NovelTrans: System for WMT24 Discourse-Level Literary Translation

Yuchen Liu[1], **Yutong Yao**[1], **Runzhe Zhan**[1], **Yuchu Lin**[2], **Derek F. Wong**[1*]
[1]NLP^{2}CT Lab, Department of Computer and Information Science, University of Macau
nlp2ct.{yuchen, yutong, runzhe}@gmail.com; derekfw@um.edu.mo
[2]DeepTranx, Zhuhai, China
yuchulin@deeptran.com

Abstract

This paper describes our submission system, NovelTrans, from NLP^{2}CT and DeepTranx for the WMT24 Discourse-Level Literary Translation Task in Chinese-English, Chinese-German, and Chinese-Russian language pairs under unconstrained conditions. For our primary system, three translations are done by GPT4o using three different settings of additional information and a terminology table generated by online models. The final result is composed of sentences that have the highest xCOMET score compared with the corresponding sentences in other results. Our system achieved an xCOMET score of 79.14 which is higher than performing a direct chapter-level translation on our dataset.

1 Introduction

In the rapidly evolving field of natural language processing (NLP), discourse-level literary machine translation remains a challenging task. It involves not only complex semantic phenomena but also long-term dependency, rare or new terminologies, and cultural background (Pang et al., 2024; Liu et al., 2023). These factors pose a high requirement for the translation model. Training or fine-tuning such a model is extremely costly. To address this, pretrained large language models (LLMs) and training-free methods like in-context learning (Brown et al., 2020) are widely used. Up to now, significant advancements have been made in sentence-level machine translation using training-free methods. These methods, such as TEaR (Feng et al., 2024), DUAL-REFLECT (Chen et al., 2024), Multi-Aspect Prompting and Selection (He et al., 2024), and Multi-Agent Debate (Liang et al., 2024), have proven effective. However, few studies have been conducted on the document level.

This paper presents our submission to the WMT24 Discourse-Level Literary Translation shared task. We utilize online commercial general-purpose LLMs, DeepSeek (DeepSeek-AI et al., 2024) and GPT4o (OpenAI et al., 2024), to perform the translation with the help of techniques including *Document-level Multi-Aspect Prompting and Selection* (*d-MAPS*), *LLM-generated terminology table* and *dynamic retrieval of in-context learning examples using Reranked BM25* (R-BM25; Agrawal et al. 2023). We also explore the potential of *post-correction of punctuation errors* in LLMs' translation results. Using the above method, NovelTrans achieves an xCOMET score of 79.14, 0.68 points higher than the GPT4o baseline. Moreover, the consistency of rare or unseen terminologies has significantly improved and the number of mistranslated or awkwardly translated phrases is greatly reduced. The remaining part of this paper is structured as follows. Section 2 contains an overview of our pipelines and detailed descriptions of each procedure in the pipelines. Experiments and results analysis of our method are given in Section 3. Finally, the conclusion is presented in Section 4.

2 System Overview

2.1 Pipeline

For our pipeline, we implemented three variants which were named Primary, Contrastive-1, and Contrastive-2. The Primary system has a pipeline shown in Figure 1. For each input document, we first generate a terminology table and then replace all terminologies in the document with their corresponding translations, ensuring the consistency of terminology translation throughout the document. Then the document is split into chapters using regular expressions. Each chapter is divided into 20-line segments. Each segment is translated using GPT4o, with MAPS and R-BM25 enhancing the translation quality. The translated text will then proceed to the post-correction stage, where the GPT4o model will detect and resolve punctuation errors. For the

*Corresponding Author.

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 980–986
November 15–16, 2024.

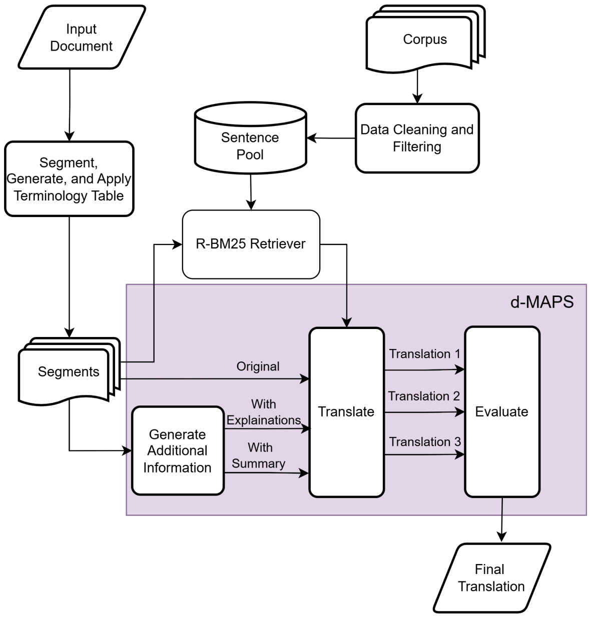

Figure 1: The translation flowchart of our NovelTrans system where post-correction is not included.

