# Automatic Evaluation of the WMT23 General Machine Translation Task

**WMT23 Organizers**

In this document, we present an automatic evaluation of the systems submitted to the general machine translation task. Please keep in mind that these rankings are not official. WMT only uses human evaluation for the official rankings.

We ran three different automatic metrics:

- chrF (Popović, 2015): A tokenization independent metric operating at character-level with a higher correlation with human judgments than BLEU.
- BLEU (Papineni et al., 2002): The standard BLEU.
- COMET (Rei et al., 2020): A state-of-the-art metric based on a pre-trained language model. We used the default model "Unbabel/wmt22-comet-da."

chrF and BLEU scores are computed with SacreBLEU (Post, 2018).[1]

We also tested whether the difference between systems' metric scores is statistically significant. We used the default parameters of "comet-compare" for paired bootstrap resampling (Koehn, 2004). In the tables reporting on statistical significant testing, the background color is darker for more significant differences (lower p-value) and the score difference is underlined if the p-value is below 0.05.

[1] https://github.com/mjpost/sacrebleu

| System | Rank | chrF |
|---|---|---|
| GPT4-5shot | 1 | 61.0 |
| CUNI-GA | 2 | 57.9 |
| GTCOM_Peter | 3 | 57.6 |
| CUNI-Transformer | 4 | 57.4 |
| MUNI-NLP | 5 | 57.0 |
| Lan-BridgeMT | 6 | 55.7 |
| ONLINE-W | 7 | 55.0 |
| ONLINE-B | 8 | 54.7 |
| ONLINE-A | 9 | 54.4 |
| ONLINE-G | 10 | 53.7 |
| ONLINE-Y | 11 | 53.4 |
| NLLB_Greedy | 12 | 52.5 |
| NLLB_MBR_BLEU | 13 | 52.3 |

| System | Rank | BLEU |
|---|---|---|
| GPT4-5shot | 1 | 32.8 |
| CUNI-Transformer | 2 | 30.2 |
| GTCOM_Peter | 3 | 29.8 |
| CUNI-GA | 4 | 29.5 |
| MUNI-NLP | 5 | 28.3 |
| Lan-BridgeMT | 6 | 27.5 |
| ONLINE-W | 7 | 26.8 |
| ONLINE-B | 8 | 25.7 |
| ONLINE-A | 9 | 25.4 |
| NLLB_MBR_BLEU | 10 | 25.1 |
| NLLB_Greedy | 11 | 24.9 |
| ONLINE-G | 12 | 24.8 |
| ONLINE-Y | 13 | 24.2 |

| System | Rank | COMET |
|---|---|---|
| CUNI-GA | 1 | 90.9 |
| GPT4-5shot | 2 | 90.8 |
| ONLINE-W | 3 | 89.4 |
| GTCOM_Peter | 4 | 88.9 |
| ONLINE-B | 5 | 88.8 |
| ONLINE-A | 6 | 88.2 |
| CUNI-Transformer | 7 | 88.0 |
| ONLINE-G | 8 | 87.7 |
| MUNI-NLP | 9 | 87.0 |
| ONLINE-Y | 10 | 86.5 |
| NLLB_Greedy | 11 | 86.3 |
| NLLB_MBR_BLEU | 12 | 86.3 |
| Lan-BridgeMT | 13 | 86.0 |

Table 1: Scores for the cs→uk translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF |
|---|---|---|
| ONLINE-W | 1 | 76.3 |
| ONLINE-B | 2 | 70.4 |
| ZengHuiMT | 3 | 67.5 |
| ONLINE-A | 4 | 66.3 |
| CUNI-GA | 5 | 65.9 |
| GTCOM_Peter | 6 | 65.4 |
| CUNI-DocTransformer | 7 | 65.1 |
| ONLINE-Y | 8 | 64.6 |
| CUNI-Transformer | 9 | 63.9 |
| Lan-BridgeMT | 10 | 63.8 |
| ONLINE-G | 11 | 63.7 |
| ONLINE-M | 12 | 63.2 |
| GPT4-5shot | 13 | 62.3 |
| NLLB_Greedy | 14 | 60.0 |
| NLLB_MBR_BLEU | 15 | 59.1 |

| System | Rank | BLEU |
|---|---|---|
| ONLINE-W | 1 | 59.4 |
| ONLINE-B | 2 | 50.1 |
| ONLINE-A | 3 | 43.4 |
| CUNI-GA | 4 | 43.3 |
| ZengHuiMT | 5 | 43.1 |
| CUNI-DocTransformer | 6 | 42.5 |
| GTCOM_Peter | 7 | 42.3 |
| CUNI-Transformer | 8 | 41.4 |
| ONLINE-Y | 9 | 40.8 |
| Lan-BridgeMT | 10 | 40.7 |
| ONLINE-G | 11 | 39.6 |
| ONLINE-M | 12 | 39.6 |
| GPT4-5shot | 13 | 37.8 |
| NLLB_Greedy | 14 | 35.9 |
| NLLB_MBR_BLEU | 15 | 35.1 |

| System | Rank | COMET |
|---|---|---|
| ONLINE-W | 1 | 91.8 |
| CUNI-GA | 2 | 90.8 |
| ONLINE-B | 3 | 89.9 |
| GPT4-5shot | 4 | 89.4 |
| ONLINE-A | 5 | 88.4 |
| CUNI-DocTransformer | 6 | 88.3 |
| GTCOM_Peter | 7 | 87.7 |
| ONLINE-M | 8 | 87.4 |
| Lan-BridgeMT | 9 | 87.3 |
| CUNI-Transformer | 10 | 87.2 |
| NLLB_Greedy | 11 | 87.1 |
| ONLINE-Y | 12 | 87.0 |
| NLLB_MBR_BLEU | 13 | 86.9 |
| ONLINE-G | 14 | 85.9 |
| ZengHuiMT | 15 | 85.4 |

Table 2: Scores for the en→cs translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF |
|---|---|---|
| ONLINE-W | 1 | 72.1 |
| ONLINE-A | 2 | 70.0 |
| GPT4-5shot | 3 | 69.8 |
| ONLINE-B | 4 | 69.1 |
| ONLINE-G | 5 | 69.1 |
| ONLINE-Y | 6 | 68.4 |
| ZengHuiMT | 7 | 67.6 |
| Lan-BridgeMT | 8 | 66.7 |
| GTCOM_Peter | 9 | 66.6 |
| ONLINE-M | 10 | 66.5 |
| NLLB_MBR_BLEU | 11 | 57.6 |
| NLLB_Greedy | 12 | 57.3 |
| AIRC | 13 | 57.2 |

| System | Rank | BLEU |
|---|---|---|
| ONLINE-W | 1 | 51.8 |
| GPT4-5shot | 2 | 47.9 |
| ONLINE-A | 3 | 47.9 |
| ONLINE-B | 4 | 46.3 |
| ONLINE-G | 5 | 46.0 |
| ONLINE-Y | 6 | 43.9 |
| GTCOM_Peter | 7 | 42.2 |
| Lan-BridgeMT | 8 | 42.1 |
| ONLINE-M | 9 | 41.3 |
| ZengHuiMT | 10 | 40.8 |
| NLLB_Greedy | 11 | 33.1 |
| AIRC | 12 | 32.4 |
| NLLB_MBR_BLEU | 13 | 32.4 |

| System | Rank | COMET |
|---|---|---|
| GPT4-5shot | 1 | 86.3 |
| ONLINE-W | 2 | 86.0 |
| ONLINE-B | 3 | 85.6 |
| ONLINE-A | 4 | 85.5 |
| ONLINE-Y | 5 | 84.9 |
| ONLINE-M | 6 | 84.8 |
| ONLINE-G | 7 | 84.6 |
| GTCOM_Peter | 8 | 82.7 |
| NLLB_MBR_BLEU | 9 | 81.4 |
| ZengHuiMT | 10 | 81.1 |
| Lan-BridgeMT | 11 | 80.9 |
| NLLB_Greedy | 12 | 79.9 |
| AIRC | 13 | 78.7 |

