No Context Needed: Contextual Quandary In Idiomatic Reasoning With Pre-Trained Language Models

Kellen Tan Cheng[1], **Suma Bhat**[1, 2]
[1]Princeton University
[2]University of Illinois, Urbana-Champaign
[1]{kellentan, sumabhat}@princeton.edu

Abstract

Reasoning in the presence of idiomatic expressions (IEs) remains a challenging frontier in natural language understanding (NLU). Unlike standard text, the non-compositional nature of an IE makes it difficult for model comprehension, as their figurative or non-literal meaning usually cannot be inferred from the constituent words alone. It stands to reason that in these challenging circumstances, pre-trained language models (PTLMs) should make use of the surrounding context to infer additional information about the IE. In this paper, we investigate the utilization of said context for idiomatic reasoning tasks, which is under-explored relative to arithmetic or commonsense reasoning (Liu et al., 2022; Yu et al., 2023). Preliminary findings point to a surprising observation: general purpose PTLMs are actually negatively affected by the context, as performance almost always increases with its removal. In these scenarios, models may see gains of up to 3.89%. As a result, we argue that only IE-aware models remain suitable for idiomatic reasoning tasks, given the unexpected and unexplainable manner in which general purpose PTLMs reason over IEs. Additionally, we conduct studies to examine how models utilize the context in various situations, as well as an in-depth analysis on dataset formation and quality.[1] Finally, we provide some explanations and insights into the reasoning process itself based on our results.

1 Introduction

In natural language, there are many methods to express canonical stories, ideas, or scenarios in a succinct and fluent manner. One such method is through the use of idiomatic expressions (IEs), which are a form of multi-word expression (MWE) that carry a figurative, or non-compositional, meaning (Moon, 1998; Cacciari and Tabossi, 2014). IE comprehension remains a challenging frontier in natural language processing (NLP), with the challenge arising from this non-compositional aspect of an IE (Stowe et al., 2022; Tayyar Madabushi et al., 2021; Sag et al., 2002). This refers to the fact that the figurative meanings of many IEs cannot be inferred from the constituent words alone. One such example would be the IE “squaring the circle", where its meaning “to handle things/resolve" cannot be inferred from the IE’s individual words.

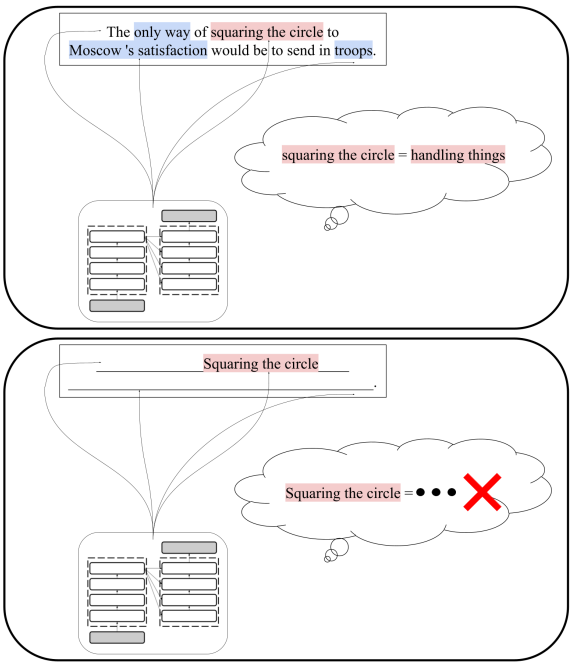

Figure 1: Visualization of the IE reasoning scenario with context (top) and without context (bottom). We hypothesize that models should be making use of context hints (highlighted in blue) to help comprehend the IE (highlighted in red).

Given the recent advances in PTLMs (Vaswani et al., 2017; Almazrouei et al., 2023; Touvron et al., 2023; OpenAI, 2023), it was commonly understood that transformer models’ excellent performance on a variety of natural language understanding (NLU) tasks was attributable to an improved reasoning process, specifically one that is more aligned with

[1]https://github.com/kellentan/No_Contexts.

humans (Schramowski et al., 2022; Dasgupta et al., 2023; Wei et al., 2022). Indeed, recent work on explanatory reasoning and chain-of-thought only served to cement this notion for many PTLMs (Wei et al., 2022; Wang et al., 2023; Shi et al., 2023). Thus, it stands to reason that for IE-related tasks, PTLMs would also comprehend idioms in a manner similar to humans (i.e. it should utilize the context to help build a better understanding of the IE (Levorato and Cacciari, 1992; Cain et al., 2009; Ortony et al., 1978; Chiara Levorato et al., 2004)). This paradigm is represented in the top half of Figure 1, where we would expect a model to gain a better understanding of the idiom "squaring the circle" by looking at the surrounding context. Specifically, phrases like "the only way" and "send in troops", would help indicate that a situation is being resolved or handled. For such a difficult IE, removing the contexts, as seen in the bottom half of Figure 1, would make it nearly impossible for a model to understand what "squaring the circle" actually means. Theoretically, reasoning without the context should be much more difficult, as there are no contextual clues to help build its understanding. This logic is also motivated by our understanding of the distributional hypothesis, i.e. words with similar meanings should appear in similar contexts (Harris, 1954), and thus provide an aid for IE comprehension.

Several aspects of IE reasoning have been the focus of recent studies, which investigated techniques for IE comprehension via idiomaticity detection (Zeng and Bhat, 2021; Zhou et al., 2023a; Liu, 2019), representation (Zeng and Bhat, 2022, 2023; Škvorc et al., 2022; Hashimoto and Tsuruoka, 2016; Adewumi et al., 2022; Liu and Hwa, 2017), and the use of structured knowledge such as knowledge graphs and knowledge bases (Zeng et al., 2023; Wang et al., 2019).

Our work uses a first-principles approach to examine context utilization in the presence of IEs. We chose to proceed with a data-driven approach, keeping models constant while only focusing on changes to the data samples themselves, thus guaranteeing that any differences in performance are directly attributable to how we modified the data. Note that our study is focused on analyzing the linguistic reasoning capabilities, specifically figurative reasoning, of models, which is an underexplored aspect of reasoning compared to arithmetic or commonsense reasoning (Liu et al., 2022; Yu et al., 2023). Our contributions are as follows:

(1) We demonstrate that for IE-related tasks, PTLMs surprisingly perform *better* in the absence of some, if not all, of the context, reaching gains of up to 3.89%. This refutes the idea that PTLM performance, at least in the IE setting, is attributable to a more human-like reasoning process.

(2) Naturally, this leads to subsequent avenues of thought concerned with understanding what exact factors these models use when reasoning over IEs. We perform additional analyses in order to understand how PTLMs perform better without context.

(3) We argue that general PTLMs may struggle for idiomatic tasks, and that IE-aware models exhibit behaviors more consistent with a human-like reasoning process. We also call for better dataset formation in idiomatic reasoning.

2 Experimental Setup

In our work, we investigate how PTLMs utilize the context for two distinct tasks: idiomatic NLI and idiomatic continuation acceptability. Note that the number of tasks at our disposal is limited due to the dearth of available IE reasoning datasets. The tasks, setups, and models are described in the following subsections.