Contrastive-2 system, the MAPS uses a different way to determine the quality of translation and will be discussed in Section 2.2. The Contrastive-1 system is the same as the primary system except for the removal of the post-correction stage. As the API service for GPT4o we used contains a content filter, if a segment's translation is filtered by the content filter, the process will be handled using the DeepSeek API.

2.2 Document-level Multi-Aspect Prompting and Selection

Multi-Aspect Prompting and Selection (MAPS) is a powerful prompting strategy that can help a model understand the complicated relationships in discourse-level corpus better. Inspired by the MAPS, we chose to transfer MAPS to the document-level (d-MAPS). Considering both resource limitations and characteristics of web novels, we implemented d-MAPS as follows. We first acquire explanations for colloquialisms and the segment summary through the cooperation of DeepSeek and GPT4o. Then, three different translations are produced by GPT4o: one with explanations, one with the summary, and one without any extra information. Afterward, the COMET-22-kiwi reference-free translation quality evaluation model (Rei et al., 2022) is applied to obtain the quality score of each sentence in these three results. To select the final translation result, we employ two different strategies. In the Primary and Contrastive-1 system, the final result is composed of sentences that have the highest xCOMET score compared to the corresponding sentences in other translations. In Contrastive-2, the final translation is determined by choosing the result with the highest average xCOMET score.

2.3 LLM-generated Terminology Table

In the traditional novel translation pipeline, it is crucial to set up a terminology table before the translation to unify the translations of those rare terms throughout the corpus. To generate the terminology table, we use the DeepSeek API which has better knowledge of Chinese cultural backgrounds to retrieve proper nouns and then translate these words into the target language considering their context. With the terminology table acquired, we then replace all the terms in the source corpus with their corresponding translations to ensure consistency. The consistency mentioned above refers to the uniformity of special terminology translation.

2.4 Re-ranked BM25

Re-ranked BM25 (R-BM25; Agrawal et al. 2023) is an in-context example retriever that can ensure both sample quality and retrieving speed. After 100 sentences are retrieved by a normal BM25 retriever, a score will be computed for each sentence using the following formula, in which S and Q denote the source and retrieved sentence's n-grams separately.

$$R_n = \frac{\sum_{\text{ngram} \in S \cap Q} \text{Count}_{\text{matched}}(\text{ ngram })}{\sum_{\text{ngram} \in S} \text{Count}_S(\text{ ngram })} \quad (1)$$

$$\text{Score } = \exp\left(\frac{1}{n} \sum_{n} \log\left(R_n\right)\right) \quad (2)$$

Then these sentences are re-ranked using these scores to solve the problems that BM25 favors rare words (Robertson and Zaragoza, 2009). To form the sentence pool for the R-BM25 to search, we utilize the GuoFeng Webnovel Corpus[1] (Wang et al., 2023) which has three subsets named TRAIN, VALID1, and VALID2. By combining all three subsets, we formed a large dataset and then filtered out sentences with low xCOMET scores. During the experiment, VALID2 is not included because our valid set is sampled from VALID2. To generate the in-context learning examples for a particular segment, we retrieve three samples for each sentence

[1]http://www2.statmt.org/wmt23/literary-translation-task.html

	Zh-En		Zh-Ru		Zh-De	
	xCOMET	d-BLEU	xCOMET	d-BLEU	xCOMET	d-BLEU
DeepSeek	76.58	18.03	-	-	-	-
GPT3.5-Turbo-16k	77.33	17.92	-	-	-	-
GPT4o baseline	78.46	**18.85**	83.74	**26.51**	80.69	38.33
NovelTrans (Ours)	**79.14**	18.69	**84.42**	26.44	**80.85**	**39.78**

Table 1: Experiment result compared with other models. Results listed here expect NovelTrans are all generated by direct chapter-level translation. xCOMET scores in this and tables below are all computed using XCOMET-XL.