Table 3: Scores for the de→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF |
|---|---|---|
| ONLINE-W | 1 | 71.8 |
| ONLINE-A | 2 | 69.7 |
| ZengHuiMT | 3 | 69.4 |
| GPT4-5shot | 4 | 69.1 |
| ONLINE-B | 5 | 69.1 |
| ONLINE-Y | 6 | 69.1 |
| ONLINE-G | 7 | 69.0 |
| ONLINE-M | 8 | 66.9 |
| Lan-BridgeMT | 9 | 66.1 |
| NLLB_Greedy | 10 | 56.2 |
| NLLB_MBR_BLEU | 11 | 55.4 |
| AIRC | 12 | 52.2 |

| System | Rank | BLEU |
|---|---|---|
| ONLINE-W | 1 | 47.8 |
| ONLINE-A | 2 | 43.7 |
| GPT4-5shot | 3 | 43.6 |
| ONLINE-Y | 4 | 43.6 |
| ONLINE-G | 5 | 43.2 |
| ONLINE-B | 6 | 42.7 |
| ONLINE-M | 7 | 40.5 |
| ZengHuiMT | 8 | 40.5 |
| Lan-BridgeMT | 9 | 39.4 |
| NLLB_Greedy | 10 | 31.1 |
| NLLB_MBR_BLEU | 11 | 29.6 |
| AIRC | 12 | 26.5 |

| System | Rank | COMET |
|---|---|---|
| ONLINE-W | 1 | 85.5 |
| GPT4-5shot | 2 | 85.0 |
| ONLINE-B | 3 | 84.8 |
| ONLINE-Y | 4 | 84.1 |
| ONLINE-A | 5 | 83.7 |
| ONLINE-G | 6 | 82.5 |
| ONLINE-M | 7 | 81.7 |
| Lan-BridgeMT | 8 | 80.4 |
| ZengHuiMT | 9 | 79.4 |
| NLLB_MBR_BLEU | 10 | 78.0 |
| NLLB_Greedy | 11 | 77.9 |
| AIRC | 12 | 72.9 |

Table 4: Scores for the en→de translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF | System | Rank | BLEU | System | Rank | COMET |
|---|---|---|---|---|---|---|---|---|
| ONLINE-B | 1 | 87.5 | ONLINE-B | 1 | 76.5 | ONLINE-B | 1 | 89.9 |
| ZengHuiMT | 2 | 76.3 | GTCOM_Peter | 2 | 59.2 | ONLINE-A | 2 | 87.0 |
| GTCOM_Peter | 3 | 76.2 | ZengHuiMT | 3 | 56.6 | GPT4-5shot | 3 | 86.9 |
| ONLINE-A | 4 | 73.3 | ONLINE-A | 4 | 53.9 | GTCOM_Peter | 4 | 86.7 |
| GPT4-5shot | 5 | 71.4 | GPT4-5shot | 5 | 51.2 | ONLINE-G | 5 | 85.6 |
| UvA-LTL | 6 | 70.9 | UvA-LTL | 6 | 51.0 | ZengHuiMT | 6 | 85.6 |
| ONLINE-Y | 7 | 70.5 | ONLINE-Y | 7 | 49.8 | ONLINE-Y | 7 | 84.9 |
| ONLINE-G | 8 | 69.8 | ONLINE-G | 8 | 49.3 | UvA-LTL | 8 | 84.7 |
| NLLB_Greedy | 9 | 64.4 | NLLB_Greedy | 9 | 42.5 | NLLB_MBR_BLEU | 9 | 82.9 |
| Lan-BridgeMT | 10 | 63.5 | Lan-BridgeMT | 10 | 41.4 | NLLB_Greedy | 10 | 82.8 |
| NLLB_MBR_BLEU | 11 | 63.0 | NLLB_MBR_BLEU | 11 | 40.7 | Samsung_Research_Philippines | 11 | 82.6 |
| Samsung_Research_Philippines | 12 | 55.5 | Samsung_Research_Philippines | 12 | 34.0 | Lan-BridgeMT | 12 | 82.4 |

Table 5: Scores for the he→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF | System | Rank | BLEU | System | Rank | COMET |
|---|---|---|---|---|---|---|---|---|
| ONLINE-B | 1 | 66.4 | ONLINE-B | 1 | 47.8 | ONLINE-B | 1 | 86.4 |
| ZengHuiMT | 2 | 62.1 | ONLINE-A | 2 | 38.9 | ONLINE-A | 2 | 85.7 |
| ONLINE-A | 3 | 61.7 | GTCOM_Peter | 3 | 37.2 | GPT4-5shot | 3 | 84.9 |
| GTCOM_Peter | 4 | 61.1 | ONLINE-Y | 4 | 37.2 | GTCOM_Peter | 4 | 84.7 |
| ONLINE-Y | 5 | 60.4 | ZengHuiMT | 5 | 36.5 | ONLINE-Y | 5 | 84.7 |
| UvA-LTL | 6 | 59.0 | UvA-LTL | 6 | 35.0 | UvA-LTL | 6 | 84.2 |
| ONLINE-G | 7 | 58.1 | Samsung_Research_Philippines | 7 | 33.3 | Samsung_Research_Philippines | 7 | 83.7 |
| Samsung_Research_Philippines | 8 | 57.3 | ONLINE-G | 8 | 33.2 | Lan-BridgeMT | 8 | 83.0 |
| Lan-BridgeMT | 9 | 54.9 | NLLB_MBR_BLEU | 9 | 30.8 | NLLB_Greedy | 9 | 82.9 |
| NLLB_Greedy | 10 | 54.8 | Lan-BridgeMT | 10 | 30.5 | ZengHuiMT | 10 | 82.7 |
| NLLB_MBR_BLEU | 11 | 54.3 | NLLB_Greedy | 11 | 30.3 | NLLB_MBR_BLEU | 11 | 82.5 |
| GPT4-5shot | 12 | 54.0 | GPT4-5shot | 12 | 27.0 | ONLINE-G | 12 | 82.2 |

Table 6: Scores for the en→he translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF |
|---|---|---|
| ONLINE-W | 1 | 51.4 |
| GPT4-5shot | 2 | 51.2 |
| SKIM | 3 | 51.1 |
| ONLINE-A | 4 | 49.6 |
| NAIST-NICT | 5 | 49.5 |
| ONLINE-Y | 6 | 49.5 |
| ZengHuiMT | 7 | 49.5 |
| ONLINE-B | 8 | 49.3 |
| GTCOM_Peter | 9 | 48.7 |
| Lan-BridgeMT | 10 | 47.3 |
| ANVITA | 11 | 46.7 |
| ONLINE-G | 12 | 45.5 |
| KYB | 13 | 43.9 |
| ONLINE-M | 14 | 43.9 |
| AIRC | 15 | 40.5 |
| NLLB_MBR_BLEU | 16 | 39.2 |
| NLLB_Greedy | 17 | 39.0 |

| System | Rank | BLEU |
|---|---|---|
| ONLINE-W | 1 | 25.9 |
| SKIM | 2 | 24.8 |
| GPT4-5shot | 3 | 24.1 |
| ONLINE-B | 4 | 23.9 |
| NAIST-NICT | 5 | 23.0 |
| ONLINE-A | 6 | 23.0 |
| ZengHuiMT | 7 | 22.6 |
| GTCOM_Peter | 8 | 22.3 |
| ONLINE-Y | 9 | 22.3 |
| ANVITA | 10 | 20.9 |
| Lan-BridgeMT | 11 | 20.2 |
| ONLINE-G | 12 | 18.3 |
| KYB | 13 | 17.6 |
| ONLINE-M | 14 | 17.2 |
| AIRC | 15 | 14.9 |
| NLLB_MBR_BLEU | 16 | 14.7 |
| NLLB_Greedy | 17 | 14.2 |

| System | Rank | COMET |
|---|---|---|
| SKIM | 1 | 84.0 |
| GPT4-5shot | 2 | 83.4 |
| ONLINE-W | 3 | 82.3 |
| NAIST-NICT | 4 | 81.9 |
| ONLINE-Y | 5 | 81.6 |
| ONLINE-B | 6 | 81.5 |
| ONLINE-A | 7 | 81.0 |
| GTCOM_Peter | 8 | 80.2 |
| ANVITA | 9 | 79.5 |
| Lan-BridgeMT | 10 | 79.3 |
| ZengHuiMT | 11 | 79.2 |
| ONLINE-G | 12 | 77.8 |
| ONLINE-M | 13 | 77.5 |
| KYB | 14 | 76.6 |
| NLLB_MBR_BLEU | 15 | 75.2 |
| AIRC | 16 | 74.5 |
| NLLB_Greedy | 17 | 74.3 |