2.1 Tasks

Idiomatic NLI is very similar to conventional NLI, with the only difference being the presence of an IE within each sample. Recall that in NLI, the objective for a PTLM, given two texts (a premise and hypothesis), is to determine whether the meaning of both texts are in entailment or in contradiction (Williams et al., 2018). The correct inference in this task is thus contingent upon a solid comprehension of both texts, as well as the IE itself. For our investigations, we utilize the IMPLI dataset for the idiomatic NLI task (Stowe et al., 2022). After balancing, the IMPLI dataset contains 13,650 training samples split evenly between both classes, taken over the silver split of the entire dataset (which in turn draws from IE-focused corpora) (Haagsma et al., 2020; Zhou et al., 2021; Korkontzelos et al., 2013). Inference is done on the gold split, which consists of 1,157 handcrafted samples, of which 528 are entailment and 629 are non-entailment. Specifically, the 629 non-entailment samples are further categorized into 254 normal non-entailment samples and 375 antonym non-entailment samples (where the meanings of both texts are direct opposites of each other).

Idiomatic continuation acceptability requires slightly more than just comprehension of the IE,

but also an understanding of how the IE interacts more deeply within a given context. In this case, the model receives two texts per sample. The first text is known as the narrative, which contains multiple sentences to build up a story or setting, and also contains a target IE. The other text is known as the continuation, and is a candidate sentence that would supposedly appear next in the narrative. The objective of any model in this setting is to determine whether the candidate is an acceptable continuation sentence for the given narrative. As an example, consider the following narrative: "It's not a bad sensation, but the type I experience when Noah rests an arm around my shoulder when we're walking down the street, or when he places a hand on the small of my back when he guides me through a crowded room. It's like a large cape drawn around me, making me feel safe and wanted. Making me feel included. I stagger back. My legs hit the stool, and I lower myself down onto it. Scanning the room, I see people from every *walk of life*." The model should then understand that a sentence such as "There were people clearly from a high class, laden in lavish clothes, as well as young college-aged students with frayed t shirts" is an acceptable continuation, as it reflects the meaning of the IE accurately and fits the narrative context of a crowded room. In our paper, we take these samples from the FigurativeNarrativeBenchmark dataset (Chakrabarty et al., 2022a). Note that each sample of the FigurativeNarrativeBenchmark dataset contains three texts: a narrative, a correct candidate continuation, and an incorrect candidate continuation. We take each sample from the dataset and transform it into two idiomatic continuation acceptability examples: one that pairs the narrative with the correct continuation, and another that pairs the same narrative with the incorrect continuation. After dataset transformation, we end up with 6,408 training samples and 3,084 test samples, where both sets of samples are evenly split between the two classes. Note that we chose these datasets as they represent the newest state-of-the-art benchmarks for their tasks, and thus their results should be representative of PTLM performance in IE reasoning scenarios. Results for other IE reasoning datasets, such as FLUTE (Chakrabarty et al., 2022b), can be found in Appendix C.

2.2 Setup

The pertinent question in both scenarios is given as follows: what constitutes "context" in each task?

For idiomatic NLI, we define the context in a sample as the set of common words that are shared between both the premise and the hypothesis, *and also* are not part of the IE itself. Referring to the example in Figure 1, our context would be the following: the only way of _____ to Moscow's satisfaction would be to send in troops. Obviously the non-context words are those that constitute the IE in the premise, and the corresponding words in the hypothesis. Context removal would simply discard the context from both the premise and the hypothesis, with our new sample defined as follows: <Premise: Squaring the circle. Hypothesis: Handling things.> In addition to context removal, we may also define context shuffling, which simply takes all the context words in each sample and randomly permutes them (while leaving the placement and order of the IE's words intact).

For idiomatic continuation acceptability, the length of the narrative text in each sample lends itself towards different types of context removal. We define the context here in two manners: sentential context and extra-sentential context. Sentential context refers to the non-IE words that are a part of the sentence containing the IE. Extra-sentential context refers to the other sentences in the narrative that do not contain the targeted IE. Thus, context removal here refers to the removal of extra-sentential context, while total context removal refers to the removal of both the extra-sentential context as well as the sentential context. Similar to the idiomatic NLI task, context shuffling simply takes all the context words (both sentential and extra-sentential context) and randomly permutes them (while leaving the placement and order of the IE's words intact). We also perform studies examining percentage removal, which simply removes a percentage of words from the narrative. This may be done by removing words from the front of the narrative, as typically the IE appears near the end of the narrative (thus we remove the furthest words first). Note that we do not examine random removal, as it could introduce additional grammatical errors during the inference process.

2.3 Models

BART denotes the pre-trained BART-large model, with 400M parameters (Lewis et al., 2020).

BART-IEKG denotes an IE-aware BART-large model, which was injected with knowledge from the IEKG knowledge base. We take this model from (Zeng et al., 2023).

BART-MNLI denotes a BART-large model that has been fine-tuned on the MNLI dataset (Williams et al., 2018). We derive this model in the manner described by (Zeng et al., 2023).
BART-MNLI-IEKG denotes the state-of-the-art BART-large model. It takes the **BART-MNLI** model and injects it with knowledge from the IEKG knowledge base. We take this model from (Zeng et al., 2023).
Mistral denotes a pre-trained Mistral-7B model, with 7B parameters (Jiang et al., 2023).
Mistral-FT takes the **Mistral** model and fine-tunes the classification head.
PIER+ denotes the best performing checkpoint of an IE-aware BART-base model (140M parameters) (Lewis et al., 2020). This model has been optimized to learn the best embeddings for IE representation. We take this model from (Zeng and Bhat, 2023).

The BART models were chosen to provide a comparison against IE-aware models, as seen in (Zeng et al., 2023; Zeng and Bhat, 2023, 2022). We chose Mistral due to its superior performance against Llama-2 (Jiang et al., 2023).

As a quick note, models that were trained on MNLI (i.e. BART-MNLI and BART-MNLI-IEKG) were not evaluated on the FigurativeNarrativeBenchmark, as that task is different from NLI.

3 Results & Discussion

Our results are demonstrated for the idiomatic NLI setting as well as the idiomatic narrative continuation acceptability task. Note that individual model hyperparameters are described in detail in Appendix D. Additionally, please refer to Appendix A to see the full results on each dataset per model.

3.1 Effect of Context Removal

Model	Acc.	Context Kept	Gain
BART	77.79%	60%	+0.59%
BART-IEKG	78.02%	90%	+0.17%
Mistral	49.64%	20%	+0.55%
Mistral-FT	63.20%	80%	+0.42%
PIER+	65.50%	90%	+0.42%

Table 1: A comparison of different models' best performance on the FigurativeNarrativeBenchmark dataset with a percentage of the original context.

We find that general PTLMs tend to perform better with some form of context removal, while IE-aware models typically degrade without contexts. Figure 2 illustrates this dichotomy on the IMPLI dataset, as general PTLMs such as BART, BART-MNLI, and Mistral-FT demonstrate significant gains in performance without the context, reaching as high as 3.89%. The gains for BART and BART-MNLI are statistically significant (90% confidence). Conversely, IE-aware models such as BART-IEKG and PIER+ exhibit the opposite behavior, losing as much as 2.24% in performance. This phenomenon seems to disappear with regards to the FigurativeNarrativeBenchmark dataset, where according to Figure 3, almost all models perform worse without context. Without extra-sentential context, the performance drops of IE-aware models (BART-IEKG, PIER+) are statistically significant (90% confidence). Without any context, the performance drops further. Even in this scenario though, IE-aware models such as BART-IEKG exhibit the largest degradation in performance, which suggests that the model possesses the highest utilization of context for its reasoning capabilities. Additionally, these results do not paint a full picture on this dataset, as we see from Table 1 that performance gains can still be observed with partial removal of the context. Once again, even in this perspective, IE-aware models exhibit the lowest performance gain (0.17% for BART-IEKG) while typically keeping a higher percentage of the original text (90% for both BART-IEKG and PIER+).

