# Multifaceted Challenge Set for Evaluating Machine Translation Performance

**Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang**
**Zongyao Li, Jiaxin Guo, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, Yanfei Jiang**
Huawei Translation Services Center, Beijing, China
{chenxiaoyu35,weidaimeng,wuzhanglin2,zhuting20,shanghengchao,
lizongyao,guojiaxin1,nicolas.xie,leilizhi,
yanghao30,jiangyanfei}@huawei.com

## Abstract

Machine Translation Evaluation is critical to Machine Translation researches, as the evaluation results reflect the effectiveness of training strategies. As a result, a fair and efficient evaluation method is necessary. Many researchers have raised questions about currently available evaluation metrics from various perspectives, and propose suggestions accordingly. However, to our knowledge, few researchers has analyzed the difficulty level of source sentence and its influence on evaluation results. This paper presents HW-TSC's submission to the WMT23 MT Test Suites shared task. We propose a systematic approach for construing challenge sets from four aspects: word difficulty, length difficulty, grammar difficulty and model learning difficulty. We open-source two Multifaceted Challenge Sets for Zh→En and En→Zh. We also present results of participants in this year's General MT shared task on our test sets.

## 1 Introduction

Machine Translation (MT) Evaluation is an indispensable part of MT research, helping researchers verify the effectiveness of proposed training strategies and offering suggestions for future researches. However, automatic machine evaluation has raised a lot of concerns during decades of practices. One research direction is to explore the weakness of available evaluation metrics (Koehn and Monz, 2006; Callison-Burch et al., 2006; Post, 2018; Chen et al., 2022). Another direction is to analyze the soundness of test sets. For example, Freitag et al. (2020) discuss the impact of reference translationese on the evaluation results.

However, to our knowledge, few researches (Ahrenberg, 2018; Isabelle et al., 2017) has been done to discuss the influence of source sentences on the evaluation results. With the advancement of machine translations in recent years, we think that randomly sampled test sets may not be able to reflect the true gaps among models, as you can't test freshman's capability with grade-1 quiz. So we propose a strategy to collect test sentences with high-level of difficulty. The strategy considers a sentence's difficulty level from four dimensions, including word difficulty, length difficulty, grammar difficulty and model learning difficulty.

This paper presents our constructed Multifaceted Challenge Sets[1] for Zh→En and En→Zh language pairs using the strategy mentioned above. Each of the test set contains 2,000 sentences. The source sentences are from the open-sourced English Wikipedia corpus[2] while the translations are provided by our in-house translators. We report the results of participants in this year's General MT shared task on our test sets and hope to gain some insight by comparing our results with the official evaluation results.

## 2 Challenge Set Construction

### 2.1 Measuring Difficulty Level of a Test Set

We propose four indexes to measure the difficulty level of a sentence: word difficulty, length difficulty, grammar difficulty and model learning difficulty.

**Word Difficulty** Word difficulty is measured based on the frequency of a word appeared in the parallel training corpus. In general, the lower the frequency of a word in the training data, the more challenging for neural machine translation (NMT) to translate the word correctly.

We calculate the frequency of all words in the officially provided parallel data for the General MT shared task, and select words with frequency of more than 10 times and less than 99 times as the low-frequency word list. It should be noted that although some words fall into this frequency

[1] The test sets are open-sourced at: https://github.com/HwTsc/Multifaceted_Challenge_Set_for_MT

[2] https://dumps.Wikipediamedia.org/enWikipedia/, version 20230520 is used

*Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 217–223
December 6–7, 2023. ©2023 Association for Computational Linguistics

| system | BLEU | chrF | COMET22 | $\text{Rank}_{BLEU}$ | $\text{Rank}_{chrF}$ | $\text{Rank}_{COMET}$ |
|---|---|---|---|---|---|---|
| GPT4-5SHOT | 31.01 | 59.19 | 82.75 | 1 | 2 | 1 |
| Lan-BridgeMT | 29.81 | 59.45 | 82.16 | 2 | 1 | 2 |
| ONLINE-B | 29.67 | 57.60 | 80.32 | 3 | 3 | 3 |
| ZengHuiMT | 28.68 | 55.66 | 79.14 | 4 | 6 | 11 |
| Yishu | 27.64 | 54.95 | 80.14 | 5 | 8 | 5 |
| ONLINE-G | 27.15 | 57.37 | 80.00 | 6 | 4 | 6 |
| ONLINE-A | 27.08 | 56.25 | 79.99 | 7 | 5 | 7 |
| ONLINE-Y | 25.05 | 54.54 | 79.61 | 8 | 9 | 10 |
| **IOL_Research** | **24.95** | **52.53** | **80.21** | **9** | **11** | **4** |
| **HW-TSC** | **24.90** | **52.56** | **79.75** | **10** | **10** | **8** |
| ONLINE-W | 23.58 | 55.18 | 79.68 | 11 | 7 | 9 |
| ONLINE-M | 20.92 | 51.00 | 76.50 | 12 | 12 | 13 |
| NLLB_Greedy | 18.27 | 45.63 | 76.35 | 13 | 13 | 14 |
| NLLB_MBR_BLEU | 17.92 | 45.50 | 76.86 | 14 | 14 | 12 |
| **ANVITA** | **16.78** | **40.85** | **75.43** | **15** | **15** | **15** |