Table 7: Scores for the ja→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF |
|---|---|---|
| ONLINE-B | 1 | 35.2 |
| ONLINE-Y | 2 | 34.1 |
| ONLINE-W | 3 | 33.5 |
| SKIM | 4 | 33.5 |
| ZengHuiMT | 5 | 32.9 |
| NAIST-NICT | 6 | 32.0 |
| ONLINE-A | 7 | 31.4 |
| GPT4-5shot | 8 | 31.0 |
| Lan-BridgeMT | 9 | 30.4 |
| ONLINE-M | 10 | 29.6 |
| ANVITA | 11 | 29.3 |
| KYB | 12 | 27.7 |
| AIRC | 13 | 27.6 |
| ONLINE-G | 14 | 27.3 |
| NLLB_Greedy | 15 | 20.9 |
| NLLB_MBR_BLEU | 16 | 18.7 |

| System | Rank | BLEU |
|---|---|---|
| ONLINE-B | 1 | 25.3 |
| ONLINE-W | 2 | 24.5 |
| ONLINE-Y | 3 | 24.5 |
| SKIM | 4 | 24.3 |
| NAIST-NICT | 5 | 22.6 |
| ZengHuiMT | 6 | 22.6 |
| ONLINE-A | 7 | 21.4 |
| GPT4-5shot | 8 | 21.3 |
| Lan-BridgeMT | 9 | 20.5 |
| ONLINE-M | 10 | 19.8 |
| ANVITA | 11 | 19.4 |
| KYB | 12 | 17.8 |
| AIRC | 13 | 17.6 |
| ONLINE-G | 14 | 17.2 |
| NLLB_Greedy | 15 | 11.3 |
| NLLB_MBR_BLEU | 16 | 9.0 |

| System | Rank | COMET |
|---|---|---|
| ONLINE-B | 1 | 88.2 |
| ONLINE-W | 2 | 87.5 |
| ONLINE-Y | 3 | 87.3 |
| GPT4-5shot | 4 | 87.0 |
| SKIM | 5 | 86.6 |
| NAIST-NICT | 6 | 86.2 |
| ZengHuiMT | 7 | 85.3 |
| ONLINE-A | 8 | 85.2 |
| Lan-BridgeMT | 9 | 84.5 |
| ONLINE-M | 10 | 83.3 |
| ANVITA | 11 | 82.7 |
| KYB | 12 | 80.8 |
| AIRC | 13 | 80.7 |
| ONLINE-G | 14 | 80.4 |
| NLLB_Greedy | 15 | 79.3 |
| NLLB_MBR_BLEU | 16 | 77.7 |

Table 8: Scores for the en→ja translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:ja-mecab-0.996-IPA|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF |
|---|---|---|
| GPT4-5shot | 1 | 60.4 |
| ONLINE-G | 2 | 59.6 |
| ONLINE-A | 3 | 59.4 |
| ONLINE-B | 4 | 59.4 |
| ZengHuiMT | 5 | 58.9 |
| ONLINE-Y | 6 | 58.6 |
| PROMT | 7 | 58.4 |
| ONLINE-W | 8 | 58.3 |
| Lan-BridgeMT | 9 | 57.4 |
| ONLINE-M | 10 | 56.7 |
| NLLB_MBR_BLEU | 11 | 55.8 |
| NLLB_Greedy | 12 | 55.5 |

| System | Rank | BLEU |
|---|---|---|
| ONLINE-B | 1 | 34.5 |
| GPT4-5shot | 2 | 34.4 |
| ONLINE-G | 3 | 34.0 |
| ONLINE-A | 4 | 33.8 |
| ONLINE-Y | 5 | 33.2 |
| ONLINE-W | 6 | 33.1 |
| PROMT | 7 | 32.8 |
| Lan-BridgeMT | 8 | 31.8 |
| ZengHuiMT | 9 | 31.3 |
| NLLB_MBR_BLEU | 10 | 31.0 |
| ONLINE-M | 11 | 30.7 |
| NLLB_Greedy | 12 | 30.3 |

| System | Rank | COMET |
|---|---|---|
| GPT4-5shot | 1 | 83.5 |
| ONLINE-Y | 2 | 82.5 |
| ONLINE-B | 3 | 82.3 |
| ONLINE-W | 4 | 82.2 |
| ONLINE-G | 5 | 82.0 |
| ONLINE-A | 6 | 81.9 |
| PROMT | 7 | 80.9 |
| ONLINE-M | 8 | 80.7 |
| NLLB_MBR_BLEU | 9 | 80.5 |
| NLLB_Greedy | 10 | 80.1 |
| Lan-BridgeMT | 11 | 79.9 |
| ZengHuiMT | 12 | 79.5 |

Table 9: Scores for the ru→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF |
|---|---|---|
| ONLINE-B | 1 | 61.9 |
| ONLINE-A | 2 | 59.0 |
| ONLINE-G | 3 | 58.9 |
| ZengHuiMT | 4 | 58.8 |
| ONLINE-W | 5 | 56.6 |
| ONLINE-Y | 6 | 56.4 |
| GPT4-5shot | 7 | 56.2 |
| Lan-BridgeMT | 8 | 55.7 |
| PROMT | 9 | 55.4 |
| ONLINE-M | 10 | 55.1 |
| NLLB_Greedy | 11 | 53.3 |
| NLLB_MBR_BLEU | 12 | 53.1 |

| System | Rank | BLEU |
|---|---|---|
| ONLINE-B | 1 | 40.4 |
| ONLINE-A | 2 | 34.8 |
| ONLINE-G | 3 | 32.9 |
| ONLINE-Y | 4 | 32.0 |
| ZengHuiMT | 5 | 31.6 |
| ONLINE-W | 6 | 31.4 |
| ONLINE-M | 7 | 30.9 |
| Lan-BridgeMT | 8 | 30.7 |
| GPT4-5shot | 9 | 30.6 |
| PROMT | 10 | 30.5 |
| NLLB_MBR_BLEU | 11 | 28.4 |
| NLLB_Greedy | 12 | 28.2 |

| System | Rank | COMET |
|---|---|---|
| ONLINE-G | 1 | 86.6 |
| ONLINE-W | 2 | 86.6 |
| ONLINE-B | 3 | 86.2 |
| GPT4-5shot | 4 | 86.1 |
| ONLINE-Y | 5 | 85.5 |
| ONLINE-A | 6 | 85.3 |
| ONLINE-M | 7 | 83.2 |
| Lan-BridgeMT | 8 | 83.1 |
| NLLB_Greedy | 9 | 82.9 |
| NLLB_MBR_BLEU | 10 | 82.7 |
| PROMT | 11 | 82.3 |
| ZengHuiMT | 12 | 81.3 |