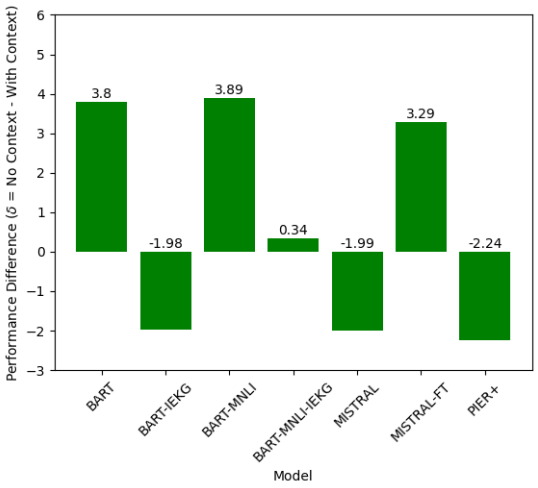

Figure 2: Visualization of the performance gain without context for various models on the IMPLI dataset.

3.2 Effect of Shuffled Context

As another method of analyzing context utilization, we were interested in observing the sensitivity of models towards shuffled context. In this regard, we see that models do display some sensitivity to context shuffling, with mixed results. General PTLMs

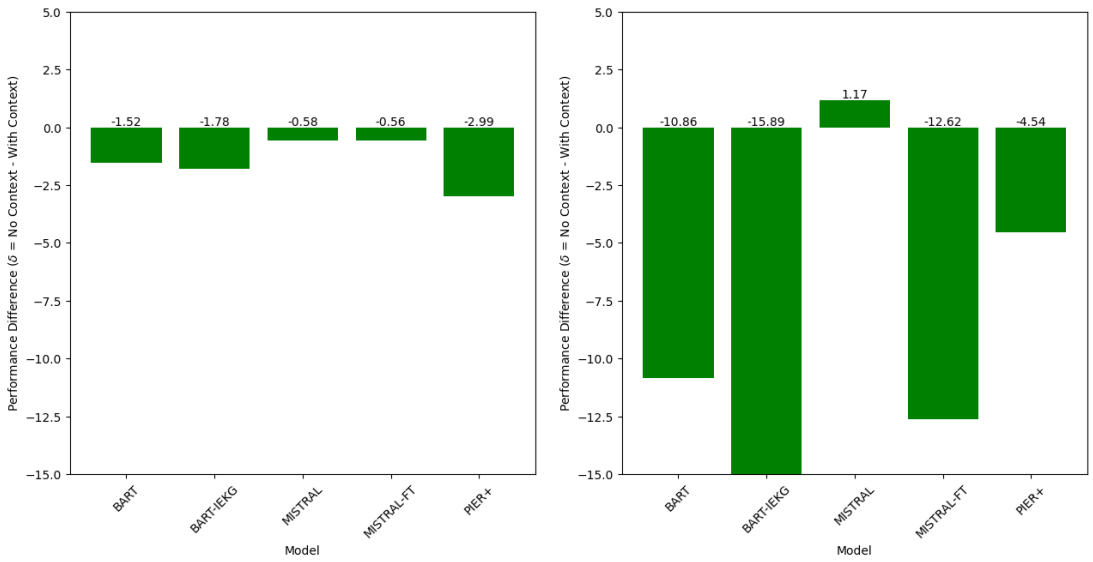

Figure 3: Visualization of the performance gain **without extra-sentential context** (left) and **without any context** (right) for various models on the FigurativeNarrativeBenchmark dataset.

Model	IMPLI	FigurativeNarrativeBenchmark
BART	3.19%	-7.19%
BART-IEKG	-1.98%	-8.01%
BART-MNLI	-8.55%	N/A
BART-MNLI-IEKG	-6.05%	N/A
Mistral	0.60%	1.23%
Mistral-FT	0.61%	-6.75%
PIER+	-1.29%	-3.37%

Table 2: A comparison of different models' performance gain when the context is shuffled.

such as BART, Mistral, and Mistral-FT still display performance gains when the context is shuffled for the IMPLI dataset, as seen in Table 2, with Mistral exhibiting this behavior on the FigurativeNarrativeBenchmark dataset as well. However, IE-aware models such as BART-IEKG and PIER+ still display much higher sensitivity towards context shuffling, with both models illustrating performance drops on both datasets. Additionally, BART-IEKG exhibits the largest drop on the FigurativeNarrativeBenchmark at over 8%.

3.3 Replacing IEs with Randomly Generated Strings

Additionally, we were interested in testing how well models utilized the context when faced with unknown words. The motivation here is that with randomly generated strings, this ensures that models have no pre-existing understanding of the string. As a result, PTLMs must reason over the context in order to comprehend this randomly generated string. From Figure 4, it is interesting to note that the general trends observed in Figure 2 still hold in this setting. Only IE-aware models (BART, BART-MNLI-IEKG, PIER+) exhibit performance drops without the context, which suggests that these models relied on the context to help build an understanding of these unknown, randomly generated strings. On the other hand, general PTLMs continue to exhibit performance gains without the context, even though they could not possibly have any pre-existing understanding of the string itself. While most of these gains are moderate, it is surprising that Mistral-FT can achieve a 5.10% increase in performance (statistically significant at 95% confidence). Model behavior on the FigurativeNarrativeBenchmark dataset is more consistent with our expectation, as seen in Figure 5. Here, most models demonstrate some usage of the context, as the performance drops without any context are much larger compared to Figure 2, with the results for BART, BART-IEKG, and PIER+ being statistically significant (99% confidence). Once again, however, IE-aware models like BART-IEKG displayed the highest drops in performance, losing up to 17.77% with total context removal. Even

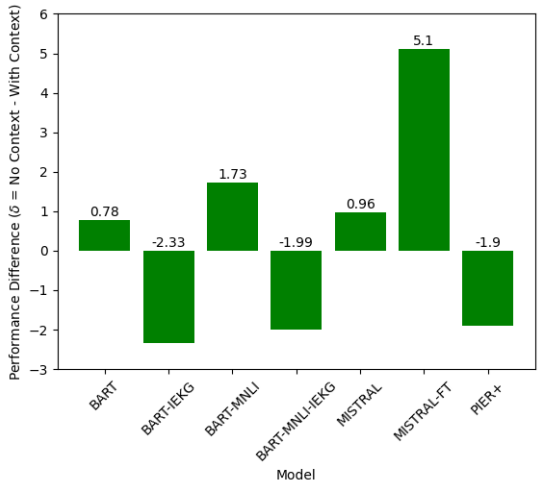

Figure 4: Visualization of the performance gain without context for various models on the IMPLI dataset. Note that the IEs in the dataset have been replaced with randomly generated strings.

when replacing the IEs with randomly generated strings, our results illustrate that IE-aware models display a much higher utilization of context compared to general PTLMs.

3.4 Effect of Idiomaticity

Another point of interest was determining the non-compositional nature of the idioms in our datasets. One plausible hypothesis for the under-utilization of context is that the meaning of the IEs in our data is easily inferred from just its constituent words. If that was the case, then general PTLMs might ignore the context since the expression itself provides a sufficient understanding of the idiom. However, this scenario does not appear to be the case. We found that the average idiomaticity of IEs in IMPLI to be around 84.64%, while for FigurativeNarrativeBenchmark this number is even higher at 88.31%. We were able to extract an IE's idiomaticity score from the MAGPIE dataset, where larger values indicated that the IE was more figurative (Haagsma et al., 2020). As a result, the high idiomaticity of the idioms would indicate that the IEs are highly figurative, and so their meaning cannot be reasonably inferred from just a literal reading of the expression. Therefore, even in situations which would necessitate a general PTLM to use contextual information, the model still neglects to utilize the context.

