IIIT HYD's Submission for WMT23 Test-suite task

Ananya Mukherjee and **Manish Shrivastava**
Machine Translation - Natural Language Processing Lab
Language Technologies Research Centre
International Institute of Information Technology - Hyderabad
ananya.mukherjee@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

This paper summarizes the results of our test suite evaluation on 12 machine translation systems submitted at the Shared Task of the 8th Conference of Machine Translation (WMT23) for English-German (en-de) language pair. Our test suite covers five specific **domains** (*entertainment, environment, health, science, legal*) and spans five distinct **writing styles** (*descriptive, judgments, narrative, reporting, technical-writing*). We present our analysis through automatic evaluation methods, conducted with a focus on domain-specific and writing style-specific evaluations. Our test-suite is available at https://github.com/wmt-conference/wmt23-testsuites/tree/main/submissions/en-de/IIITHYD_TestSuite

1 Introduction

Neural Machine Translation has made significant strides and has achieved a level of quality that proves valuable in numerous everyday scenarios. Nonetheless, various assessment methods for Machine Translation suggest that there is still ample room for enhancement. One such evaluation approach, geared towards identifying translation deficiencies in a more systematic manner, involves the utilization of test suites or challenge sets. Unlike conventional evaluations that draw test sets from random everyday texts, test suites comprise sentences that are carefully curated or selected to assess the MT systems' competence in translating specific linguistic phenomena. In this context, we present the results obtained from applying these test suites, analyzing the performance of state-of-the-art systems concerning numerous linguistically-driven phenomena. These test suites were administered to 12 MT systems submitted during the 8th Conference of Machine Translation (WMT23) (Kocmi et al., 2023) for English–German language pair.

We have developed a comprehensive test suite that encompasses five distinct domains (entertainment, environment, health, science, legal) and spans five different writing styles (descriptive, judgments, narrative, reporting, technical writing). The primary objective of the test suite is not to gauge a system's overall translation performance, as this aspect is already evaluated through manual assessment and various additional metrics within the primary shared task. Instead, the test suite focuses on assessing the translational proficiency across diverse domains and writing styles.

2 Test suite details

Table 1 illustrates the distribution of sentences per domain and per writing style, with a total of 2268 sentences.

2.1 Sentence Selection

In order to ensure diversity and robustness in our test suite, we collected English sentences from a wide array of sources, including BBC NEWS, Children's Stories, Textbooks, Journals, and Legal Datasets. These sentences were then categorized into clusters based on several criteria, such as the count of Noun Phrases (NP), Verb Phrases (VP), Named Entities (NE), Subordinate Clauses (SC), Discourse Markers (DM), Punctuation (P), and Sentence Length (SL).

Within each domain, we chose to include 70% of the sentences from each cluster in our dataset, thereby augmenting the diversity and comprehensiveness of our test suite.

2.2 Evaluation

Our automatic evaluation process for the 12 systems is conducted in three phases. The first phase assesses the overall test suite, the second phase focuses on specific domains, and the third phase examines various writing styles. In addition to these automatic evaluations, we conducted manual

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 246–251
December 6–7, 2023.

Writing Style	Domain					Total
	Entertainment	*Environment*	*Health*	*Science*	*Legal*	
Descriptive		27	39	33		427
Judgements					348	449
Narrative		38	33	61		492
Reporting	427	374	399	458		552
Technical-writing		10	21			348
Total	99	348	132	1658	31	**2268**

Table 1: Test-suite statistics (Count of sentences in each domain per writing-style)

MT systems	COMETKIWI
ONLINE-B	0.847 (1)
ONLINE-Y	0.847 (1)
ONLINE-W	0.846 (3)
ONLINE-A	0.845 (4)
GPT4-5shot (Hendy et al., 2023)	0.842 (5)
ONLINE-G	0.841 (6)
ONLINE-M	0.839 (7)
Lan-BridgeMT (Wu and Hu, 2023)	0.833 (8)
NLLB_Greedy (NLLB Team et al., 2022)	0.831 (9)
NLLB_MBR_BLE	0.831 (9)
ZengHuiMT (Zeng, 2023)	0.815 (11)
AIRC (Rikters and Miwa, 2023)	0.809 (12)

Table 2: System-wise ranking based on COMETKIWI scores. Top five systems are highlighted in bold. Ranks are mentioned in brackets

analyses with the assistance of professional German speakers who aided us in identifying the errors made by the systems, providing valuable insights into their translation quality.

2.3 Experiment Setup

In this paper, we present the evaluation of 12 sysems with our test suite. The systems are part of the news translation task of the Eighth Conference on Machine Translation (WMT23). We cover the system outputs for English-German (en-de) language pair.

