Automatic Evaluation of the WMT23 General Machine Translation Task

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 ranked the systems according to their scores. Unconstrained systems are indicated with a grey background in the tables.

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	COMET
CUNI-GA	90.9
GPT4-5shot	90.8
ONLINE-W	89.4
GTCOM_Peter	88.9
ONLINE-B	88.8
ONLINE-A	88.2
CUNI-Transformer	88.0
ONLINE-G	87.7
MUNI-NLP	87.0
ONLINE-Y	86.5
NLLB_Greedy	86.3
NLLB_MBR_BLEU	86.3
Lan-BridgeMT	86.0

System	chrF
GPT4-5shot	61.0
CUNI-GA	57.9
GTCOM_Peter	57.6
CUNI-Transformer	57.4
MUNI-NLP	57.0
Lan-BridgeMT	55.7
ONLINE-W	55.0
ONLINE-B	54.7
ONLINE-A	54.4
ONLINE-G	53.7
ONLINE-Y	53.4
NLLB_Greedy	52.5
NLLB_MBR_BLEU	52.3

System	BLEU
GPT4-5shot	32.8
CUNI-Transformer	30.2
GTCOM_Peter	29.8
CUNI-GA	29.5
MUNI-NLP	28.3
Lan-BridgeMT	27.5
ONLINE-W	26.8
ONLINE-B	25.7
ONLINE-A	25.4
NLLB_MBR_BLEU	25.1
NLLB_Greedy	24.9
ONLINE-G	24.8
ONLINE-Y	24.2

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	COMET
ONLINE-W	91.8
CUNI-GA	90.8
ONLINE-B	89.9
GPT4-5shot	89.4
ONLINE-A	88.4
CUNI-DocTransformer	88.3
GTCOM_Peter	87.7
ONLINE-M	87.4
Lan-BridgeMT	87.3
CUNI-Transformer	87.2
NLLB_Greedy	87.1
ONLINE-Y	87.0
NLLB_MBR_BLEU	86.9
ONLINE-G	85.9
ZengHuiMT	85.4

System	chrF
ONLINE-W	76.3
ONLINE-B	70.4
ZengHuiMT	67.5
ONLINE-A	66.3
CUNI-GA	65.9
GTCOM_Peter	65.4
CUNI-DocTransformer	65.1
ONLINE-Y	64.6
CUNI-Transformer	63.9
Lan-BridgeMT	63.8
ONLINE-G	63.7
ONLINE-M	63.2
GPT4-5shot	62.3
NLLB_Greedy	60.0
NLLB_MBR_BLEU	59.1

System	BLEU
ONLINE-W	59.4
ONLINE-B	50.1
ONLINE-A	43.4
CUNI-GA	43.3
ZengHuiMT	43.1
CUNI-DocTransformer	42.5
GTCOM_Peter	42.3
CUNI-Transformer	41.4
ONLINE-Y	40.8
Lan-BridgeMT	40.7
ONLINE-G	39.6
ONLINE-M	39.6
GPT4-5shot	37.8
NLLB_Greedy	35.9
NLLB_MBR_BLEU	35.1

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	COMET
GPT4-5shot	86.3
ONLINE-W	86.0
ONLINE-B	85.6
ONLINE-A	85.5
ONLINE-Y	84.9
ONLINE-M	84.8
ONLINE-G	84.6
GTCOM_Peter	82.7
NLLB_MBR_BLEU	81.4
ZengHuiMT	81.1
Lan-BridgeMT	80.9
NLLB_Greedy	79.9
AIRC	78.7

System	chrF
ONLINE-W	72.1
ONLINE-A	70.0
GPT4-5shot	69.8
ONLINE-B	69.1
ONLINE-G	69.1
ONLINE-Y	68.4
ZengHuiMT	67.6
Lan-BridgeMT	66.7
GTCOM_Peter	66.6
ONLINE-M	66.5
NLLB_MBR_BLEU	57.6
NLLB_Greedy	57.3
AIRC	57.2

