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.

https://github.com/mjpost/sacrebleu

System	Rank	chrF	System	Rank	BLEU	System	Rank	COMET
GPT4-5shot	1	61.0	GPT4-5shot	1	32.8	CUNI-GA	1	90.9
CUNI-GA	2	57.9	CUNI-Transformer	2	30.2	GPT4-5shot	2	90.8
GTCOM_Peter	3	57.6	GTCOM_Peter	3	29.8	ONLINE-W	3	89.4
CUNI-Transformer	4	57.4	CUNI-GA	4	29.5	GTCOM_Peter	4	88.9
MUNI-NLP	5	57.0	MUNI-NLP	5	28.3	ONLINE-B	5	88.8
Lan-BridgeMT	6	55.7	Lan-BridgeMT	6	27.5	ONLINE-A	6	88.2
ONLINE-W	7	55.0	ONLINE-W	7	26.8	CUNI-Transformer	7	88.0
ONLINE-B	8	54.7	ONLINE-B	8	25.7	ONLINE-G	8	87.7
ONLINE-A	9	54.4	ONLINE-A	9	25.4	MUNI-NLP	9	87.0
ONLINE-G	10	53.7	NLLB_MBR_BLEU	10	25.1	ONLINE-Y	10	86.5
ONLINE-Y	11	53.4	NLLB_Greedy	11	24.9	NLLB_Greedy	11	86.3
NLLB_Greedy	12	52.5	ONLINE-Ğ	12	24.8	NLLB_MBR_BLEU	12	86.3
NLLB_MBR_BLEU	13	52.3	ONLINE-Y	13	24.2	Lan-BridgeMT	13	86.0

Table 1: Scores for the cs \rightarrow uk translation task: chrF (nrefs:1|case:mixedleff:yeslnc:6|nw:0|space:nolversion:2.2.1), BLEU (nrefs:1|case:mixedleff:noltok:13a|smooth:explversion:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	Rank	chrF
System	Kank	CIIII
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

Table 2: Scores for the en→cs translation task: chrF (nrefs:1|case:mixedleff:yeslnc:6|nw:0|space:nolversion:2.2.1), BLEU (nrefs:1|case:mixedleff:noltok:13a|smooth:explversion:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	Rank	chrF	System	Rank	BLEU	System	Rank	COMET
ONLINE-W	1	72.1	ONLINE-W	1	51.8	GPT4-5shot	1	86.3
ONLINE-A	2	70.0	GPT4-5shot	2	47.9	ONLINE-W	2	86.0
GPT4-5shot	3	69.8	ONLINE-A	3	47.9	ONLINE-B	3	85.6
ONLINE-B	4	69.1	ONLINE-B	4	46.3	ONLINE-A	4	85.5
ONLINE-G	5	69.1	ONLINE-G	5	46.0	ONLINE-Y	5	84.9
ONLINE-Y	6	68.4	ONLINE-Y	6	43.9	ONLINE-M	6	84.8
ZengHuiMT	7	67.6	GTCOM_Peter	7	42.2	ONLINE-G	7	84.6
Lan-BridgeMT	8	66.7	Lan-BridgeMT	8	42.1	GTCOM_Peter	8	82.7
GTCOM_Peter	9	66.6	ONLINE-M	9	41.3	NLLB_MBR_BLEU	9	81.4
ONLINE-M	10	66.5	ZengHuiMT	10	40.8	ZengHuiMT	10	81.1
NLLB_MBR_BLEU	11	57.6	NLLB_Greedy	11	33.1	Lan-BridgeMT	11	80.9
NLLB_Greedy	12	57.3	AIRC	12	32.4	NLLB_Greedy	12	79.9
AIRC	13	57.2	NLLB_MBR_BLEU	13	32.4	AIRC	13	78.7

