Do GPTs Produce Less Literal Translations?

Vikas Raunak **Arul Menezes** **Matt Post** **Hany Hassan Awadalla**

Microsoft Azure AI
Redmond, Washington
{viraunak,arulm,mpost,hanyh}@microsoft.com

Abstract

Large Language Models (LLMs) such as GPT-3 have emerged as general-purpose language models capable of addressing many natural language generation or understanding tasks. 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 differ *qualitatively* from the translations generated by standard Neural Machine Translation (NMT) models. In this work, we investigate these differences in terms of the literalness of translations produced by the two systems. Using literalness measures involving word alignment and monotonicity, we find that translations out of English (E→X) from GPTs tend to be less literal, while exhibiting similar or better scores on MT quality metrics. We demonstrate that this finding is borne out in human evaluations as well. We then show that these differences are especially pronounced when translating sentences that contain idiomatic expressions.

1 Introduction

Despite training only on a language-modeling objective, with no *explicit* supervision on aligned parallel data (Briakou et al., 2023), 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; Hendy et al., 2023). Work investigating the output of these models has noted that the gains in performance are not visible when using older surface-based metrics such as BLEU (Papineni et al., 2002a), which typically show large losses against NMT systems. This raises a question: How do these LLM translations differ *qualitatively* from those of traditional NMT systems?

We explore this question using the property of translation *literalness*. Machine translation systems have long been noted for their tendency to produce overly-literal translations (Dankers et al., 2022b), and we have observed anecdotally that LLMs seem less susceptible to this problem (Table 1). We investigate whether these observations can be validated quantitatively. First, we use measures based on word alignment and monotonicity to quantify whether LLMs produce less literal translations than NMT systems, and ground these numbers in human evaluation (§ 2). Next, we look specifically at idioms, comparing how literally they are translated under both natural and synthetic data settings (§ 3).

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 .

Table 1: An example where GPT-3 produces a more natural (non-literal) translation of an English idiom. When word-aligning these sentences, the source word *skin* remains unaligned for the GPT-3 translation.

Our investigations focus on the translation between English and German, Chinese, and Russian, three typologically diverse languages. Our findings are summarized as follows: (1) We find that translations from two LLMs from the GPT series of LLMs are indeed generally less literal than those of their NMT counterparts when translating *out* of English, and (2) that this is particularly true in the case of sentences with idiomatic expressions.

2 Quantifying Translation Literalness

We compare the state-of-the-art NMT systems against the most capable publicly-accessible GPT models (at the time of writing) across measures designed to capture 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

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1041–1050
July 9-14, 2023 ©2023 Association for Computational Linguistics

System	Source	Translation
MS	Time is running out for Iran nuclear deal, Germany says,	Die Zeit für das Atomabkommen mit dem Iran läuft ab, sagt Deutschland
GPT	Time is running out for Iran nuclear deal, Germany says,	Deutschland sagt, die Zeit für das iranische Atomabkommen läuft ab.
MS	You're welcome, one moment please.	Sie sind willkommen, einen Moment bitte.
GPT	You're welcome, one moment please.	Bitte sehr, einen Moment bitte.

Table 2: Translation examples with different Non-Monotonicity (NM) and Unaligned Source Word (USW) scores for MS-Translator (lower) and text-davinci-003 translations (higher) from the WMT-22 En-De test set, for illustration.

for evaluation (Barrault et al., 2021).

Measures of Quality We use COMET-QE[1] (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, COMET-QE as a metric suffers from an apparent blindness to copy errors (i.e., cases in which the model produces output in the source language) (He et al., 2022). To mitigate this, we apply a language identifier (Joulin et al., 2017) 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 unaligned source words (as suggested 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).

We use the multilingual-BERT-based awesome-aligner (Devlin et al., 2019; Dou and Neubig, 2021) to obtain the word to word alignments between the source and the translation. Table 2 presents an illustration of translations with different USW and NM scores[2], obtained from different systems.

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. Microsoft-Translator: MS-Translator is one of the strongest publicly 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 family (Brown et al., 2020). It represents one of the strongest publicly-accessible LLMs (Liang et al., 2022).

4. text-davinci-003: The text-davinci-003 model further improves upon text-davinci-002 for many tasks[3] (Liang et al., 2022).

For both the GPT models, we randomly select eight samples from the corresponding WMT-21 development set, and use these in the prompt as demonstrations for obtaining all translations from GPTs.