Table 10: Scores for the en→ru translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF |
|---|---|---|
| GTCOM_Peter | 1 | 69.3 |
| ONLINE-W | 2 | 69.2 |
| ONLINE-B | 3 | 69.0 |
| ZengHuiMT | 4 | 68.5 |
| ONLINE-A | 5 | 68.3 |
| ONLINE-Y | 6 | 68.2 |
| GPT4-5shot | 7 | 68.1 |
| ONLINE-G | 8 | 68.0 |
| Lan-BridgeMT | 9 | 66.2 |
| NLLB_Greedy | 10 | 62.4 |
| NLLB_MBR_BLEU | 11 | 62.4 |

| System | Rank | BLEU |
|---|---|---|
| ONLINE-W | 1 | 47.4 |
| GTCOM_Peter | 2 | 46.4 |
| ONLINE-B | 3 | 46.0 |
| ONLINE-A | 4 | 45.9 |
| ONLINE-Y | 5 | 45.7 |
| ONLINE-G | 6 | 44.9 |
| GPT4-5shot | 7 | 43.9 |
| ZengHuiMT | 8 | 43.5 |
| Lan-BridgeMT | 9 | 42.3 |
| NLLB_MBR_BLEU | 10 | 38.1 |
| NLLB_Greedy | 11 | 37.8 |

| System | Rank | COMET |
|---|---|---|
| ONLINE-W | 1 | 87.5 |
| GPT4-5shot | 2 | 87.1 |
| ONLINE-B | 3 | 86.8 |
| GTCOM_Peter | 4 | 86.3 |
| ONLINE-A | 5 | 86.3 |
| ONLINE-G | 6 | 86.2 |
| ONLINE-Y | 7 | 85.8 |
| Lan-BridgeMT | 8 | 84.8 |
| ZengHuiMT | 9 | 84.4 |
| NLLB_MBR_BLEU | 10 | 84.3 |
| NLLB_Greedy | 11 | 84.2 |

Table 11: Scores for the uk→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF |
|---|---|---|
| ONLINE-B | 1 | 61.7 |
| ONLINE-W | 2 | 59.2 |
| ZengHuiMT | 3 | 56.4 |
| ONLINE-G | 4 | 56.1 |
| ONLINE-A | 5 | 55.8 |
| ONLINE-Y | 6 | 55.4 |
| GTCOM_Peter | 7 | 54.4 |
| GPT4-5shot | 8 | 53.0 |
| Lan-BridgeMT | 9 | 51.9 |
| NLLB_Greedy | 10 | 50.8 |
| NLLB_MBR_BLEU | 11 | 50.5 |

| System | Rank | BLEU |
|---|---|---|
| ONLINE-B | 1 | 39.8 |
| ONLINE-W | 2 | 34.9 |
| ONLINE-A | 3 | 30.3 |
| ONLINE-Y | 4 | 29.5 |
| ONLINE-G | 5 | 28.6 |
| ZengHuiMT | 6 | 27.8 |
| GTCOM_Peter | 7 | 27.5 |
| GPT4-5shot | 8 | 25.2 |
| NLLB_MBR_BLEU | 9 | 24.9 |
| Lan-BridgeMT | 10 | 24.6 |
| NLLB_Greedy | 11 | 24.5 |

| System | Rank | COMET |
|---|---|---|
| ONLINE-W | 1 | 86.7 |
| ONLINE-B | 2 | 85.6 |
| GPT4-5shot | 3 | 85.3 |
| ONLINE-G | 4 | 85.3 |
| ONLINE-A | 5 | 83.2 |
| ONLINE-Y | 6 | 82.9 |
| GTCOM_Peter | 7 | 82.1 |
| NLLB_Greedy | 8 | 82.1 |
| NLLB_MBR_BLEU | 9 | 81.7 |
| Lan-BridgeMT | 10 | 80.4 |
| ZengHuiMT | 11 | 79.0 |

Table 12: Scores for the en→uk translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF |
|---|---|---|
| HW-TSC | 1 | 57.5 |
| ONLINE-B | 2 | 57.5 |
| Yishu | 3 | 57.4 |
| ZengHuiMT | 4 | 54.6 |
| ONLINE-G | 5 | 53.9 |
| ONLINE-A | 6 | 53.4 |
| GPT4-5shot | 7 | 53.1 |
| Lan-BridgeMT | 8 | 53.1 |
| ONLINE-W | 9 | 52.5 |
| IOL_Research | 10 | 52.4 |
| ONLINE-Y | 11 | 52.3 |
| ONLINE-M | 12 | 49.7 |
| ANVITA | 13 | 47.1 |
| NLLB_Greedy | 14 | 46.1 |
| NLLB_MBR_BLEU | 15 | 45.8 |

| System | Rank | BLEU |
|---|---|---|
| HW-TSC | 1 | 33.6 |
| ONLINE-B | 2 | 33.5 |
| Yishu | 3 | 33.4 |
| ONLINE-A | 4 | 28.3 |
| Lan-BridgeMT | 5 | 27.3 |
| IOL_Research | 6 | 27.2 |
| ZengHuiMT | 7 | 27.0 |
| GPT4-5shot | 8 | 26.8 |
| ONLINE-G | 9 | 26.6 |
| ONLINE-W | 10 | 26.4 |
| ONLINE-Y | 11 | 25.0 |
| ONLINE-M | 12 | 23.5 |
| ANVITA | 13 | 21.8 |
| NLLB_Greedy | 14 | 20.5 |
| NLLB_MBR_BLEU | 15 | 19.8 |

| System | Rank | COMET |
|---|---|---|
| HW-TSC | 1 | 82.8 |
| ONLINE-B | 2 | 82.7 |
| Yishu | 3 | 82.7 |
| GPT4-5shot | 4 | 81.6 |
| Lan-BridgeMT | 5 | 81.2 |
| ONLINE-G | 6 | 80.9 |
| ONLINE-Y | 7 | 80.6 |
| ONLINE-A | 8 | 80.3 |
| ZengHuiMT | 9 | 79.6 |
| ONLINE-W | 10 | 79.3 |
| IOL_Research | 11 | 79.2 |
| ONLINE-M | 12 | 77.7 |
| NLLB_MBR_BLEU | 13 | 76.8 |
| ANVITA | 14 | 76.6 |
| NLLB_Greedy | 15 | 76.4 |

Table 13: Scores for the zh→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| System | Rank | chrF |
|---|---|---|
| HW-TSC | 1 | 53.8 |
| Yishu | 2 | 53.0 |
| ONLINE-B | 3 | 52.9 |
| ONLINE-A | 4 | 52.8 |
| IOL_Research | 5 | 51.9 |
| ONLINE-M | 6 | 50.6 |
| ONLINE-Y | 7 | 49.8 |
| ONLINE-G | 8 | 49.4 |
| ONLINE-W | 9 | 47.3 |
| ZengHuiMT | 10 | 47.0 |
| Lan-BridgeMT | 11 | 46.8 |
| GPT4-5shot | 12 | 46.5 |
| ANVITA | 13 | 36.9 |
| NLLB_Greedy | 14 | 26.3 |
| NLLB_MBR_BLEU | 15 | 21.1 |

| System | Rank | BLEU |
|---|---|---|
| HW-TSC | 1 | 58.6 |
| ONLINE-A | 2 | 58.5 |
| Yishu | 3 | 57.6 |
| ONLINE-B | 4 | 57.5 |
| IOL_Research | 5 | 56.9 |
| ONLINE-M | 6 | 54.9 |
| ONLINE-Y | 7 | 54.2 |
| ONLINE-G | 8 | 54.1 |
| ZengHuiMT | 9 | 52.9 |
| ONLINE-W | 10 | 52.1 |
| Lan-BridgeMT | 11 | 50.2 |
| GPT4-5shot | 12 | 49.6 |
| ANVITA | 13 | 38.9 |
| NLLB_Greedy | 14 | 27.4 |
| NLLB_MBR_BLEU | 15 | 19.1 |

| System | Rank | COMET |
|---|---|---|
| ONLINE-B | 1 | 88.1 |
| Yishu | 2 | 88.1 |
| HW-TSC | 3 | 87.3 |
| GPT4-5shot | 4 | 87.1 |
| ONLINE-W | 5 | 86.8 |
| Lan-BridgeMT | 6 | 86.6 |
| ONLINE-Y | 7 | 86.5 |
| ONLINE-A | 8 | 86.2 |
| IOL_Research | 9 | 85.3 |
| ZengHuiMT | 10 | 84.3 |
| ONLINE-M | 11 | 84.2 |
| ONLINE-G | 12 | 83.8 |
| NLLB_Greedy | 13 | 75.7 |
| ANVITA | 14 | 75.6 |
| NLLB_MBR_BLEU | 15 | 71.5 |

