Rolling the DICE on Idiomaticity: How LLMs Fail to Grasp Context

Maggie Mi[1] **Aline Villavicencio**[1,2] **Nafise Sadat Moosavi**[1]
[1]University of Sheffield [2]University of Exeter
zmi1@sheffield.ac.uk a.villavicencio@exeter.ac.uk
n.s.moosavi@sheffield.ac.uk

Abstract

Human processing of idioms heavily depends on interpreting the surrounding context in which they appear. While large language models (LLMs) have achieved impressive performance on idiomaticity detection benchmarks, this success may be driven by reasoning shortcuts present in existing datasets. To address this, we introduce a novel, controlled contrastive dataset (DICE) specifically designed to assess whether LLMs can effectively leverage context to disambiguate idiomatic meanings. Furthermore, we investigate the influence of collocational frequency and sentence probability—proxies for human processing known to affect idiom resolution—on model performance. Our results show that LLMs frequently fail to resolve idiomaticity when it depends on contextual understanding, and they perform better on sentences deemed more likely by the model. Additionally, idiom frequency influences performance but does not guarantee accurate interpretation. Our findings emphasize the limitations of current models in grasping contextual meaning and highlight the need for more context-sensitive evaluation.

https://github.com/mi-m1/dice

1 Introduction

Idiomatic expressions (IEs) are strange birds whose meaning may not be straightforwardly related to the meaning of the component words in isolation. For example, proficient English speakers understand "*kick the bucket*" not as "*striking a metal container*", but as "*to die*". Estimates suggest that there are 25,000 fixed expressions in English alone (Wèinreich, 1969), and a similar estimate is quoted for French (Gross, 1982). Notably, this figure is comparable to the order of magnitude of individual words in the lexicon (Jackendoff, 1997). This suggests that idioms are not mere linguistic curiosities but fundamental components of language.

Additionally, some of these expressions are ambiguous, "potentially idiomatic expressions" (PIEs) 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 often depends on its context and 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, understanding idiomatic expressions in context is also crucial to grasp the underlying meaning of the text.

Existing datasets that feature expressions with both literal and idiomatic usages often fail to rigorously examine the role of context in disambiguating meaning. This shortcoming arises because the literal meanings in such datasets typically result from syntactic modifications or semantic shifts, which inherently disrupt the idiomaticity of the expression. In fact, these changes in form have been deliberately used as signals to disentangle and differentiate literal from figurative meaning (Fazly et al., 2009). For example, passivisation (2) or modification to the expression (3) often strip the expression of its idiomatic meaning, as shown in comparison to the idiomatic usage in (1) (Jackendoff, 1997; Gibbs and Gonzales, 1985; Langlotz, 2006; Kyriacou et al., 2020)[1].

(1) He **kicked the bucket.**

(2) The **bucket** was **kicked** by him.

(3) He **kicked** the tin **bucket** angrily.

Consequently, this allows models to exploit surface-level differences, such as changes in gram-

[1]"*Kick the bucket*" is a classic example of a rigid idiom, but expressions vary in flexibility; some can better tolerate changes without losing their idiomatic meaning than others.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7314–7332
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

matical structure or insertions, as shortcuts for determining literalness, rather than relying on their understanding of idiomaticity in context as the task intends. This undermines the evaluation's effectiveness by encouraging models to use superficial cues instead of deeper contextual comprehension.

In this paper, we address this gap by introducing *DICE* (Dataset for Idiomatic Contrastive Evaluation), a benchmark designed to evaluate the ability of LLMs to interpret idiomatic expressions in context. We focus on idioms, as these figurative expressions may serve as key indicators of a model's linguistic understanding. Given the possibility of a predominant sense, DICE presents idiomatic expressions in both literal and figurative contexts and challenges models to rely on context for the correct interpretation. This contrastive evaluation forces models to distinguish between meanings based solely on context, preventing them from relying on memorized idioms. We hypothesize that if models do not depend on memorization, they will perform equally well in both senses of a PIE.

Moreover, in human language processing, factors such as frequency (Swinney and Cutler, 1979) and familiarity (Nippold and Rudzinski, 1993), alongside context (Estill and Kemper, 1982; Gibbs Jr et al., 1989; Gibbs and Nayak, 1989), are known to influence idiom comprehension and processing speed (Tabossi et al., 2009). We investigate whether similar effects hold for language models. Specifically, beyond contextual understanding, we examine the influence of a language-intrinsic feature (expression frequency) and a model-intrinsic feature (sentence likelihood) on model performance.

Contributions (1) We present DICE, a comprehensive and robust evaluation dataset, containing PIEs that occur in the same grammatical form across both figurative and literal contexts. (2) Through fine-grained evaluations, we find that models struggle to use context for idiomaticity processing. (3) Based on frequency estimates, we find that frequency is not a "free lunch": whilst highly frequent idioms may be more likely to be disambiguated correctly, there is a trade-off in model performance between literal and figurative settings. (4) For models that demonstrate some capacity for idiomatic understanding, we observe a strong relationship between the likelihood of the contextual sentence and idiomaticity detection performance.

2 Related Works

Idiomaticity Detection. The task of idiomaticity sense disambiguation (ISD), 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). This task is typically framed as binary classification. While large language models have achieved strong performance on existing ISD benchmarks (Phelps et al., 2024; Zeng and Bhat, 2021), it remains unclear whether this reflects true contextual understanding or reliance on memorized surface forms (Garcia et al., 2021b). Given the crucial role of context in resolving idiomaticity, it is essential to evaluate whether models are genuinely interpreting surrounding text or simply exploiting distributional cues.

Existing Datasets. The biggest dataset for idiomatic sense disambiguation is MAGPIE (Haagsma et al., 2020). MAGPIE contains a total of 56,622 PIE instances, across 1,756 idioms. However, a large amount of deviation of the form of an expression was allowed when MAGPIE was curated. As a result, most of the literal uses of PIEs involve modifications to the form of the expression (similar to the example in § 1). Other large datasets targeting various types of IEs have been released: The VNC-Tokens dataset focusing on verb-noun combinations (Cook et al., 2008), IDIX on verb-noun phrase or verb-prepositional phrase expressions (Sporleder et al., 2010). SemEval-2013 has unrestricted expressions (Korkontzelos et al., 2013), AStitchInLanguageModels contains noun compounds (Tayyar Madabushi et al., 2021) and more recently, IdioTS contains a mixture of expressions changed and unchanged to support the literal meaning (De Luca Fornaciari et al., 2024).

To address these problems: (1) we propose a novel evaluation set (DICE), which strictly controls the form of idiomatic expressions. This design eliminates the possibility that models rely on grammatical variations for idiomaticity disambiguation. By maintaining consistent forms across literal and figurative contexts, DICE ensures that the challenge lies in understanding contextual nuances, thereby providing a more accurate assessment of a model's idiomatic comprehension. (2) Existing datasets typically focus on a single type of expression. We address this limitation by including both phrasal idioms and noun compounds to provide broader coverage of idiomatic phenomena.