3.5 Case Studies

To understand why some models displayed such low context utilization, we performed a few case studies into the model behavior as well as dataset quality.

One potential explanation for why general PTLMs perform better without context is that these models are defaulting to their pre-training knowledge, rather than using the surrounding context to actually reason (Longpre et al., 2021). Perhaps in these scenarios, the model would be able to focus purely on the IE in order to recall a better understanding of it. If this were true, we would expect that models perform better without context for higher frequency IEs (those that occur more often in pre-training corpora). We were able to collect frequency counts for about 200 samples of the IMPLI dataset. These frequencies are noted from the Corpus of Contemporary American English (COCA), which contains approximately 1.1B words (Davies, 2010). From Figure 6, our findings show that models actually perform better with context on higher frequency IEs, while performing better without context for lower frequency IEs. Clearly then, it is not the case that these models have a pre-existing understanding of the IE.

Another potential explanation could be due to dataset artifacts. In this case, perhaps these artifacts are biasing models towards the incorrect inference, such that removing the context would remove these artifacts, hence increasing performance. We perform a study similar to that of (McCoy et al., 2019; Poliak et al., 2018), and find that for certain labels of the IMPLI dataset, it could be a case of particular words biasing model inference. From Table 3, words such as "thumb" and "tune" may well be biasing model predictions for the entailment class. However, the evidence remains far from conclusive, as this still does not provide an explanation for the performance on other datasets, such as FigurativeNarrativeBenchmark and FLUTE, which do not have skewed artifacts in the data (as these datasets are balanced and well-formed). For additional dataset analyses, please refer to the Appendix B.

4 Related Work

4.1 Context Usage in Transformers

Prior work has demonstrated that transformer-based models do not make optimal use of the context in a variety of tasks. In long context processing, even state-of-the-art models such as GPT-3.5 show poor context utilization in tasks such as question-answering retrieval, especially when the pertinent

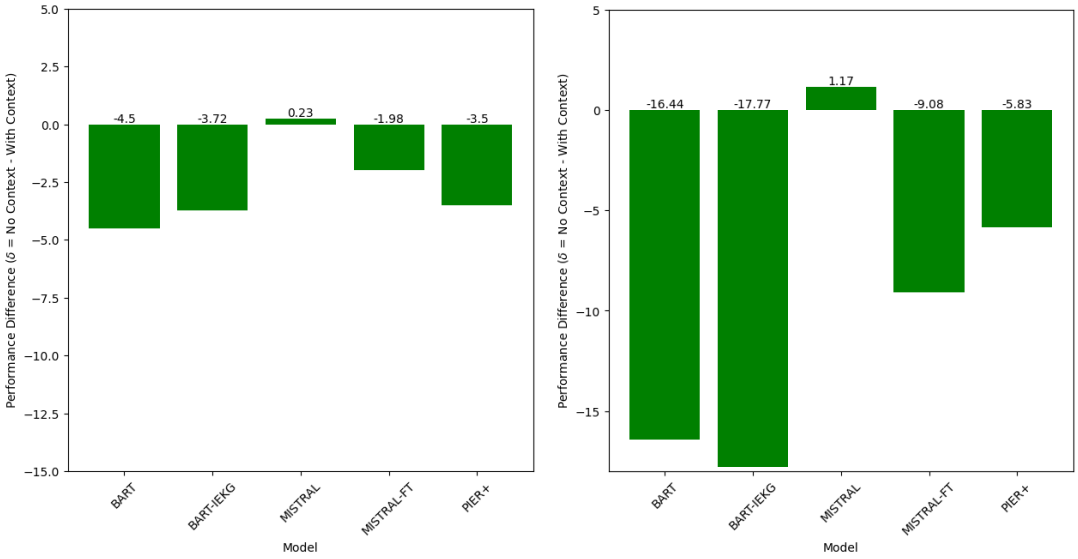

Figure 5: Visualization of the performance gain **without extra-sentential context** (left) and **without any context** (right) for various models on the FigurativeNarrativeBenchmark dataset. Note that the IEs in the dataset have been replaced with randomly generated strings.

Entailment			Non-Entailment			Antonym		
Word	P(L | W)	Frequency	Word	P(L | W)	Frequency	Word	P(L | W)	Frequency
thumb	1.00	4	turn	0.36	4	light	0.46	6
tune	0.86	6	break	0.35	8	see	0.44	4
round	0.80	4	give	0.35	9	terms	0.43	6
rule	0.75	6	board	0.33	4	stand	0.42	5
mind	0.60	6	throw	0.33	4	under	0.42	5
open	0.57	4	someone	0.32	11	over	0.42	5
set	0.57	4	as	0.31	4	behind	0.42	5
have	0.57	4	make	0.30	7	hand	0.41	9
time	0.57	8	do	0.29	5	strength	0.40	4
fall	0.57	8	face	0.29	5	play	0.40	4

Table 3: An overview of how indicative words within the IMPLI dataset correspond to particular class labels. These results are displayed for the top 10 words, in terms of highest probabilities with class label P(L | W), for each class. The results have been filtered to exclude common words such as “a" and “the", and are lower-bounded by a frequency of at least 4.

information cannot easily be found at the beginning or end of the context (Liu et al., 2023). Generally these models do not utilize this context unless the exact answer can be found (Sun et al., 2021). Models may only make some use of sentential context, as studies shown for the minimal-pair paradigm (MPP) acceptability task demonstrate that models are sensitive to only a few select contextual features (Sinha et al., 2023). Other studies have demonstrated the fact that PTLMs are surprisingly invariant towards context perturbations such as shuffling, or deleting all words except for nouns, in some instances these perturbations remove less than 15% of usable information for these models (Papadimitriou et al., 2022; O’Connor and Andreas, 2021). Our approach extends upon prior work, as we demonstrate scenarios where models can improve performance by making no utilization of the context, as we remove it entirely. We use these results to then argue that only IE-aware models remain suitable for IE reasoning tasks. Our additional studies affirm the poor sensitivity of these general PTLMs to context perturbations, such as shuffling, and we go further by also examining whether dataset features may be an indirect cause of this peculiar reasoning behavior.

4.2 Idiomatic Expression Reasoning

Previous research has utilized many techniques towards improving PTLMs’ IE reasoning capabilities. These include better datasets, embedding representations, and training schemes. Traditional IE datasets include MAGPIE (Haagsma