Results We compare the performance of the four systems on the WMT-21 test sets. Figure 1 shows the results of this comparison. A key observation is that while the GPT based translations achieve superior COMET-QE+LID scores than Microsoft Translator across the language pairs (except En-Ru), they

[1]wmt20-comet-qe-da

[2]Metrics: https://github.com/vyraun/literalness

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

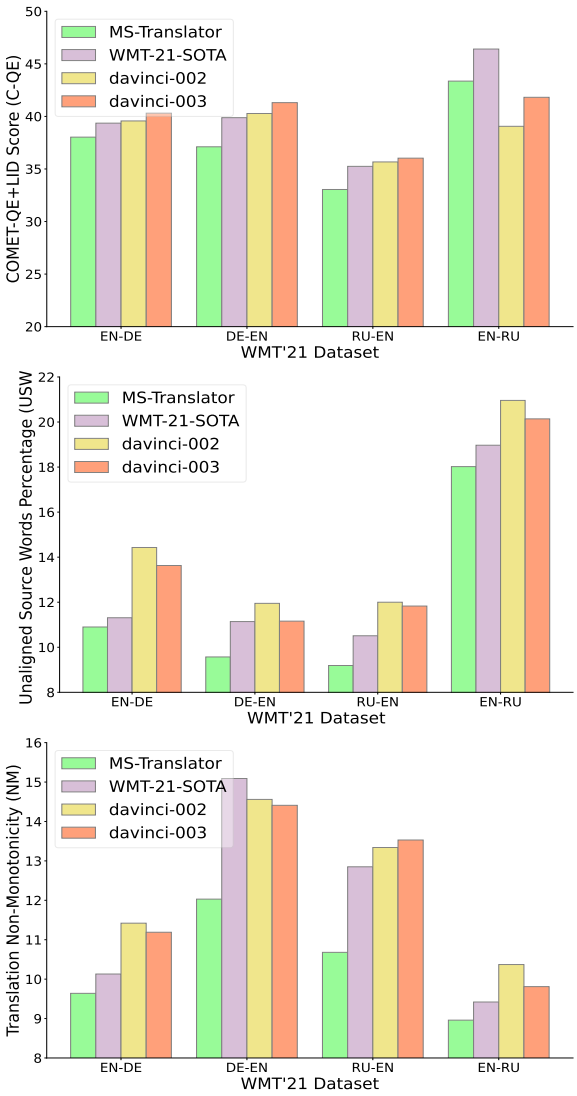

Figure 1: Measurements: The NMT Systems and GPT models achieve similar COMET-QE+LID Scores (Top), there exists a significant gap in the number of unaligned source words (USW) across the datasets (Bottom). Further, GPT translations obtain higher non-monotonicity scores for E-X translations (Middle).

also consistently obtain considerably higher number of unaligned source words. This result holds for the comparison between the WMT-21-SOTA and GPT systems as well. Further, GPT 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 GPTs do produce less literal E→X translations.

Human Evaluation We verify the conclusion from the results in Figure 1 by conducting a human evaluation of translation literalness on 6 WMT-22 language pairs: En-De, En-Ru, En-Zh and De-En, Ru-En, Zh-En. For each language pair, we randomly sample 100 source-translation pairs, with translations obtained from MS-Translator (a strong commercial NMT system) and text-davinci-003 (a strong commercial LLM) (Hendy et al., 2023). We used zero-shot text-davinci-003 translations for human evaluations in order to eliminate any biases through the use of specific demonstration examples. In each case, we ask a human annotator (bilingual speaker for Zh-En, target-language native plus bilingual speaker otherwise) to annotate 100 translations from both GPT and MS-Translator and select which of the two translations is more literal. The human annotation interface is described in Appendix A. The results in Table 3 show that the annotators rate the GPT translations as less literal.

Lang-Pair	MS-Translator	Davinci-003	Equal	Diff
En-De	52	32	16	+20
En-Zh	42	32	24	+10
En-Ru	41	37	22	+ 4
De-En	48	26	26	+12
Zh-En	42	38	20	+ 4
Ru-En	52	28	20	+24

Table 3: Human Evaluation Results across different language pairs on which is the *more literal translation*: the numbers are from annotations done on 100 translations obtained from both MS-Translator and Davinci-003.