System	BLEU
ONLINE-W	51.8
GPT4-5shot	47.9
ONLINE-A	47.9
ONLINE-B	46.3
ONLINE-G	46.0
ONLINE-Y	43.9
GTCOM_Peter	42.2
Lan-BridgeMT	42.1
ONLINE-M	41.3
ZengHuiMT	40.8
NLLB_Greedy	33.1
AIRC	32.4
NLLB_MBR_BLEU	32.4

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	COMET
ONLINE-W	85.5
GPT4-5shot	85.0
ONLINE-B	84.8
ONLINE-Y	84.1
ONLINE-A	83.7
ONLINE-G	82.5
ONLINE-M	81.7
Lan-BridgeMT	80.4
ZengHuiMT	79.4
NLLB_MBR_BLEU	78.0
NLLB_Greedy	77.9
AIRC	72.9

System	chrF
ONLINE-W	71.8
ONLINE-A	69.7
ZengHuiMT	69.4
GPT4-5shot	69.1
ONLINE-B	69.1
ONLINE-Y	69.1
ONLINE-G	69.0
ONLINE-M	66.9
Lan-BridgeMT	66.1
NLLB_Greedy	56.2
NLLB_MBR_BLEU	55.4
AIRC	52.2

System	BLEU
ONLINE-W	47.8
ONLINE-A	43.7
GPT4-5shot	43.6
ONLINE-Y	43.6
ONLINE-G	43.2
ONLINE-B	42.7
ONLINE-M	40.5
ZengHuiMT	40.5
Lan-BridgeMT	39.4
NLLB_Greedy	31.1
NLLB_MBR_BLEU	29.6
AIRC	26.5

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	COMET
ONLINE-B	89.9
ONLINE-A	87.0
GPT4-5shot	86.9
GTCOM_Peter	86.7
ONLINE-G	85.6
ZengHuiMT	85.6
ONLINE-Y	84.9
UvA-LTL	84.7
NLLB_MBR_BLEU	82.9
NLLB_Greedy	82.8
Samsung_Research_Philippines	82.6
Lan-BridgeMT	82.4

System	chrF
ONLINE-B	87.5
ZengHuiMT	76.3
GTCOM_Peter	76.2
ONLINE-A	73.3
GPT4-5shot	71.4
UvA-LTL	70.9
ONLINE-Y	70.5
ONLINE-G	69.8
NLLB_Greedy	64.4
Lan-BridgeMT	63.5
NLLB_MBR_BLEU	63.0
Samsung_Research_Philippines	55.5

System	BLEU
ONLINE-B	76.5
GTCOM_Peter	59.2
ZengHuiMT	56.6
ONLINE-A	53.9
GPT4-5shot	51.2
UvA-LTL	51.0
ONLINE-Y	49.8
ONLINE-G	49.3
NLLB_Greedy	42.5
Lan-BridgeMT	41.4
NLLB_MBR_BLEU	40.7
Samsung_Research_Philippines	34.0

Table 5: Scores for the he→en (refA) 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	COMET
GPT4-5shot	86.4
ONLINE-B	85.6
ONLINE-A	85.3
GTCOM_Peter	84.5
ONLINE-G	84.0
UvA-LTL	83.3
ZengHuiMT	83.3
ONLINE-Y	82.9
NLLB_MBR_BLEU	81.8
NLLB_Greedy	81.7
Lan-BridgeMT	81.3
Samsung_Research_Philippines	81.3

System	chrF
GPT4-5shot	69.5
ONLINE-B	66.5
ONLINE-A	65.6
GTCOM_Peter	65.3
ZengHuiMT	65.1
UvA-LTL	63.3
ONLINE-G	62.8
ONLINE-Y	62.0
NLLB_Greedy	59.6
Lan-BridgeMT	59.0
NLLB_MBR_BLEU	58.6
Samsung_Research_Philippines	51.3

System	BLEU
GPT4-5shot	50.4
ONLINE-B	45.0
GTCOM_Peter	44.4
ONLINE-A	44.4
UvA-LTL	41.7
ZengHuiMT	41.7
ONLINE-G	40.9
ONLINE-Y	38.5
NLLB_Greedy	37.1
Lan-BridgeMT	36.2
NLLB_MBR_BLEU	36.2
Samsung_Research_Philippines	29.8