Table 14: Scores for the en→zh translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

| | CUNI-GA. | GPT4-5sh. | ONLINE-W. | GTCOM_Pe. | ONLINE-B. | ONLINE-A. | CUNI-Tra. | ONLINE-G. | MUNI-NLP. | ONLINE-Y. | NLLB_Gre. | NLLB_MBR. | Lan-Brid. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CUNI-GA | 0.0 | 0.1 | 1.5 | 2.0 | 2.1 | 2.7 | 2.9 | 3.2 | 3.9 | 4.4 | 4.6 | 4.6 | 4.9 |
| GPT4-5shot | | 0.0 | 1.4 | 1.9 | 2.0 | 2.6 | 2.8 | 3.1 | 3.8 | 4.3 | 4.5 | 4.5 | 4.8 |
| ONLINE-W | | | 0.0 | 0.5 | 0.6 | 1.2 | 1.4 | 1.7 | 2.4 | 2.9 | 3.1 | 3.1 | 3.4 |
| GTCOM_Peter | | | | 0.0 | 0.1 | 0.7 | 0.9 | 1.2 | 1.9 | 2.4 | 2.6 | 2.6 | 2.9 |
| ONLINE-B | | | | | 0.0 | 0.6 | 0.8 | 1.1 | 1.8 | 2.3 | 2.5 | 2.5 | 2.8 |
| ONLINE-A | | | | | | 0.0 | 0.2 | 0.5 | 1.2 | 1.7 | 1.9 | 1.9 | 2.2 |
| CUNI-Transformer | | | | | | | 0.0 | 0.3 | 1.0 | 1.5 | 1.7 | 1.7 | 2.0 |
| ONLINE-G | | | | | | | | 0.0 | 0.7 | 1.2 | 1.4 | 1.4 | 1.7 |
| MUNI-NLP | | | | | | | | | 0.0 | 0.5 | 0.7 | 0.7 | 1.0 |
| ONLINE-Y | | | | | | | | | | 0.0 | 0.2 | 0.2 | 0.5 |
| NLLB_Greedy | | | | | | | | | | | 0.0 | 0.0 | 0.3 |
| NLLB_MBR_BLEU | | | | | | | | | | | | 0.0 | 0.3 |

Table 15: Statistical significance testing of the COMET score difference for each system pair for the cs→uk.

| | ONLINE-W. | CUNI-GA. | ONLINE-B. | GPT4-5sh. | ONLINE-A. | CUNI-Doc. | GTCOM_Pe. | ONLINE-M. | Lan-Brid. | CUNI-Tra. | NLLB_Gre. | ONLINE-Y. | NLLB_MBR. | ONLINE-G. | ZengHuiM. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ONLINE-W | 0.0 | 1.0 | 1.9 | 2.4 | 3.4 | 3.5 | 4.1 | 4.4 | 4.5 | 4.6 | 4.7 | 4.8 | 4.9 | 5.9 | 6.4 |
| CUNI-GA | | 0.0 | 0.9 | 1.4 | 2.4 | 2.5 | 3.1 | 3.4 | 3.5 | 3.6 | 3.7 | 3.8 | 3.9 | 4.9 | 5.4 |
| ONLINE-B | | | 0.0 | 0.5 | 1.5 | 1.6 | 2.2 | 2.5 | 2.6 | 2.7 | 2.8 | 2.9 | 3.0 | 4.0 | 4.5 |
| GPT4-5shot | | | | 0.0 | 1.0 | 1.1 | 1.7 | 2.0 | 2.1 | 2.2 | 2.3 | 2.4 | 2.5 | 3.5 | 4.0 |
| ONLINE-A | | | | | 0.0 | 0.1 | 0.7 | 1.0 | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 2.5 | 3.0 |
| CUNI-DocTransformer | | | | | | 0.0 | 0.6 | 0.9 | 1.0 | 1.1 | 1.2 | 1.3 | 1.4 | 2.4 | 2.9 |
| GTCOM_Peter | | | | | | | 0.0 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 1.8 | 2.3 |
| ONLINE-M | | | | | | | | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 1.5 | 2.0 |
| Lan-BridgeMT | | | | | | | | | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 1.4 | 1.9 |
| CUNI-Transformer | | | | | | | | | | 0.0 | 0.1 | 0.2 | 0.3 | 1.3 | 1.8 |
| NLLB_Greedy | | | | | | | | | | | 0.0 | 0.1 | 0.2 | 1.2 | 1.7 |
| ONLINE-Y | | | | | | | | | | | | 0.0 | 0.1 | 1.1 | 1.6 |
| NLLB_MBR_BLEU | | | | | | | | | | | | | 0.0 | 1.0 | 1.5 |
| ONLINE-G | | | | | | | | | | | | | | 0.0 | 0.5 |

Table 16: Statistical significance testing of the COMET score difference for each system pair for the en→cs.

| | GPT4-5sh. | ONLINE-W. | ONLINE-B. | ONLINE-A. | ONLINE-Y. | ONLINE-M. | ONLINE-G. | GTCOM_Pe. | NLLB_MBR. | ZengHuiM. | Lan-Brid. | NLLB_Gre. | AIRC. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT4-5shot | 0.0 | 0.3 | 0.7 | 0.8 | 1.4 | 1.5 | 1.7 | 3.6 | 4.9 | 5.2 | 5.4 | 6.4 | 7.6 |
| ONLINE-W | | 0.0 | 0.4 | 0.5 | 1.1 | 1.2 | 1.4 | 3.3 | 4.6 | 4.9 | 5.1 | 6.1 | 7.3 |
| ONLINE-B | | | 0.0 | 0.1 | 0.7 | 0.8 | 1.0 | 2.9 | 4.2 | 4.5 | 4.7 | 5.7 | 6.9 |
| ONLINE-A | | | | 0.0 | 0.6 | 0.7 | 0.9 | 2.8 | 4.1 | 4.4 | 4.6 | 5.6 | 6.8 |
| ONLINE-Y | | | | | 0.0 | 0.1 | 0.3 | 2.2 | 3.5 | 3.8 | 4.0 | 5.0 | 6.2 |
| ONLINE-M | | | | | | 0.0 | 0.2 | 2.1 | 3.4 | 3.7 | 3.9 | 4.9 | 6.1 |
| ONLINE-G | | | | | | | 0.0 | 1.9 | 3.2 | 3.5 | 3.7 | 4.7 | 5.9 |
| GTCOM_Peter | | | | | | | | 0.0 | 1.3 | 1.6 | 1.8 | 2.8 | 4.0 |
| NLLB_MBR_BLEU | | | | | | | | | 0.0 | 0.3 | 0.5 | 1.5 | 2.7 |
| ZengHuiMT | | | | | | | | | | 0.0 | 0.2 | 1.2 | 2.4 |
| Lan-BridgeMT | | | | | | | | | | | 0.0 | 1.0 | 2.2 |
| NLLB_Greedy | | | | | | | | | | | | 0.0 | 1.2 |

Table 17: Statistical significance testing of the COMET score difference for each system pair for the de→en.

| | ONLINE-W. | GPT4-5sh. | ONLINE-B. | ONLINE-Y. | ONLINE-A. | ONLINE-G. | ONLINE-M. | Lan-Brid. | ZengHuiM. | NLLB_MBR. | NLLB_Gre. | AIRC. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ONLINE-W | 0.0 | 0.5 | 0.7 | 1.4 | 1.8 | 3.0 | 3.8 | 5.1 | 6.1 | 7.5 | 7.6 | 12.6 |
| GPT4-5shot | | 0.0 | 0.2 | 0.9 | 1.3 | 2.5 | 3.3 | 4.6 | 5.6 | 7.0 | 7.1 | 12.1 |
| ONLINE-B | | | 0.0 | 0.7 | 1.1 | 2.3 | 3.1 | 4.4 | 5.4 | 6.8 | 6.9 | 11.9 |
| ONLINE-Y | | | | 0.0 | 0.4 | 1.6 | 2.4 | 3.7 | 4.7 | 6.1 | 6.2 | 11.2 |
| ONLINE-A | | | | | 0.0 | 1.2 | 2.0 | 3.3 | 4.3 | 5.7 | 5.8 | 10.8 |
| ONLINE-G | | | | | | 0.0 | 0.8 | 2.1 | 3.1 | 4.5 | 4.6 | 9.6 |
| ONLINE-M | | | | | | | 0.0 | 1.3 | 2.3 | 3.7 | 3.8 | 8.8 |
| Lan-BridgeMT | | | | | | | | 0.0 | 1.0 | 2.4 | 2.5 | 7.5 |
| ZengHuiMT | | | | | | | | | 0.0 | 1.4 | 1.5 | 6.5 |
| NLLB_MBR_BLEU | | | | | | | | | | 0.0 | 0.1 | 5.1 |
| NLLB_Greedy | | | | | | | | | | | 0.0 | 5.0 |