Experiments on Best WMT-22 NMT Systems Further, we also experiment with the WMT-Best systems on the WMT-22 General Machine Translation task (Kocmi et al., 2022). We evaluate USW and NM on De-En, Ja-En, En-Zh and Zh-En, since on each of these language pairs, text-davinci-003's few-shot performance is very close to that of the WMT-Best system as per COMET-22 (Rei et al., 2022), based on the evaluation done in Hendy et al. (2023). We report our results in Table 4, which shows our prior findings replicated across the language pairs. For example, text-davinci-003, despite obtaining a 0.2 to 0.6 *higher* COMET-22 score than the best WMT systems on these language pairs, consistently obtains a *higher* USW score and a higher NM score in all but one comparison (NM for En-De). Note that the NM score differences for Chinese and Japanese are larger in magnitude owing to alignment deviations measured over character-level alignments. Further, we refer the reader to Hendy et al. (2023) for similar USW and NM comparisons of translations from text-davinci-003 and MS-Translator.

Language Pair	USW Diff	NM Diff
En-Zh	+ 4.93	+ 12.94
De-En	+ 1.04	- 0.10
Zh-En	+ 4.93	+ 13.06
Ja-En	+ 6.10	+ 11.13

Table 4: USW and NM score differences of text-davinci-003 relative to WMT-Best on the WMT-22 test sets.

MT System	C-QE ↑	USW ↓	NM ↓
MS-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 5: Natural Idiomatic Sentences: Combined Results over MAGPIE, EPIE, PIE (5,712 sentences).

3 Effects On Figurative Compositionality

In this section, we explore whether the less literal nature of E→X translations produced by GPT models 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., 2022a) 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 different from the *desired* figurative translation.

3.1 Translation with Idiomatic Datasets

In this section, we quantify the differences in the translation of sentences with idioms between traditional NMT systems and a GPT model. There do not exist any English-centric parallel corpora dedicated to sentences with idioms. 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:

- *MAGPIE* (Haagsma et al., 2020) contains a set of sentences annotated with their idiomaticity, alongside a confidence score. We use the sentences pertaining to the news domain which are marked as idiomatic with cent percent annotator confidence (totalling 3,666 sentences).

- *EPIE* (Saxena and Paul, 2020) contains idioms and example sentences demonstrating their usage. We use the sentences available for static idioms (totalling 1,046 sentences).

- The *PIE dataset* (Zhou et al., 2021) contains idioms along with their usage. We randomly sample 1K sentences from the corpus.

Results The results are presented in Table 5. 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, even though they are less likely to do this than surface-level metrics such as BLEU or ChrF (Papineni et al., 2002b; Popović, 2015). 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.

3.2 Synthetic Experiments

For our next experiments, we generate synthetic English sentences, each containing expressions of specific *type(s)*: (i) names, (ii) random descriptive phrases, and (iii) idioms. We prompt text-davinci-002 in a zero-shot manner, asking it to generate a sentence with different *instantiations* of each of these types (details are in appendix B). We then translate these sentences using the different systems, in order to investigate the relative effects on our literalness metrics between systems and across types. In each of the control experiments, we translate the synthetic English sentences to German.

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

Synthetic Dataset 1 As described, we generate sentences containing expressions of the 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 B. Results are in Table 6.

Num Idioms	1	2	3	4
USW	17.58	18.39	18.28	18.99

Table 7: Synthetic sentences with multiple idioms (1-4): Increasing the number of idioms increases the number of unaligned source words in text-davinci-002 translations.

Synthetic Dataset 2 We generate sentences containing *multiple* idioms (varying from 1 to 4). The prompts & examples are presented in appendix B. The results are presented in Table 7.

Results Table 6 shows that the percentage of unaligned source words is highest in the case of idioms, followed by random descriptive phrases & named entities. The results are consistent with the hypothesis that the explored GPT models produce less literal E→X translations, since named entities or descriptive phrases in a sentence would admit more literal translations as acceptable, unlike sentences with idioms. Davinci-002 obtains a much higher COMET-QE score in the case of translations of sentences with idioms, yet obtains a higher percentage of unaligned source words. Similarly, the difference in non-monotonicity scores is also considerably higher for the case of idioms. These results provide some evidence that the improved results of the GPT model, together with the *lower literalness* numbers, stem from correct translation of idiomatic expressions. Table 7 shows that this effect only increases with the number of idioms.

4 Discussion

In our experiments conducted across different NMT systems and GPT models, we find evidence that GPTs produce translations with greater non-literalness for E→X in general. There could be a number of potential causes for this; we list two plausible hypotheses below:

Parallel Data Bias NMT models are trained on parallel data, which often contains very literal web-collected outputs. Some of this may even be the output of previous-generation MT systems, which is highly adopted and hard to detect. In addition, even high quality target text in parallel data always contains artifacts that distinguishes it from text originally written in that language, i.e. the 'translationese' effect (Gellerstam, 2005). These factors could likely contribute to making NMT translations comparatively more literal.