Table 6: Scores for the he→en (refB) translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-B	86.4
ONLINE-A	85.7
GPT4-5shot	84.9
GTCOM_Peter	84.7
ONLINE-Y	84.7
UvA-LTL	84.2
Samsung_Research_Philippines	83.7
Lan-BridgeMT	83.0
NLLB_Greedy	82.9
ZengHuiMT	82.7
NLLB_MBR_BLEU	82.5
ONLINE-G	82.2

System	chrF
ONLINE-B	66.4
ZengHuiMT	62.1
ONLINE-A	61.7
GTCOM_Peter	61.1
ONLINE-Y	60.4
UvA-LTL	59.0
ONLINE-G	58.1
Samsung_Research_Philippines	57.3
Lan-BridgeMT	54.9
NLLB_Greedy	54.8
NLLB_MBR_BLEU	54.3
GPT4-5shot	54.0

System	BLEU
ONLINE-B	47.8
ONLINE-A	38.9
GTCOM_Peter	37.2
ONLINE-Y	37.2
ZengHuiMT	36.5
UvA-LTL	35.0
Samsung_Research_Philippines	33.3
ONLINE-G	33.2
NLLB_MBR_BLEU	30.8
Lan-BridgeMT	30.5
NLLB_Greedy	30.3
GPT4-5shot	27.0

Table 7: 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	COMET
SKIM	84.0
GPT4-5shot	83.4
ONLINE-W	82.3
NAIST-NICT	81.9
ONLINE-Y	81.6
ONLINE-B	81.5
ONLINE-A	81.0
GTCOM_Peter	80.2
ANVITA	79.5
Lan-BridgeMT	79.3
ZengHuiMT	79.2
ONLINE-G	77.8
ONLINE-M	77.5
KYB	76.6
NLLB_MBR_BLEU	75.2
AIRC	74.5
NLLB_Greedy	74.3

System	chrF
ONLINE-W	51.4
GPT4-5shot	51.2
SKIM	51.1
ONLINE-A	49.6
NAIST-NICT	49.5
ONLINE-Y	49.5
ZengHuiMT	49.5
ONLINE-B	49.3
GTCOM_Peter	48.7
Lan-BridgeMT	47.3
ANVITA	46.7
ONLINE-G	45.5
KYB	43.9
ONLINE-M	43.9
AIRC	40.5
NLLB_MBR_BLEU	39.2
NLLB_Greedy	39.0

System	BLEU
ONLINE-W	25.9
SKIM	24.8
GPT4-5shot	24.1
ONLINE-B	23.9
NAIST-NICT	23.0
ONLINE-A	23.0
ZengHuiMT	22.6
GTCOM_Peter	22.3
ONLINE-Y	22.3
ANVITA	20.9
Lan-BridgeMT	20.2
ONLINE-G	18.3
KYB	17.6
ONLINE-M	17.2
AIRC	14.9
NLLB_MBR_BLEU	14.7
NLLB_Greedy	14.2

Table 8: 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	COMET
ONLINE-B	88.2
ONLINE-W	87.5
ONLINE-Y	87.3
GPT4-5shot	87.0
SKIM	86.6
NAIST-NICT	86.2
ZengHuiMT	85.3
ONLINE-A	85.2
Lan-BridgeMT	84.5
ONLINE-M	13.3
ANVITA	82.7
KYB	80.8
AIRC	80.7
ONLINE-G	80.4
NLLB_Greedy	79.3
NLLB_MBR_BLEU	77.7

System	chrF
ONLINE-B	35.2
ONLINE-Y	34.1
ONLINE-W	33.5
SKIM	33.5
ZengHuiMT	32.9
NAIST-NICT	32.0
ONLINE-A	31.4
GPT4-5shot	31.0
Lan-BridgeMT	30.4
ONLINE-M	29.6
ANVITA	29.3
KYB	27.7
AIRC	27.6
ONLINE-G	27.3
NLLB_Greedy	20.9
NLLB_MBR_BLEU	18.7