Table 1: BLEU, chrF and COMET Scores for the Zh→En translation task. Constrained systems are indicated in bold.

range, they can be divided into high-frequency subwords (e.g. newsagent = news + agent), which certainly does not meet the difficulty requirement. So we manually check the English and Chinese word lists and remove words that are consisted of high-frequency subwords. Finally we use the word lists to match the Wikipedia corpus to collect test sentences.

**Length Difficulty** Extremely Long and short sentences can be challenging for NMT models. In our daily practice, we find that omission and logic errors are more frequently seen in extremely long sentences. Meanwhile, due to the lack of enough context information, extremely short sentences are also error-prone.

We calculate the length (the number of English words/Chinese characters) of each sentence in the Wikipedia corpus and select 1,000 longest and shortest sentences respectively. We manually check semantics of each sentence and finally select 250 extremely long and 250 extremely short sentences as the test cases. The removed sentences include those that are incomplete, or contains obvious translationese (probably back-translation results from other languages).

**Grammar Difficulty** Kauchak et al. (2017) propose measuring the grammar difficulty of a sentence using the frequency of the 3rd level sentence parse tree. They employ Berkeley Parser to parse the 5.4M Wikipedia corpus and create 11 frequency bins.

Inspired by their strategy, we use Berkeley Parser to parse all sentences in the Wikipedia corpus and calculate the frequency of each 3rd level parse tree pattern. We exclude patterns that appear only once, which are highly possible to be noisy data. Then we select 1,000 sentences of which their grammar pattern has the lowest frequency as the candidate pool. Finally we manually check the semantics of each candidate and select 500 test sentences.

**Model Learning Difficulty** Zhao et al. (2019) observe that the translation quality is related to the entropy of the source sentence. The higher the source sentence entropy, the more likely the sentence is under-translated. They propose a formula to calculate entropy of the source sentence: Assume a word $s$ contains $K$ candidate translations, each of which has a probability $p_k$, the translation entropy for this word can be calculated by:

$$E(s) = -\sum_{k=1}^{k} p_k * \log p_k \qquad (1)$$

Using this formula, we calculate the entropy of each sentence in the Wikipedia corpus and select 1,000 sentences with the highest entropy as the candidate pool. Then we manually check the semantics of each sentence and finally select 500 test cases.

| system | BLEU | chrF | COMET22 | Rank$_{BLEU}$ | Rank$_{chrF}$ | Rank$_{COMET}$ |
|---|---|---|---|---|---|---|
| Yishu | 48.74 | 45.18 | 86.47 | 1 | 1 | 2 |
| ONLINE-B | 48.72 | 45.17 | 86.47 | 2 | 2 | 2 |
| ONLINE-W | 45.99 | 42.89 | 86.55 | 3 | 3 | 1 |
| **IOL_Research** | **45.28** | **41.17** | **85.29** | **4** | **4** | **6** |
| ONLINE-A | 44.92 | 40.72 | 84.82 | 5 | 5 | 8 |
| **HW-TSC** | **44.29** | **39.91** | **85.11** | **6** | **7** | **7** |
| ONLINE-Y | 43.72 | 40.03 | 84.51 | 7 | 6 | 9 |
| ONLINE-M | 41.85 | 39.24 | 82.1 | 8 | 8 | 10 |
| GPT4-5shot | 41.73 | 38.61 | 85.64 | 9 | 9 | 4 |
| LAN-BRIDGEMT | 39.89 | 37.83 | 85.52 | 10 | 10 | 5 |
| ONLINE-G | 39.77 | 37.09 | 81.63 | 11 | 11 | 11 |
| ZengHuiMT | 35.34 | 31.6 | 81.24 | 12 | 13 | 12 |
| **ANVITA** | **35.28** | **34.02** | **78.99** | **13** | **12** | **14** |
| NLLB_Greedy | 30.12 | 27.98 | 79 | 14 | 14 | 13 |
| NLLB_MBR_BLEU | 25.84 | 26.02 | 76.62 | 15 | 15 | 15 |

Table 2: BLEU, chrF and COMET Scores for the En→Zh translation task. Constrained systems are indicated in bold.