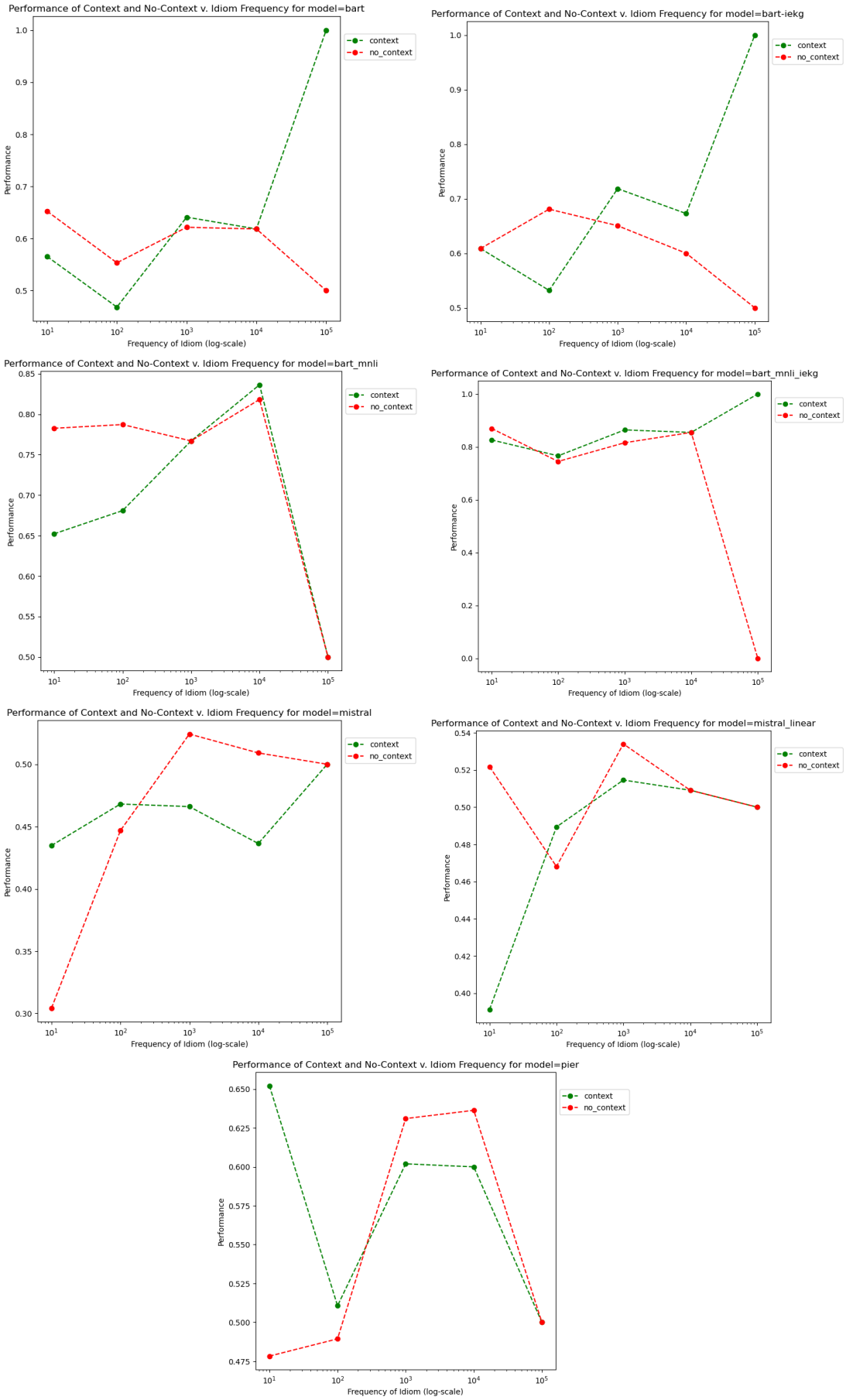

Figure 6: Visualizations of the context and no-context performance splits versus the idiom frequency for all models on a subset of the IMPLI dataset (∼200 samples).

et al., 2020) and PIE (Zhou et al., 2021), as well as older datasets such as SemEval5b (Korkontzelos et al., 2013), which all provide the backbone for many IE-related tasks. More recent datasets include IMPLI (Stowe et al., 2022) and Figurative-NarrativeBenchmark (Chakrabarty et al., 2022a), amongst others (Chakrabarty et al., 2022b; Tayyar Madabushi et al., 2021). Structured knowledge bases have also been proposed to help improve PTLMs' IE capabilities, most recently with IEKG (Zeng et al., 2023). Other approaches examine and improve upon IE embedding representations (Zeng and Bhat, 2022, 2023; Tan and Jiang, 2021; Adewumi et al., 2022; Liu and Hwa, 2017; Škvorc et al., 2022), which lead to improvements in downstream tasks. Different training methods may also be proposed for better IE representations (Hashimoto and Tsuruoka, 2016) or for downstream performance (Zhou et al., 2023a,b). Our approach is not directly concerned with modifying or improving the IE reasoning process, but attempting to understand how these models utilize context in the presence of IEs. We believe that our results can be complementary to current efforts to improve IE reasoning, and can serve as an informative aid for future research.

4.3 Transformer Attention Utilization

Previous work has usually taken an alternative avenue when investigating the inference process, specifically through observing a model's attention mechanism. Indeed, there is a long debate about the explanatory abilities of attention (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Bibal et al., 2022), and whether researchers can learn anything about the inference process by examining attention weights and maps. The consensus appears to be that attention does reflect the inference procedure in some manner, but the specific effects are unknown (Serrano and Smith, 2019). Unlike prior research, our work examines context utilization through a data-driven perspective, by altering the data samples themselves and keeping models fixed. In this manner, we can ensure that observed changes in performance arise from a direct result of how we perturbed the data. We believe our context utilization studies can provide a complementary viewpoint on interpreting the reasoning process.

5 Conclusions

In this paper, we refute the hypothesis that PTLMs utilize the context when reasoning in the presence of IEs. We extend upon the work of others, and demonstrate that not only are models not sensitive to contexts, but in fact, models do not need any context at all. For difficult IE reasoning tasks, we demonstrate this peculiarity in the idiomatic NLI and narrative continuation acceptability tasks, where we see performance gains despite removing context words from the data samples. We perform our experiments across a range of model complexities from 140M to 7B parameters. Our results, which showcase poor context utilization for general PTLMs, demonstrate a need for a closer examination of the inference process. Future work should examine additional factors within the model or datasets that may cause low context utilization, methods to improve context utilization, and also extend our analysis to other domains within NLU.

6 Limitations

The scope of our paper focuses on the reasoning behavior of PTLMs for English IE reasoning tasks. We examined context utilization only for the IE setting, as opposed to generally as a whole. This is because in non-IE settings the formulation becomes less well-defined. Indeed, in general usage it is less clear what should constitute as context in those scenarios. Are we talking about the most informative words in a sample, and if so, how would one go about defining informativeness (since this notion will vary depending on the downstream application)? While there have been some attempts at formulating this notion (Montariol et al., 2019; Harutyunyan et al., 2021; Schick and Schütze, 2019), this is an open question that we believe to be best left for future work.

Another limitation of our study was the lack of compute resources, which inhibited us from fully fine-tuning the largest PTLMs. However, we attempted to address this issue by including the newest state-of-the-art models such as Mistral-7B (Jiang et al., 2023). However, with our current resources, we were unable to fully fine-tune the model for our task. Nevertheless, we believe its out-of-the-box performance, as well as its performance with a fine-tuned classification head, should serve as an important baseline indicator of why even state-of-the-art PTLMs remain inadequate in the IE reasoning setting. Of course, future work would also showcase our results for a larger number of state-of-the-art PTLMs in addition to Mistral.

Finally, for our study we utilized classification-style tasks in English. Due to the scarce nature

of IE datasets, we stuck to English datasets. We believe that due to the universal nature of IE reasoning required for any IE-related task, our results are sufficient to demonstrate the inconsistencies between expected PTLM reasoning behavior and their actual behavior. Nonetheless, it would be beneficial for future work to include a wider suite of tasks and languages, and investigate the context utilization phenomenon for NLU tasks as a whole.

7 Ethics Statement

Note that we do not create any new models or datasets in our work, nor did we collect any data from users, as we are focused instead on analyzing current model behavior on downstream applications. Our studies are conducted on IE reasoning tasks for transformer models, which are intended to measure semantic analysis. We made sure to utilize only publicly available and peer-reviewed datasets to ensure quality control and safeguard against inputting data that contains toxicity or sensitive information. The results from these studies are intended to help guide future research and further analyses into how transformer models utilize the context.