Table 3: Scores for the de \rightarrow en translation task: chrF (nrefs:1|case:mixedleff:yeslnc:6|nw:0|space:nolversion:2.2.1), BLEU (nrefs:1|case:mixedleff:noltok:13a|smooth:explversion:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	Rank	chrF	System	Rank	BLEU	System	Ra
ONLINE-W	1	71.8	ONLINE-W	1	47.8	ONLINE-W	1
ONLINE-A	2	69.7	ONLINE-A	2	43.7	GPT4-5shot	2
ZengHuiMT	3	69.4	GPT4-5shot	3	43.6	ONLINE-B	3
GPT4-5shot	4	69.1	ONLINE-Y	4	43.6	ONLINE-Y	4
ONLINE-B	5	69.1	ONLINE-G	5	43.2	ONLINE-A	5
ONLINE-Y	6	69.1	ONLINE-B	6	42.7	ONLINE-G	6
ONLINE-G	7	69.0	ONLINE-M	7	40.5	ONLINE-M	7
ONLINE-M	8	66.9	ZengHuiMT	8	40.5	Lan-BridgeMT	8
Lan-BridgeMT	9	66.1	Lan-BridgeMT	9	39.4	ZengHuiMT	9
NLLB_Greedy	10	56.2	NLLB_Greedy	10	31.1	NLLB_MBR_BLEU	10
NLLB_MBR_BLEU	11	55.4	NLLB_MBR_BLEU	11	29.6	NLLB_Greedy	11
AIRC	12	52.2	AIRC	12	26.5	AIRC	12

Table 4: Scores for the en \rightarrow 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:mixedleff:yeslnc:6|nw:0|space:nolversion:2.2.1), BLEU (nrefs:1|case:mixedleff:noltok:13a|smooth:explversion: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 \rightarrow 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	System	Rank	BLEU	System	Rank	
ONLINE-W	1	51.4	ONLINE-W	1	25.9	SKIM	1	
GPT4-5shot	2	51.2	SKIM	2	24.8	GPT4-5shot	2	
SKIM	3	51.1	GPT4-5shot	3	24.1	ONLINE-W	3	
ONLINE-A	4	49.6	ONLINE-B	4	23.9	NAIST-NICT	4	
NAIST-NICT	5	49.5	NAIST-NICT	5	23.0	ONLINE-Y	5	
ONLINE-Y	6	49.5	ONLINE-A	6	23.0	ONLINE-B	6	
ZengHuiMT	7	49.5	ZengHuiMT	7	22.6	ONLINE-A	7	
ONLINE-B	8	49.3	GTCOM_Peter	8	22.3	GTCOM_Peter	8	
GTCOM_Peter	9	48.7	ONLINE-Y	9	22.3	ANVITA	9	
Lan-BridgeMT	10	47.3	ANVITA	10	20.9	Lan-BridgeMT	10	
ANVITA	11	46.7	Lan-BridgeMT	11	20.2	ZengHuiMT	11	
ONLINE-G	12	45.5	ONLINE-G	12	18.3	ONLINE-G	12	
KYB	13	43.9	KYB	13	17.6	ONLINE-M	13	
ONLINE-M	14	43.9	ONLINE-M	14	17.2	KYB	14	
AIRC	15	40.5	AIRC	15	14.9	NLLB_MBR_BLEU	15	
ILLB_MBR_BLEU	16	39.2	NLLB_MBR_BLEU	16	14.7	AIRC	16	
NLLB_Greedy	17	39.0	NLLB_Greedy	17	14.2	NLLB_Greedy	17	