### 2.2 Test Set Composition

Our Zh→En and En→Zh test sets each contains 2,000 sentences, 500 sentences per category. The source sentences are selected from the open-source Englisjh Wikipedia corpus (version 20230520)[3], using the strategy we mentioned above. The target sentences are translated by our in-house translators, without referring to any machine translation models. We recruit 10 translators whose average working experience in the translation field exceed 5 years.

## 3 Results and Discussions

### 3.1 Results on the Multifaceted Challenge Set

Table 1 and Table 2 present the Zh→En and En→Zh results, including sacreBLEU (Post, 2018), chrF (Popović, 2015), and COMET-22 (Rei et al., 2022), as well as corresponding ranks. The ranks are quite different from the official results. However, as we are unable to keep the domain distribution of our test set the same as that of the official test set, we cannot draw a conclusion of whether the ranking difference is due to different levels of source sentence difficulty or domain difference.

If the ranking difference is caused by the different difficulty levels, we can conclude that systems that perform well on average test sets may not perform as well on challenge sets. So we may need a set of test sets at different difficulty levels to comprehensively evaluate model performance. Or if the ranking difference is caused by domain issues, the top-ranked systems on the official test sets may not be so general as the task name, General MT, suggests.

We also report COMET results on each subset (see table 3 and table 4) and try to understand model performance on each dimension. According to table 3, performances of Zh→En systems vary greater under the Word and Length dimensions, as the standard deviation scores are greater than that of other dimensions and the overall result. The result indicates that incorporating low-frequent words and extremely long/short sentences into the test set may better help to significantly differ model performances. The result is similar for En→Zh translation. As shown in table 4, the standard deviation under the Word dimension is much greater than that of the overall result and other dimensions. The standard deviation under the Length category is second largest, although a little bit lower than that of the overall result.

### 3.2 Towards More Sound Evaluation

Automatic evaluation is still the first option for MT researchers considering its speed and cost. More reliable evaluation metrics, e.g. COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), now provide more reliable evaluation results that more align with human evaluations. Meanwhile, we believe

[3] https://dumps.Wikipediamedia.org/enWikipedia/, version 20230520 is used.

| System | Vocab | Grammar | Length | Learning | overall |
|---|---|---|---|---|---|
| **ANVITA** | **72.32** | **77.72** | **73.52** | **78.33** | **75.43** |
| GPT4-5SHOT | 80.84 | 82.65 | 83.81 | 83.67 | 82.75 |
| **HW-TSC** | **77.66** | **80.69** | **79.02** | **81.59** | **79.75** |
| **IOL_Research** | **77.94** | **80.39** | **80.81** | **81.65** | **80.21** |
| Lan-BridgeMT | 80.31 | 82.05 | 83.24 | 83.02 | 82.16 |
| NLLB_Greedy | 73.41 | 78.33 | 74.81 | 78.84 | 76.35 |
| NLLB_MBR_BLEU | 73.59 | 78.41 | 76.05 | 79.4 | 76.86 |
| ONLINE-A | 77.97 | 80.12 | 80.9 | 80.92 | 79.99 |
| ONLINE-B | 78.4 | 80.53 | 80.98 | 81.32 | 80.32 |
| ONLINE-G | 78.03 | 80.46 | 81.09 | 80.39 | 80 |
| ONLINE-M | 73.7 | 77.52 | 76.18 | 78.56 | 76.5 |
| ONLINE-W | 77.4 | 80.36 | 79.71 | 81.26 | 79.68 |
| ONLINE-Y | 77.21 | 80.43 | 80.07 | 80.7 | 79.61 |
| Yishu | 78.49 | 80.58 | 80.1 | 81.34 | 80.14 |
| ZengHuiMT | 77.6 | 79.22 | 80.42 | 79.29 | 79.14 |
| Standard Deviation | 2.56 | 1.47 | 2.97 | 1.57 | 2.10 |

