Prompting Large Language Models for Idiomatic Translation

Antonio Castaldo
University of Naples L'Orientale, Italy
University of Pisa, Italy
antonio.castaldo@phd.unipi.it

Johanna Monti
University of Naples L'Orientale, Italy
jmonti@unior.it

Abstract

Large Language Models (LLMs) have demonstrated impressive performance in translating content across different languages and genres. Yet, their potential in the creative aspects of machine translation has not been fully explored. In this paper, we seek to identify the strengths and weaknesses inherent in different LLMs when applied to one of the most prominent features of creative works: the translation of idiomatic expressions. We present an overview of their performance in the EN→IT language pair, a context characterized by an evident lack of bilingual data tailored for idiomatic translation. Lastly, we investigate the impact of prompt design on the quality of machine translation, drawing on recent findings which indicate a substantial variation in the performance of LLMs depending on the prompts utilized.

1 Introduction

Recent advancements in the field of artificial intelligence, particularly with the emergence of Generative Pre-trained Transformer (GPT) models, have prompted the beginning of a new era of exploration into the applicability of large language models (LLMs) for machine translation tasks. The recent development and refinement of LLMs, such as GPT-3.5 and GPT-4 (Brown et al., 2020), have demonstrated their remarkable performance in understanding and generating natural language (Ahuja et al., 2023), thus positioning these models at the forefront of research into the translation of creative textual genres, including the nuanced task of translating idiomatic expressions. Traditional neural machine translation (NMT) systems often falter in accurately capturing the essence of idiomatic expressions, tending towards translations that are either overly literal or misinterpret the intended meaning. In contrast, recent studies have illustrated the ability of GPT models to adopt less literal translation approaches, especially in handling idiomatic expressions, leveraging an enhanced understanding of context and figurative language. This contribution will evaluate various large language models (LLMs) to establish benchmarks for their effectiveness in translating idiomatic expressions in the English-Italian language pair. The objective is to identify the strengths and weaknesses inherent in different LLMs when applied to machine translation (MT) tasks, particularly focusing on the nuanced aspect of creative language. Furthermore, the study will explore the impact of prompt design on MT quality, drawing on the findings of Ahuja et al. (2023) that suggest that the performance of LLMs in multilingual tasks can vary significantly with the prompts used. Through these evaluations, we seek to contribute to the improvement of machine translation technologies, highlighting the potential of LLMs to make creative works more accessible across languages, enriching cultural exchange and overcoming language barriers.

2 Related Work

Research in the use of large language models for machine translation has been pursued following two main axes. The first involves issues specific to LLMs, such as the influence that prompt templates

Proceedings of the 1st Workshop on Creative-text Translation and Technology, pages 32–39
June 27, 2024 ©2024 European Association for Machine Translation

may have on the model output (Zhang et al., 2023; Lu et al., 2023; Peng et al., 2023). The second line focuses on the evaluations of LLMs in various translation scenarios, covering multilingual (Jiao et al., 2023b; Hendy et al., 2023; Zhu et al., 2023), document-level (Wang et al., 2023; Karpinska and Iyyer, 2023; Wu et al., 2024), low-resource translation (Moslem et al., 2023a; Mao and Yu, 2024), hallucination (Guerreiro et al., 2023) and domain adaptation (Hendy et al., 2023). This study positions itself within the second research axis, concentrating on the evaluation of LLMs in specialized translation scenarios. Despite the large body of research currently being conducted on LLMs performance, research to date has not yet fully explored their application to the translation of creative texts. This study does not aim to provide a comprehensive overview of the topic, but we seek to evaluate the intricate task of translating idiomatic expressions, a critical aspect that challenges the adaptability and understanding of these models.

3 Experimental Setup

In this section, we describe the methodology used in our experiments, including the translation process and the evaluation metrics employed. We initiated the translation process leveraging OpenAI API and the HuggingFace library (Wolf et al., 2020) in Python, generating four batches of translations using four distinct prompts applied to the models `gpt-3.5-turbo` and `Mistral-7B-v0.1`.