Table 7: Scores for the ja \rightarrow en translation task: chrF (nrefs:1|case:mixedleff:yeslnc:6|nw:0|space:nolversion:2.2.1), BLEU (nrefs:1|case:mixedleff:noltok:13a|smooth:explversion:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	Rank	chrF	System	Rank	BLEU	System	Rank	COMET
ONLINE-B	1	35.2	ONLINE-B	1	25.3	ONLINE-B	1	88.2
ONLINE-Y	2	34.1	ONLINE-W	2	24.5	ONLINE-W	2	87.5
ONLINE-W	3	33.5	ONLINE-Y	3	24.5	ONLINE-Y	3	87.3
SKIM	4	33.5	SKIM	4	24.3	GPT4-5shot	4	87.0
ZengHuiMT	5	32.9	NAIST-NICT	5	22.6	SKIM	5	86.6
NAIST-NICT	6	32.0	ZengHuiMT	6	22.6	NAIST-NICT	6	86.2
ONLINE-A	7	31.4	ONLINE-A	7	21.4	ZengHuiMT	7	85.3
GPT4-5shot	8	31.0	GPT4-5shot	8	21.3	ONLINE-A	8	85.2
Lan-BridgeMT	9	30.4	Lan-BridgeMT	9	20.5	Lan-BridgeMT	9	84.5
ONLINE-M	10	29.6	ONLINE-M	10	19.8	ONLINE-M	10	83.3
ANVITA	11	29.3	ANVITA	11	19.4	ANVITA	11	82.7
KYB	12	27.7	KYB	12	17.8	KYB	12	80.8
AIRC	13	27.6	AIRC	13	17.6	AIRC	13	80.7
ONLINE-G	14	27.3	ONLINE-G	14	17.2	ONLINE-G	14	80.4
NLLB_Greedy	15	20.9	NLLB_Greedy	15	11.3	NLLB_Greedy	15	79.3
NLLB_MBR_BLEU	16	18.7	NLLB_MBR_BLEU	16	9.0	NLLB_MBR_BLEU	16	77.7

Table 8: Scores for the en \rightarrow ja translation task: chrF (nrefs:1|case:mixedleff:yeslnc:6|nw:0|space:nolversion:2.2.1), BLEU (nrefs:1|case:mixedleff:noltok:ja-mecab-0.996-IPA|smooth:explversion:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	Rank	chrF	_	System	Rank	BLEU	System	Rank	COMET
GPT4-5shot	1	60.4		ONLINE-B	1	34.5	GPT4-5shot	1	83.5
ONLINE-G	2	59.6		GPT4-5shot	2	34.4	ONLINE-Y	2	82.5
ONLINE-A	3	59.4		ONLINE-G	3	34.0	ONLINE-B	3	82.3
ONLINE-B	4	59.4		ONLINE-A	4	33.8	ONLINE-W	4	82.2
ZengHuiMT	5	58.9		ONLINE-Y	5	33.2	ONLINE-G	5	82.0
ONLINE-Y	6	58.6		ONLINE-W	6	33.1	ONLINE-A	6	81.9
PROMT	7	58.4		PROMT	7	32.8	PROMT	7	80.9
ONLINE-W	8	58.3		Lan-BridgeMT	8	31.8	ONLINE-M	8	80.7
Lan-BridgeMT	9	57.4		ZengHuiMT	9	31.3	NLLB_MBR_BLEU	9	80.5
ONLINE-M	10	56.7		NLLB_MBR_BLEU	10	31.0	NLLB_Greedy	10	80.1
NLLB_MBR_BLEU	11	55.8		ONLINE-M	11	30.7	Lan-BridgeMT	11	79.9
NLLB_Greedy	12	55.5		NLLB_Greedy	12	30.3	ZengHuiMT	12	79.5