Contrastive Evaluation. This evaluation paradigm involves comparing model outputs on carefully constructed input pairs that differ in minimal, controlled ways, typically along a dimension relevant to a specific linguistic phenomenon (Prasad et al., 2019; Garcia et al., 2021a). Such comparisons have been found to be particularly effective at isolating specific linguistic competencies or revealing systematic weaknesses in generalization and robustness (Linzen et al., 2016; Sennrich, 2017; Robertson, 2019). Our work adopts this approach to assess contextual comprehension in idiomaticity detection. Specifically, DICE presents potentially idiomatic expressions in both literal and figurative uses while holding surface form constant. This setup requires models to rely entirely on surrounding context to disambiguate meaning, eliminating the possibility of exploiting shallow lexical or syntactic cues.

Memorization and Context. Pretrained language models tend to handle IEs mainly through memorization of stored expressions and token distributions rather than reasoning about the meaning in context. This reliance on memorization becomes evident when models face novel noun compounds. For instance, Li et al. (2022) found that while GPT-3's interpretations of novel compounds closely matched human responses, its performance faltered when tasked with interpreting nonsensical strings, which suggests limited contextual flexibility.

Similarly, Coil and Shwartz (2023) explored noun compound interpretation using GPT-3 and found that, although the model performs well with common noun compounds, its performance drops significantly with novel compounds. This suggests that the model relies on pre-existing knowledge for familiar compounds, but struggles when required to reason about unseen combinations. Supporting this view, Cheng and Bhat (2024) found that models sometimes perform better on idiomatic reasoning tasks when contextual information is removed, implying that context may occasionally mislead rather than assist the model. Sun et al. (2021) further showed that LLMs tend to use context only when the relevant information is explicitly present. These findings collectively raise questions about whether current models truly engage in contextual reasoning or primarily depend on surface-level cues.

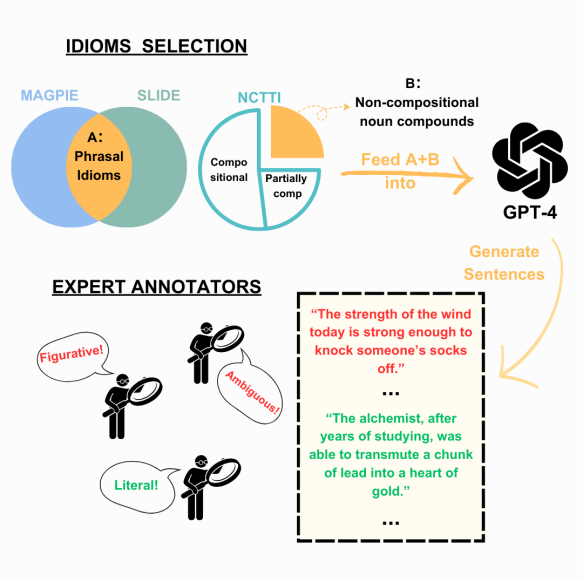


Figure 1: Overview of the DICE curation process. A list of idioms is extracted from existing datasets, and GPT-4 is prompted to generate sentences using each idiom in a literal context. Experts annotate these to form the literal subset. Figurative examples are drawn directly from existing datasets.

3 DICE: Dataset for Idiomatic Contrastive Evaluation

We now describe the construction of DICE. Figure 1 shows an overview of the curation process.

Expression Selection. To build our dataset, we compiled a comprehensive list of idiomatic expressions, focusing on both phrasal idioms and noun compound idioms. For phrasal idioms, we identified overlapping expressions from two established resources: MAGPIE (Haagsma et al., 2020) and SLIDE (Jochim et al., 2018). For noun compound idioms, we selected idiomatic expressions that appeared in both the NCTTI dataset (Garcia et al., 2021b) and AStitchInLanguageModels (Tayyar Madabushi et al., 2021). We specifically focused on idiomatic noun compounds, as the meaning of these expressions cannot be directly derived from the meanings of their individual words.

We excluded compositional and partially compositional compounds, as these tend to have dominant literal meanings that are difficult to override in context (e.g., "*skin tone*" or "*noble gas*"). Such expressions do not present the same interpretive challenge for models, as their meanings are more directly tied to their components. By focusing solely on non-compositional idioms, we ensure that the

	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	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	-

Table 1: Summary statistics of the DICE dataset.

dataset tests the model's ability to interpret figurative language based on context rather than relying on straightforward lexical composition. This selection process resulted in a total of 783 unique idiomatic expressions, comprising 680 phrasal idioms and 103 idiomatic noun compounds.

Sentence Generation. We used GPT-4 (OpenAI et al., 2024) to generate sentences where a given idiom appears in a literal context, intentionally suppressing its figurative interpretation. The specific prompting setup we used for sentence generation is provided in Appendix C. To determine the best model for sentence generation, we piloted the process using three versions: GPT-4o, GPT-4, and GPT-3.5. The models were prompted to produce three different sentences, where the form of the idiom must be kept the same. Overall, we found GPT-4 to perform the best in 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. In total, we obtained 2,349 sentences from GPT-4.

Expert Annotations. We recruited four experts with at least three years of university-level experience in Linguistics, compensated at a rate of £15/hour. The annotators reviewed each sentence and decided whether to accept it unconditionally, skip it, or reject it if the idiom's figurative meaning could not be fully suppressed. In cases of rejection, annotators provided reasons such as ambiguity, figurative interpretation, change of form, or other issues. We provide information about this subset in Appendix B along with the briefing given to annotators in Appendix D. 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 3. The inter-rater agreement was high, as indicated by a Cohen's kappa coefficient of 0.95.

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. This ensures that the dataset is balanced with regards to the idiomatic and literal interpretation of each expression.

3.1 Dataset Statistics

In total, DICE consists of 2,066 sentences, featuring 402 expressions. A summary of its statistics is presented in Table 1. In this paper, we only use the subset whose literal meaning was confirmed by the annotators. However, we release the complete dataset, along with an additional subset that can serve as a resource for additional analyses or for creating even more challenging evaluation settings (see Appendix B).

4 How Well Do LLMs Use Context for Idiomaticity?

Using DICE, 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 containing an idiom and asked it to classify the idiom as either "literal" or "figurative" based on its usage in context.

4.1 Experimental Setup.

Models. We evalute 13 models on the task of idiomatic sense disambiguation. These models are: GPT-4o, GPT-3.5-Turbo (Brown et al., 2020), FLAN-T5 models in the XXL (11B), XL (3B), Large (780M), Small (80M) sizes (Chung et al.,

2023), Llama 2 (7B, 13B, 70B) (Touvron et al., 2023), and Llama 3 models (incl. Llama 3.1 (405B, 8B, 70B) (Dubey et al., 2024). Additionally, we evaluated GPT-4 (OpenAI et al., 2024), which was used to generate the sentences. The computational resources used are reported in Appendix E.

Prompts. We used three prompt variations (Table 5 in Appendix E.2) that are semantically equivalent but differ in surface form. This allows us to evaluate the robustness of model predictions to prompt phrasing. We report the mean and standard deviation across these prompts to assess performance stability under variation. We also run experiments in a few-shot setting, where an annotated example (shown in the middle part of Table 5) is prepended to each prompt.

