Does GPT-3 Produce Less Literal Translations?

Anonymous ACL submission

Abstract

Large Language Models (LLMs) such as GPT-3 have emerged as general purpose language models capable of addressing any natural language generation or understanding task. On the task of Machine Translation (MT), multiple works have investigated few-shot prompting mechanisms to elicit better translations from LLMs. However, there has been relatively little investigation on how such translations qualitatively differ from the translations generated by Neural Machine Translation (NMT) models. In this work, we focus on translation literalness as a property to better differentiate the characteristics of translations from LLMs in the GPT family. We show that E-X translations from GPT-3, even though achieving similar (or better) quality estimates than NMT models, incur a significantly higher number of unaligned source words as well as higher non-monotonicity, which indicates a bias towards less literal translations. We show that this effect also becomes apparent in human evaluations of translation literalness. We further investigate this hypothesis by conducting experiments on sentences with idioms (both natural as well as synthetic), wherein the desired translations themselves admit greater figurativeness.

1 Introduction

Contrary to traditional sequence-to-sequence Neural Machine Translation (NMT) models trained on parallel corpora (Sutskever et al., 2014; Vaswani et al., 2017), the translation abilities in Large Language Models (LLMs) is an emergent phenomenon that arises in the absence of any explicit supervision for the task during training. Despite training only on a language modeling objective, without any aligned parallel data, LLMs such as GPT-3 or PaLM (Brown et al., 2020; Chowdhery et al., 2022) achieve close to state-of-the-art translation performance under few-shot prompting (Vilar et al., 2022). As such, a natural question to ask is: *at the same levels of quality, how do the translations from LLMs qualitatively differ from the translations produced using NMT systems?*

Source: He survived by the skin of his teeth .
NMT: Il a survécu par la peau de ses dents .
GPT-3: Il a survécu de justesse .

Figure 1: An example illustrating the difference in translation literalness between NMT models and GPT-3 (text-davinci-002). GPT-3 produces a less literal translation of the source ('He barely survived.'). Further, generating a word-by-word alignment between the source and LLM output leaves the source word '*skin*' unaligned.

We posit that the property of translation literalness (an example is presented in Figure 1[1]) could serve as a useful differentiation axis. Besides being an interesting scientific inquiry, investigating the differences in translation literalness between LLMs and NMT systems could have direct applications. For example, determining NMT and LLM differences on this axis would be useful for applying them selectively on inputs wherein desired translations themselves admit less (or more) literalness. In this work, we explore these questions quantitatively. Our contributions are as follows:

1. We quantitatively explore the differences in translation literalness between translations produced by GPT-3 and NMT systems. We demonstrate that even when translations from GPT-3 achieve higher quality estimates than NMT systems, they exhibit a greater bias towards non-literalness for E-X translations.

2. Through controlled experiments on both natural and synthetic sentences, we demonstrate that sentences containing idioms (figurative language) represents a partition of the input space which could directly benefit from the increased non-literal expressivity of the E-X translations produced by GPT-3.

[1]Both Bing Translator and Google Translator produce the literal translation (public APIs accessed on Jan 10, 2023).

2 Quantifying Translation Literalness

Experiment: We compare the state-of-the-art NMT systems against the most capable publically accessible GPT-3 models across a number of measures designed to elicit differences in translation literalness. We conduct both automatic metric-based as well as human evaluations. We explain the evaluation and experimental details below.

Datasets: We use the official WMT21 En-De, De-En, En-Ru and Ru-En News Translation test sets for evaluation (Barrault et al., 2021). Appendix A also presents the results for En-Cs.

Measures of Quality: We use COMET-QE (Rei et al., 2020) as the Quality Estimation (QE) measure (Fomicheva et al., 2020) to quantify the fluency and adequacy of translations. Using QE as a metric presents the advantage that it precludes the presence of any reference bias, which has been shown to be detrimental in estimating the LLM output quality in related sequence transduction tasks (Goyal et al., 2022). On the other hand, QE as a metric suffers from an apparent blindness to copy errors (i.e., cases in which the model produces a translation in the source language). To mitigate this, we apply a language identifier on the translation output and set the translation to null if the translation language is the same as the source language. Therefore, we name this metric COMET-QE + LID.