System	BLEU
ONLINE-B	25.3
ONLINE-W	24.5
ONLINE-Y	24.5
SKIM	24.3
NAIST-NICT	22.6
ZengHuiMT	22.6
ONLINE-A	21.4
GPT4-5shot	21.3
Lan-BridgeMT	20.5
ONLINE-M	19.8
ANVITA	19.4
KYB	17.8
AIRC	17.6
ONLINE-G	17.2
NLLB_Greedy	11.3
NLLB_MBR_BLEU	9.0

Table 9: 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	COMET
GPT4-5shot	83.5
ONLINE-Y	82.5
ONLINE-B	82.3
ONLINE-W	82.2
ONLINE-G	82.0
ONLINE-A	81.9
PROMT	80.9
ONLINE-M	80.7
NLLB_MBR_BLEU	80.5
NLLB_Greedy	80.1
Lan-BridgeMT	79.9
ZengHuiMT	79.5

System	chrF
GPT4-5shot	60.4
ONLINE-G	59.6
ONLINE-A	59.4
ONLINE-B	59.4
ZengHuiMT	58.9
ONLINE-Y	58.6
PROMT	58.4
ONLINE-W	58.3
Lan-BridgeMT	57.4
ONLINE-M	56.7
NLLB_MBR_BLEU	55.8
NLLB_Greedy	55.5

System	BLEU
ONLINE-B	34.5
GPT4-5shot	34.4
ONLINE-G	34.0
ONLINE-A	33.8
ONLINE-Y	33.2
ONLINE-W	33.1
PROMT	32.8
Lan-BridgeMT	31.8
ZengHuiMT	31.3
NLLB_MBR_BLEU	31.0
ONLINE-M	30.7
NLLB_Greedy	30.3

Table 10: 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	COMET
ONLINE-G	86.6
ONLINE-W	86.6
ONLINE-B	86.2
GPT4-5shot	86.1
ONLINE-Y	85.5
ONLINE-A	85.3
ONLINE-M	83.2
Lan-BridgeMT	83.1
NLLB_Greedy	82.9
NLLB_MBR_BLEU	82.7
PROMT	82.3
ZengHuiMT	81.3

System	chrF
ONLINE-B	61.9
ONLINE-A	59.0
ONLINE-G	58.9
ZengHuiMT	58.8
ONLINE-W	56.6
ONLINE-Y	56.4
GPT4-5shot	56.2
Lan-BridgeMT	55.7
PROMT	55.4
ONLINE-M	55.1
NLLB_Greedy	53.3
NLLB_MBR_BLEU	53.1

System	BLEU
ONLINE-B	40.4
ONLINE-A	34.8
ONLINE-G	32.9
ONLINE-Y	32.0
ZengHuiMT	31.6
ONLINE-W	31.4
ONLINE-M	30.9
Lan-BridgeMT	30.7
GPT4-5shot	30.6
PROMT	30.5
NLLB_MBR_BLEU	28.4
NLLB_Greedy	28.2

Table 11: 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	COMET
ONLINE-W	87.5
GPT4-5shot	87.1
ONLINE-B	86.8
GTCOM_Peter	86.3
ONLINE-A	86.3
ONLINE-G	86.2
ONLINE-Y	85.8
Lan-BridgeMT	84.8
ZengHuiMT	84.4
NLLB_MBR_BLEU	84.3
NLLB_Greedy	84.2

System	chrF
GTCOM_Peter	69.3
ONLINE-W	69.2
ONLINE-B	69.0
ZengHuiMT	68.5
ONLINE-A	68.3
ONLINE-Y	68.2
GPT4-5shot	68.1
ONLINE-G	68.0
Lan-BridgeMT	66.2
NLLB_Greedy	62.4
NLLB_MBR_BLEU	62.4

System	BLEU
ONLINE-W	47.4
GTCOM_Peter	46.4
ONLINE-B	46.0
ONLINE-A	45.9
ONLINE-Y	45.7
ONLINE-G	44.9
GPT4-5shot	43.9
ZengHuiMT	43.5
Lan-BridgeMT	42.3
NLLB_MBR_BLEU	38.1
NLLB_Greedy	37.8