Table 3: COMET22 results of Zh→En systems on each subset and on the overall challenge set, as well as the standard deviation of all systems' COMET22 scores under the category.

| System | Vocab | Grammar | Length | Learning | Overall |
|---|---|---|---|---|---|
| **ANVITA** | **77.84** | **79.46** | **77.95** | **80.9** | **79.0** |
| GPT4-5SHOT | 83.88 | 85.78 | 86.3 | 86.7 | 85.6 |
| **HW-TSC** | **83.53** | **84.99** | **86.39** | **85.78** | **85.1** |
| **IOL_Research** | **83.63** | **85.03** | **86.87** | **85.87** | **85.3** |
| Lan-BridgeMT | 84.27 | 84.94 | 86.76 | 86.18 | 85.5 |
| NLLB_Greedy | 74.98 | 79.53 | 81.18 | 80.34 | 79.0 |
| NLLB_MBR_BLEU | 71.62 | 77.33 | 79.99 | 77.59 | 76.6 |
| ONLINE-A | 83.25 | 84.33 | 86.22 | 85.68 | 84.8 |
| ONLINE-B | 85.94 | 85.58 | 87.21 | 87.37 | 86.5 |
| ONLINE-G | 79.43 | 81.18 | 83.41 | 82.68 | 81.6 |
| ONLINE-M | 80.32 | 82.57 | 82.05 | 83.74 | 82.1 |
| ONLINE-W | 85.52 | 85.99 | 87.66 | 87.18 | 86.6 |
| ONLINE-Y | 82.93 | 84.47 | 85.3 | 85.55 | 84.5 |
| Yishu | 85.98 | 85.56 | 87.22 | 87.37 | 86.5 |
| ZengHuiMT | 79.38 | 81.76 | 81.19 | 82.9 | 81.2 |
| Standard Deviation | 4.20 | 2.76 | 3.14 | 2.95 | 3.19 |

Table 4: COMET22 results of En→Zh systems on each subset and on the overall challenge set, as well as the standard deviation of all systems' COMET22 scores under the category.

there should be a more systematic approach to construct test sets. In addition to domains, we should also put difficulty level into consideration. The randomly sampled test sets represent the average difficulty level in a certain domain, which can reflect the general capability of models. However, to learn the current weakness of MT and push further researches, we need challenge sets.

## 4 Conclusion and Limitations

This paper presents HW-TSC's submission to the WMT23 MT Test Suites shared task. We propose increasing the test set difficulty level to better measure model performances. We propose a strategy to collect test sets with high difficulty level: word difficulty, length difficulty, grammar difficulty and model learning difficulty. We construct two multifaceted Challenge Sets for Zh→En and En→Zh directions using this strategy and report automatic evaluations of participants in this year's General MT shared task on our test sets.

However, due to time constraints, we do not perform human evaluations on the test results, which we believe will offer more insights on the performance of our challenge sets. For future researches, we will conduct direct assessment (DA) and error annotations to explore the performance of each participants on the challenge sets and compare the result with the official test sets. In addition, we will construct relatively simple test sets in the same domain, and compare the results with these challenge sets, hoping to gain more insights on the role of source sentence difficulty level.

## References

Lars Ahrenberg. 2018. A challenge set for english-swedish machine translation.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, et al. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the wmt22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. *arXiv preprint arXiv:2004.06063*.

Pierre Isabelle, Colin Cherry, and George F. Foster. 2017. A challenge set approach to evaluating machine translation. In *Conference on Empirical Methods in Natural Language Processing*.

David Kauchak, Gondy Leroy, and Alan Hogue. 2017. Measuring text difficulty using parse-tree frequency. *Journal of the Association for Information Science and Technology*, 68(9):2088–2100.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121.

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.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT 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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur P Parikh. 2020. Learning to evaluate translation beyond english: Bleurt submissions to the wmt metrics 2020 shared task. *arXiv preprint arXiv:2010.04297*.

Yang Zhao, Jiajun Zhang, Chengqing Zong, Zhongjun He, and Hua Wu. 2019. Addressing the under-translation problem from the entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 451–458.