In our paper, we included several studies into dataset quality, and investigated whether artifacts in the data could be a potential explanation for poor context utilization. Thus, our results do investigate various potential artifacts in the dataset, specifically how it influences the model inference process. We do not explicitly study the biases present however, as we were interested specifically in the context utilization of these models in IE reasoning scenarios.

It is imperative that the findings from our paper are not misused for models and datasets deployed in the real-world. Techniques such as context removal and context shuffling should never be implemented in actuality, as these methods only serve to provide an insight and analysis towards a better understanding of how models perform IE reasoning. They should instead be used to help guide future analyses and studies into model behavior.

Finally, we did not perform extensive pretraining or fine-tuning of models that would result in a sizeable environmental impact. Please refer to Appendix E for a description of our compute resources and the scale of our experiments.

Acknowledgements

This research was supported in part by the National Science Foundation under Grant No. IIS 2230817.

References

Tosin Adewumi, Foteini Liwicki, and Marcus Liwicki. 2022. Vector representations of idioms in conversational systems. *Sci*, 4(4).

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

C. Cacciari and P. Tabossi. 2014. *Idioms: Processing, Structure, and Interpretation*. Taylor & Francis.

Kate Cain, Andrea S. Towse, and Rachael S. Knight. 2009. The development of idiom comprehension: An investigation of semantic and contextual processing skills. *Journal of Experimental Child Psychology*, 102(3):280–298.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maria Chiara Levorato, Barbara Nesi, and Cristina Cacciari. 2004. Reading comprehension and understanding idiomatic expressions: A developmental study. *Brain and Language*, 91(3):303–314.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2023. Language models show human-like content effects on reasoning tasks.

Mark Davies. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, 25(4):447–464.

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.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

Hrayr Harutyunyan, Alessandro Achille, Giovanni Paolini, Orchid Majumder, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. 2021. Estimating informativeness of samples with smooth unique information. In *International Conference on Learning Representations*.

Kazuma Hashimoto and Yoshimasa Tsuruoka. 2016. Adaptive joint learning of compositional and non-compositional phrase embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 205–215, Berlin, Germany. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Changsheng Liu. 2019. Toward robust and efficient interpretations of idiomatic expressions in context.

Changsheng Liu and Rebecca Hwa. 2017. Representations of context in recognizing the figurative and literal usages of idioms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Syrielle Montariol, Aina Garí Soler, and Alexandre Allauzen. 2019. Exploring sentence informativeness. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019. Volume II : Articles courts*, pages 303–312, Toulouse, France. ATALA.

R. Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford Studies in Lexicography. Clarendon Press.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? 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 851–864, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Andrew Ortony, Diane L. Schallert, Ralph E. Reynolds, and Stephen J. Antos. 1978. Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of Verbal Learning and Verbal Behavior*, 17(4):465–477.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying grammatical role, BERT doesn't care about word order... except when it matters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–643, Dublin, Ireland. Association for Computational Linguistics.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Timo Schick and Hinrich Schütze. 2019. Attentive mimicking: Better word embeddings by attending to informative contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 489–494, Minneapolis, Minnesota. Association for Computational Linguistics.

Patrick Schramowski, Cigdem Turan-Schwiewager, Nico Andersen, Constantin Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4:258–268.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

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.

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

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.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Meiling Wang, Min Xiao, Changliang Li, Yu Guo, Zhixin Zhao, and Xiaonan Liu. 2019. STAC: Science toolkit based on Chinese idiom knowledge graph. In *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*, pages 57–61, Minneapolis, Minnesota. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2023. Natural language reasoning, a survey.

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

Ziheng Zeng and Suma Bhat. 2022. Getting BART to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

Ziheng Zeng and Suma Bhat. 2023. Unified representation for non-compositional and compositional expressions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11696–11710, Singapore. Association for Computational Linguistics.

Ziheng Zeng, Kellen Cheng, Srihari Nanniyur, Jianing Zhou, and Suma Bhat. 2023. IEKG: A commonsense knowledge graph for idiomatic expressions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14243–14264, Singapore. Association for Computational Linguistics.

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

Jianing Zhou, Ziheng Zeng, and Suma Bhat. 2023a. CLCL: Non-compositional expression detection with contrastive learning and curriculum learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 730–743, Toronto, Canada. Association for Computational Linguistics.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2023b. Non-compositional expression generation based on curriculum learning and continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4320–4335, Singapore. Association for Computational Linguistics.

Tadej Škvorc, Polona Gantar, and Marko Robnik-Šikonja. 2022. Mice: Mining idioms with contextual embeddings. *Knowledge-Based Systems*, 235:107606.

A Full Task Results

Note that any statistically significant results are noted in the caption for each figure/table. Additionally, we also report the confidence level.

A.1 Effect of Context Removal

Tables 8 and 9 present the complete breakdown of our results for each model. We see that general PTLMs typically gain performance without the context, while the opposite is usually true for IE-aware models (for the IMPLI dataset). Recall that while models appear to lose performance with context removal on the FigurativeNarrativeBenchmark dataset, we know from Table 1 that these models can still perform better with partial removal of the context. Interestingly, we note that the performance across classes for the IMPLI dataset appears more even after removing the context, whilst the opposite is true for the FigurativeNarrativeBenchmark dataset.

A.2 FigurativeNarrativeBenchmark Percentage Removal Results

We provide the full results for the FigurativeNarrativeBenchmark dataset with a percentage removal of the context words. The results for BART, BART-IEKG, Mistral, Mistral-FT, and PIER+ can be found in Tables 10, 11, 12, 13, and 14, respectively. Importantly, IE-aware models must typically retain a larger portion of the original context than general purpose PTLMs.

A.3 Effect of Shuffled Context

We provide the comprehensive performance on each dataset when the context has been shuffled, seen from Table 16. As we saw, for the most part performance increases in the shuffled context setting for general PTLMs, with BART-MNLI being an exception. After shuffling, general PTLMs perform more evenly across each class, while for the FigurativeNarrativeBenchmark this is dependent on the model.

A.4 Replacing IEs with Randomly Generated Strings

Here we see the full table of results when replacing the IE in the dataset with a randomly generated string. For the IMPLI dataset, from Table 17, not only do general PTLMs like BART, BART-MNLI, Mistral, and Mistral-FT perform better without the context, but their performance splits across each class somehow become more balanced without context. While most models exhibit a performance drop for the FigurativeNarrativeBenchmark, Mistral surprisingly displays performance gain, as seen in Table 18. It is also interesting to note how the performance across each class changes without the context, with most models generally heavily biased towards predicting that a continuation is acceptable. This runs contradictory with the IMPLI results, where models tend to perform more evenly across classes without context. This behavior represents a potentially interesting avenue for further investigation.

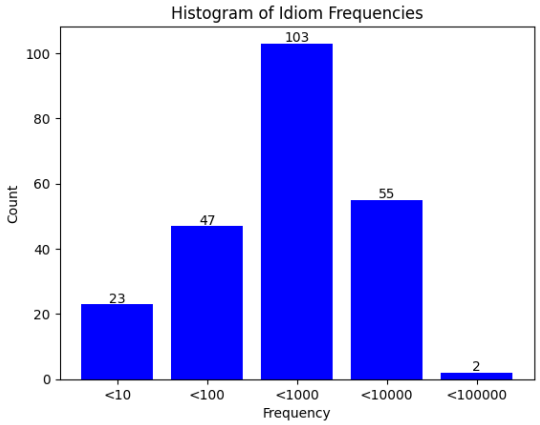

Figure 7: A histogram of frequency counts for idioms found in the IMPLI dataset (∼200 samples).