Table 12: 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	COMET
ONLINE-W	86.7
ONLINE-B	85.6
GPT4-5shot	85.3
ONLINE-G	85.3
ONLINE-A	83.2
ONLINE-Y	82.9
GTCOM_Peter	82.1
NLLB_Greedy	82.1
NLLB_MBR_BLEU	81.7
Lan-BridgeMT	80.4
ZengHuiMT	79.0

System	chrF
ONLINE-B	61.7
ONLINE-W	59.2
ZengHuiMT	56.4
ONLINE-G	56.1
ONLINE-A	55.8
ONLINE-Y	55.4
GTCOM_Peter	54.4
GPT4-5shot	53.0
Lan-BridgeMT	51.9
NLLB_Greedy	50.8
NLLB_MBR_BLEU	50.5

System	BLEU
ONLINE-B	39.8
ONLINE-W	34.9
ONLINE-A	30.3
ONLINE-Y	29.5
ONLINE-G	28.6
ZengHuiMT	27.8
GTCOM_Peter	27.5
GPT4-5shot	25.2
NLLB_MBR_BLEU	24.9
Lan-BridgeMT	24.6
NLLB_Greedy	24.5

Table 13: 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	COMET
HW-TSC	82.8
ONLINE-B	82.7
Yishu	82.7
GPT4-5shot	81.6
Lan-BridgeMT	81.2
ONLINE-G	80.9
ONLINE-Y	80.6
ONLINE-A	80.3
ZengHuiMT	79.6
ONLINE-W	79.3
IOL_Research	79.2
ONLINE-M	77.7
NLLB_MBR_BLEU	76.8
ANVITA	76.6
NLLB_Greedy	76.4

System	chrF
HW-TSC	57.5
ONLINE-B	57.5
Yishu	57.4
ZengHuiMT	54.6
ONLINE-G	53.9
ONLINE-A	53.4
GPT4-5shot	53.1
Lan-BridgeMT	53.1
ONLINE-W	52.5
IOL_Research	52.4
ONLINE-Y	52.3
ONLINE-M	49.7
ANVITA	47.1
NLLB_Greedy	46.1
NLLB_MBR_BLEU	45.8

System	BLEU
HW-TSC	33.6
ONLINE-B	33.5
Yishu	33.4
ONLINE-A	28.3
Lan-BridgeMT	27.3
IOL_Research	27.2
ZengHuiMT	27.0
GPT4-5shot	26.8
ONLINE-G	26.6
ONLINE-W	26.4
ONLINE-Y	25.0
ONLINE-M	23.5
ANVITA	21.8
NLLB_Greedy	20.5
NLLB_MBR_BLEU	19.8

Table 14: 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	COMET
ONLINE-B	88.1
Yishu	88.1
HW-TSC	87.3
GPT4-5shot	87.1
ONLINE-W	86.8
Lan-BridgeMT	86.6
ONLINE-Y	86.5
ONLINE-A	86.2
IOL_Research	85.3
ZengHuiMT	84.3
ONLINE-M	84.2
ONLINE-G	83.8
NLLB_Greedy	75.7
ANVITA	75.6
NLLB_MBR_BLEU	71.5

System	chrF
HW-TSC	53.8
Yishu	53.0
ONLINE-B	52.9
ONLINE-A	52.8
IOL_Research	51.9
ONLINE-M	50.6
ONLINE-Y	49.8
ONLINE-G	49.4
ONLINE-W	47.3
ZengHuiMT	47.0
Lan-BridgeMT	46.8
GPT4-5shot	46.5
ANVITA	36.9
NLLB_Greedy	26.3
NLLB_MBR_BLEU	21.1

System	BLEU
HW-TSC	58.6
ONLINE-A	58.5
Yishu	57.6
ONLINE-B	57.5
IOL_Research	56.9
ONLINE-M	54.9
ONLINE-Y	54.2
ONLINE-G	54.1
ZengHuiMT	52.9
ONLINE-W	52.1
Lan-BridgeMT	50.2
GPT4-5shot	49.6
ANVITA	38.9
NLLB_Greedy	27.4
NLLB_MBR_BLEU	19.1

Table 15: 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).