For the machine translation (MT) evaluation, we used the online evaluation platform MATEO (Vanroy et al., 2023), which provides an easy-to-use user interface for the evaluation of translations, utilizing state-of-the-art neural and n-gram evaluation metrics. We conducted the experiment in three independent trials to ensure the reliability of the results and replicability of the experiment.

3.1 Dataset Selection

In this section, we describe the composition of the dataset used for our experiments, which comprises a set of 350 Italian-English sentence pairs, where 18 idiomatic expressions are used in both their literal and idiomatic meanings. This corpus was assembled utilizing two primary sources: the Italian Dodiom corpus (Eryiğit et al., 2023) and the Reverso Context online database. The Dodiom corpus, a curated collection of Italian and Turkish idiomatic expressions, was initially gathered using a gamified crowdsourcing bot on the Telegram platform. After being collected, the corpus underwent a rigorous annotation process by linguistic experts, as detailed in Morza et al. (2022). The revision process ensured the idioms' authenticity and their contextual relevance.

Leveraging the idiomatic expressions collected using the Dodiom corpus, we proceeded to extract corresponding bilingual sentence pairs that incorporate these idioms from the Reverso Context database. Reverso Context, known for its extensive repository of real-life language usage examples across multiple languages, served as an ideal resource for obtaining authentic usage examples of the idiomatic expressions we have collected.

3.2 Annotation

The 321 extracted sentence pairs were thoroughly evaluated and annotated. This step was crucial to verify the translation accuracy of the idioms and to confirm their relevance within the given contexts, regardless of the initial quality level of Reverso Context. The annotation process was conducted by a native Italian speaker, who had completed a Master's degree in linguistics, accumulating five years of academic education. Their linguistic proficiency and compatibility with our study is certified by English, being the primary language of their university studies. The annotation was conducted on an online platform, developed in Flask, specifically for the scope of this study.

First, the annotator was asked to conduct a binary evaluation of the adequacy of each pair of bilingual sentences, focusing on whether the translated expressions conveyed the original meaning and nuance of the idiom in the source language, and whether the translation extracted by Reverso Context was relevant to the source text. This step allowed us to exclude incorrect and irrelevant examples. Then, the annotator was asked to annotate whether the idiomatic expressions within each sentence were used in their literal or figurative sense. Finally, before beginning our experiments, we proceeded to remove every sentence pair considered unsatisfactory in their translation and relevance.

The process allowed us to obtain a curated dataset, comprised of 254 bilingual segments, on which we could conduct an evaluation of MT quality and prompting impact.

3.3 Prompt Templates

For our study, we select four prompt templates, three of which are derived from studies by Gao et al. (2024), Zhang et al. (2023), and Jiao et al. (2023b), and a five-shot prompt, developed within the scope of our current study. The prompts were chosen on the basis of the high performance reported in the relative literature. The prompt templates that we have selected differ in their length and in the information they convey to the model.

We present an overview of the prompt templates in the following table, with the following annotations: ♦ shows the presence of a line break, `[src]` stands for source language, `[tgt]` stands for target language, and `[input]` stands for the text to be translated.

Prompt ID	Prompt Template
A	[src]: [input] ♦ [tgt]:
B	Please provide the [tgt] translation for this sentence: [input] ♦ Translation:
C	This is a [src] to [tgt] translation, please provide the [tgt] translation for this sentence: [input] ♦ Translation:
D	[src]: [source$_1$] ♦ [tgt]: [target$_1$] ♦ ... [src]: [source$_k$] ♦ [tgt]: [target$_k$] ♦ [src]: [input] ♦ [tgt]:

Table 1: Overview of the prompt templates used in this study

Prompt A offers a concise structure that directly maps the source language to the target language, where brevity is exchanged for clarity of the instructions, which in this case is inferred from the context. Prompt B presents a more descriptive approach, including the target language in a clear instruction, however the source language is not included. Prompt C is the most descriptive one, presenting detailed instructions that include both the source and the target language.