Table 9: Scores for the ru—en translation task: chrF (nrefs:1|case:mixedleff:yeslnc:6|nw:0|space:nolversion:2.2.1), BLEU (nrefs:1|case:mixedleff:noltok:13a|smooth:explversion:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	Rank	chrF	System	Rank	BLEU	System	Rank	COMET
ONLINE-B	1	61.9	ONLINE-B	1	40.4	ONLINE-G	1	86.6
ONLINE-A	2	59.0	ONLINE-A	2	34.8	ONLINE-W	2	86.6
ONLINE-G	3	58.9	ONLINE-G	3	32.9	ONLINE-B	3	86.2
ZengHuiMT	4	58.8	ONLINE-Y	4	32.0	GPT4-5shot	4	86.1
ONLINE-W	5	56.6	ZengHuiMT	5	31.6	ONLINE-Y	5	85.5
ONLINE-Y	6	56.4	ONLINE-W	6	31.4	ONLINE-A	6	85.3
GPT4-5shot	7	56.2	ONLINE-M	7	30.9	ONLINE-M	7	83.2
Lan-BridgeMT	8	55.7	Lan-BridgeMT	8	30.7	Lan-BridgeMT	8	83.1
PROMT	9	55.4	GPT4-5shot	9	30.6	NLLB_Greedy	9	82.9
ONLINE-M	10	55.1	PROMT	10	30.5	NLLB_MBR_BLEU	10	82.7
NLLB_Greedy	11	53.3	NLLB_MBR_BLEU	11	28.4	PROMT	11	82.3
NLLB_MBR_BLEU	12	53.1	NLLB_Greedy	12	28.2	ZengHuiMT	12	81.3

Table 10: Scores for the en \rightarrow ru translation task: chrF (nrefs:1|case:mixedleff:yeslnc:6|nw:0|space:nolversion:2.2.1), BLEU (nrefs:1|case:mixedleff:noltok:13a|smooth:explversion:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	Rank	chrF	System	Rank	BLEU	System	Rank	CO
GTCOM_Peter	1	69.3	ONLINE-W	1	47.4	ONLINE-W	1	
ONLINE-W	2	69.2	GTCOM_Peter	2	46.4	GPT4-5shot	2	:
ONLINE-B	3	69.0	ONLINE-B	3	46.0	ONLINE-B	3	
ZengHuiMT	4	68.5	ONLINE-A	4	45.9	GTCOM_Peter	4	8
ONLINE-A	5	68.3	ONLINE-Y	5	45.7	ONLINE-A	5	
ONLINE-Y	6	68.2	ONLINE-G	6	44.9	ONLINE-G	6	
GPT4-5shot	7	68.1	GPT4-5shot	7	43.9	ONLINE-Y	7	8
ONLINE-G	8	68.0	ZengHuiMT	8	43.5	Lan-BridgeMT	8	8
Lan-BridgeMT	9	66.2	Lan-BridgeMT	9	42.3	ZengHuiMT	9	8
NLLB_Greedy	10	62.4	NLLB_MBR_BLEU	10	38.1	NLLB_MBR_BLEU	10	8
ILLB_MBR_BLEÚ	11	62.4	NLLB_Greedy	11	37.8	NLLB_Greedy	11	:

Table 11: Scores for the uk—en translation task: chrF (nrefs:1lcase:mixedleff:yeslnc:6lnw:0lspace:nolversion:2.2.1), BLEU (nrefs:1lcase:mixedleff:noltok:13alsmooth:explversion:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	Rank	chrF	System	Rank	BLEU	System	Rank	COMET
ONLINE-B	1	61.7	ONLINE-B	1	39.8	ONLINE-W	1	86.7
ONLINE-W	2	59.2	ONLINE-W	2	34.9	ONLINE-B	2	85.6
ZengHuiMT	3	56.4	ONLINE-A	3	30.3	GPT4-5shot	3	85.3
ONLINE-G	4	56.1	ONLINE-Y	4	29.5	ONLINE-G	4	85.3
ONLINE-A	5	55.8	ONLINE-G	5	28.6	ONLINE-A	5	83.2
ONLINE-Y	6	55.4	ZengHuiMT	6	27.8	ONLINE-Y	6	82.9
GTCOM_Peter	7	54.4	GTCOM_Peter	7	27.5	GTCOM_Peter	7	82.1
GPT4-5shot	8	53.0	GPT4-5shot	8	25.2	NLLB_Greedy	8	82.1
Lan-BridgeMT	9	51.9	NLLB MBR BLEU	9	24.9	NLLB MBR BLEU	9	81.7
NLLB Greedy	10	50.8	Lan-BridgeMT	10	24.6	Lan-BridgeMT	10	80.4
NLLB_MBR_BLEÚ	11	50.5	NLLB_Greedy	11	24.5	ZengHuiMT	11	79.0

Table 12: Scores for the en→uk translation task: chrF (nrefs:1|case:mixedleff:yeslnc:6|nw:0|space:nolversion:2.2.1), BLEU (nrefs:1|case:mixedleff:noltok:13a|smooth:explversion:2.2.1), COMET (Unbabel/wmt22-comet-da).