Table 18: Statistical significance testing of the COMET score difference for each system pair for the en→de.

| | ONLINE-B. | ONLINE-A. | GPT4-5sh. | GTCOM_Pe. | ONLINE-G. | ZengHuiM. | ONLINE-Y. | UvA-LTL. | NLLB_MBR. | NLLB_Gre. | Samsung_. | Lan-Brid. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ONLINE-B | 0.0 | 2.9 | 3.0 | 3.2 | 4.3 | 4.3 | 5.0 | 5.2 | 7.0 | 7.1 | 7.3 | 7.5 |
| ONLINE-A | | 0.0 | 0.1 | 0.3 | 1.4 | 1.4 | 2.1 | 2.3 | 4.1 | 4.2 | 4.4 | 4.6 |
| GPT4-5shot | | | 0.0 | 0.2 | 1.3 | 1.3 | 2.0 | 2.2 | 4.0 | 4.1 | 4.3 | 4.5 |
| GTCOM_Peter | | | | 0.0 | 1.1 | 1.1 | 1.8 | 2.0 | 3.8 | 3.9 | 4.1 | 4.3 |
| ONLINE-G | | | | | 0.0 | 0.0 | 0.7 | 0.9 | 2.7 | 2.8 | 3.0 | 3.2 |
| ZengHuiMT | | | | | | 0.0 | 0.7 | 0.9 | 2.7 | 2.8 | 3.0 | 3.2 |
| ONLINE-Y | | | | | | | 0.0 | 0.2 | 2.0 | 2.1 | 2.3 | 2.5 |
| UvA-LTL | | | | | | | | 0.0 | 1.8 | 1.9 | 2.1 | 2.3 |
| NLLB_MBR_BLEU | | | | | | | | | 0.0 | 0.1 | 0.3 | 0.5 |
| NLLB_Greedy | | | | | | | | | | 0.0 | 0.2 | 0.4 |
| Samsung_Research_Philippines | | | | | | | | | | | 0.0 | 0.2 |

Table 19: Statistical significance testing of the COMET score difference for each system pair for the he→en.

| | ONLINE-B. | ONLINE-A. | GPT4-5sh. | GTCOM_Pe. | ONLINE-Y. | UvA-LTL. | Samsung_. | Lan-Brid. | NLLB_Gre. | ZengHuiM. | NLLB_MBR. | ONLINE-G. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ONLINE-B | 0.0 | 0.7 | 1.5 | 1.7 | 1.7 | 2.2 | 2.7 | 3.4 | 3.5 | 3.7 | 3.9 | 4.2 |
| ONLINE-A | | 0.0 | 0.8 | 1.0 | 1.0 | 1.5 | 2.0 | 2.7 | 2.8 | 3.0 | 3.2 | 3.5 |
| GPT4-5shot | | | 0.0 | 0.2 | 0.2 | 0.7 | 1.2 | 1.9 | 2.0 | 2.2 | 2.4 | 2.7 |
| GTCOM_Peter | | | | 0.0 | 0.0 | 0.5 | 1.0 | 1.7 | 1.8 | 2.0 | 2.2 | 2.5 |
| ONLINE-Y | | | | | 0.0 | 0.5 | 1.0 | 1.7 | 1.8 | 2.0 | 2.2 | 2.5 |
| UvA-LTL | | | | | | 0.0 | 0.5 | 1.2 | 1.3 | 1.5 | 1.7 | 2.0 |
| Samsung_Research_Philippines | | | | | | | 0.0 | 0.7 | 0.8 | 1.0 | 1.2 | 1.5 |
| Lan-BridgeMT | | | | | | | | 0.0 | 0.1 | 0.3 | 0.5 | 0.8 |
| NLLB_Greedy | | | | | | | | | 0.0 | 0.2 | 0.4 | 0.7 |
| ZengHuiMT | | | | | | | | | | 0.0 | 0.2 | 0.5 |
| NLLB_MBR_BLEU | | | | | | | | | | | 0.0 | 0.3 |

Table 20: Statistical significance testing of the COMET score difference for each system pair for the en→he.

| | SKIM. | GPT4-5sh. | ONLINE-W. | NAIST-NI. | ONLINE-Y. | ONLINE-B. | ONLINE-A. | GTCOM_Pe. | ANVITA. | Lan-Brid. | ZengHuiM. | ONLINE-G. | ONLINE-M. | KYB. | NLLB_MBR. | AIRC. | NLLB_Gre. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SKIM | 0.0 | 0.6 | 1.7 | 2.1 | 2.4 | 2.5 | 3.0 | 3.8 | 4.5 | 4.7 | 4.8 | 6.2 | 6.5 | 7.4 | 8.8 | 9.5 | 9.7 |
| GPT4-5shot | | 0.0 | 1.1 | 1.5 | 1.8 | 1.9 | 2.4 | 3.2 | 3.9 | 4.1 | 4.2 | 5.6 | 5.9 | 6.8 | 8.2 | 8.9 | 9.1 |
| ONLINE-W | | | 0.0 | 0.4 | 0.7 | 0.8 | 1.3 | 2.1 | 2.8 | 3.0 | 3.1 | 4.5 | 4.8 | 5.7 | 7.1 | 7.8 | 8.0 |
| NAIST-NICT | | | | 0.0 | 0.3 | 0.4 | 0.9 | 1.7 | 2.4 | 2.6 | 2.7 | 4.1 | 4.4 | 5.3 | 6.7 | 7.4 | 7.6 |
| ONLINE-Y | | | | | 0.0 | 0.1 | 0.6 | 1.4 | 2.1 | 2.3 | 2.4 | 3.8 | 4.1 | 5.0 | 6.4 | 7.1 | 7.3 |
| ONLINE-B | | | | | | 0.0 | 0.5 | 1.3 | 2.0 | 2.2 | 2.3 | 3.7 | 4.0 | 4.9 | 6.3 | 7.0 | 7.2 |
| ONLINE-A | | | | | | | 0.0 | 0.8 | 1.5 | 1.7 | 1.8 | 3.2 | 3.5 | 4.4 | 5.8 | 6.5 | 6.7 |
| GTCOM_Peter | | | | | | | | 0.0 | 0.7 | 0.9 | 1.0 | 2.4 | 2.7 | 3.6 | 5.0 | 5.7 | 5.9 |
| ANVITA | | | | | | | | | 0.0 | 0.2 | 0.3 | 1.7 | 2.0 | 2.9 | 4.3 | 5.0 | 5.2 |
| Lan-BridgeMT | | | | | | | | | | 0.0 | 0.1 | 1.5 | 1.8 | 2.7 | 4.1 | 4.8 | 5.0 |
| ZengHuiMT | | | | | | | | | | | 0.0 | 1.4 | 1.7 | 2.6 | 4.0 | 4.7 | 4.9 |
| ONLINE-G | | | | | | | | | | | | 0.0 | 0.3 | 1.2 | 2.6 | 3.3 | 3.5 |
| ONLINE-M | | | | | | | | | | | | | 0.0 | 0.9 | 2.3 | 3.0 | 3.2 |
| KYB | | | | | | | | | | | | | | 0.0 | 1.4 | 2.1 | 2.3 |
| NLLB_MBR_BLEU | | | | | | | | | | | | | | | 0.0 | 0.7 | 0.9 |
| AIRC | | | | | | | | | | | | | | | | 0.0 | 0.2 |