Measures of Translation Literalness: There do not exist any known metrics with high correlation geared towards quantifying translation literalness. We propose and consider two automatic measures at the corpus-level:

1. **Unaligned Source Words (USW)**: Two translations with very similar fluency and adequacy could be differentiated in terms of their literalness by computing word to word alignment between the source and the translation, then measuring the number of source words left unaligned. When *controlled for quality*, a less literal translation is likely to contain more words that do not align with the words in the source sentence (e.g., in Figure 1).

2. **Translation Non-Monotonicity (NM)**: Another measure of literalness is how closely the translation tracks the word order in the source. We use the non-monotonicity metric proposed in Schioppa et al. (2021), which computes the deviation from the diagonal in the word to word alignment as the non-monotonicity measure. This can also be interpreted as (normalized) alignment crossings, which has been shown to correlate with translation non-literalness (Schaeffer and Carl, 2014).

Note that the above two measures make use of complementary information from alignments – NM is computed only using the information from aligned source words. Therefore, to adjudicate literalness we use the two metrics in combination. We use the multilingual-BERT (Devlin et al., 2019) based awesome-aligner (Dou and Neubig, 2021), a state-of-the-art aligner to obtain the word to word alignments between the source and the translation.

Systems Under Evaluation: We experiment with the below four systems (NMT and LLMs):

1. WMT-21-SOTA: The Facebook multilingual system (Tran et al., 2021) won the WMT-21 News Translation task (Barrault et al., 2021), and thereby represents the strongest NMT system on the WMT'21 test sets.

2. Bing-Translator: Bing-Translator represents one of the strongest publically available commercial NMT systems (Raunak et al., 2022).

3. text-davinci-002: The text-davinci-002 model is an instruction fine-tuned model in the GPT-3 family (Brown et al., 2020). It represents one of the strongest publically accessible LLMs (Liang et al., 2022).

4. text-davinci-003: The text-davinci-003 model further improves upon text-davinci-002 for many tasks[2] (Liang et al., 2022). For both the GPT-3 models we prompt using eight randomly sampled examples from the corresponding WMT-21 development set.

Results: We compare the performance of the four systems on the four WMT-21 test sets. Figure 2 shows the results of this comparison. A key observation is that while the GPT-3 based translations achieve superior COMET-QE+LID scores than Bing Translator across the language pairs (except En-Ru), they also consistently obtain considerably higher number of unaligned source words. This result holds for the comparison between the WMT-21-SOTA and GPT-3 systems as