Method	xCOMET	BLEU	d-BLEU
GPT4o baseline	78.46	20.17	18.85
NovelTrans	**79.14**	19.94	18.69
w/o ICL	78.85	19.67	18.63
w/o ICL & Terminology Table	78.71	20.60	18.63
w/o ICL, Terminology Table & d-MAPS	78.68	**20.80**	**18.97**

Table 2: Ablation study of our proposed pipeline. ICL examples are selected by R-BM25 score. Terminology table represents the terminology table obtained by the cooperation of GPT4o and DeepSeek. The GPT4o baseline is generated by directly translating the text at the chapter level.

in that segment using R-BM25 and then randomly sample eight sentences to form the final in-context learning example. It is tested that choosing eight examples will result in the best performance boost.

2.5 Post-Correction of Translation

After reviewing the translation results, we observed that punctuation errors, such as comma splices, appeared at a high frequency due to the inappropriate use of punctuation in the source corpus. To solve this, we employed a post-processing method that uses GPT4o to correct punctuation errors at the sentence level. Given the sentence above and below the target sentence, we asked the model to check and resolve punctuation errors. This method resulted in a better version of the target sentence.

3 Experiments

3.1 Experiments Setup

The datasets we used are GuoFeng Webnovel Corpus V1 and V2. V1 contains a Chinese-English parallel corpus while V2 contains Chinese-German and Chinese-Russian nonparallel corpus. For the Chinese-English direction, we performed experiments on 10 chapters in VALID2 of the dataset. These chapters are taken from different books to avoid bias. For Chinese-German and Chinese-Russian direction, we chose 4 chapters from different books and aligned them separately using GPT4o API before experimenting. The GPT4o API we used is provided by OpenAI. The DeepSeek API is provided by DeepSeek Open Platform[2]. Since the BLEU score faces the problem of inaccuracy in evaluating Zero Pronoun Translation tasks (Zhan et al., 2023; Xu et al., 2023), we focused more on the COMET score. To be better aligned with the human evaluation, we chose to use **XCOMET-XL** (Guerreiro et al., 2023) to compute the xCOMET score. BLEU and d-BLEU scores are all computed by **SacreBleu** (Post, 2018). To compute d-BLEU, we join all sentences in the document together and treat them as a single sentence since it is the method used to compute the d-BLEU score in the previous year's WMT literary translation task (Wang et al., 2023).

3.2 Results

Table 1 shows the comparison between our system and other online models in Chinese-English, Chinese-German, and Chinese-Russian translation direction. The result shows that our system achieves a higher xCOMET score in exchange for the d-BLUE performance.

3.3 Ablation Study

We conduct ablation study on Chinese-English direction. The result, provided in Table 2, shows that

[2]https://platform.deepseek.com/

Source	GPT4o Baseline	NovelTrans
走，全部跟我走，去破坏对方的 (rival) 世界级传送阵.	Go, all of you come with me to destroy the *other side's* (Wrong) world-class teleportation array.	Let's go, everyone follows me to destroy the *enemy's* (Correct) world-class teleportation array.
这四个字，是郑州城人类最后的绝唱 (the last song of mankind in the city of Zhengzhou).	These four words were the *last human song of Zhengzhou* (Bad Phrase Translation).	These four words were the *last elegy of humanity in Zhengzhou city* (Correct).

Table 3: Case study where examples are taken from different pipeline methods.

Source	Without Correction	With Correction
"别紧张，自己人。"	"Don't be nervous, I'm one of you."	"Don't be nervous; I'm one of you."
他们打开背后的涡旋引擎跳了下去	They activated the vortex engine on their backs, jumping down	They activated the vortex engine on their backs before jumping down.

Table 4: Comparison of translation results with or without post-translation correction.

Position	Source	Without Term Table	With Term Table
Near the start of a chapter	若非此刻在天渡船上,可能已经大打出手.	If they weren't on the *Tian Du ship*, he might have already started a fight.	If they weren't on the *Heavenly Ferry*, he might have already started a fight.
Near the end of the same chapter	不多时,天渡船抵达对岸.	Before long, the *Heaven Crossing Boat* (Inconsistent) reached the other side.	Before long, the *Heavenly Ferry* (Consistent) reached the opposite bank.