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

Accuracy includes two sub evaluations:

- **Figurative Accuracy**: We compute the accuracy of each model in correctly identifying the figurative uses of expressions within the figurative subset. Let F be the set of all figurative instances, y_i and $\hat{y}_i$ denote the true and predicted labels, respectively.

$$\text{Accuracy}_{\text{fig}} = \frac{1}{|F|} \sum_{i \in F} \mathbb{1}(\hat{y}_i = y_i) \quad (1)$$

where

$$\mathbb{1}(\text{condition}) = \begin{cases} 1 & \text{if condition is true} \\ 0 & \text{otherwise} \end{cases}$$

- **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. Let L be the set of all literal instances.

$$\text{Accuracy}_{\text{lit}} = \frac{1}{|L|} \sum_{i \in L} \mathbb{1}(\hat{y}_i = y_i) \quad (2)$$

Lenient Consistency: This metric rewards the model for consistently classifying all instances of an expression as either literal or figurative within a specific type. For example, if the model correctly classifies all literal variations of a given expression as "literal", it earns a point for that expression. Similarly, the model is rewarded if it correctly identifies all figurative instances of an expression as "idiomatic". Let E be the set of all expressions. Let L_e and F_e be the literal and figurative instances of expression e, respectively.

$$C_e^{\text{lit}} = \begin{cases} 1 & \text{if } \forall i \in L_e, \hat{y}_i = y_i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$C_e^{\text{fig}} = \begin{cases} 1 & \text{if } \forall i \in F_e, \hat{y}_i = y_i \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$\text{Lenient Consistency} = \frac{1}{2|E|} \sum_{e \in E} \left(C_e^{\text{lit}} + C_e^{\text{fig}} \right) \quad (5)$$

Strict Consistency: 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. Let V_e be the set of all variations of expression e.

$$S_e = \begin{cases} 1 & \text{if } \forall i \in V_e, \hat{y}_i = y_i \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

$$\text{Strict Consistency} = \frac{1}{|E|} \sum_{e \in E} S_e \quad (7)$$

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 2 presents the results of model performances on our evaluation set. It is important to note that, GPT-4 was used to generate the sentences for DICE, so we do not consider its performance to be revealing.

Degradation across evaluation levels. As expected, performance declines from Accuracy to Lenient Consistency, and further to Strict

Model	Accuracy			Lenient Consistency			Strict Consistency
	Figurative	Literal	Overall	Figurative	Literal	Overall	Both Settings
ZERO-SHOT							
GPT-4o	87.05 ± 3.62	87.30 ± 2.98	84.33 ± 4.44	69.49 ± 11.71	71.06 ± 6.68	70.32 ± 7.11	48.59 ± 9.75
GPT-3.5 Turbo	79.05 ± 5.01	70.02 ± 12.72	75.54 ± 7.81	82.59 ± 9.17	44.36 ± 22.28	63.47 ± 7.61	32.84 ± 15.81
Flan-T5-XXL (11B)	77.18 ± 1.40	74.91 ± 8.35	76.40 ± 4.49	63.93 ± 13.71	58.79 ± 23.16	61.36 ± 4.73	32.92 ± 6.80
Flan-T5-XL (3B)	70.48 ± 3.56	33.94 ± 26.91	59.65 ± 8.19	91.13 ± 6.97	13.02 ± 11.24	52.07 ± 3.58	9.95 ± 8.88
Flan-T5-Large (780M)	66.63 ± 0.10	3.45 ± 4.72	50.42 ± 0.53	97.68 ± 3.40	0.58 ± 0.80	49.13 ± 1.30	0.58 ± 0.80
Flan-T5-Small (80M)	0.51 ± 0.59	66.72 ± 0.07	50.13 ± 0.15	0.00 ± 0.00	100.00 ± 0.00	50.00 ± 0.00	0.00 ± 0.00
Llama 3.1 (405B)	88.63 ± 2.36	88.25 ± 3.93	**88.45 ± 3.10**	78.52 ± 5.61	80.02 ± 12.43	**79.27 ± 3.46**	**60.36 ± 6.61**
Llama 3 (70B)	87.72 ± 4.63	86.13 ± 7.10	87.00 ± 5.73	81.84 ± 4.00	72.64 ± 16.12	77.24 ± 7.45	57.55 ± 12.41
Llama 3 (8B)	79.27 ± 1.97	74.01 ± 2.79	76.91 ± 2.25	77.86 ± 5.18	48.76 ± 3.37	63.31 ± 1.43	33.83 ± 2.60
Llama 2 (70B)	76.28 ± 4.39	56.64 ± 17.13	69.62 ± 7.82	93.20 ± 4.75	24.54 ± 16.89	59.12 ± 5.78	21.81 ± 13.51
Llama 2 (13B)	68.99 ± 1.39	36.09 ± 3.85	58.26 ± 1.96	85.41 ± 3.56	8.37 ± 3.34	46.93 ± 2.30	5.64 ± 2.00
Llama 2 (7B)	55.51 ± 19.54	31.97 ± 24.25	51.34 ± 1.55	59.87 ± 46.26	18.08 ± 29.16	38.97 ± 8.59	1.66 ± 1.37
GPT-4	88.56 ± 2.03	88.63 ± 2.08	88.48 ± 2.18	79.02 ± 3.11	78.03 ± 4.60	78.52 ± 2.95	59.62 ± 4.67
ONE-SHOT							
GPT-4o	89.43 ± 1.23	90.15 ± 1.71	**89.72 ± 1.45**	74.63 ± 1.99	87.40 ± 5.81	**81.01 ± 1.93**	**63.52 ± 3.15**
GPT-3.5 Turbo	79.41 ± 4.19	72.69 ± 10.87	76.70 ± 6.54	78.44 ± 8.80	49.42 ± 18.96	63.93 ± 5.92	34.16 ± 12.19
Flan-T5-XXL (11B)	10.20 ± 15.69	67.90 ± 1.91	52.79 ± 4.34	1.58 ± 2.52	99.25 ± 1.29	50.41 ± 0.61	1.49 ± 2.37
Flan-T5-XL (3B)	0.64 ± 0.80	66.71 ± 0.11	50.13 ± 0.22	0.08 ± 0.14	99.83 ± 0.29	49.96 ± 0.19	0.08 ± 0.14
Flan-T5-Large (780M)	3.28 ± 3.64	66.27 ± 0.45	50.00 ± 0.00	0.66 ± 0.76	96.93 ± 3.73	48.80 ± 1.48	0.00 ± 0.00
Flan-T5-Small (80M)	45.23 ± 39.19	35.55 ± 33.55	53.03 ± 5.25	60.78 ± 53.37	37.31 ± 54.62	49.05 ± 1.65	2.40 ± 4.16
Llama 3.1 (405B)	89.57 ± 1.80	89.54 ± 2.54	89.53 ± 2.17	79.10 ± 3.26	82.01 ± 7.85	80.56 ± 2.56	63.27 ± 4.66
Llama 3 (70B)	87.75 ± 3.76	86.97 ± 5.64	87.27 ± 4.61	78.52 ± 3.59	75.62 ± 14.01	77.07 ± 6.00	57.55 ± 10.22
Llama 3 (8B)	80.32 ± 5.33	73.81 ± 11.40	77.59 ± 7.62	79.35 ± 1.08	48.01 ± 15.70	63.68 ± 7.34	34.91 ± 13.59
Llama 2 (70B)	70.40 ± 1.19	31.44 ± 6.18	58.65 ± 2.28	96.52 ± 0.66	7.55 ± 2.75	52.03 ± 1.50	6.72 ± 2.63
Llama 2 (13B)	70.64 ± 1.15	34.20 ± 6.92	59.36 ± 2.45	94.94 ± 0.52	9.54 ± 4.14	52.24 ± 1.83	8.29 ± 3.11
Llama 2 (7B)	70.26 ± 3.14	42.18 ± 26.31	61.21 ± 9.28	80.76 ± 15.43	20.73 ± 22.25	50.75 ± 3.46	11.69 ± 10.42
GPT-4	88.52 ± 1.49	88.95 ± 2.09	88.42 ± 1.73	78.44 ± 0.76	77.94 ± 5.84	78.19 ± 2.63	58.87 ± 4.86