[2]LLMs: https://beta.openai.com/docs/models/

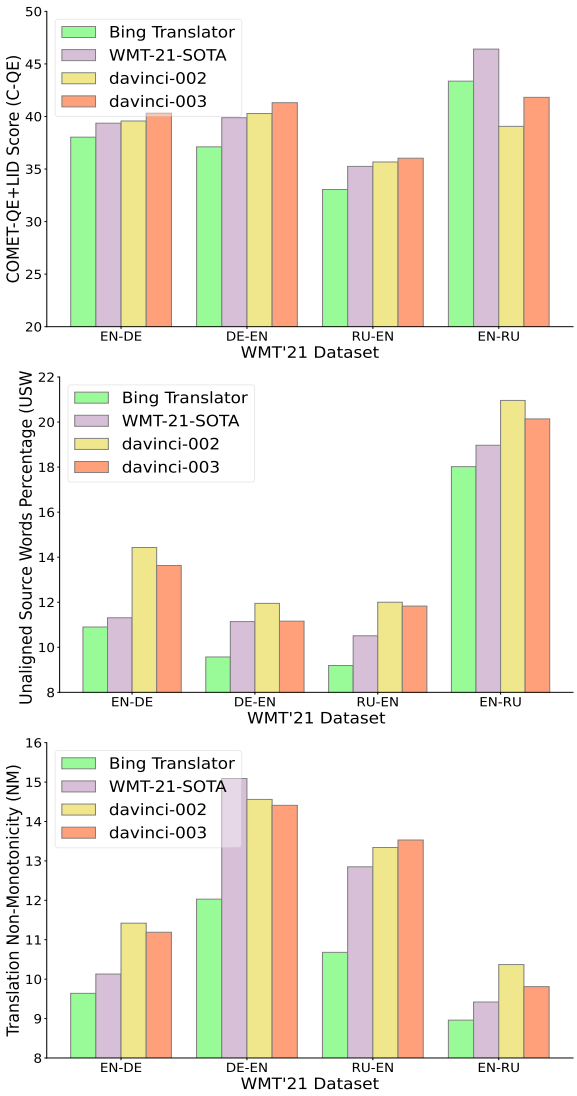

Figure 2: Measurements: While the NMT Systems and GPT-3 models achieve similar COMET-QE+LID Scores (Top), there exists a significant gap in the number of unaligned source words (USW) across the datasets (Bottom). Further, GPT-3 translations obtain higher non-monotonicity scores for E-X translations (Middle).

well. Further, GPT-3 translations also consistently show higher non-monotonicity for E-X translations. However, this is not the case for translations into English, wherein the multilingual WMT-21-SOTA system obtains very close non-monotonicity measurements. The *combined interpretation* of these measurements suggests that GPT-3 does produce less literal translations for E-X translations.

Human Evaluation: We further verify the conclusion from the results in Figure 2 by conducting human evaluation for En-De and En-Ru over 100 random samples from the test set using crowdsourced (majority vote of 3 annotations), native speaker as well as expert (linguist) annotators (details in Appendix B). The results, presented in Table 1, show that predominantly, the annotators rate the GPT-3 translations as less literal.

Lang-Pair	Expert	Native	Crowdsourced
En-De	+14 %	-3 %	+6 %
En-Ru	-4 %	+6 %	+9 %

Table 1: Human Evaluation Results with different annotators: the numbers represent the difference in percentage of translations that were labelled as less literal for text-davinci-002, when compared to Bing Translator.

3 Effects On Figurative Compositionality

In this section, we explore whether the less literal nature of E-X translations produced by GPT-3 could be leveraged to generate higher quality translations for certain inputs. We posit the phenomenon of composing the non-compositional meanings of idioms (Dankers et al., 2022) with the meanings of the compositional constituents within a sentence as figurative compositionality. Thereby, a model exhibiting greater figurative compositionality would be able to abstract the meaning of the idiomatic expression in the source sentence and express it in the target language non-literally, either through a non-literal (paraphrased) expression of the idiom's meaning or through an equivalent idiom in the target language. Note that greater non-literalness does not imply better figurative compositionality. Non-literalness in a translation could potentially be generated by variations in translation which does not conform to the *desired* figurative translation.

Experiment: In this section, we quantify the differences in the translation of sentences with idioms between traditional NMT systems and GPT-3. There do not any existing English-centric parallel corpora dedicated to sentences with idiomatic expressions. Therefore, we experiment with monolingual (English) sentences with idioms. The translations are generated with the same prompt in section 2. The datasets with *natural idiomatic sentences* are enumerated below:

MT System	C-QE ↑	USW ↓	NM ↓
Bing Translator	21.46	13.70	9.63
WMT'21 SOTA	23.25	14.47	10.21
text-davinci-002	**23.67**	**18.08**	**11.39**

Table 2: Natural Idiomatic Sentences: Combined Results over MAGPIE, EPIE, PIE (5712 sentences).

MAGPIE: MAGPIE (Haagsma et al., 2020) contains a set of sentences annotated with their id-

iomaticity, alongside a confidence score. We use the sentences pertaining to the news domain which are marked as idiomatic with cent percent annotator confidence (totalling 3666 sentences).