Table 21: Statistical significance testing of the COMET score difference for each system pair for the ja→en.

| | ONLINE-B. | ONLINE-W. | ONLINE-Y. | GPT4-5sh. | SKIM. | NAIST-NI. | ZengHuiM. | ONLINE-A. | Lan-Brid. | ONLINE-M. | ANVITA. | KYB. | AIRC. | ONLINE-G. | NLLB_Gre. | NLLB_MBR. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ONLINE-B | 0.0 | 0.7 | 0.9 | 1.2 | 1.6 | 2.0 | 2.9 | 3.0 | 3.7 | 4.9 | 5.5 | 7.4 | 7.5 | 7.8 | 8.9 | 10.5 |
| ONLINE-W | | 0.0 | 0.2 | 0.5 | 0.9 | 1.3 | 2.2 | 2.3 | 3.0 | 4.2 | 4.8 | 6.7 | 6.8 | 7.1 | 8.2 | 9.8 |
| ONLINE-Y | | | 0.0 | 0.3 | 0.7 | 1.1 | 2.0 | 2.1 | 2.8 | 4.0 | 4.6 | 6.5 | 6.6 | 6.9 | 8.0 | 9.6 |
| GPT4-5shot | | | | 0.0 | 0.4 | 0.8 | 1.7 | 1.8 | 2.5 | 3.7 | 4.3 | 6.2 | 6.3 | 6.6 | 7.7 | 9.3 |
| SKIM | | | | | 0.0 | 0.4 | 1.3 | 1.4 | 2.1 | 3.3 | 3.9 | 5.8 | 5.9 | 6.2 | 7.3 | 8.9 |
| NAIST-NICT | | | | | | 0.0 | 0.9 | 1.0 | 1.7 | 2.9 | 3.5 | 5.4 | 5.5 | 5.8 | 6.9 | 8.5 |
| ZengHuiMT | | | | | | | 0.0 | 0.1 | 0.8 | 2.0 | 2.6 | 4.5 | 4.6 | 4.9 | 6.0 | 7.6 |
| ONLINE-A | | | | | | | | 0.0 | 0.7 | 1.9 | 2.5 | 4.4 | 4.5 | 4.8 | 5.9 | 7.5 |
| Lan-BridgeMT | | | | | | | | | 0.0 | 1.2 | 1.8 | 3.7 | 3.8 | 4.1 | 5.2 | 6.8 |
| ONLINE-M | | | | | | | | | | 0.0 | 0.6 | 2.5 | 2.6 | 2.9 | 4.0 | 5.6 |
| ANVITA | | | | | | | | | | | 0.0 | 1.9 | 2.0 | 2.3 | 3.4 | 5.0 |
| KYB | | | | | | | | | | | | 0.0 | 0.1 | 0.4 | 1.5 | 3.1 |
| AIRC | | | | | | | | | | | | | 0.0 | 0.3 | 1.4 | 3.0 |
| ONLINE-G | | | | | | | | | | | | | | 0.0 | 1.1 | 2.7 |
| NLLB_Greedy | | | | | | | | | | | | | | | 0.0 | 1.6 |

Table 22: Statistical significance testing of the COMET score difference for each system pair for the en→ja.

| | GPT4-5sh. | ONLINE-Y. | ONLINE-B. | ONLINE-W. | ONLINE-G. | ONLINE-A. | PROMT. | ONLINE-M. | NLLB_MBR. | NLLB_Gre. | Lan-Brid. | ZengHuiM. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT4-5shot | 0.0 | 1.0 | 1.2 | 1.3 | 1.5 | 1.6 | 2.6 | 2.8 | 3.0 | 3.4 | 3.6 | 4.0 |
| ONLINE-Y | | 0.0 | 0.2 | 0.3 | 0.5 | 0.6 | 1.6 | 1.8 | 2.0 | 2.4 | 2.6 | 3.0 |
| ONLINE-B | | | 0.0 | 0.1 | 0.3 | 0.4 | 1.4 | 1.6 | 1.8 | 2.2 | 2.4 | 2.8 |
| ONLINE-W | | | | 0.0 | 0.2 | 0.3 | 1.3 | 1.5 | 1.7 | 2.1 | 2.3 | 2.7 |
| ONLINE-G | | | | | 0.0 | 0.1 | 1.1 | 1.3 | 1.5 | 1.9 | 2.1 | 2.5 |
| ONLINE-A | | | | | | 0.0 | 1.0 | 1.2 | 1.4 | 1.8 | 2.0 | 2.4 |
| PROMT | | | | | | | 0.0 | 0.2 | 0.4 | 0.8 | 1.0 | 1.4 |
| ONLINE-M | | | | | | | | 0.0 | 0.2 | 0.6 | 0.8 | 1.2 |
| NLLB_MBR_BLEU | | | | | | | | | 0.0 | 0.4 | 0.6 | 1.0 |
| NLLB_Greedy | | | | | | | | | | 0.0 | 0.2 | 0.6 |
| Lan-BridgeMT | | | | | | | | | | | 0.0 | 0.4 |

Table 23: Statistical significance testing of the COMET score difference for each system pair for the ru→en.

| | ONLINE-G. | ONLINE-W. | ONLINE-B. | GPT4-5sh. | ONLINE-Y. | ONLINE-A. | ONLINE-M. | Lan-Brid. | NLLB_Gre. | NLLB_MBR. | PROMT. | ZengHuiM. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ONLINE-G | 0.0 | 0.0 | 0.4 | 0.5 | 1.1 | 1.3 | 3.4 | 3.5 | 3.7 | 3.9 | 4.3 | 5.3 |
| ONLINE-W | | 0.0 | 0.4 | 0.5 | 1.1 | 1.3 | 3.4 | 3.5 | 3.7 | 3.9 | 4.3 | 5.3 |
| ONLINE-B | | | 0.0 | 0.1 | 0.7 | 0.9 | 3.0 | 3.1 | 3.3 | 3.5 | 3.9 | 4.9 |
| GPT4-5shot | | | | 0.0 | 0.6 | 0.8 | 2.9 | 3.0 | 3.2 | 3.4 | 3.8 | 4.8 |
| ONLINE-Y | | | | | 0.0 | 0.2 | 2.3 | 2.4 | 2.6 | 2.8 | 3.2 | 4.2 |
| ONLINE-A | | | | | | 0.0 | 2.1 | 2.2 | 2.4 | 2.6 | 3.0 | 4.0 |
| ONLINE-M | | | | | | | 0.0 | 0.1 | 0.3 | 0.5 | 0.9 | 1.9 |
| Lan-BridgeMT | | | | | | | | 0.0 | 0.2 | 0.4 | 0.8 | 1.8 |
| NLLB_Greedy | | | | | | | | | 0.0 | 0.2 | 0.6 | 1.6 |
| NLLB_MBR_BLEU | | | | | | | | | | 0.0 | 0.4 | 1.4 |
| PROMT | | | | | | | | | | | 0.0 | 1.0 |

Table 24: Statistical significance testing of the COMET score difference for each system pair for the en→ru.

| | ONLINE-W. | GPT4-5sh. | ONLINE-B. | GTCOM_Pe. | ONLINE-A. | ONLINE-G. | ONLINE-Y. | Lan-Brid. | ZengHuiM. | NLLB_MBR. | NLLB_Gre. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ONLINE-W | 0.0 | 0.4 | 0.7 | 1.2 | 1.2 | 1.3 | 1.7 | 2.7 | 3.1 | 3.2 | 3.3 |
| GPT4-5shot | | 0.0 | 0.3 | 0.8 | 0.8 | 0.9 | 1.3 | 2.3 | 2.7 | 2.8 | 2.9 |
| ONLINE-B | | | 0.0 | 0.5 | 0.5 | 0.6 | 1.0 | 2.0 | 2.4 | 2.5 | 2.6 |
| GTCOM_Peter | | | | 0.0 | 0.0 | 0.1 | 0.5 | 1.5 | 1.9 | 2.0 | 2.1 |
| ONLINE-A | | | | | 0.0 | 0.1 | 0.5 | 1.5 | 1.9 | 2.0 | 2.1 |
| ONLINE-G | | | | | | 0.0 | 0.4 | 1.4 | 1.8 | 1.9 | 2.0 |
| ONLINE-Y | | | | | | | 0.0 | 1.0 | 1.4 | 1.5 | 1.6 |
| Lan-BridgeMT | | | | | | | | 0.0 | 0.4 | 0.5 | 0.6 |
| ZengHuiMT | | | | | | | | | 0.0 | 0.1 | 0.2 |
| NLLB_MBR_BLEU | | | | | | | | | | 0.0 | 0.1 |