Our contribution, Prompt D, extends the concept of minimalistic mapping (as in Prompt A) through a few-shot learning approach. It involves presenting the model with five contextual examples ($k = 5$) prior to the translation task, selected for their relevance to the input text. This methodology is designed to leverage the model's in-context learning ability (Brown et al., 2020) to improve the translation performance thanks to the exposure to related translation examples (Garcia et al., 2023; Lu et al., 2023). For the implementation of this five-shot prompt, the examples were selected on the basis of their semantic similarity to the input sentence. Whereas the common procedure is to generate semantic embeddings with models such as LaBSE (Hendy et al., 2023), we provide a proof of concept using a computationally efficient and non-neural TF-IDF Vectorizer. Despite its simplicity, the vectorizer effectively represents the texts in a multidimensional space, allowing the calculation of cosine similarity to identify examples most relevant to the given input sentence. This strategy aims to provide the model with contextually pertinent examples, thereby enhancing its ability to infer and execute the translation task.

4 Evaluation

We present a comprehensive evaluation of the four prompt templates we have selected, using two models: `gpt-3.5-turbo-1106` and `Mistral-7B-v-0.1`. Our evaluation used two mainstream neural evaluation metrics: COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020). These metrics have shown a very high correlation with human judgment and are established in the evaluation of LLM-based machine translation (Moslem et al., 2023a; Hendy et al., 2023). We have decided to include BLEU (Papineni et al., 2002) in our evaluation, as it remains a widely recognized standard metric in MT evaluation, despite its limitations for our specific translation context. More specifically, in the context of accurately conveying idiomatic expressions into another language, there is frequently a mismatch between the length of the sentence in the source and target texts. Metrics such as BLEU and ChrF (Popović, 2015) may not be the most adequate for the task, as they tend to penalize length, lexical discrepancies and brevity of the translations, which are not necessarily indicative of poor translation quality, especially in the context of idiomatic expressions.

4.1 Results with GPT-3.5

When testing the model `gpt-3.5-turbo`, the five-shot template we developed, Prompt D, consistently outperformed the others in terms of BLEURT and COMET scores, displaying statistical significance (p-value < 0.05) in every evaluation instance, as shown in Table 2. Prompt C was the second best-performing prompt in BLEURT

and COMET, and the absolute best in terms of BLEU score.

Table 2: Evaluation of automated MT metrics for the selected prompts, using the model gpt-3.5-turbo-1106. Asterisks represent statistical significance (p-value < 0.05).

System	BLEURT	COMET	BLEU
Prompt A	70.09	80.78	36.39
Prompt B	70.17	81.05	37.43
Prompt C	70.54*	81.16*	**38.25***
Prompt D (k=5)	**71.17***	**81.71***	37.70*

The observed BLEU scores were found to be significantly unsatisfactory, in line with expectations. Interestingly, this shortfall cannot be attributed to a discrepancy in sentence lengths, which were quite similar to both the source text (with an average sentence length of 16.98) and the reference translations (with an average sentence length of 17.95). Instead, the limitations may stem from the brevity penalty inherent in the BLEU metric, coupled with a lack of n-gram overlap in the translations.

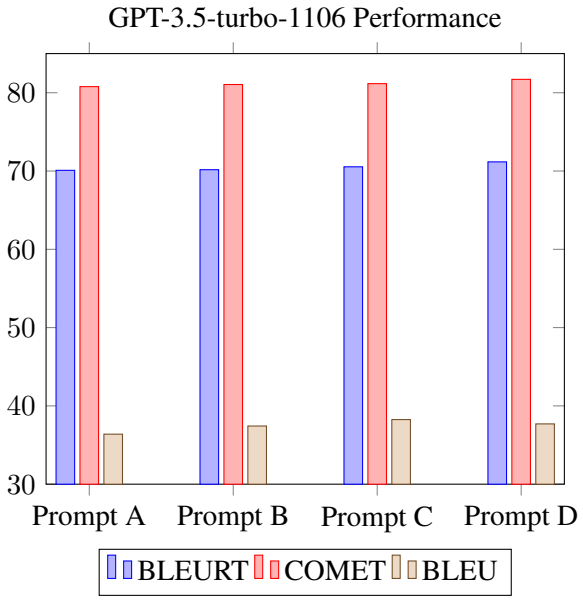

Figure 1: GPT-3.5-turbo-1106 performance per prompt template, calculated by BLEU, COMET and BLEURT.