EPIE: EPIE (Saxena and Paul, 2020) contains idioms, alongside a representative sentence that demonstrates it usage. We use the sentences available for static idioms (totalling 1046 sentences).

PIE: The PIE dataset (Zhou et al., 2021) contains a set of idioms along with its idiomatic usage. We randomly sample 1K sentences from the corpus.

Results: The results are presented in Table 2. We find that text-davinci-002 produces better quality translations than the WMT'21 SOTA system, with greater number of unaligned words as well as with higher non-monotonicity.

Further Analysis: Note that a direct attribution of the gain in translation quality to better translation of idioms specifically is challenging. Further, similarity-based quality metrics such as COMET-QE themselves might be penalizing non-literalness. Therefore, while a natural monolingual dataset presents a useful testbed for investigating figurative compositionality abilities, an explicit comparison of figurative compositionality between the systems is very difficult. Therefore, we also conduct experiments on synthetic data, where we explicitly control the fine-grained attributes of the input sentences. We do this by allocating most of the variation among the input sentences to certain constituent expressions in synthetic data generation.

Control Experiments: We generate synthetic English sentences, each containing certain expressions, using GPT-3 text-davinci-002 in a zero-shot manner (prompt details are in appendix C). In each of the control experiments, we translate the synthetic English sentences to German.

Synthetic Dataset 1: We generate sentences containing expressions of three types, namely, named entities (e.g., 'Jessica Alba'), random descriptive phrases (e.g., 'large cake on plate') and idioms (e.g., 'a shot in the dark'). Expression sources as well as further data generation details are presented in appendix C. Note that idioms, unlike the other two expressions do admit a less literal translation as the desired translation. Results are in Table 3.

Synthetic Dataset 2: We generate sentences containing multiple idioms (varying from 1 to 4). The prompts & examples are presented in appendix C. The results are presented in Table 4.

Expression	C-QE ↑	USW ↓	NM ↓
Random Phrases	-2.45	+1.62	+0.14
Named Entities	-1.50	+0.81	+0.39
Idioms	**+5.90**	**+2.82**	**+1.95**

Table 3: Synthetic sentences with Idioms vs Synthetic sentences containing other expressions: The difference between GPT-3 (text-davinci-002) performance and NMT performance (Bing Translator) is reported.

Num Idioms	1	2	3	4
USW	17.58	18.39	18.28	18.99

Table 4: Synthetic sentences with multiple idioms (1-4): Increasing the number of idioms increases the number of unaligned source words in text-davinci-002 translations.

Results: Table 3 shows that the percentage of unaligned source words is highest in the case of idioms, followed by random descriptive phrases and named entities. The results are consistent with the hypothesis that GPT-3 produces less literal E-X translations, since named entities or descriptive phrases in a sentence would admit more literal translations as acceptable, while for idioms this assertion is not true. The significantly higher percentage of unaligned source words for idioms is despite the fact that GPT-3 obtains a much higher COMET-QE score in the case of translations of sentences with idioms. Similarly, the difference in non-monotonicity scores is also considerably higher for the case of idioms. These results present evidence that non-literalness in GPT-3 translations does conform to desired figurative translations, leading to better translations for sentences with idioms. Further, Table 4 shows that with increasing number of idioms in the synthetic sentences, the percentage of unaligned source words keeps increasing, another piece of evidence consistent with our hypothesis.

4 Summary and Conclusion

We investigated how the translations obtained through LLMs from the GPT family are qualitatively different by quantifying the property of translation literalness. We find that for E-X translations, there is a greater bias towards non-literalness in GPT-3 translations. We also show that this bias does conform to desired figurativeness in the translation of sentences containing idioms. We hope that our work leads to more explorations towards better characterizing the translations from LLMs.

