurations that are not mentioned, we used the default setting.
Model: GPT-4
"role": "system", "content": "You are an expert of English"
"role": "user", "content": "Generate three sentences using the expression: 'idiom', where the expression has a literal meaning. Each sentence must contain the expression unchanged. Format these sentences as a Python list. Don't say anything that are not the sentences."
temperature = 0.8

B Participant Briefing

Upon signing up for participation, each annotator received a 30mins training session where they were shown examples, including 1. We omit the Participation Information sheet since this contains information that could break anonymity.

C Implementation Details

We ran the FLAN-T5 models, Llama-2, Llama3 and Mistral on a NVIDIA H100 GPU. Each model was evaluated with three different prompts. All of the results we report are the average across the three prompt settings. We use OpenAI's API for interactions with the GPT models, and HuggingFace for the rest of the models.

Model: **GPT-4, GPT-4o, GPT-3.5-Turbo**
Prompt 1: "Is the expression 'idiom' used figuratively or literally in the sentence: 'sentence'. Answer 'i' for figurative, 'l' for literal."
Prompt 2: "In the sentence 'sentence', is the

expression 'idiom' being used figuratively or literally? Respond with 'i' for figurative and 'l' for literal."
Prompt 3: "How is the expression 'idiom' used in this context: 'sentence'. Output 'i' if the expression holds figurative meaning, output 'l' if the expression holds literal meaning."

Models: **Flan-T5-XXL, Flan-T5-XL, Flan-T5-Large, Flan-T5-Small**
Prompt 1: "Is the meaning of expression idiomatic or literal? If used idiomatically, answer 'i', if literally, answer 'l'." "Expression: idiom" "Sentence: sentence".
Prompt 2: "In the sentence 'sentence', is the expression 'idiom' being used figuratively or literally? Respond with 'i' for figurative and 'l' for literal."
Prompt 3: "How is the expression 'idiom' used in this context: 'sentence'. Output 'i' if the expression holds figurative meaning, output 'l' if the expression holds literal meaning."

Models: **meta-llama/Meta-Llama-3-8B-Instruct, meta-llama/Llama-2-7B-chat-hf**
Prompt 1: "role": "system", "content": "You are a language expert."
"role": "user", "content": "expression: 'idiom'sentence: 'sentence' QUESTION: Is the expression figurative or literal? Generate the letter 'i' if the idiom is used figuratively, or generate 'l' if the expression is used literally. Only generate the letter."
Prompt 2: "role": "system", "content": "You are an assistant." "role": "user", "content": "expression: 'idiom'sentence: 'sentence' QUESTION: Given a contextual sentence and an expression, tell me if the expression is used figuratively or literally. Either generate the letter 'i' if figurative or generate the letter 'l' if literal."
Prompt 3: "role": "system", "content": "You are a native speaker of English." "role": "user", "content": f"expression: 'idiom'sentence: 'sentence' QUESTION: Does the expression hold a figurative or literal meaning in the contextual sentence? Generate a letter 'i' for figurative meaning, or 'l' for literal meaning."

Model: **mistralai/Mistral-7B-Instruct-v0.3**
Prompt 1: "[INST] You are a language expert who can only generate one letter. Your task is to interpret the sentence, and generate a letter "i" if the idiom is used figuratively, or generate "l" if the expression is used literally. expression: 'idiom' sentence: 'sentence' Generate a Python list containing the letter.[/INST]"
Prompt 2: "[INST] You are an assistant, who can only generate one letter. Given a contextual sentence and an expression, tell me if the expression is used figurative or literally. Either generate "i" if figurative, or generate "l" if literal. expression: 'idiom' sentence; 'sentence' Generate a Python list containing the letter.[/INST]"
Prompt 3: "[INST] You are a native speaker of English, who can only generate one letter. Does the expression hold a figurative or literal meaning in the following contextual sentence? Generate a letter "i" for figurative meaning, or "l" for literal meaning. expression: 'idiom' sentence; 'sentence' Generate a Python list containing the letter.[/INST]"

D F1 Scores across Each Class

Table 5 represents the f1 scores, each model obtained on each class (figurative and literal).