This issue is particularly pronounced in the handling of idiomatic expressions, where translations often adopt a more creative and less order-bound approach. This hypothesis is supported by a substantial difference in the average BLEU scores: sentences with idiomatic meanings scored an average of 32, while sentences with literal meanings achieved an average score of 39.7. This discrepancy is significantly less pronounced when evaluated using the COMET metric, which shows only a 3-point difference between the two scenarios. In contrast, neural metrics consistently yielded high scores, surpassing 70 across all tested prompts. This suggests that while traditional metrics like BLEU may struggle to evaluate the nuances of creative translations, particularly of idiomatic expressions, neural-based evaluation metrics such as COMET offer a more effective assessment, potentially capturing aspects of translation quality that BLEU overlooks, thanks to their use of semantic embeddings.

4.2 Results with Mistral-7B

The second model we evaluated is the open-source multilingual LLM, `Mistral-7B` (Jiang et al., 2023), developed by the homonymous French company. As reported in the release publication, Mistral has excelled on several NLP benchmarks. Remarkably, its smallest checkpoint, trained on only 7B parameters, has outperformed much larger models, such as `LLama-2-13B` and `LLama-1-34B`, developed by Meta. When fine-tuned on a downstream machine translation task, Mistral has outperformed `gpt-3.5-turbo`, as seen in Moslem et al. (2023b), demonstrating the capability of Mistral to be an effective open source asset for multilingual machine translation.

Table 3: Evaluation of automated MT metrics for the selected prompts, using the model Mistral-7B-v0.1. Asterisks represent statistical significance (p-value < 0.05).

System	BLEURT	COMET	BLEU
Prompt A	64.85	76.55	33.84
Prompt B	64.21	75.76*	32.91
Prompt C	64.98	76.78	33.11
Prompt D (k=5)	**68.60***	**79.56***	**36.57***

Upon testing Mistral on the same set of prompt templates from our preceding experiment, it was observed that Mistral's adherence to given instructions was not as precise as the model developed by OpenAI. Prompt A and C were the worst performing templates, whereas the more informative Prompt C scored better than the others. Even in this case, five-shot prompting (Prompt D) displayed the best results in every evaluation instance.

The translations generated by Mistral were found to include numerous unnecessary excerpts and hallucinations. The core issue identified was

not the quality of the translation per se, but the format of the responses. These did not align with the expected format derived from the reference and the source texts. Instead, Mistral introduced extraneous phrases like "Perhaps, you would..." or "I think an accurate translation would be..." which inevitably led to lower scores on evaluative metrics, especially n-gram based ones (Table 3 and Figure 2).

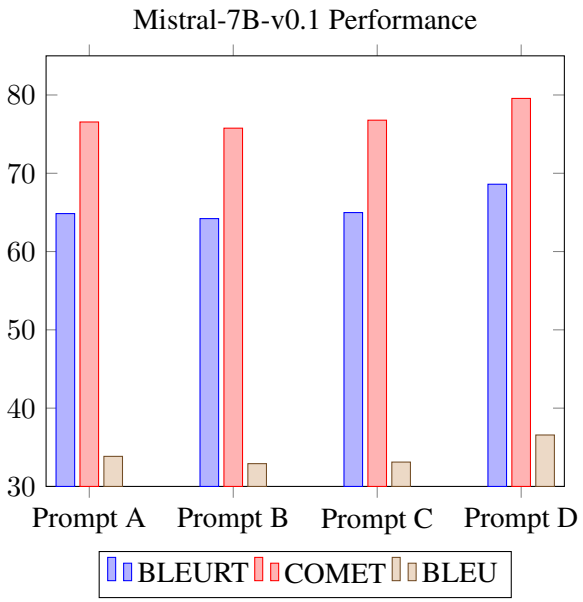

Figure 2: Mistral-7b-v0.1 presents very low BLEU scores for Prompts B and C.

Contrary to previous expectations, our findings reveal that a single machine translation prompt can indeed deliver optimal performance across both ChatGPT and Mistral-7B models. This challenges the conventional wisdom that prompts need to be meticulously customized for each model to elicit the best output. Our analysis suggests that the shared architecture and similar training datasets of these models might allow them to understand and process instructions in a comparable manner, thus offering similarly effective performance to the same prompt. The successful application of one prompt across these models highlights a potential universality in prompt effectiveness.