B Dataset Exploration

In this section we take a closer examination of various dataset properties, and observe whether there may be additional artifacts and noise present that could bias the model away from utilizing the context correctly. Note that we present these results for the IMPLI dataset, as both the FLUTE and FigurativeNarrativeBenchmark datasets do not exhibit these imbalances between the labels.

From Table 5, we see that generally the non-entailment samples outnumber the entailment samples at all settings, when observing the length of the idiom (in characters). These trends are also seen in Table 6, where we compared how the sample labels were distributed versus the idiom multiplicity, i.e. the number of times that idiom appears in the dataset. Additionally, Table 4 illustrates that even narrative length is skewed towards the non-entailment samples. This in and of itself is quite peculiar, given that in our studies, entailment performance tends to be better than non-entailment performance regardless of the task. These results also conflict with Table 3, which suggest that dataset

Label	≤ 10	≤ 20	≤ 30	≤ 40	≤ 50	≤ 60	≤ 70	≤ 80	≤ 90	≤ 100	>100
E	45	145	148	103	41	27	12	3	2	1	1
NE	64	174	181	117	50	27	11	3	1	1	0

Table 4: A comparison of the distribution of labels on the IMPLI dataset versus the sample length (in terms of characters).

Label	≤ 10	≤ 20	≤ 30	≤ 40
E	103	382	42	1
NE	118	451	59	1

Table 5: A comparison of the distribution of labels on the IMPLI dataset versus the idiom length (in terms of characters).

Label	≤ 2	≤ 4	≤ 6	≤ 8	≤ 10	>10
E	126	178	111	32	8	12
NE	122	242	159	41	11	17

Table 6: A comparison of the distribution of labels on the IMPLI dataset versus the idiom multiplicity. Note that idiom multiplicity refers to the number of times the idiom appears in the dataset. For example, an idiom with multiplicity 2 means that two different samples contained that same idiom.

Model	LR	Batch Size	Weight Decay
BART	1.37e-5	32	0.01
BART-IEKG	1.31	32	0.01
Mistral	N/A	8	N/A
Mistral-FT	2e-5	8	0.01
PIER+	2e-5	32	0.01

Table 7: A listing of the specific set of hyperparameters utilized by each model for the FigurativeNarrativeBenchmark dataset (to two significant figures). Note that LR denotes learning rate.

artifacts cannot adequately explain general PTLMs' low context utilization in these scenarios. It may be the case that there are other, less obvious, spurious features that the model may be influenced by when making inference.

C FLUTE Dataset

In this section we performed our context utilization experiments on the FLUTE dataset (Chakrabarty et al., 2022b). From Table 15, it seems that most models do exhibit a drop in performance without context, with the notable exception being Mistral-FT, which improves by 3.20%. However, note that the tiny size of the dataset, which only contains 1,768 training samples and 250 test samples, indicates that these results may not be generalizable. Nonetheless, the peculiar performance of models such as Mistral-FT indicates that there may still be sup-optimal context utilization exhibited by these models.

D Hyperparameter Setup

For the IMPLI experiments, we ran all non-Mistral models with the same set of hyperparameters as found in (Zeng et al., 2023), in order to provide a direct and easy comparison between different context settings. Specifically, we utilized a learning rate of 2e-5, with a batch size of 32, and a weight decay of 0.01. For the Mistral models on the IMPLI dataset, we simply changed the batch size to 8, in order to accommodate our compute requirements. Note that we used these same parameters for our results on the FLUTE dataset as well, with the exception of BART, which used a batch size of 8 instead of 32. Peculiarly, we found that BART fails with a batch size of 32, as it would always predict entailment (this is especially strange since FLUTE is a perfectly balanced dataset across all splits).

For the FigurativeNarrativeBenchmark dataset, we chose specialized optimal hyperparameters for each non-Mistral model, as we found that certain models were liable to degenerate into a majority classifier under particular hyperparameter settings (i.e. predicting 100% on one class and 0% on another class), which is unexpected, given that Figura-

	With Context				Without Context			
Model	Acc.	E Acc.	NE Acc.	ANT Acc.	Acc.	E Acc.	NE Acc.	ANT Acc.
BART	57.22%	94.89%	46.85%	11.20%	61.02%	82.01%	70.47%	24.80%
BART-IEKG	64.56%	97.35%	52.36%	26.67%	62.58%	86.36%	68.50%	25.07%
BART-MNLI	77.87%	96.78%	34.65%	80.53%	81.76%	87.50%	53.94%	92.53%
BART-MNLI-IEKG	81.16%	96.59%	65.75%	69.87%	81.50%	84.47%	76.77%	80.53%
Mistral	46.50%	81.06%	15.75%	18.67%	44.51%	74.43%	16.93%	21.07%
Mistral-FT	49.09%	82.01%	18.90%	23.20%	52.38%	34.47%	61.81%	71.12%
PIER+	56.78%	89.39%	53.54%	13.07%	54.54%	85.23%	46.06%	17.07%

Table 8: A comparison of different models on the IMPLI dataset with and without context. Note that E stands for entailment, NE stands for non-entailment, and ANT stands for antonym non-entailment. The performance gains for BART and BART-MNLI are statistically significant (90% confidence).

	With Full Context			With Sentential Context			With No Context		
Model	Acc.	CC Acc.	IC Acc.	Acc.	CC Acc.	IC Acc.	Acc.	CC Acc.	IC Acc.
BART	77.20%	80.03%	74.38%	75.68%	84.50%	66.86%	66.34%	79.12%	53.63%
BART-IEKG	77.85%	78.60%	77.11%	76.07%	82.49%	69.71%	61.96%	92.41%	31.45%
Mistral	49.09%	76.91%	21.27%	48.51%	79.31%	17.77%	50.26%	96.37%	4.15%
Mistral-FT	62.78%	66.34%	59.39%	62.22%	56.81%	67.57%	50.16%	99.09%	1.23%
PIER+	65.08%	66.80%	63.36%	62.09%	73.22%	51.04%	60.54%	74.90%	46.24%

Table 9: A comparison of different models on the FigurativeNarrativeBenchmark dataset with full context, with sentential context only, and removing all context. Note that CC stands for the correct class, and IC for the incorrect class. The performance drops for BART-IEKG and PIER+ with only sentential context are statistically significant (90% confidence). The performance drops without any context are all statistically significant (99% confidence).

Acc.	CC Acc.	IC Acc.	Percentage Kept
77.27%	80.22%	74.32%	90%
77.66%	80.54%	74.77%	80%
77.76%	80.87%	74.71%	70%
77.79%	81.13%	74.45%	60%
76.52%	81.13%	71.92%	50%
76.65%	81.78%	71.53%	40%
75.81%	83.33%	68.22%	30%
71.85%	82.43%	61.22%	20%
66.47%	80.16%	52.79%	10%

Table 10: A comparison of performance on the FigurativeNarrativeBenchmark dataset for the BART-large model.

Acc.	CC Acc.	IC Acc.	Percentage Kept
78.02%	78.99%	77.04%	90%
77.27%	78.60%	76.01%	80%
77.59%	79.44%	75.88%	70%
77.04%	79.51%	74.71%	60%
76.46%	80.80%	74.12%	50%
76.72%	80.54%	72.83%	40%
75.71%	82.23%	69.20%	30%
71.98%	81.97%	62.00%	20%
64.69%	82.68%	46.69%	10%

Table 11: A comparison of performance on the FigurativeNarrativeBenchmark dataset for the BART-IEKG model.