2.4 Automatic Evaluation

To evaluate the performance of the 12 submitted MT systems, we utilize COMETKIWI (Rei et al., 2022) scores, which offer quality estimation scores derived from the source sentence and MT output. Using these scores, we determine the system rankings, as outlined in Table 2. We chose COMETKIWI because it performed best among the other reference-free metrics in the recent WMT22 Metrics Shared Task (Freitag et al., 2022).

2.4.1 Domain-wise Evaluation

We have calculated COMETKIWI scores for each domain and presented them in Figure 1.

From this figure, we can deduce that ONLINE-B, ONLINE-Y, ONLINE-W, and ONLINE-A exhibit a high degree of consistency in their performance across all five domains.

However, it is worth noting that GPT4-5shot displayed subpar performance when applied to legal data, while NLLB_Greedy demonstrated comparatively lower performance in the context of environmental data.

Another important evident observation is that the machine translation (MT) systems exhibit a similar trend in both the health and science domains. This similarity may be attributed to the interconnected nature of these domains.

Notably, both ZengHuiMT and AIRC displayed consistently poor performance across all domains.

2.4.2 Writing-Style-wise Evaluation

We have computed COMETKIWI scores for sentence belonging to various writing styles and visualized the results in Figure 2.

ONLINE-W excels in narrative writing style sentences, but its performance declines significantly for technical writing style. In contrast, NLLB_Greedy performs poorly across descriptive, reporting, and technical writing styles.

Both ZengHuiMT and AIRC exhibit subpar performance across all the writing-styles. Additionally, GPT4-5Shot experiences a decline in its performance when it comes to judgments.

ONLINE-G, on the other hand, demonstrates better performance in technical writing and reporting styles.

Indeed, based on COMETKIWI scores, it is clear that both ONLINE-B and ONLINE-Y consistently outperformed other MT systems across a diverse array of writing styles and domains. This consistent