4.3 Prompts and Models Comparison

As we underlined in the previous sections, the choice of prompt critically influences a model's comprehension and translation of idiomatic expressions. We present an overview of selected translation output, illustrating how influential the prompt choice is on the model's performance. In Table 4, the examples focus on the translation of the Italian idiom "attaccare bottone" (literally translated as "buttoning up") which commonly translates to "striking up a conversation" in English.

Table 4: Translations of "attaccare bottone" using different prompts on the Mistral model, ordered from A to D

Prompt	Translation	COMET
A	You know, he could hit on any girl, anywhere.	76.54
B	You know, he could buttonhole every girl, everywhere and always.	75.76
C	Do you know, he could buttonhole every girl, always and everywhere.	76.77
D	You know, he could get any girl he wanted, anywhere.	**79.55**

While Prompts B and C mistake the intended meaning and generate a literal translation, Prompts A and D align closely to the reference translation and the intended meaning of the Italian idiom. The results we obtain clearly showcase how impactful the prompt choice is on the model's understanding and translation performance.

Building on this, in the following table, we extend the analysis to the OpenAI model, comparing how GPT-3.5 and Mistral handle the same idiomatic input, in their best or worst performance scenario. We display the translation outputs for two Italian idioms: "prendere con le pinze" (literally translated as "to take with tweezers") which idiomatically translates to "to take with a grain of salt" and "avere le mani lunghe" (literally translated as "to have long hands") which translates to "to have sticky fingers".

For the idiom "prendere con le pinze", the Mistral model produced an inaccurate translation, where the subject is missing and the idiomatic expression is translated literally, failing to convey the exact meaning of the input sentence. In contrast, even the least effective prompt with GPT-3.5 provides an accurate and contextually appropriate translation. Mistral is able to accurately translate the idiom, only when prompted by Prompt D. Finally, with the idiom "avere le mani lunghe", both Mistral and GPT-3.5's accurately translate the idiom into two possible correct meanings: Mistral

Table 5: Translations of "Prendere con le pinze" and "Avere le mani lunghe" using different prompts on Mistral and GPT-3.5

Model	Prompt	Translation
Mistral	Worst	Terry, is to be taken with the pliers, ok?
GPT	Worst	Terry, it's to be taken with a grain of salt, okay?
Mistral	Best	You know me, Watson, I'm handsy...
GPT	Best	You know me, Watson, I have sticky fingers...

translates it as being inclined to violence, while GPT-3.5's translation conveys the concept of being inclined to steal with "having sticky fingers".

5 Conclusions

This work presents a preliminary analysis on the use of LLMs for the translation of idiomatic expressions. We find that, given the same dataset and task, identical prompts may have optimal efficacy across various models, as seen for Mistral-7B and GPT-3.5, and that it is possible to optimize the model's performance by choosing an optimal prompt. In our experiments, the five-shot prompt (Prompt D) consistently outperformed other prompts in terms of BLEURT and COMET across both models, over three independent trails, demonstrating the efficacy of leveraging in-context learning ability to improve the model's understanding of idiomatic expressions. As for zero-shot prompting, Prompt C consistently performed the best. We find that GPT-3.5 consistently outperforms Mistral-7B which, on the other hand, can come close to GPT's performance when prompted correctly. Finally, we underline the limitations of traditional metrics based on n-grams, such as BLEU, in evaluating the translation of idiomatic expressions, advocating the use of neural-based evaluation metrics that better capture semantic nuances. Overall, our findings promote a more strategic approach to prompt selection and model use in machine translation, pointing towards a future where LLMs can be used effectively for nuanced and culturally-specific translation tasks. As the field of MT continues to evolve, so too will the strategies for leveraging the full potential of large language models in understanding and translating the rich nuances of human language.

5.1 Limitations

As a preliminary study, there are several aspects that should be improved to make it more comprehensive and reliable. Currently, due to the very specific nature of our task, our evaluation is conducted on a self-compiled dataset of 254 bilingual sentences, presenting only a limited number of idiomatic expressions. For resource and time constraints, the evaluation was only conducted using automated evaluation metrics. Finally, while we have identified that for a given dataset there is an optimal prompt for different models, the underlying factors determining an optimal prompt's performance, given the same task, remain unclear. It is worth noting that our findings are specific to the linguistic context of this evaluation, and the results may differ when applied to other language pairs.