Table 2: Results are reported as mean ± standard deviation over 3 different prompt variants, under zero-shot (top) and one-shot (bottom) conditions. GPT-4 is shown separately, as it generated the evaluation sentences.

Consistency. The results from the strictest evaluation show that, only three models—Llama 3.1, Llama 3, and GPT4o—achieve an accuracy above 40%, with 60.36%, 57.55%, and 48.59% respectively. This pattern highlights that, while the current LLMs can correctly classify individual instances, they often fail to do so consistently for both literal and figurative uses of the same idiomatic expression. Compounding this, we observe substantial standard deviations across prompts, particularly for smaller models. If the models truly relied on contextual understanding, we would expect them to perform consistently across varying levels of consistency and remain robust to prompt variations; however, the results suggest that the performance is highly inconsistent, and thus, models are not effectively leveraging context.

Preference towards figurative. The general trend for based on lenient consistency aligns with our observation on base accuracy: 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. We observe the largest aggregate drop for GPT-4o which indicates that GPT-4o's high performance (evidenced by a overall accuracy of 84.33 ± 4.44 in zershot evaluations) stems from its consistency across a broad range of idioms. However, the model lacks a deep understanding of these idioms, and it is susceptible to variations. This is illustrated by an overall Lenient Consistency score of 70.32 ± 7.1, which indicates that the model can only accurately interpret a subset of idioms consistently across different texts. Llama 3.1 is the model with the least performance difference across the two subsets, indicating a more balanced understanding of both figurative and literal contexts.

Mixed impact of one-shot prompting. Introducing an example improves performance for some models—most notably GPT-4o, which shows gains across all metrics and achieves the highest strict consistency (63.5%). However, this improvement is not consistent: for many models, particularly the Flan-T5 variants, one-shot prompting offers little benefit or even degrades performance. This likely stems from

a task-specific issue: instead of clarifying the task requirements, showing examples where idiomatic expressions appear literally clashes with the model's prior expectations. Consequently, the one-shot example creates uncertainty instead of enhancing contextual understanding.

Summary. While larger models like GPT-4o and Llama 3.1 (405B) show stronger overall performance, the results reveal that idiomaticity disambiguation, especially under consistency-based evaluation, remains a challenging task. The substantial drop from accuracy to strict consistency, coupled with high variation across prompts, suggests that models still rely heavily on shallow heuristics rather than robust contextual understanding.

5 Impact of Frequency and Sentence Likelihood on Model Performance

This section examines how idiom frequency and sentence likelihood impact LLM performance. In human processing, idiom frequency and familiarity, alongside context, influence comprehension (Cronk et al., 1993; Levorato and Cacciari, 1992; Schweigert, 1986; Brysbaert et al., 2018). Similarly, in LLMs, expression frequency and sentence probability may shape idiom detection. Exploring these factors helps clarify how language-intrinsic and model-intrinsic features affect performance.

As shown in Figure 2, the frequency of idioms in DICE varies, with the majority of expressions occurring fewer than 200,000 times. The highest concentration of idioms falls within the 0 to 100,000 range. To focus on the most relevant portion of the dataset, we limit our analysis to this range to avoid skewed results.

5.1 Frequency Estimation

We use the English Web Corpus (enTenTen) (Jakubíček et al., 2013) to approximate the frequency of idioms in our dataset. This was for two key reasons: (1) it is parsed and tagged which allows us to query all morphological forms of the expressions using lemmas, ensuring comprehensive frequency counts, and (2) its large scale 52 billion words across diverse genres offers a robust generalization of natural language usage. While enTenTen is not identical to the pretraining datasets of LLMs, it provides a reasonable proxy for estimating expression frequency, as high-frequency terms in enTenTen are likely to appear frequently in similar web-based pretraining corpora. This analysis helps us understand whether model performance is affected by how often they may have encountered each expression during training. See Appendix F for further details on how we obtain frequency counts.

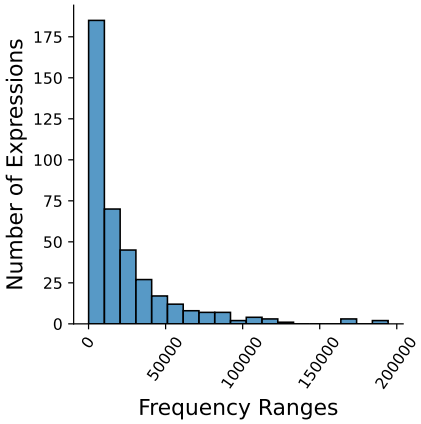


Figure 2: Frequency distribution of idioms in DICE below 200,000 counts.

5.2 Likelihood Scores

LLMs are trained to maximize the likelihood of their training data, learning to assign higher probabilities to sequences that resemble those seen during training. At inference time, these models can assign likelihood scores to input sentences, reflecting how typical or expected a sentence appears under the model's internal distribution. Recent studies have shown that models tend to perform better, or assign higher evaluation scores, to sentences they deem more likely, regardless of the specific target task or evaluation criteria (Ohi et al., 2024; McCoy et al., 2024). In our setting, we use sentence-level likelihood as a proxy to test whether model performance on idiomaticity disambiguation is influenced by how probable a sentence appears to the model. Specifically, we ask whether models are more accurate on DICE when the input context has high likelihood, potentially indicating a preference for familiar or prototypical constructions over true contextual reasoning.