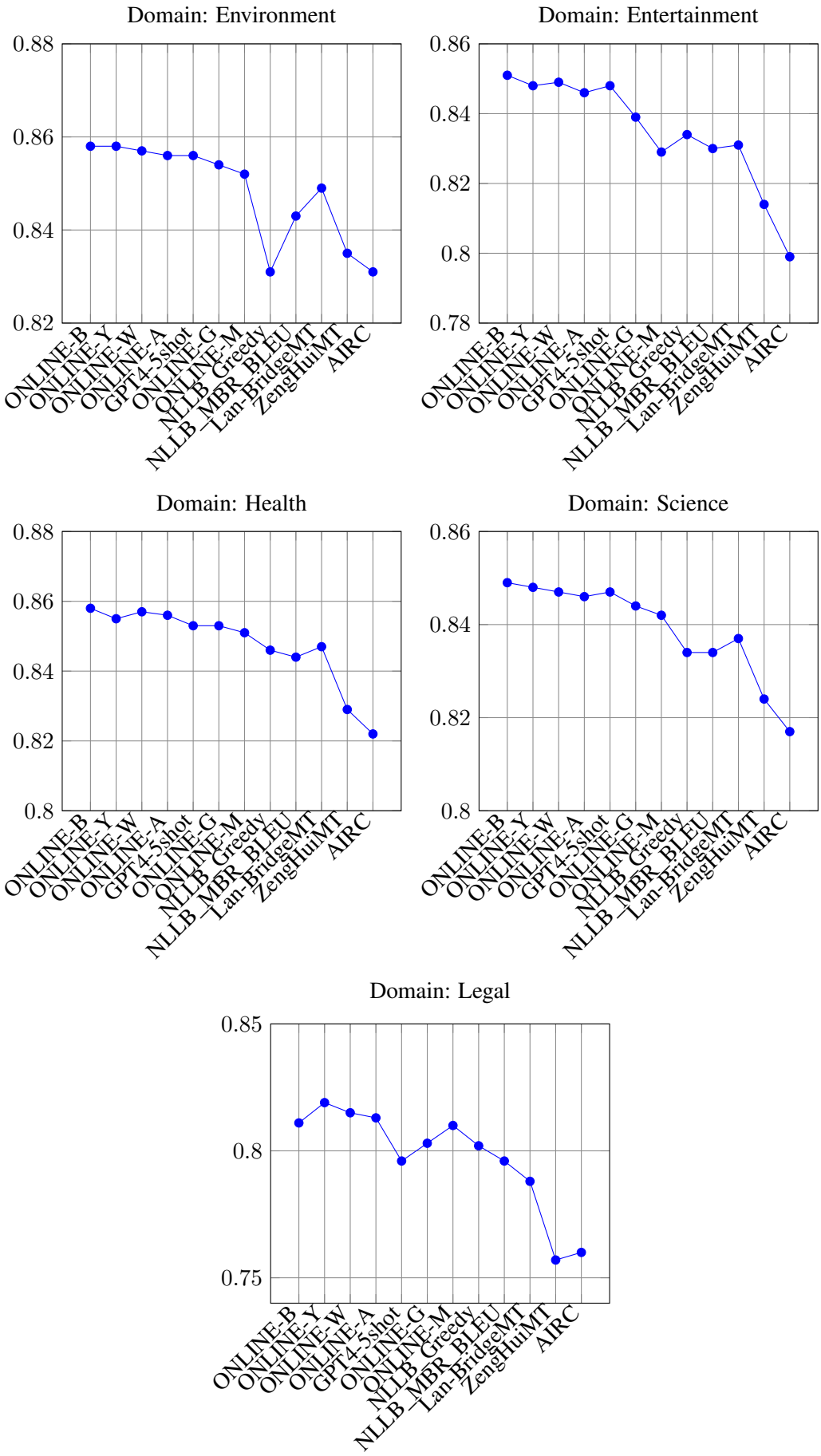

Figure 1: COMETKIWI scores of the systems with respect to domains

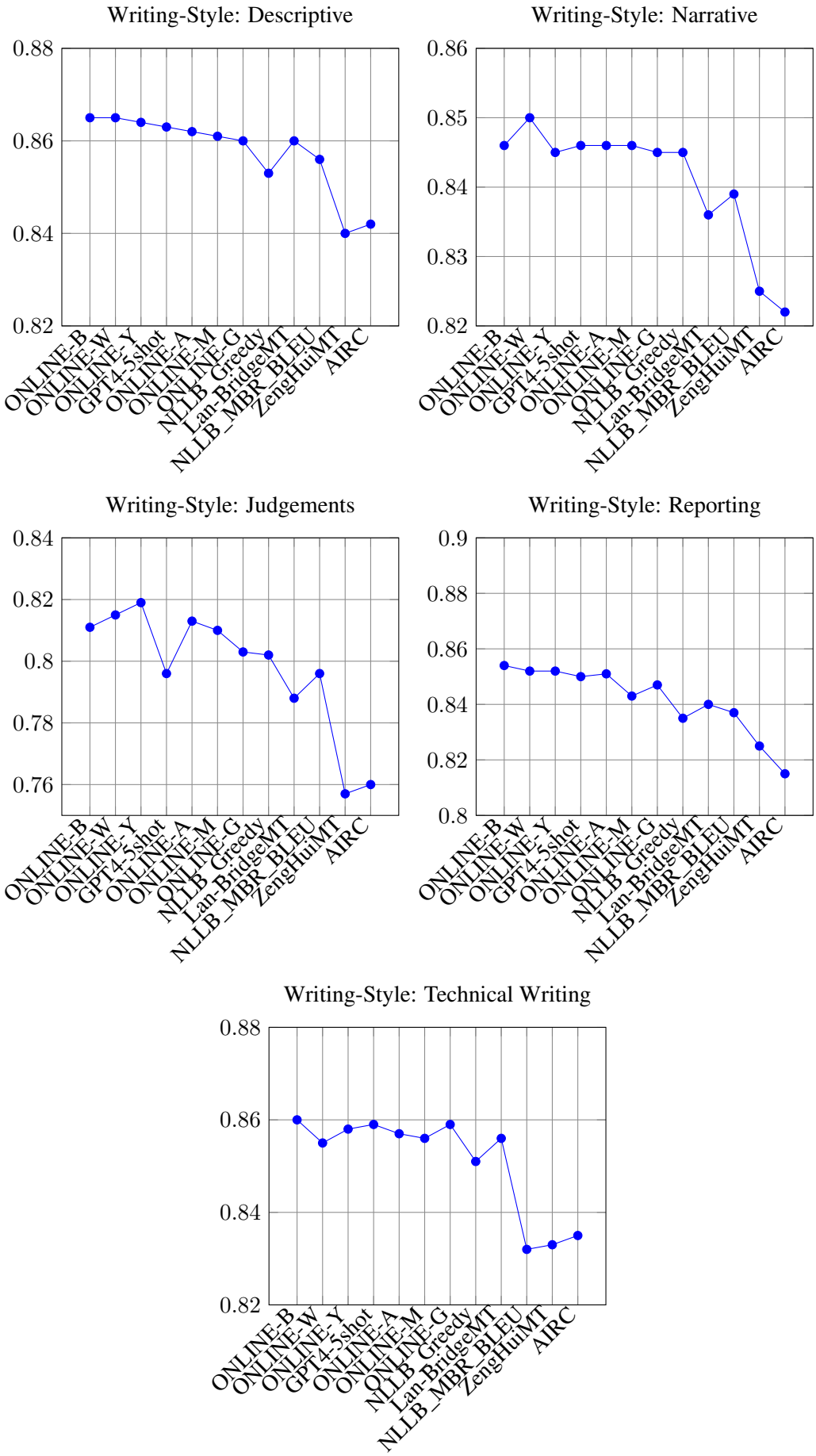

Figure 2: COMETKIWI scores of the systems with respect to writing-styles

superiority in performance suggests that these two MT systems are more robust and versatile, making them strong contenders for a wide range of translation tasks and scenarios.

2.5 Manual Assessments

These manual assessments are carried out voluntarily by professional German speakers who hold graduate-level qualifications and possess good knowledge in the domains covered by our test suite.

2.5.1 Gender-Neutral Pronouns

Machine translation (MT) systems often ascribe gender (sein/ihr $\sim$ his/her) to gender-neutral pronouns (it) in English. For instance, in the sentence '*Its age is not too dissimilar*,' ONLINE-B, ONLINE-M, ONLINE-G, ONLINE-A, ONLINE-W, Lan-BridgeMT, GPT4-5shot, and ZengHuiMT tended to assign the masculine gender 'Sein,' while the remaining systems ONLINE-Y, NLLB_Greedy, NLLB_MBR_BLEU, and AIRC preferred the feminine gender 'Ihr.' However, it's worth noting that in German, 'Sein' is typically used for neutral gender, thus introducing an intriguing linguistic nuance.

2.5.2 Repetition

Another intriguing factor is the phenomenon of Repetition, which is evident in cases like ZengHuiMT, where the translation includes additional information.

English source: a) Doing that amount is enough to reduce the risk of developing heart disease and stroke by 17% and cancer by 7%, the findings suggest.
b) While all living elements — the birds, animals and plants, forests, fisheries etc.— are biotic elements, abiotic elements include air, water, land etc.

Translation by ZengHuiMT: a) Die Ergebnisse deuten darauf hin, dass diese Menge ausreicht, um das Risiko für Herzerkrankungen und Schlaganfälle um 17 % und für Krebs um 7 % zu senken, so die Ergebnisse.
b) Während alle lebenden Elemente - Vögel, Tiere und Pflanzen, Wälder, Fischerei usw. - sind. Sie sind biotische Elemente, abiotische Elemente umfassen Luft, Wasser, Land usw.

Comment: a) The German translation is clear but includes an unnecessary repetition of **so die Ergebnisse** (the findings suggest) at the end.