5.2 Future Work

In our future work, we aim to address the current limitations of our study, to make it more reliable and comprehensive by focusing on different areas. First of all, we would like to expand the scope of our research, building a more comprehensive dataset, for a better representation not only of italian idiomatic expressions but also of other features specific of creative text. Regarding prompts, we find it necessary to continue exploring the several prompts that are being researched, such as pivot prompting (Jiao et al., 2023a) and chain-of-dictionary (Lu et al., 2023), and also prompt ensembles, such as those seen in Feng et al. (2024). We deem it necessary to research the best prompting techniques, in order to achieve the very best performance from the models at our disposal, contributing especially to the use of small-scale open-source models, such as the Mistral-7B model we have used in this study. By pursuing these directions, we aim to improve our understanding of how LLMs can be more effectively utilized for the task of translating idiomatic expressions, and more broadly, creative works.

6 Acknowledgements

This work has been funded by the National PhD programme in Artificial Intelligence, partnered by University of Pisa and University of Naples "L'Orientale", through a doctoral grant established by Ex DM 318, of type 4.1, co-financed by the National Recovery and Resilience Plan.

References

Ahuja, Kabir, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. March. arXiv: 2303.12528.

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

Eryiğit, GülŞen, Ali Şentaş, and Johanna Monti. 2023. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, 29(4):909–941, July. Number: 4.

Feng, Zhaopeng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Improving LLM-based Machine Translation with Systematic Self-Correction, March.

Gao, Pengzhi, Zhongjun He, Hua Wu, and Haifeng Wang. 2024. Towards Boosting Many-to-Many Multilingual Machine Translation with Large Language Models, January. Issue: arXiv:2401.05861 arXiv:2401.05861 [cs].

Garcia, Xavier, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. February. arXiv: 2302.01398.

Guerreiro, Nuno M., Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. March. arXiv: 2303.16104.

Hendy, Amr, 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. February. arXiv: 2302.09210.

Jiang, Albert Q., 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, October. Issue: arXiv:2310.06825 arXiv:2310.06825 [cs].

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine, November.

Jiao, Wenxiang, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023b. Is ChatGPT A Good Translator? A Preliminary Study. *arXiv*.

Karpinska, Marzena and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist, May. Issue: arXiv:2304.03245 arXiv:2304.03245 [cs].

Lu, Hongyuan, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-Dictionary Prompting Elicits Translation in Large Language Models, May. Issue: arXiv:2305.06575 arXiv:2305.06575 [cs].

Mao, Zhuoyuan and Yen Yu. 2024. Tuning LLMs with Contrastive Alignment Instructions for Machine Translation in Unseen, Low-resource Languages, January. Issue: arXiv:2401.05811 arXiv:2401.05811 [cs].

Morza, Giuseppina, Raffaele Manna, and Johanna Monti. 2022. Assessing the Quality of an Italian Crowdsourced Idiom Corpus:the Dodiom Experiment. pages 4205–4211, Marseille, France, June. European Language Resources Association.

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. Adaptive Machine Translation with Large Language Models. January. arXiv: 2301.13294.

Moslem, Yasmin, Rejwanul Haque, and Andy Way. 2023b. Fine-tuning Large Language Models for Adaptive Machine Translation, December. Issue: arXiv:2312.12740 arXiv:2312.12740 [cs].

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Peng, Keqin, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation. *SSRN Electronic Journal*.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth*

Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July. Association for Computational Linguistics.

Vanroy, Bram, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere, Finland, June. European Association for Machine Translation.

Wang, Longyue, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-Level Machine Translation with Large Language Models. April. arXiv: 2304.02210.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Wu, Minghao, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting Large Language Models for Document-Level Machine Translation, January. Issue: arXiv:2401.06468 arXiv:2401.06468 [cs].

Zhang, Biao, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A Case Study, January. Issue: arXiv:2301.07069 arXiv:2301.07069 [cs].

Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. April. arXiv: 2304.04675.