Model	Figurative F1	Literal F1
GPT-4o	**68.6 ± 14.0**	39.2 ± 35.1
GPT-3.5-Turbo	**78.4 ± 3.6**	69.8 ± 11.8
Flan-T5-XXL	**77.2 ± 1.4**	74.9 ± 8.4
Flan-T5-XL	**70.5 ± 3.6**	33.9 ± 26.9
Flan-T5-Large	**66.6 ± 0.1**	3.5 ± 4.7
Flan-T5-Small	0.5 ± 0.6	**66.7 ± 0.1**
Llama-3	22.5 ± 30.2	**65.6 ± 2.9**
LlAMA-2	**50.0 ± 18.5**	34.8 ± 30.1
Mistral-7B	**72.8 ± 4.1**	42.0 ± 21.6
GPT-4	**88.5 ± 1.7**	88.1 ± 1.8

Table 5: Mean F1 score ± 1 std (over 3 runs). **Bold** values denote the best performance across each class for each model.

E Additional Results for Frequency Analysis

We present supplementary results we obtained. Table 6 shows the F1 scores across each frequency bin, for the figurative and literal subsets. Table 7 shows the Consistency scores for each frequency group. Table 8 shows the expressions on which the top 3 models achieved the highest robustness score.

Model	Rare		Low		Moderate		High	
	Figurative	Literal	Figurative	Literal	Figurative	Literal	Figurative	Literal
GPT-4o	67.3 ± 23.6	45.7 ± 43.6	67.6 ± 15.2	39.2 ± 26.2	69.9 ± 9.7	37.2 ± 33.9	84.9 ± 9.5	46.4 ± 43.5
GPT-3.5-Turbo	78.5 ± 6.6	62.1 ± 19.2	78.4 ± 3.5	69.5 ± 21.1	77.3 ± 4.2	70.7 ± 12.6	83.6 ± 2.6	80.4 ± 10.6
Flan-T5-XXL	81.2 ± 2.4	79.3 ± 8.1	76.6 ± 1.3	74.1 ± 4.8	77.5 ± 2.1	76.5 ± 6.6	83.0 ± 5.5	82.3 ± 6.8
Flan-T5-XL	72.9 ± 4.4	38.1 ± 21.8	70.2 ± 3.4	32.7 ± 30.1	70.7 ± 3.5	36.3 ± 30.4	78.4 ± 9.1	54.7 ± 35.1
Flan-T5-Large	67.9 ± 1.1	10.3 ± 8.9	66.7 ± 0.1	3.2 ± 4.1	66.3 ± 1.0	3.6 ± 3.5	65.1 ± 1.5	6.5 ± 11.2
Flan-T5-Small	0.0 ± 0.0	66.7 ± 0.0	0.5 ± 0.5	66.7 ± 34.7	0.8 ± 1.5	66.8 ± 0.2	0.0 ± 0.0	66.7 ± 0.0
Llama-3	41.6 ± 38.9	61.0 ± 15.4	43.1 ± 37.3	62.5 ± 0.1	42.2 ± 36.5	64.2 ± 3.0	47.7 ± 42.0	70.5 ± 9.0
Llama-2	46.7 ± 18.5	32.7 ± 28.3	49.8 ± 18.5	35.0 ± 11.5	51.6 ± 18.1	34.3 ± 29.9	49.0 ± 24.0	32.3 ± 28.5
Mistral-7B	76.1 ± 6.2	50.8 ± 22.6	72.5 ± 3.9	41.3 ± 1.4	73.5 ± 4.7	42.8 ± 23.7	79.0 ± 7.4	59.7 ± 24.2
GPT-4	91.5 ± 4.5	91.8 ± 3.8	88.4 ± 1.3	88.0 ± 8.9	87.0 ± 3.2	87.1 ± 3.7	95.4 ± 2.9	93.8 ± 2.2

Table 6: Mean F1 scores $\pm$ 1 std (over 3 prompts)..