Table 25: Statistical significance testing of the COMET score difference for each system pair for the uk→en.

| | ONLINE-W. | ONLINE-B. | GPT4-5sh. | ONLINE-G. | ONLINE-A. | ONLINE-Y. | GTCOM_Pe. | NLLB_Gre. | NLLB_MBR. | Lan-Brid. | ZengHuiM. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ONLINE-W | 0.0 | 1.1 | 1.4 | 1.4 | 3.5 | 3.8 | 4.6 | 4.6 | 5.0 | 6.3 | 7.7 |
| ONLINE-B | | 0.0 | 0.3 | 0.3 | 2.4 | 2.7 | 3.5 | 3.5 | 3.9 | 5.2 | 6.6 |
| GPT4-5shot | | | 0.0 | 0.0 | 2.1 | 2.4 | 3.2 | 3.2 | 3.6 | 4.9 | 6.3 |
| ONLINE-G | | | | 0.0 | 2.1 | 2.4 | 3.2 | 3.2 | 3.6 | 4.9 | 6.3 |
| ONLINE-A | | | | | 0.0 | 0.3 | 1.1 | 1.1 | 1.5 | 2.8 | 4.2 |
| ONLINE-Y | | | | | | 0.0 | 0.8 | 0.8 | 1.2 | 2.5 | 3.9 |
| GTCOM_Peter | | | | | | | 0.0 | 0.0 | 0.4 | 1.7 | 3.1 |
| NLLB_Greedy | | | | | | | | 0.0 | 0.4 | 1.7 | 3.1 |
| NLLB_MBR_BLEU | | | | | | | | | 0.0 | 1.3 | 2.7 |
| Lan-BridgeMT | | | | | | | | | | 0.0 | 1.4 |

Table 26: Statistical significance testing of the COMET score difference for each system pair for the en→uk.

| | HW-TSC. | ONLINE-B. | Yishu. | GPT4-5sh. | Lan-Brid. | ONLINE-G. | ONLINE-Y. | ONLINE-A. | ZengHuiM. | ONLINE-W. | IOL_Rese. | ONLINE-M. | NLLB_MBR. | ANVITA. | NLLB_Gre. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| HW-TSC | 0.0 | 0.1 | 0.1 | 1.2 | 1.6 | 1.9 | 2.2 | 2.5 | 3.2 | 3.5 | 3.6 | 5.1 | 6.0 | 6.2 | 6.4 |
| ONLINE-B | | 0.0 | 0.0 | 1.1 | 1.5 | 1.8 | 2.1 | 2.4 | 3.1 | 3.4 | 3.5 | 5.0 | 5.9 | 6.1 | 6.3 |
| Yishu | | | 0.0 | 1.1 | 1.5 | 1.8 | 2.1 | 2.4 | 3.1 | 3.4 | 3.5 | 5.0 | 5.9 | 6.1 | 6.3 |
| GPT4-5shot | | | | 0.0 | 0.4 | 0.7 | 1.0 | 1.3 | 2.0 | 2.3 | 2.4 | 3.9 | 4.8 | 5.0 | 5.2 |
| Lan-BridgeMT | | | | | 0.0 | 0.3 | 0.6 | 0.9 | 1.6 | 1.9 | 2.0 | 3.5 | 4.4 | 4.6 | 4.8 |
| ONLINE-G | | | | | | 0.0 | 0.3 | 0.6 | 1.3 | 1.6 | 1.7 | 3.2 | 4.1 | 4.3 | 4.5 |
| ONLINE-Y | | | | | | | 0.0 | 0.3 | 1.0 | 1.3 | 1.4 | 2.9 | 3.8 | 4.0 | 4.2 |
| ONLINE-A | | | | | | | | 0.0 | 0.7 | 1.0 | 1.1 | 2.6 | 3.5 | 3.7 | 3.9 |
| ZengHuiMT | | | | | | | | | 0.0 | 0.3 | 0.4 | 1.9 | 2.8 | 3.0 | 3.2 |
| ONLINE-W | | | | | | | | | | 0.0 | 0.1 | 1.6 | 2.5 | 2.7 | 2.9 |
| IOL_Research | | | | | | | | | | | 0.0 | 1.5 | 2.4 | 2.6 | 2.8 |
| ONLINE-M | | | | | | | | | | | | 0.0 | 0.9 | 1.1 | 1.3 |
| NLLB_MBR_BLEU | | | | | | | | | | | | | 0.0 | 0.2 | 0.4 |
| ANVITA | | | | | | | | | | | | | | 0.0 | 0.2 |

Table 27: Statistical significance testing of the COMET score difference for each system pair for the zh→en.

| | ONLINE-B. | Yishu. | HW-TSC. | GPT4-5sh. | ONLINE-W. | Lan-Brid. | ONLINE-Y. | ONLINE-A. | IOL_Rese. | ZengHuiM. | ONLINE-M. | ONLINE-G. | NLLB_Gre. | ANVITA. | NLLB_MBR. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ONLINE-B | 0.0 | 0.0 | 0.8 | 1.0 | 1.3 | 1.5 | 1.6 | 1.9 | 2.8 | 3.8 | 3.9 | 4.3 | 12.4 | 12.5 | 16.6 |
| Yishu | | 0.0 | 0.8 | 1.0 | 1.3 | 1.5 | 1.6 | 1.9 | 2.8 | 3.8 | 3.9 | 4.3 | 12.4 | 12.5 | 16.6 |
| HW-TSC | | | 0.0 | 0.2 | 0.5 | 0.7 | 0.8 | 1.1 | 2.0 | 3.0 | 3.1 | 3.5 | 11.6 | 11.7 | 15.8 |
| GPT4-5shot | | | | 0.0 | 0.3 | 0.5 | 0.6 | 0.9 | 1.8 | 2.8 | 2.9 | 3.3 | 11.4 | 11.5 | 15.6 |
| ONLINE-W | | | | | 0.0 | 0.2 | 0.3 | 0.6 | 1.5 | 2.5 | 2.6 | 3.0 | 11.1 | 11.2 | 15.3 |
| Lan-BridgeMT | | | | | | 0.0 | 0.1 | 0.4 | 1.3 | 2.3 | 2.4 | 2.8 | 10.9 | 11.0 | 15.1 |
| ONLINE-Y | | | | | | | 0.0 | 0.3 | 1.2 | 2.2 | 2.3 | 2.7 | 10.8 | 10.9 | 15.0 |
| ONLINE-A | | | | | | | | 0.0 | 0.9 | 1.9 | 2.0 | 2.4 | 10.5 | 10.6 | 14.7 |
| IOL_Research | | | | | | | | | 0.0 | 1.0 | 1.1 | 1.5 | 9.6 | 9.7 | 13.8 |
| ZengHuiMT | | | | | | | | | | 0.0 | 0.1 | 0.5 | 8.6 | 8.7 | 12.8 |
| ONLINE-M | | | | | | | | | | | 0.0 | 0.4 | 8.5 | 8.6 | 12.7 |
| ONLINE-G | | | | | | | | | | | | 0.0 | 8.1 | 8.2 | 12.3 |
| NLLB_Greedy | | | | | | | | | | | | | 0.0 | 0.1 | 4.2 |
| ANVITA | | | | | | | | | | | | | | 0.0 | 4.1 |

Table 28: Statistical significance testing of the COMET score difference for each system pair for the en→zh.

## References


Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

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

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

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, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.