To analyze the relationship between sentence

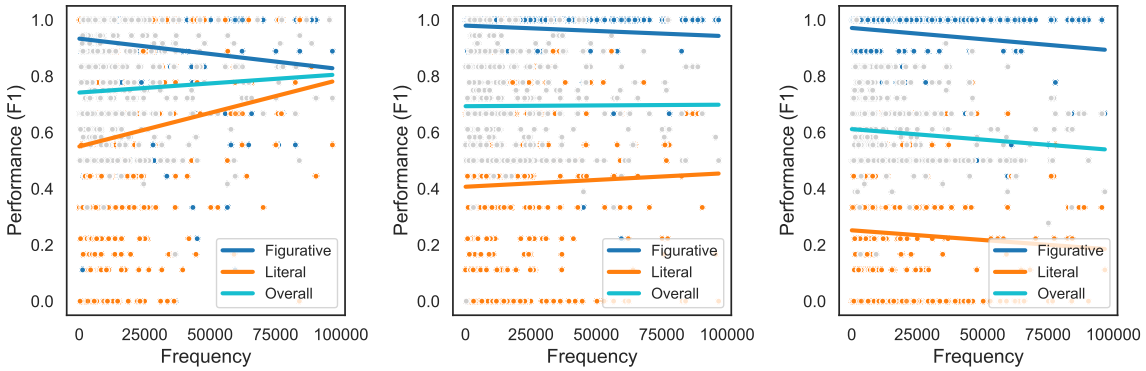


Figure 3: Frequency results of GPT-3.5 Turbo, Llama 2 (70B), Flan-T5 XL (left to right).

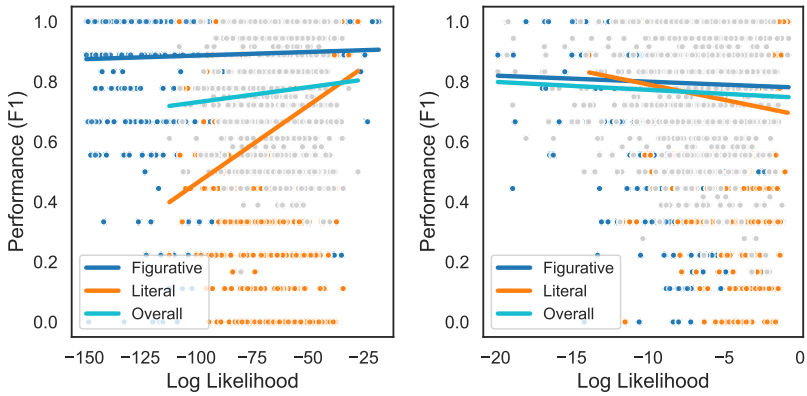


Figure 4: Likelihood results from Llama 3 (8B) and Flan-T5 XXL (left to right).

likelihood and model performance, we compute the likelihood of each sentence in DICE using the standard language modeling formulation. For a sentence $y = [y^{(1)}, \ldots, y^{(T)}]$, the likelihood is given by:

$$P(y) = \prod_{t=1}^{T} P(y^{(t)} \mid y^{(<t)}),$$

where $y^{(<t)}$ denotes the preceding tokens. In practice, we compute the log-likelihood (negative cross-entropy loss) for each sentence and use it as a proxy for model-assigned likelihood (Appendix G).

5.3 Results

To explore the impact of these two features on model performance, we fit a linear regression model, with idiom frequency/likelihood as the independent variable and the model's accuracy as the dependent variable. This approach allows us to estimate how these two variations influence the ability of language models to distinguish between literal and idiomatic uses.

We present the frequency results for GPT-3.5 Turbo, Llama 2 (70B) and Flan-T5 XL in Figure 3. For the likelihood analysis, we present the results for Llama 3 (8B) and Flan-T5 XXL in Figure 4. Results for the other models can be found in Appendix H. These models were chosen as they are representative of the general patterns observed across all the models evaluated.

Frequency is not a free lunch. In general, we do not observe a consistent pattern that accuracy correlates with expression frequency, suggesting that while models may have encountered certain expressions frequently during pretraining, they may still struggle to properly interpret their idiomaticity in new contexts. Notably, this trend was observed in 9 out of the 13 models we analyzed. Additionally, as the frequency of an expression increases, the models tend to perform better at identifying its literal occurrences, but their accuracy in recognizing idiomatic uses declines. One possible explanation is that for high-frequency expressions, the model may have seen both literal and idiomatic usages during pretraining. However, due to

limited contextual understanding, the model may default to interpreting these expressions literally more often, regardless of the actual usage; however, this hypothesis requires further investigation. The contrasting relationship between the two settings explains why no overall correlation is observed between accuracy and frequency. Therefore, frequency does not guarantee performance.

Likelihood $\neq$ Understanding. Our analysis of sentence likelihoods reveals contrasting trends across model families. In the case of the Llama models, performance correlates positively with sentence probability. In both settings, the model perform better on sentences on which it has a higher likelihood. This is particularly the case on the literal subsets, as indicated by Llama 2 (13B) and Llama 3 (8B). For the Flan-T5 models, we see a negative or negligible correlation between frequency/likelihood and performance. As seen in § 4.2, Small and Large do not appear to have effectively utilized the context to learn meanings as successfully as the other models. The counter-intuitive pattern observed in XL and XXL models could be due to models being over-confident in their wrong prediction. This leads to situations where model would assign high probabilities to idiomatic sentences but performs poorly on the idiomaticity detection task, where surface-level fluency inflates likelihood scores without supporting deeper semantic resolution. These results show that high likelihood does not necessarily imply correct contextual understanding, especially for figurative language.

6 Conclusion

In this work, we contribute to idiomaticity detection in NLP with several key findings. First, we introduce DICE, a challenging dataset for context-dependent idiom detection, where distinguishing figurative from literal meanings relies heavily on understanding context. Second, we propose an evaluation framework that measures both overall accuracy and strict consistency, requiring models to correctly identify all figurative and literal instances of an expression across different contexts. Third, we show that current LLMs struggle to use context effectively, highlighting the need for models that better capture contextual nuances.

We also investigate the effects of expression frequency and sentence likelihood. While frequency can correlate with performance in some settings, it does not guarantee accurate interpretation, highlighting that surface-level exposure is not a substitute for contextual reasoning. Similarly, while some models tend to perform better on sentences with higher likelihood, this correlation is inconsistent across models. Overall, our findings highlight the limitations of existing models in comprehending idiomatic language and highlight the need for evaluation settings, and model architectures, that emphasize deep contextual understanding over memorization or distributional familiarity.

7 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, this relies on the existence of idiomaticity datasets in the target languages. Moreover, the idea of making idioms literal might not be translatable to other languages, where the expression's domainant, figurative meaning cannot be overridden.

Acknowledgements

We thank the anonymous reviewers and metareviewer for their feedback on this paper. This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1], UK EPSRC

grantEP/T02450X/1, the Turing Institute Fellowship and the UniDive COST Action. We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing. Additional thanks to our annotators for their hard work. A further acknowledgement to Lily Zeng, Thomas Pickard and members of the MWE Group for their helpful discussions.

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.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. 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.

Marc Brysbaert, Paweł Mandera, and Emmanuel Keuleers. 2018. The word frequency effect in word processing: An updated review. *Current directions in psychological science*, 27(1):45–50.