			_							
System	Rank	chrF		System	Rank	BLEU		System	Rank	COME
HW-TSC	1	57.5	-	HW-TSC	1	33.6	•	HW-TSC	1	82.8
ONLINE-B	2	57.5		ONLINE-B	2	33.5		ONLINE-B	2	82.7
Yishu	3	57.4		Yishu	3	33.4		Yishu	3	82.7
ZengHuiMT	4	54.6		ONLINE-A	4	28.3		GPT4-5shot	4	81.6
ONLINE-G	5	53.9		Lan-BridgeMT	5	27.3		Lan-BridgeMT	5	81.2
ONLINE-A	6	53.4		IOL_Research	6	27.2		ONLINE-G	6	80.9
GPT4-5shot	7	53.1		ZengHuiMT	7	27.0		ONLINE-Y	7	80.6
Lan-BridgeMT	8	53.1		GPT4-5shot	8	26.8		ONLINE-A	8	80.3
ONLINE-W	9	52.5		ONLINE-G	9	26.6		ZengHuiMT	9	79.6
IOL_Research	10	52.4		ONLINE-W	10	26.4		ONLINE-W	10	79.3
ONLINE-Y	11	52.3		ONLINE-Y	11	25.0		IOL_Research	11	79.2
ONLINE-M	12	49.7		ONLINE-M	12	23.5		ONLINE-M	12	77.7
ANVITA	13	47.1		ANVITA	13	21.8		NLLB_MBR_BLEU	13	76.8
NLLB_Greedy	14	46.1		NLLB_Greedy	14	20.5		ANVITA	14	76.6
NLLB_MBR_BLEU	15	45.8		NLLB_MBR_BLEÚ	15	19.8		NLLB_Greedy	15	76.4

Table 13: Scores for the zh—en translation task: chrF (nrefs:1|case:mixedleff:yeslnc:6|nw:0|space:nolversion:2.2.1), BLEU (nrefs:1|case:mixedleff:noltok:13a|smooth:explversion:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	Rank	chrF	System	Rank	BLEU	_	System	Ranl	k		
HW-TSC	1	53.8	HW-TSC	1	58.6	_	ONLINE-B	1			
Yishu	2	53.0	ONLINE-A	2	58.5		Yishu	2			
ONLINE-B	3	52.9	Yishu	3	57.6		HW-TSC	3			
ONLINE-A	4	52.8	ONLINE-B	4	57.5		GPT4-5shot	4			
IOL_Research	5	51.9	IOL_Research	. 5	56.9		ONLINE-W	5			
ONLINE-M	6	50.6	ONLINE-M	6	54.9		Lan-BridgeMT	6			
ONLINE-Y	7	49.8	ONLINE-Y	7	54.2		ONLINE-Y	7			
ONLINE-G	8	49.4	ONLINE-G	8	54.1		ONLINE-A	8			
ONLINE-W	9	47.3	ZengHuiMT	9	52.9		IOL_Research	9			
ZengHuiMT	10	47.0	ONLINE-W	10	52.1		ZengHuiMT	10			
Lan-BridgeMT	11	46.8	Lan-BridgeMT	11	50.2		ONLINE-M	11			
GPT4-5shot	12	46.5	GPT4-5shot	12	49.6		ONLINE-G	12			
ANVITA	13	36.9	ANVITA	13	38.9		NLLB_Greedy	13			
NLLB_Greedy	14	26.3	NLLB_Greedy	14	27.4		ĀNVITĀ	14			
NLLB_MBR_BLEÚ	15	21.1	NLLB_MBR_BLEŬ	15	19.1		NLLB_MBR_BLEU	15			

Table 14: Scores for the en \rightarrow zh translation task: chrF (nrefs:1lcase:mixedleff:yeslnc:6lnw:0lspace:nolversion:2.2.1), BLEU (nrefs:1lcase:mixedleff:noltok:zhlsmooth:explversion:2.2.1), COMET (Unbabel/wmt22-comet-da).