Table 5: Comparison of translation results with or without LLM-generated terminology table.

removal of component in our system will result in a performance drop on xCOMET.

3.4 Analysis

Table 3 shows two examples taken from our experiment. In the first example, the direct translation of GPT4o uses an ambiguous phrase, "other side", which can mean both an enemy and a geographically opposite side. However, with the context, we can easily determine that the "other side" here conveys only the meaning of "rival". In the second example, the Chinese word "绝唱" which means the best art piece an artist has ever made is misused as "last song before their death" in the source sentence. Our system understood what the author wanted to convey and chose a suitable word, "elegy", rather than doing a literal translation. These examples show that, compared with the baseline, our method has a stronger understanding of the context and Chinese cultural background. Table 4 demonstrates the effect of post-correction. The GPT4o model can detect and correct punctuation errors, especially comma splices that occur at high frequency, in various ways. Table 5 shows an example of inconsistency in the translation of special terms and our method can greatly reduce this type of problem.

4 Conclusion

We successfully deployed a discourse-level translation pipeline using online language models and adapted several sentence-level techniques for discourse-level translation. Our system achieved a higher xCOMET score than direct translation using GPT-4o. However, our research has some limitations. Adapting MAPS to discourse-level translation may disrupt long-term dependencies, indicating a need for further investigation in this area. Additionally, our method utilizes significantly more tokens than direct translation, necessitating further discussion on how to reduce token usage.

Acknowledgement

This work was supported in part by the Science and Technology Development Fund, Macau SAR (Grant Nos. FDCT/060/2022/AFJ, FDCT/0070/2022/AMJ), the Research Program of Guangdong Province (Grant No. 2220004002576, EF2023-00090-FST), Tencent AI Lab Rhino-Bird Gift Fund (Grant No. EF2023-00151-FST), and the Multi-year Research Grant from the University of Macau (Grant No. MYRG-GRG2023-00006-FST-UMDF).

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. DUAL-REFLECT: Enhancing large language models for reflective translation through dual learning feedback mechanisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–704, Bangkok, Thailand. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Yang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jin Chen, Jingyang Yuan, Junjie Qiu, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruizhe Pan, Runxin Xu, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Size Zheng, T. Wang, Tian Pei, Tian Yuan, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Liu, Xin Xie, Xingkai Yu, Xinnan Song, Xinyi Zhou, Xinyu Yang, Xuan Lu, Xuecheng Su, Y. Wu, Y. K. Li, Y. X. Wei, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Zheng, Yichao Zhang, Yiliang Xiong, Yilong Zhao, Ying He, Ying Tang, Yishi Piao, Yixin Dong, Yixuan Tan, Yiyuan Liu, Yongji Wang, Yongqiang Guo, Yuchen Zhu, Yuduan Wang, Yuheng Zou, Yukun Zha, Yunxian Ma, Yuting Yan, Yuxiang You, Yuxuan Liu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhewen Hao, Zhihong Shao, Zhiniu Wen, Zhipeng Xu, Zhongyu Zhang, Zhuoshu Li, Zihan Wang, Zihui Gu, Zilin Li, and Ziwei Xie. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Tear: Improving llm-based machine translation with systematic self-refinement.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2024. Encouraging divergent thinking in large language models through multi-agent debate.

Xuebo Liu, Yutong Wang, Derek F. Wong, Runzhe Zhan, Liangxuan Yu, and Min Zhang. 2023. Revisiting commonsense reasoning in machine translation: Training, evaluation and challenge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15536–15550, Toronto, Canada. Association for Computational Linguistics.

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

Jianhui Pang, Fanghua Ye, Longyue Wang, Dian Yu, Derek F Wong, Shuming Shi, and Zhaopeng Tu. 2024. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *arXiv preprint arXiv:2401.08350*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

Mingzhou Xu, Longyue Wang, Siyou Liu, Derek F Wong, Shuming Shi, and Zhaopeng Tu. 2023. A benchmark dataset and evaluation methodology for chinese zero pronoun translation. *Language Resources and Evaluation*, 57(3):1263–1293.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, Cuilian Zhang, Lidia S. Chao, and Min Zhang. 2023. Test-time adaptation for machine translation evaluation by uncertainty minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 807–820, Toronto, Canada. Association for Computational Linguistics.