## A Appendix

### Statistical Significance for Zh→En and En→Zh

| . | Lan-BridgeMT | ONLINE-B | ZenghuiMT | Yishu | ONLINE-G | ONLINE-A | ONLINE-Y | IOL_Research | HW-TSC | ONLINE-W | ONLINE-M | NLLB_Greedy | NLLB_MBR_BLEU | ANVITA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT4-5SHOT | 1.2 | 1.3 | 2.3 | 3.4 | 3.9 | 3.9 | 6.0 | 6.1 | 6.1 | 7.4 | 10.1 | 12.7 | 13.1 | 14.2 |
| Lan-BridgeMT | 0.0 | 0.1 | 1.1 | 2.2 | 2.7 | 2.7 | 4.8 | 4.9 | 4.9 | 6.2 | 8.9 | 11.5 | 11.9 | 13.0 |
| ONLINE-B |  | 0.0 | 1.0 | 2.0 | 2.5 | 2.6 | 4.6 | 4.7 | 4.8 | 6.1 | 8.8 | 11.4 | 11.8 | 12.9 |
| ZenghuiMT |  |  | 0.0 | 1.0 | 1.5 | 1.6 | 3.6 | 3.7 | 3.8 | 5.1 | 7.8 | 10.4 | 10.8 | 11.9 |
| Yishu |  |  |  | 0.0 | 0.5 | 0.6 | 2.6 | 2.7 | 2.7 | 4.1 | 6.7 | 9.4 | 9.7 | 10.9 |
| ONLINE-G |  |  |  |  | 0.0 | 0.1 | 2.1 | 2.2 | 2.3 | 3.6 | 6.2 | 8.9 | 9.2 | 10.4 |
| ONLINE-A |  |  |  |  |  | 0.0 | 2.0 | 2.1 | 2.2 | 3.5 | 6.2 | 8.8 | 9.2 | 10.3 |
| ONLINE-Y |  |  |  |  |  |  | 0.0 | 0.1 | 0.2 | 1.5 | 4.1 | 6.8 | 7.1 | 8.3 |
| IOL_Research |  |  |  |  |  |  |  | 0.0 | 0.1 | 1.4 | 4.0 | 6.7 | 7.0 | 8.2 |
| HW-TSC |  |  |  |  |  |  |  |  | 0.0 | 1.3 | 4.0 | 6.6 | 7.0 | 8.1 |
| ONLINE-W |  |  |  |  |  |  |  |  |  | 0.0 | 2.7 | 5.3 | 5.7 | 6.8 |
| ONLINE-M |  |  |  |  |  |  |  |  |  |  | 0.0 | 2.7 | 3.0 | 4.1 |
| NLLB_Greedy |  |  |  |  |  |  |  |  |  |  |  | 0.0 | 0.3 | 1.5 |
| NLLB_MBR_BLEU |  |  |  |  |  |  |  |  |  |  |  |  | 0.0 | 1.1 |
| ANVITA |  |  |  |  |  |  |  |  |  |  |  |  |  | 0.0 |

height

Table 5: statistical significance testing of the BLEU score difference for each system pair for Zh→En. Score difference is in gray if the p-value is above 0.05

|  | GPT45SHOT | ONLINE-B | ONLINE-G | ONLINE-A | ZenghuiMT | ONLINE-W | Yishu | ONLINE-Y | HWTSC | IOL_Research | ONLINE-M | NLLB_Greedy | NLLB_MBR_BLEU | ANVITA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lan-BridgeMT | 0.3 | 1.9 | 2.1 | 3.2 | 3.8 | 4.3 | 4.5 | 4.9 | 6.9 | 6.9 | 8.5 | 13.8 | 14.0 | 18.6 |
| GPT4-5SHOT | 0.0 | 1.6 | 1.8 | 2.9 | 3.5 | 4.0 | 4.2 | 4.7 | 6.6 | 6.7 | 8.2 | 13.6 | 13.7 | 18.3 |
| ONLINE-B |  | 0.0 | 0.2 | 1.4 | 1.9 | 2.4 | 2.7 | 3.1 | 5.0 | 5.1 | 6.6 | 12.0 | 12.1 | 16.8 |
| ONLINE-G |  |  | 0.0 | 1.1 | 1.7 | 2.2 | 2.4 | 2.8 | 4.8 | 4.8 | 6.4 | 11.7 | 11.9 | 16.5 |
| ONLINE-A |  |  |  | 0.0 | 0.6 | 1.1 | 1.3 | 1.7 | 3.7 | 3.7 | 5.3 | 10.6 | 10.8 | 15.4 |
| ZenghuiMT |  |  |  |  | 0.0 | 0.5 | 0.7 | 1.1 | 3.1 | 3.1 | 4.7 | 10.0 | 10.2 | 14.8 |
| ONLINE-W |  |  |  |  |  | 0.0 | 0.2 | 0.6 | 2.6 | 2.7 | 4.2 | 9.6 | 9.7 | 14.3 |
| Yishu |  |  |  |  |  |  | 0.0 | 0.4 | 2.4 | 2.4 | 4.0 | 9.3 | 9.5 | 14.1 |
| ONLINE-Y |  |  |  |  |  |  |  | 0.0 | 2.0 | 2.0 | 3.5 | 8.9 | 9.0 | 13.7 |
| HW-TSC |  |  |  |  |  |  |  |  | 0.0 | 0.0 | 1.6 | 6.9 | 7.1 | 11.7 |
| IOL_Research |  |  |  |  |  |  |  |  |  | 0.0 | 1.5 | 6.9 | 7.0 | 11.7 |
| ONLINE-M |  |  |  |  |  |  |  |  |  |  | 0.0 | 5.4 | 5.5 | 10.2 |
| NLLB_Greedy |  |  |  |  |  |  |  |  |  |  |  | 0.0 | 0.1 | 4.8 |
| NLLB_MBR_BLEU |  |  |  |  |  |  |  |  |  |  |  |  | 0.0 | 4.7 |
| ANVITA |  |  |  |  |  |  |  |  |  |  |  |  |  | 0.0 |