Model	Rare		Low		Moderate		High	
	Figurative	Literal	Figurative	Literal	Figurative	Literal	Figurative	Literal
GPT-4o	33.3 ± 28.9	25 ± 28.7	32.3 ± 26.4	12.3 ± 26.4	32.2 ± 19.2	9.44 ± 19.2	48.1 ± 25.7	18.5± 25.7
GPT-3.5-Turbo	91.7 ± 14.4	16.7 ± 14.4	79.4 ± 10.3	42.1 ± 10.3	73.9 ± 10.2	48.9 ± 10.2	81.5 ± 17.0	63.0 ± 17.0
Flan-T5-XXL	58.3 ± 14.4	41.7 ± 14.4	62.3 ± 15.2	59.0± 15.2	68.3 ± 8.33	57.2 ± 8.33	74.1 ± 6.42	51.9 ± 6.42
Flan-T5-XL	100.0 ± 0.0	8.33 ± 0.0	90.9 ± 7.06	12.4 ± 7.06	89.4 ± 9.18	16.1 ± 9.18	100.0 ± 0.0	18.5 ± 0.0
Flan-T5-Large	100.0 ± 0.0	0.0 ± 0.0	98.2 ± 2.55	0.418 ± 2.55	95.6 ± 7.70	1.11 ± 7.70	92.6 ± 6.42	3.70 ± 6.41
Flan-T5-Small	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0	100.0 ± 0.0
Llama-3	16.7 ± 28.9	50.0 ± 28.9	26.1 ± 23.04	56.3 ± 23.0	22.8 ± 19.7	57.8 ± 19.7	25.94 ± 23.1	63.0 ± 23.1
Llama-2	33.3 ± 57.7	0.0 ± 57.7	43.3 ± 49.7	21.1 ± 49.7	44.4 ± 48.3	15.56 ± 48.3	51.9 ± 50.1	11.1 ± 50.1
Mistral-7B	100 ± 0.0	8.33± 0.0	93.1 ± 4.42	13.1 ± 4.42	95.6 ± 2.55	15.0 ± 2.55	100.0 ± 0.0	11.1 ± 0.0
GPT-4_	33.3 ± 28.9	25.0 ± 28.9	32.3 ± 26.4	12.3 ± 26.4	32.2 ± 19.2	9.44± 19.2	48.1 ± 25.7	18.5 ± 25.7

Table 7: Mean Consistency scores $\pm$ 1 std (over 3 prompts).

Model	Expressions	Total
Flan-T5-XXL	on the shelf, on a shoestring, break the ice, poison pill, turn the tables, cut and dried, pass the buck, closed book, acid test, out of the loop, hit the jackpot, pick someones brain, on the ropes, rock bottom, full of beans, melting pot, turn the screw, the bees knees, get under someones skin, in the raw, muddy the waters, rocket science, carrot and stick, in a nutshell, cut both ways, on the ball, hold the line, run out of steam, nest egg, raise the roof, get a rise out of, on the same page, push the envelope, add fuel to the fire, down the tubes, fly off the handle, in the bag, joined at the hip, eat humble pie, fire in the belly, on the horn, busy bee, big fish, heart of gold, night owl, cut the mustard, rat run, sitting duck, on the rocks, cook the books, fill someones shoes, drop the ball, swings and roundabouts, glass ceiling	54
GPT-3.5-Turbo	on a shoestring, blue blood, in the doghouse, cut and dried, dig up dirt, on the ropes, get off the ground, run a mile, go to the wall, circle the wagons, spit it out, to the bone, put the boot in, on the cards, take a dive, in a nutshell, hold the line, raise the roof, under the sun, on the same page, low profile, joined at the hip, carry the can, big fish, touch and go, draw a line in the sand, apples and oranges, cut the mustard, toe the line, rat run, on the rocks, hit the bottle, brass ring, fill someones shoes, ring a bell, grind to a halt, in the hole, over the top, pour cold water on	39
Mistral-7B	hold the line, toe the line, goose egg, to the bone, ring a bell, big fish, over the top	7

Table 8: Top 3 performing models and the expressions, which they have successfully understood in all senses (figurative and literal), across all sentences in the dataset.