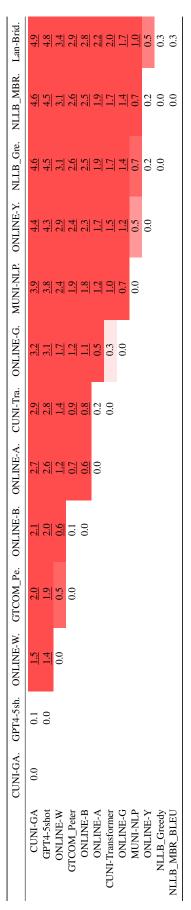


Table 15: Statistical significance testing of the COMET score difference for each system pair for the cs→uk.

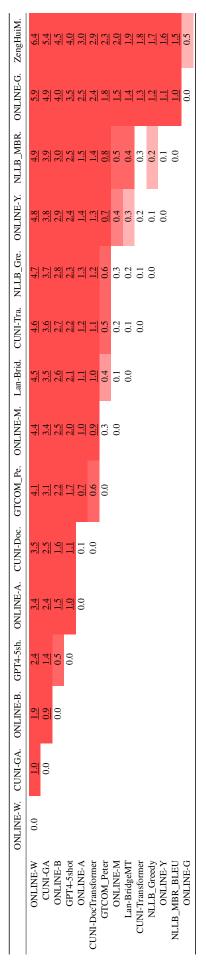


Table 16: Statistical significance testing of the COMET score difference for each system pair for the en→cs.

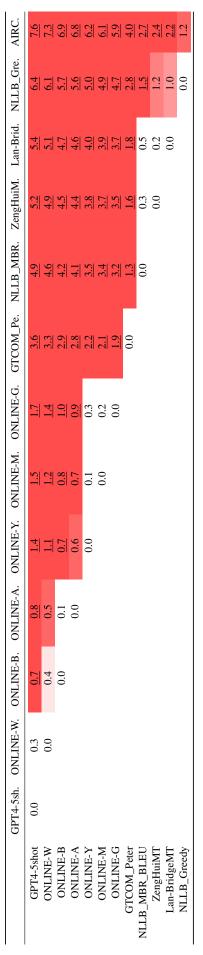


Table 17: Statistical significance testing of the COMET score difference for each system pair for the de→en.

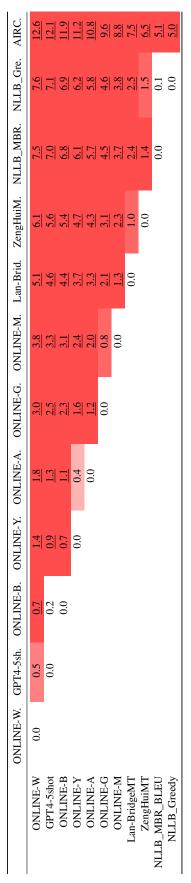


Table 18: Statistical significance testing of the COMET score difference for each system pair for the en→de.

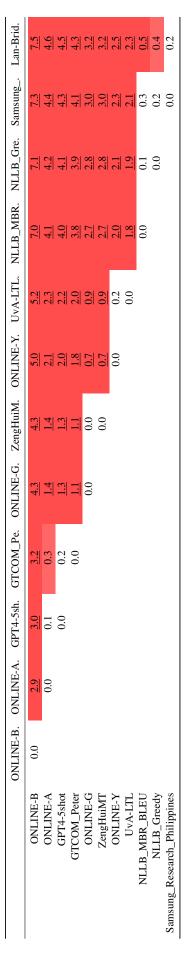


Table 19: Statistical significance testing of the COMET score difference for each system pair for the he→en.