height

Table 6: statistical significance testing of the chrF score difference for each system pair for Zh→En. Score difference is in gray if the p-value is above 0.05

|  | Lan-BridgeMT | ONLINE-B | IOL_Research | Yishu | ONLINE-G | ONLINE-A | HW-TSC | ONLINE-W | ONLINE-Y | ZenghuiMT | NLLB_MBR_BLEU | ONLINE-M | NLLB_Greedy | ANVITA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT4-5SHOT | 0.6 | 2.4 | 2.5 | 2.6 | 2.8 | 2.8 | 3.0 | 3.1 | 3.1 | 3.6 | 5.9 | 6.3 | 6.4 | 7.3 |
| Lan-BridgeMT | 0.0 | 1.8 | 2.0 | 2.0 | 2.2 | 2.2 | 2.4 | 2.5 | 2.6 | 3.0 | 5.3 | 5.7 | 5.8 | 6.7 |
| ONLINE-B |  | 0.0 | 0.1 | 0.2 | 0.3 | 0.3 | 0.6 | 0.6 | 0.7 | 1.2 | 3.5 | 3.8 | 4.0 | 4.9 |
| IOL_Research |  |  | 0.0 | 0.1 | 0.2 | 0.2 | 0.5 | 0.5 | 0.6 | 1.1 | 3.3 | 3.7 | 3.9 | 4.8 |
| Yishu |  |  |  | 0.0 | 0.1 | 0.2 | 0.4 | 0.5 | 0.5 | 1.0 | 3.3 | 3.6 | 3.8 | 4.7 |
| ONLINE-G |  |  |  |  | 0.0 | 0.0 | 0.3 | 0.3 | 0.4 | 0.9 | 3.1 | 3.5 | 3.7 | 4.6 |
| ONLINE-A |  |  |  |  |  | 0.0 | 0.2 | 0.3 | 0.4 | 0.8 | 3.1 | 3.5 | 3.6 | 4.6 |
| HW-TSC |  |  |  |  |  |  | 0.0 | 0.1 | 0.1 | 0.6 | 2.9 | 3.3 | 3.4 | 4.3 |
| ONLINE-W |  |  |  |  |  |  |  | 0.0 | 0.1 | 0.5 | 2.8 | 3.2 | 3.3 | 4.3 |
| ONLINE-Y |  |  |  |  |  |  |  |  | 0.0 | 0.5 | 2.8 | 3.1 | 3.3 | 4.2 |
| ZenghuiMT |  |  |  |  |  |  |  |  |  | 0.0 | 2.3 | 2.6 | 2.8 | 3.7 |
| NLLB_MBR_BLEU |  |  |  |  |  |  |  |  |  |  | 0.0 | 0.4 | 0.5 | 1.4 |
| ONLINE-M |  |  |  |  |  |  |  |  |  |  |  | 0.0 | 0.2 | 1.1 |
| NLLB_Greedy |  |  |  |  |  |  |  |  |  |  |  |  | 0.0 | 0.9 |
| ANVITA |  |  |  |  |  |  |  |  |  |  |  |  |  | 0.0 |

height

Table 7: statistical significance testing of the COMET score difference for each system pair for Zh→En. Score difference is in gray if the p-value is above 0.05