Acc.	CC Acc.	IC Acc.	Percentage Kept
49.35%	77.89%	20.82%	90%
48.74%	77.76%	19.65%	80%
49.38%	79.38%	19.33%	70%
49.19%	80.48%	18.03%	60%
49.22%	80.35%	17.96%	50%
49.25%	80.29%	18.29%	40%
49.29%	81.52%	17.06%	30%
49.64%	82.88%	16.41%	20%
48.80%	87.48%	10.12%	10%

Table 12: A comparison of performance on the FigurativeNarrativeBenchmark dataset for the Mistral model.

Acc.	CC Acc.	IC Acc.	Percentage Kept
62.35%	65.37%	59.40%	90%
63.20%	65.82%	60.44%	80%
62.91%	65.11%	60.89%	70%
62.29%	63.04%	61.74%	60%
61.64%	60.70%	62.58%	50%
61.84%	58.43%	65.18%	40%
60.83%	57.52%	64.20%	30%
60.18%	63.55%	56.74%	20%
54.02%	73.48%	34.50%	10%

Table 13: A comparison of performance on the FigurativeNarrativeBenchmark dataset for the Mistral-FT model.

Acc.	CC Acc.	IC Acc.	Percentage Kept
65.50%	67.38%	63.62%	90%
64.69%	67.06%	62.32%	80%
64.27%	67.12%	61.54%	70%
64.23%	69.00%	59.47%	60%
63.62%	70.43%	56.87%	50%
63.88%	73.15%	54.60%	40%
62.52%	73.15%	51.88%	30%
62.22%	73.99%	50.45%	20%
60.02%	73.67%	46.30%	10%

Table 14: A comparison of performance on the FigurativeNarrativeBenchmark dataset for the PIER+ model.

tiveNarrativeBenchmark is a fully balanced dataset. These hyperparameters can be found in Table 7. We computed these hyperparameters by performing a grid search, constraining the learning rate between 1e-5 and 5e-5, choosing between a batch size of 4, 8, 16, 32, and 64.

Note that for all experiments, we set the random seeds to 42.

E Software & Model Implementation

Note that the models we utilized are mostly found and implemented in the Transformers library from Huggingface (Wolf et al., 2020). For a proprietary model such as PIER+, we utilized the code and implementation found in their paper and repository.

Most of our experiments used a single Nvidia A100 GPU, which had 80 GB of GPU memory. Note that as the focus of our paper was not on training/fine-tuning models or achieving state-of-the-art, the majority of our experiments were not compute intensive, and were capable of running within several hours.

Finally, we recognize that some of the techniques and methods we used in our study may yet prove fruitful for future studies and analyses. As a result, we have made our code publicly available, which implements our techniques.

	With Context			Without Context		
Model	Acc.	E Acc.	NE Acc.	Acc.	E Acc.	NE Acc.
BART	94.00%	92.80%	95.20%	88.00%	88.00%	88.00%
BART-IEKG	95.20%	95.20%	95.20%	91.60%	96.00%	87.20%
BART-MNLI	95.20%	92.80%	97.60%	92.80%	91.20%	94.40%
BART-MNLI-IEKG	97.60%	96.00%	99.20%	94.00%	94.40%	93.60%
Mistral	48.80%	95.20%	2.40%	47.20%	84.80%	9.60%
Mistral-FT	58.40%	55.20%	61.60%	62.00%	45.60%	78.40%
PIER+	73.60%	80.00%	67.20%	74.00%	84.80%	63.20%

Table 15: A comparison of different models on the FLUTE dataset. Note that E stands for entailment samples, and NE for the non-entailment samples. The performance drops for BART and BART-MNLI-IEKG are statistically significant (95% confidence).

	IMPLI				FigurativeNarrativeBenchmark		
Model	Acc.	E Acc.	NE Acc.	ANT Acc.	Acc.	CC Acc.	IC Acc.
BART	60.41%	91.48%	48.43%	24.53%	70.01%	74.25%	65.76%
BART-IEKG	62.58%	95.27%	52.78%	27.20%	69.84%	70.30%	69.39%
BART-MNLI	69.32%	95.83%	33.46%	78.40%	N/A	N/A	N/A
BART-MNLI-IEKG	75.11%	93.75%	64.96%	68.27%	N/A	N/A	N/A
Mistral	47.10%	70.45%	16.93%	19.47%	50.32%	60.44%	40.21%
Mistral-FT	49.70%	60.42%	24.41%	26.40%	56.03%	36.32%	75.75%
PIER+	55.49%	87.31%	52.36%	11.73%	61.71%	55.51%	68.03%

Table 16: A comparison of different models' performance when the context is shuffled. Note that E stands for entailment, NE stands for non-entailment, and ANT stands for antonym non-entailment (IMPLI dataset). Note that CC stands for the correct class, and IC for the incorrect class (FigurativeNarrativeBenchmark dataset). The performance drops for BART-MNLI and BART-MNLI-IEKG on IMPLI are statistically significant (99% confidence). All performance drops on FigurativeNarrativeBenchmark are significant (99% confidence).

	With Context				Without Context			
Model	Acc.	E Acc.	NE Acc.	ANT Acc.	Acc.	E Acc.	NE Acc.	ANT Acc.
BART	58.25%	90.34%	54.33%	15.73%	59.03%	75.57%	71.65%	27.20%
BART-IEKG	59.55%	91.10%	50.39%	21.33%	57.22%	83.90%	66.14%	13.33%
BART-MNLI	55.66%	83.33%	21.65%	39.73%	57.39%	55.68%	53.54%	62.40%
BART-MNLI-IEKG	62.49%	78.22%	57.87%	43.47%	60.50%	65.53%	75.98%	42.93%
Mistral	47.10%	80.30%	16.54%	21.33%	48.06%	69.70%	26.77%	32.00%
Mistral-FT	48.57%	88.26%	15.35%	15.20%	53.67%	42.99%	63.78%	61.87%
PIER+	57.56%	81.44%	64.17%	19.47%	55.66%	75.95%	61.02%	23.47%

Table 17: A comparison of different models on the IMPLI dataset. Note that E stands for entailment, NE stands for non-entailment, and ANT stands for antonym non-entailment. The IEs in the dataset have all been replaced with randomly generated strings. The performance gain for Mistral-FT is statistically significant (95% confidence).

	With Full Context			With Sentential Context			With No Context		
Model	Acc.	CC Acc.	IC Acc.	Acc.	CC Acc.	IC Acc.	Acc.	CC Acc.	IC Acc.
BART	71.04%	73.35%	68.74%	66.54%	82.10%	50.97%	54.60%	90.34%	18.87%
BART-IEKG	69.91%	73.54%	66.28%	65.99%	80.61%	51.36%	52.14%	95.27%	9.01%
Mistral	48.67%	77.82%	19.46%	48.90%	81.39%	16.34%	49.84%	95.14%	4.60%
Mistral-FT	59.24%	58.75%	59.79%	57.26%	45.53%	69.13%	50.16%	98.90%	1.43%
PIER+	63.94%	66.21%	61.61%	60.44%	73.93%	46.95%	58.11%	75.29%	40.99%

Table 18: A comparison of different models on the FigurativeNarrativeBenchmark dataset. Note that CC stands for the correct class, and IC for the incorrect class. The IEs in the dataset have all been replaced with randomly generated strings. The performance drops for BART, BART-IEKG, and PIER+ are all statistically significant (99% confidence).