b) Introduces an unnecessary repetition with **Sie sind biotische Elemente**.

2.5.3 Retention

Retention is another aspect that MT evaluation must consider. When it comes to challenging or complex words, retaining them might be permissible. However, for common or simpler words, retention should be heavily penalized.

Consider an example, "*These issues rarely have simple, single-discipline solutions that can be identified in one-off events or meetings.*" where ONLINE-B, ONLINE-M, GPT4-5shot, Lan-BridgeMT and AIRC MT systems retained the word *meetings* instead of translating it to *treffen*. This highlights the importance of addressing word retention in MT evaluation.

Manual assessments are indeed valuable for identifying gaps in machine translation quality. However, they come with significant drawbacks, including the need for extensive, non-reproducible human effort, time consumption, and high costs. Therefore, in addition to diverse test sets, it is crucial to develop robust automatic evaluation metrics capable of detecting and quantifying translation flaws efficiently and consistently.

3 Conclusion

This paper provides a comprehensive overview of our evaluation of 12 machine translation systems designed for the English-German language pair, all of which were submitted to the Shared Task during the 8th Conference on Machine Translation (WMT23). Our evaluation comprises a robust and diverse test-suite covering five distinct domains and encompassing five diverse writing styles. We conduct our analysis through a combination of automated assessments and manual evaluations, with a particular focus on domain-specific and writing style-specific performance. Based on our automatic evaluation, it is evident that both ONLINE-B and ONLINE-Y consistently surpassed other MT systems in performance across a diverse array of writing styles and domains.

References

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task.

Matīss Rikters and Makoto Miwa. 2023. Aist airc submissions to the wmt23 shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with gpt language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hui Zeng. 2023. Achieving state-of-the-art multilingual translation model with minimal data and parameters. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.