|  | ONLINE-B | ONLINE-W | IOL-Research | ONLINE-A | HW-TSC | ONLINE-Y | ONLINE-M | GPT4-5shot | LAN-BRIDGEMT | ONLINE-G | ZenghuiMT | ANVITA | NLLB_Greedy | NLLB_MBR_BLEU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| yishu | 0.0 | 2.8 | 3.5 | 3.8 | 4.5 | 5.0 | 6.9 | 7.0 | 8.9 | 9.0 | 13.4 | 13.5 | 18.6 | 22.9 |
| ONLINE-B | 0.0 | 2.7 | 3.4 | 3.8 | 4.4 | 5.0 | 6.9 | 7.0 | 8.8 | 9.0 | 13.4 | 13.4 | 18.6 | 22.9 |
| ONLINE-W |  | 0.0 | 0.7 | 1.1 | 1.7 | 2.3 | 4.1 | 4.3 | 6.1 | 6.2 | 10.7 | 10.7 | 15.9 | 20.2 |
| IOL-Research |  |  | 0.0 | 0.4 | 1.0 | 1.6 | 3.4 | 3.6 | 5.4 | 5.5 | 9.9 | 10.0 | 15.2 | 19.4 |
| ONLINE-A |  |  |  | 0.0 | 0.6 | 1.2 | 3.1 | 3.2 | 5.0 | 5.2 | 9.6 | 9.6 | 14.8 | 19.1 |
| HW-TSC |  |  |  |  | 0.0 | 0.6 | 2.4 | 2.6 | 4.4 | 4.5 | 9.0 | 9.0 | 14.2 | 18.5 |
| ONLINE-Y |  |  |  |  |  | 0.0 | 1.9 | 2.0 | 3.8 | 4.0 | 8.4 | 8.4 | 13.6 | 17.9 |
| ONLINE-M |  |  |  |  |  |  | 0.0 | 0.1 | 2.0 | 2.1 | 6.5 | 6.6 | 11.7 | 16.0 |
| GPT4-5shot |  |  |  |  |  |  |  | 0.0 | 1.8 | 2.0 | 6.4 | 6.5 | 11.6 | 15.9 |
| LAN-BRIDGEMT |  |  |  |  |  |  |  |  | 0.0 | 0.1 | 4.6 | 4.6 | 9.8 | 14.1 |
| ONLINE-G |  |  |  |  |  |  |  |  |  | 0.0 | 4.4 | 4.5 | 9.7 | 13.9 |
| ZenghuiMT |  |  |  |  |  |  |  |  |  |  | 0.0 | 0.1 | 5.2 | 9.5 |
| ANVITA |  |  |  |  |  |  |  |  |  |  |  | 0.0 | 5.2 | 9.4 |
| NLLB_Greedy |  |  |  |  |  |  |  |  |  |  |  |  | 0.0 | 4.3 |
| NLLB_MBR_BLEU |  |  |  |  |  |  |  |  |  |  |  |  |  | 0.0 |

height

Table 8: statistical significance testing of the BLEU score difference for each system pair for En→Zh. Score difference is in gray if the p-value is above 0.05