Kellen Cheng and Suma Bhat. 2024. No context needed: Contextual quandary in idiomatic reasoning with pre-trained 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.

Brian C. Cronk, Susan D. Lima, and Wendy A. Schweigert. 1993. Idioms in sentences: Effects of frequency, literalness, and familiarity. *Journal of Psycholinguistic Research*, 22(1):59–82.

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.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher,

Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu,

Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Robert B Estill and Susan Kemper. 1982. Interpreting idioms. *Journal of psycholinguistic research*, 11(6):559–568.

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).

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. 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.

Raymond W. Gibbs and Gayle P. Gonzales. 1985. Syntactic frozenness in processing and remembering idioms. *Cognition*, 20(3):243–259.

Raymond W Gibbs and Nandini P Nayak. 1989. Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology*, 21(1):100–138.

Raymond W Gibbs Jr, Nandini P Nayak, and Cooper Cutting. 1989. How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of memory and language*, 28(5):576–593.

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.

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 Language 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.

Marianna Kyriacou, Kathy Conklin, and Dominic Thompson. 2020. Passivizability of idioms: Has the wrong tree been barked up? *Language and Speech*, 63(2):404–435. PMID: 31106699.

Andreas Langlotz. 2006. *Idiomatic Creativity*. John Benjamins.

Maria Chiara Levorato and Cristina Cacciari. 1992. Children's comprehension and production of idioms: the role of context and familiarity. *Journal of Child Language*, 19(2):415–433.

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.

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.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Marilyn A Nippold and Mishelle Rudzinski. 1993. Familiarity and transparency in idiom explanation: A developmental study of children and adolescents. *Journal of Speech, Language, and Hearing Research*, 36(4):728–737.

Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki. 2024. Likelihood-based mitigation of evaluation bias in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3237–3245, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Confer-*

ence on Computational Natural Language Learning (CoNLL), pages 66–76, Hong Kong, China. Association for Computational Linguistics.

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.

Wendy A Schweigert. 1986. The comprehension of familiar and less familiar idioms. *Journal of Psycholinguistic Research*, 15:33–45.

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.

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.

David A Swinney and Anne Cutler. 1979. The access and processing of idiomatic expressions. *Journal of verbal learning and verbal behavior*, 18(5):523–534.

Patrizia Tabossi, Rachele Fanari, and Kinou Wolf. 2009. Why are idioms recognized fast? *Memory & cognition*, 37:529–540.

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 Examples of Expert Annotations

We provid an example of expert annotations in Table 3.

B Additional Annotations of DICE

The analysis we focus on in this paper uses the main subsets of DICE. There are additional sentences we have collected that can be used for further analysis in the domain of idiomaticity. A summary is presented in Table 4. We make all annotations collected publicly available.

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 3: Examples of expert annotations. Definitions are taken from Ayto (2020). "N" and "Y" stands for "No" and "Yes", respectively.

	Counts	Examples and Remarks
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 4: Properties of the additional annotations that we have collected.

C Sentence Generation Prompt

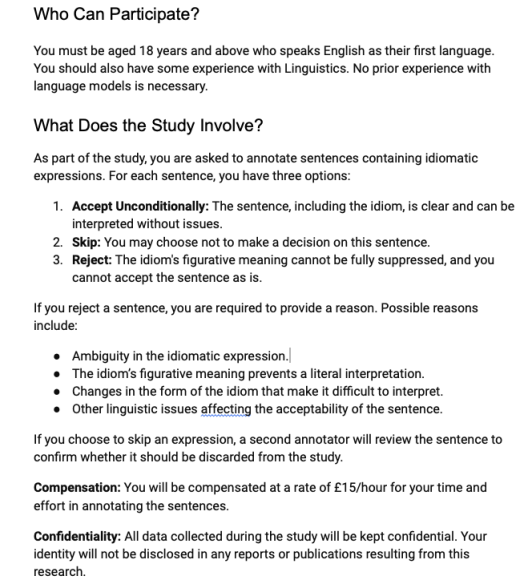
The prompt we used for generating the sentences is shown here. For other configurations that are not mentioned, we used the default setting.

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."

The temperature used was 0.8.

D Participant Briefing

Upon signing up for participation, each annotator received a 30mins training session where they were shown examples, including Table 3.



Who Can Participate?

You must be aged 18 years and above who speaks English as their first language. You should also have some experience with Linguistics. No prior experience with language models is necessary.

What Does the Study Involve?

As part of the study, you are asked to annotate sentences containing idiomatic expressions. For each sentence, you have three options:

1. **Accept Unconditionally:** The sentence, including the idiom, is clear and can be interpreted without issues.
2. **Skip:** You may choose not to make a decision on this sentence.
3. **Reject:** The idiom's figurative meaning cannot be fully suppressed, and you cannot accept the sentence as is.

If you reject a sentence, you are required to provide a reason. Possible reasons include:

- Ambiguity in the idiomatic expression.
- The idiom's figurative meaning prevents a literal interpretation.
- Changes in the form of the idiom that make it difficult to interpret.
- Other linguistic issues affecting the acceptability of the sentence.

If you choose to skip an expression, a second annotator will review the sentence to confirm whether it should be discarded from the study.

Compensation: You will be compensated at a rate of £15/hour for your time and effort in annotating the sentences.

Confidentiality: All data collected during the study will be kept confidential. Your identity will not be disclosed in any reports or publications resulting from this research.

Figure 5: Parts of the briefing annotators received.

E Evaluation Implementation Details

E.1 ChatGPT Versions

We evaluate the following GPT models: GPT-4o (gpt-4o-2024-08-06) [2], GPT-3.5-Turbo (gpt-3_5-turbo-0125)[3] and GPT-4 (gpt-4-0613)[4].

E.2 Prompting Paradigm

We ran the FLAN-T5 models on a NVIDIA H100 GPU. We use OpenAI's API for interactions with the GPT models, HuggingFace for Flan-T5 models and Replicate [5] for the Llama models. Each model was evaluated with three different prompts. All of the results we report

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

[3] https://platform.openai.com/docs/models/gpt-3.5-turbo

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

[5] https://replicate.com/

are the average across the three prompt settings.

E.3 Hyper-parameters

The hyper-parameters we used for running the evaluation are provided in Table 6.

	GPT Models	Flan-T5 Models	Llama 2s	Llama 3s	Llama 3.1
Temperature	1	-	0.7	0.7	0.6
max_tokens	Default	512	512	512	512
top_p	1	-	0.95	0.95	0.9

Table 6: Hyper-parameters used for model evaluation, sorted by model family

F Frequency Counts

We access the enTenTen corpus using SketchEngine and employ Corpus Query Language (CQL) to find concordances that match specific lexical patterns.

For each target expression, we determine its frequency in the corpus by accounting for all lemma-based forms of the expression. We simply slot in the lemmas of the target expression directly into the following CQL query pattern. For example, for the expression "*spill the beans*", the CQL would be: `[lemma="spill"][lemma="the"][lemma="bean"]`. This would capture occurrences such as "*spilled the beans*", "*spilling the beans*", and other morphological variants. We utilize NLTK (Bird and Loper, 2004) to perform lemmatisation and acquire the lemmas needed for the CQL query.