5 Limitations

One of the main hindrances in our investigation of the hypothesis that GPT-3 produces less literal translations has been the problem of measurement. We rely on a combined interpretation of multiple measurements to investigate this hypothesis and its implications. This limits the extent to which we can make strong claims about the validity of the hypothesis, since in the absence of a highly correlated metric for translation literalness, it is hard to compare systems or claim that a clear validation exists for our hypothesis. We could only claim that our investigation indicates the presence of a bias towards non-literalness in GPT-3 translations, but a stronger result would have been preferred to further disambiguate GPT-3 translation characteristics.

References

Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

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.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. Salted: A framework for salient long-tail translation error detection.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions.

Moritz Schaeffer and Michael Carl. 2014. Measuring the cognitive effort of literal translation processes. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 29–37, Gothenburg, Sweden. Association for Computational Linguistics.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions.

In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.

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.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. 2020. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225.

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.

A Further WMT-21 Comparisons

The results on WMT-21 En-Cs are presented below. We find that for E-X translations, GPT-3 exhibits greater USW and NM whether it achieves higher quality estimates or not.

MT System	C-QE ↑	USW ↓	NM ↓
Bing Translator	54.65	19.07	8.41
text-davinci-002	50.97	21.12	9.34

Table 5: text-davinci-002 vs Bing Translator on En-Cs

B Human Annotation Interface

We use annotation interface in Figure 3, in which the annotators are asked to rate the two translations. The crowdsourced annotators were recruited using a crowdsourcing website, with 0.05 Dollars paid per annotation, while expert (linguist) and native speaker annotators were recruited in-house.

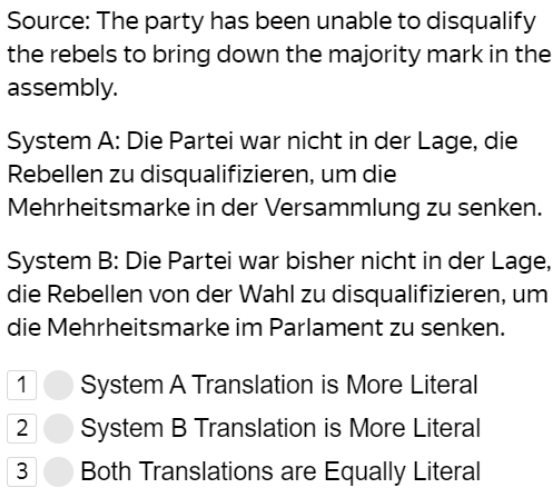

Figure 3: Annotation Interface Screenshot for Human Evaluation. The translations are randomized b/w the systems under evaluation to prevent biased evaluation.

C Synthetic Dataset Details

Synthetic Dataset 1: For each of the three expression types, 1K synthetic sentences are generated. Figures 4, 5, 6 present examples. The sources of the named entities and descriptive phrase expressions are MultiNERD (Tedeschi and Navigli, 2022) and PhraseCut (Wu et al., 2020) datasets respectively.

Synthetic Dataset 2: Figure 7 presents an example of the prompt as well as for a generated synthetic sentence containing two idioms.

Prompt: Q: Generate a sentence containing the idiom: a short fuse, in the form of a news article sentence. \n A:
Output: The man was known to have a short fuse, and often exploded into fits of anger without warning.

Figure 4: Synthetic Dataset 1: Example (Idiom)

Prompt: Q: Generate a sentence containing the entity: Wolfgang Amadeus Mozart, in the form of a news article sentence. \n A:
Output: Wolfgang Amadeus Mozart, a world renowned composer, died today at the age of 35.

Figure 5: Synthetic Dataset 1: Example (Entity)

Prompt: Q: Generate a sentence containing the phrase: white chair, in the form of a news article sentence. \n A:
Output: The white chair was found to be comfortable by the majority of the participants in the study.

Figure 6: Synthetic Dataset 1: Example (Phrase)

Prompt: Q: Generate a sentence using the two idioms: off the wall, claim to fame in the form of a news article sentence. \n A:
Output: The company's off-the-wall marketing campaign was its claim to fame.

Figure 7: Synthetic Dataset 2: Example (2 Idioms)