Language Modeling Bias Translation capability in GPTs arises in the absence of any *explicit* supervision for the task during the pre-training stage. Therefore, the computational mechanism that GPTs leverage for producing translations might be different from NMT models, imparting them greater abstractive abilities. This could have some measurable manifestation in the translations produced, e.g., in the literalness of the translations.

Differences in E→X and X→E In E→X, we consistently find that GPT translations of similar quality are less literal and in the X→E direction, we observe a few anomalies. For X→E, in Figure 1, in all but one comparison (WMT-21-SOTA vs GPTs for De-En) GPTs obtain higher measures for non-literalness. On the other hand, we did not see anomalies in the trend for E→X directions.

Variations in Experimental Setup We also experimented with a variant of USW and NM which doesn't use the alignments pertaining to stopwords. Each of our findings remain the same, with relatively minor changes in magnitudes but not in system rankings. Similarly, we observed a greater tendency towards less literalness in GPT translations in both few-shot and zero-shot settings, when compared across a range of NMT systems.

5 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 tendency towards non-literalness in GPT translations. In particular, this tendency becomes evident in GPT systems' ability to figuratively translate idioms.

6 Acknowledgements

We thank Hitokazu Matsushita for help in conducting human evaluations.

7 Limitations

Measurement of translation literalness is neither well studied nor well understood. We rely on a combined interpretation of multiple measurements to investigate our hypothesis and its implications. This limits the extent to which we can make strong claims, since in the absence of a highly correlated metric for translation literalness, it is hard to compare systems. We could only claim that our investigation indicates the presence of a tendency towards non-literalness in GPT translations, but a stronger result would have been preferred to further disambiguate the translation characteristics. Further, we only compare GPT translations in the standard zero-shot and few-shot settings and it is quite conceivable that more specific & verbose instructions could steer the LLMs to produce translations with different 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.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in palm's translation capability.

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, Elia Bruni, and Dieuwke Hupkes. 2022a. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4154–4175, Dublin, Ireland. Association for Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022b. 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.

Martin Gellerstam. 2005. *Chapter 13. Fingerprints in Translation*, pages 201–213. Multilingual Matters, Bristol, Blue Ridge Summit.

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.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2022. On the blind spots of model-based evaluation metrics for text generation. *arXiv preprint arXiv:2212.10020*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

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.

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.

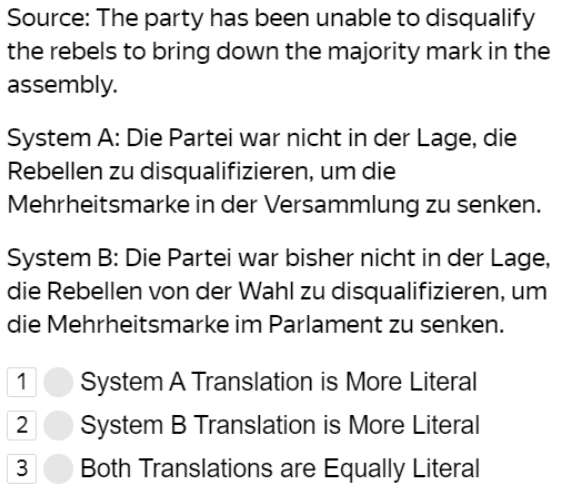

Figure 2: Annotation Interface Screenshot for Human Evaluation. The translations are randomized b/w the systems under evaluation to prevent biased evaluation.

A Human Annotation Interface

We use the annotation interface in Figure 2, in which the annotators are asked to rate the two translations. The bilingual and native speaker annotators were recruited in-house.

B Synthetic Dataset Details

Synthetic Dataset 1: For each of the three expression types, 100 synthetic sentences are generated. Figures 3, 4, 5 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.

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 3: **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 4: **Synthetic Dataset 1**: Example (**Entity**)

Synthetic Dataset 2: Figure 6 presents an example of the prompt as well as for a generated synthetic sentence containing two idioms.

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 5: **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 6: **Synthetic Dataset 2**: Example (2 **Idioms**)

C Alignments and Literalness

Schaeffer and Carl (2014) find that more alignment crossings (which is measured by the non-monotonicity metric) between the source and translations are proportional to the extra cognitive effort (measured using gazing time of human translators) required by human translators in processing non-literal translations. This links alignment crossings (the non-monotonicity measure is normalized alignment crossing) with greater non-literalness.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 5

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 3, 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3, 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3, 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3, 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3, 4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Appendix B

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix B

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.