| | ONLINE-B | ONLINE-W | IOL-Research | ONLINE-A | ONLINE-Y | HW-TSC | ONLINE-M | GPT4-5shot | LAN-BRIDGEMT | ONLINE-G | ANVITA | ZenghuiMT | NLLB_Greedy | NLLB_MBR_BLEU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| yishu | 0.0 | 2.3 | 4.0 | 4.5 | 5.2 | 5.3 | 5.9 | 6.6 | 7.4 | 8.1 | 11.2 | 13.6 | 17.2 | 19.2 |
| ONLINE-B | 0.0 | 2.3 | 4.0 | 4.5 | 5.1 | 5.3 | 5.9 | 6.6 | 7.3 | 8.1 | 11.2 | 13.6 | 17.2 | 19.2 |
| ONLINE-W | | 0.0 | 1.7 | 2.2 | 2.9 | 3.0 | 3.7 | 4.3 | 5.1 | 5.8 | 8.9 | 11.3 | 14.9 | 16.9 |
| IOL-Research | | | 0.0 | 0.5 | 1.1 | 1.3 | 1.9 | 2.6 | 3.3 | 4.1 | 7.2 | 9.6 | 13.2 | 15.2 |
| ONLINE-A | | | | 0.0 | 0.7 | 0.8 | 1.5 | 2.1 | 2.9 | 3.6 | 6.7 | 9.1 | 12.7 | 14.7 |
| ONLINE-Y | | | | | 0.0 | 0.1 | 0.8 | 1.4 | 2.2 | 2.9 | 6.0 | 8.4 | 12.1 | 14.0 |
| HW-TSC | | | | | | 0.0 | 0.7 | 1.3 | 2.1 | 2.8 | 5.9 | 8.3 | 11.9 | 13.9 |
| ONLINE-M | | | | | | | 0.0 | 0.6 | 1.4 | 2.2 | 5.2 | 7.6 | 11.3 | 13.2 |
| GPT4-5shot | | | | | | | | 0.0 | 0.8 | 1.5 | 4.6 | 7.0 | 10.6 | 12.6 |
| LAN-BRIDGEMT | | | | | | | | | 0.0 | 0.7 | 3.8 | 6.2 | 9.9 | 11.8 |
| ONLINE-G | | | | | | | | | | 0.0 | 3.1 | 5.5 | 9.1 | 11.1 |
| ANVITA | | | | | | | | | | | 0.0 | 2.4 | 6.0 | 8.0 |
| ZenghuiMT | | | | | | | | | | | | 0.0 | 3.6 | 5.6 |
| NLLB_Greedy | | | | | | | | | | | | | 0.0 | 2.0 |
| NLLB_MBR_BLEU | | | | | | | | | | | | | | 0.0 |

height

Table 9: statistical significance testing of the chrF score difference for each system pair for En→Zh. Score difference is in gray if the p-value is above 0.05

| | ONLINE-B | ONLINE-W | IOL-Research | ONLINE-A | HW-TSC | ONLINE-Y | ONLINE-M | GPT4-5shot | LAN-BRIDGEMT | ONLINE-G | ZenghuiMT | ANVITA | NLLB_Greedy | NLLB_MBR_BLEU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| yishu | 0.1 | 0.1 | 0.9 | 1.0 | 1.3 | 1.4 | 1.7 | 2.0 | 4.5 | 4.9 | 5.3 | 7.6 | 7.6 | 9.9 |
| ONLINE-B | 0.0 | 0.0 | 0.8 | 1.0 | 1.2 | 1.4 | 1.7 | 2.0 | 4.4 | 4.8 | 5.2 | 7.5 | 7.5 | 9.8 |
| ONLINE-W | | 0.0 | 0.8 | 1.0 | 1.2 | 1.4 | 1.7 | 2.0 | 4.4 | 4.8 | 5.2 | 7.5 | 7.5 | 9.8 |
| IOL-Research | | | 0.0 | 0.1 | 0.3 | 0.5 | 0.8 | 1.1 | 3.5 | 4.0 | 4.4 | 6.6 | 6.7 | 9.0 |
| ONLINE-A | | | | 0.0 | 0.2 | 0.4 | 0.7 | 1.0 | 3.4 | 3.9 | 4.3 | 6.5 | 6.5 | 8.9 |
| HW-TSC | | | | | 0.0 | 0.2 | 0.5 | 0.8 | 3.2 | 3.7 | 4.1 | 6.3 | 6.3 | 8.7 |
| ONLINE-Y | | | | | | 0.0 | 0.3 | 0.6 | 3.0 | 3.5 | 3.9 | 6.1 | 6.1 | 8.5 |
| ONLINE-M | | | | | | | 0.0 | 0.3 | 2.7 | 3.2 | 3.6 | 5.8 | 5.8 | 8.2 |
| GPT4-5shot | | | | | | | | 0.0 | 2.4 | 2.9 | 3.3 | 5.5 | 5.5 | 7.9 |
| LAN-BRIDGEMT | | | | | | | | | 0.0 | 0.5 | 0.9 | 3.1 | 3.1 | 5.5 |
| ONLINE-G | | | | | | | | | | 0.0 | 0.4 | 2.6 | 2.6 | 5.0 |
| ZenghuiMT | | | | | | | | | | | 0.0 | 2.2 | 2.3 | 4.6 |
| ANVITA | | | | | | | | | | | | 0.0 | 0.0 | 2.4 |
| NLLB_Greedy | | | | | | | | | | | | | 0.0 | 2.4 |
| NLLB_MBR_BLEU | | | | | | | | | | | | | | 0.0 |

height

Table 10: statistical significance testing of the COMET score difference for each system pair for En→Zh. Score difference is in gray if the p-value is above 0.05