.	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	0.1	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 16: Statistical significance testing of the COMET score difference for each system pair for the cs→uk.

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

	ONLINE-W.	ONLINE-B.	ONLINE-A.	ONLINE-Y.	ONLINE-M.	ONLINE-G.	GTCOM_Pe.	NLLB_MBR.	ZengHuiM.	Lan-Brid.	NLLB_Gre.	AIRC.
GPT4-5shot	0.3	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 18: Statistical significance testing of the COMET score difference for each system pair for the de→en.

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

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

	ONLINE-B.	ONLINE-A.	GTCOM_Pe.	ONLINE-G.	UvA-LTL.	ZengHuiM.	ONLINE-Y.	NLLB_MBR.	NLLB_Gre.	Lan-Brid.
GPT4-5shot	0.8	1.1	1.9	2.4	3.1	3.1	3.5	4.6	4.7	5.1
ONLINE-B	0.0	0.3	1.1	1.6	2.3	2.3	2.7	3.8	3.9	4.3
ONLINE-A		0.0	0.8	1.3	2.0	2.0	2.4	3.5	3.6	4.0
GTCOM_Peter			0.0	0.5	1.2	1.2	1.6	2.7	2.8	3.2
ONLINE-G				0.0	0.7	0.7	1.1	2.2	2.3	2.7
UvA-LTL					0.0	0.0	0.4	1.5	1.6	2.0
ZengHuiMT						0.0	0.4	1.5	1.6	2.0
ONLINE-Y							0.0	1.1	1.2	1.6
NLLB_MBR_BLEU								0.0	0.1	0.5
NLLB_Greedy									0.0	0.4
Lan-BridgeMT										0.0

Table 21: Statistical significance testing of the COMET score difference for each system pair for the he→en (refB).

	ONLINE-A.	GPT4-5sh.	GTCOM_Pe.	ONLINE-Y.	UvA-LTL.	Samsung_.	Lan-Brid.	NLLB_Gre.	ZengHuiM.	NLLB_MBR.	ONLINE-G.
ONLINE-B	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 22: Statistical significance testing of the COMET score difference for each system pair for the en→he.

	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	0.6	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 23: Statistical significance testing of the COMET score difference for each system pair for the ja→en.

	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	0.7	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 24: Statistical significance testing of the COMET score difference for each system pair for the en→ja.

	ONLINE-Y.	ONLINE-B.	ONLINE-W.	ONLINE-G.	ONLINE-A.	PROMT.	ONLINE-M.	NLLB_MBR.	NLLB_Gre.	Lan-Brid.	ZengHuiM.
GPT4-5shot	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 25: Statistical significance testing of the COMET score difference for each system pair for the ru→en.

	ONLINE-W.	ONLINE-B.	GPT4-5sh.	ONLINE-Y.	ONLINE-A.	ONLINE-M.	Lan-Brid.	NLLB_Gre.	NLLB_MBR.	PROMT.	ZengHuiM.
ONLINE-G	0.0	0.4	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 26: Statistical significance testing of the COMET score difference for each system pair for the en→ru.

	GPT4-5sh.	ONLINE-B.	GTCOM_Pe.	ONLINE-A.	ONLINE-G.	ONLINE-Y.	Lan-Brid.	ZengHuiM.	NLLB_MBR.	NLLB_Gre.
ONLINE-W	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 27: Statistical significance testing of the COMET score difference for each system pair for the uk→en.

	ONLINE-B.	GPT4-5sh.	ONLINE-G.	ONLINE-A.	ONLINE-Y.	GTCOM_Pe.	NLLB_Gre.	NLLB_MBR.	Lan-Brid.	ZengHuiM.
ONLINE-W	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 28: Statistical significance testing of the COMET score difference for each system pair for the en→uk.

	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	0.1	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 29: Statistical significance testing of the COMET score difference for each system pair for the zh→en.

	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	0.8	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 30: 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.