It is important to note that CQL query we use does not account for more syntactically flexible realizations, such as passive constructions (e.g., "the beans were spilled"), as these deviate from the fixed linear ordering captured by the CQL query. Consequently, the resulting frequency estimates represent a conservative measure. Nonetheless, for the purposes of correlating expression frequency with downstream model performance, we consider this approximation to be sufficiently informative.

G Sentence Likelihood

We derive the sentence likelihood by using the cross-entropy loss. For a sequence of tokens $y = [y^{(1)}, y^{(2)}, \ldots, y^{(T)}]$, the sentence-level cross-entropy loss L_{sentence} is defined as:

$$L_{\text{sentence}} = -\sum_{t=1}^{T} \log P(y^{(t)} \mid y^{(<t)}) \quad (8)$$

where $y^{<t}$ represents all tokens preceding the token at position t.

Recognizing the relationship between the cross-entropy loss and the sequence probability, we observe:

$$L_{\text{sentence}} = -\log \left(\prod_{t=1}^{T} P(y^{(t)} \mid y^{(<t)}) \right) \quad (9)$$

$$= -\log P(y) \quad (10)$$

Thus, the log likelihood of the sentence is

$$\log P(y) = -L_{\text{sentence}} \quad (11)$$

We calculate the log likelihood for all sentences in DICE.

H Additional Results for Frequency and Likelihood Analysis

We present the results we obtained across all the models. This includes both frequency analysis per setting, and across both settings.

Due to resource limitations we could not obtain probabilities for the larger models, which are Llama 3.1 (405B), Llama 3 (70B) and Llama 2 (70B). As a result, we only conducted the frequency analysis on these models, see Figure 6. Figure 9 presents side-by-side regression plots of the frequency and likelihood analysis. Similarly, Figure 8 presents the plots for the smaller Llama models. Finally, Figure 7 shows the frequency analysis plots for the GPT models.

Method	Prompt No.	Prompt Design
Zero-shot	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.
Example	–	The expression 'play with fire' is used figuratively in the sentence: 'The war took away the unfortunate necessity, as Unionists saw it, to play with fire in the national interest, but it did not materially alter their view of themselves.' → Output: i The expression 'play with fire' is used literally in the sentence: 'Despite the danger, he decided to play with fire, poking the embers with a stick.' → Output: l"
Few-shot	Prompt 1	Example + Prompt 1
	Prompt 2	Example + Prompt 2
	Prompt 3	Example + Prompt 3

Table 5: We use three prompts for our experiments. The top panel of the table shows the prompts used in the zero-shot setting, while the bottom panel displays the few-shot prompts. For the few-shot setting, we prepend the same example (middle panel) to each of the zero-shot prompts.

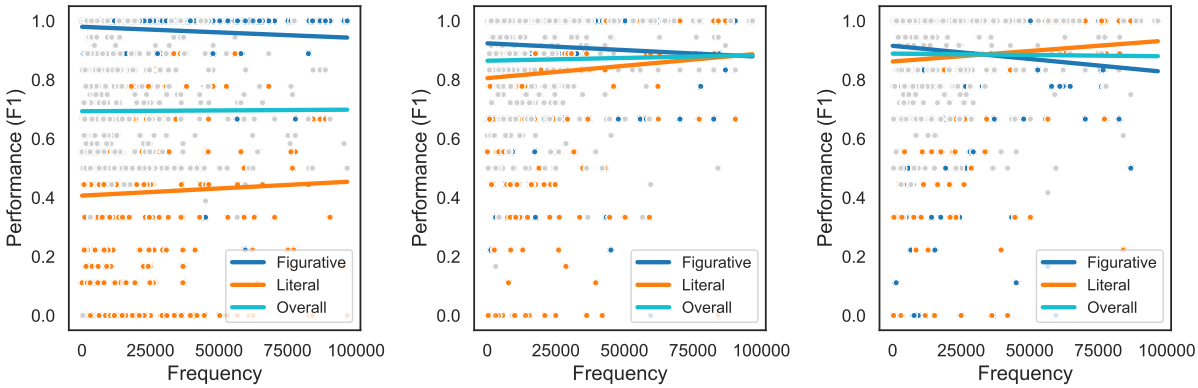


Figure 6: Left to right: Frequency analysis for Llama 2 (70B), Llama 3 (70B), and Llama 3.1 (405B)

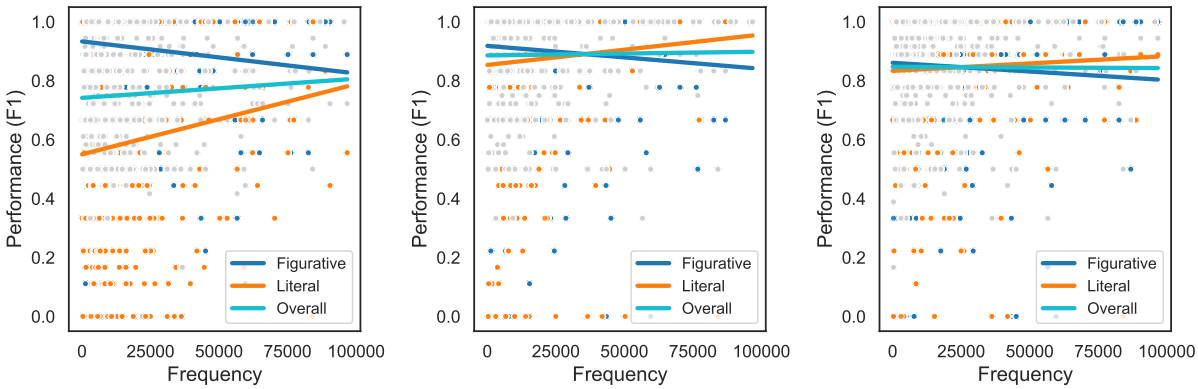


Figure 7: Left to right: Frequency analysis for GPT-3.5 Turbo, GPT-4 and GPT-4o.

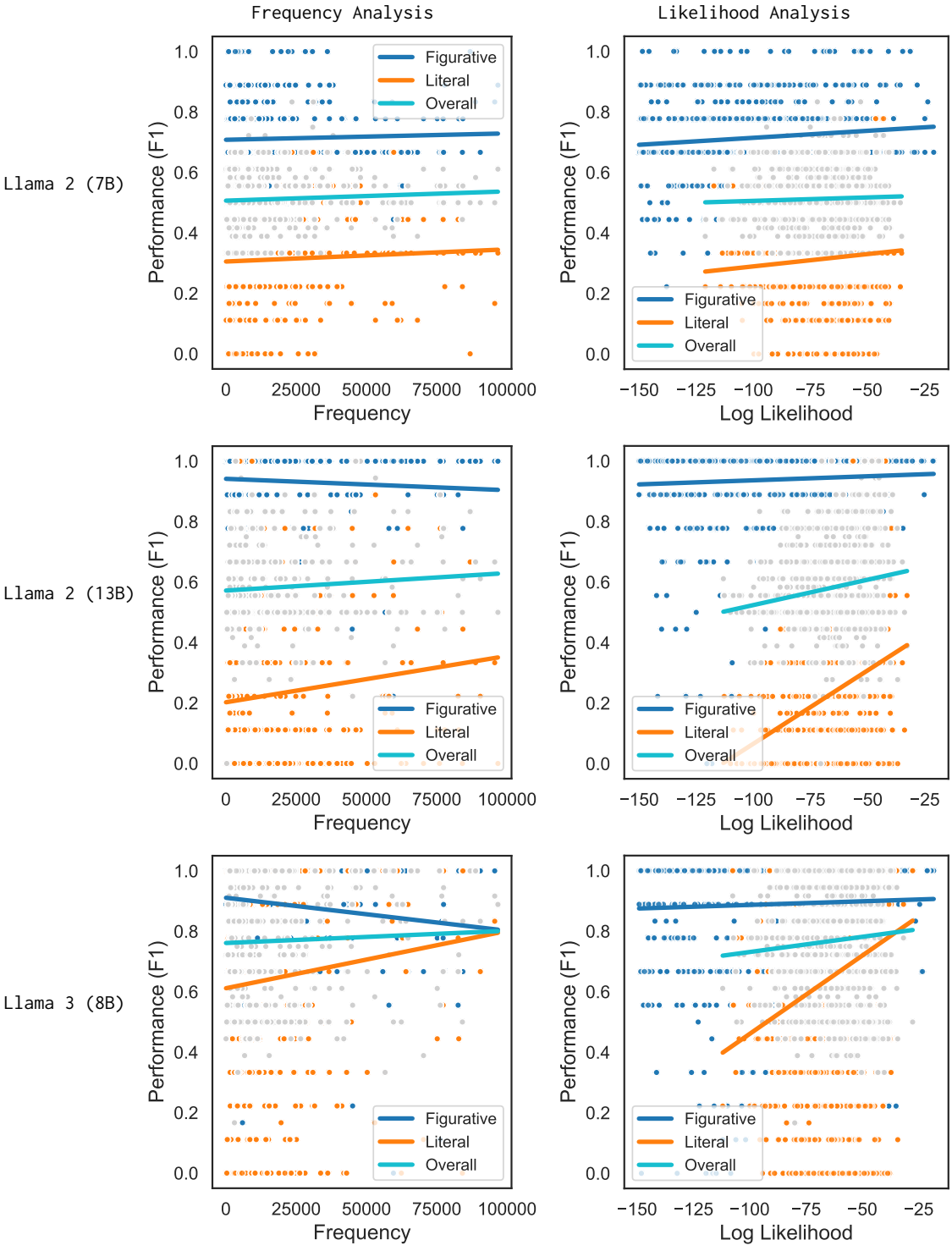


Figure 8: Visualisations of the frequency and likelihood. Smaller Llama models only.

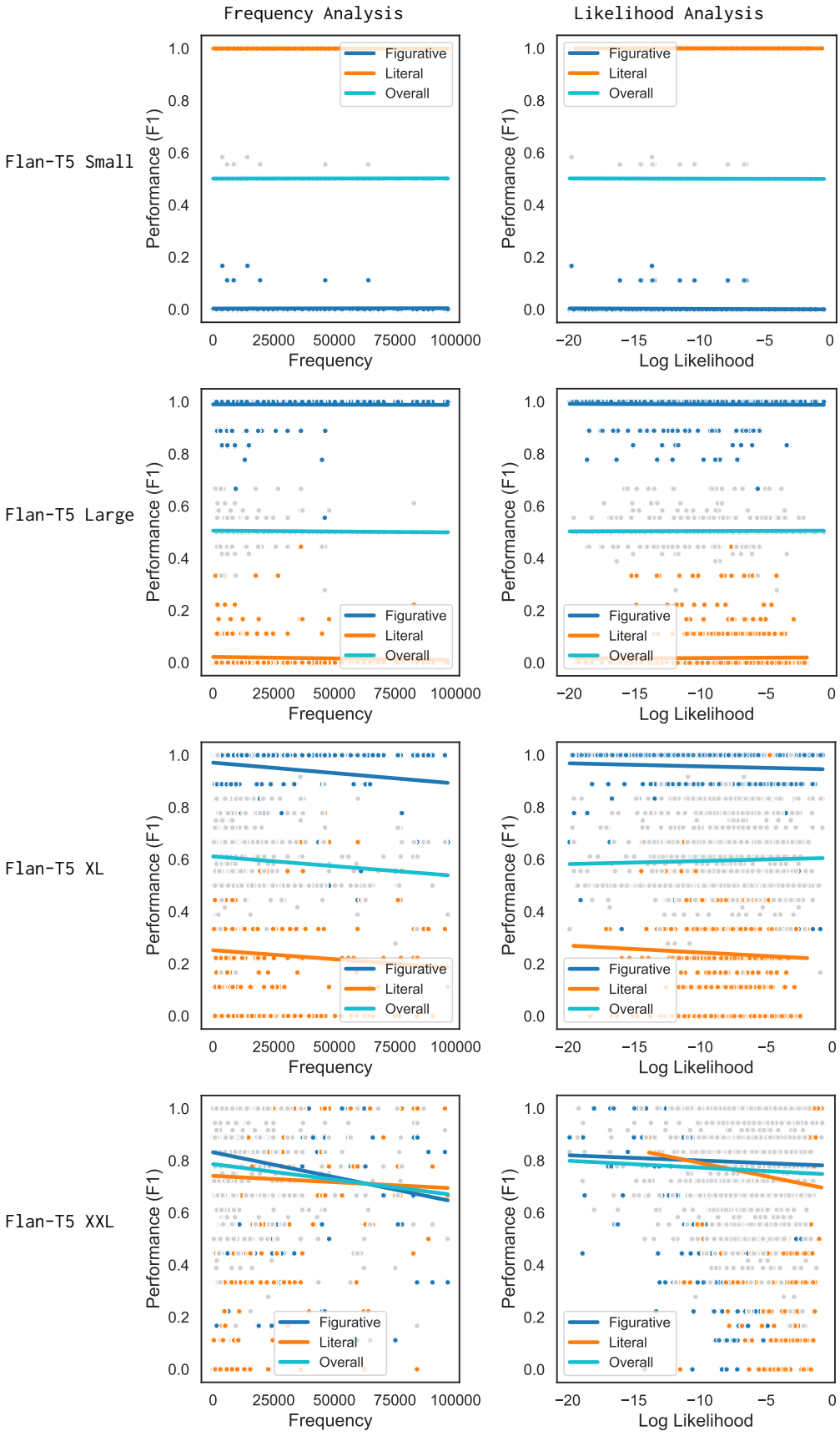


Figure 9: Visualisations of the frequency and likelihood analysis. Flan-T5 models only.