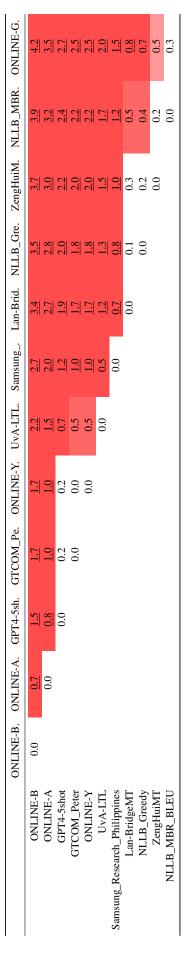


Table 20: Statistical significance testing of the COMET score difference for each system pair for the en→he.

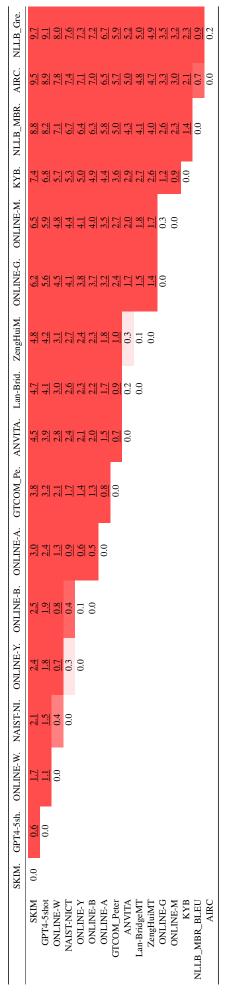


Table 21: Statistical significance testing of the COMET score difference for each system pair for the ja→en.

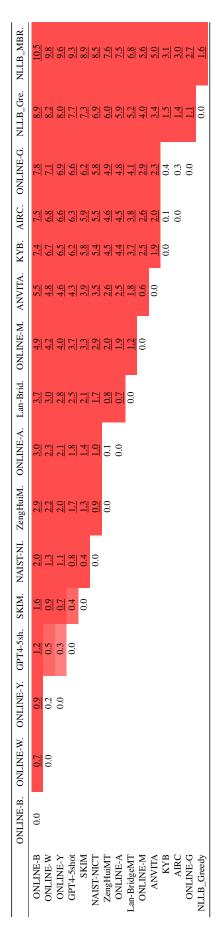


Table 22: Statistical significance testing of the COMET score difference for each system pair for the en→ja.

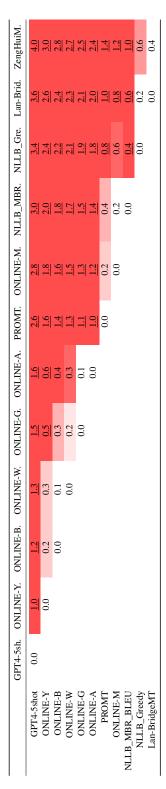


Table 23: Statistical significance testing of the COMET score difference for each system pair for the ru→en.

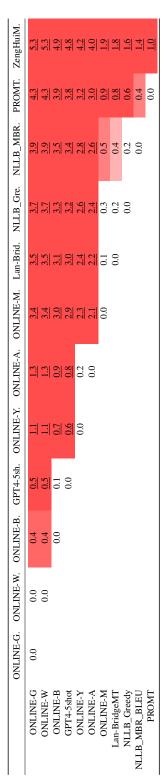


Table 24: Statistical significance testing of the COMET score difference for each system pair for the en→ru.

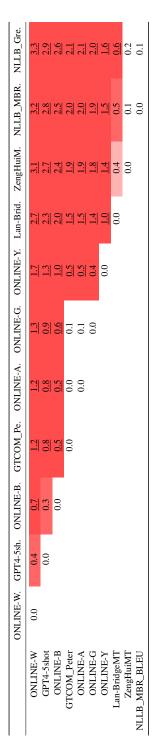


Table 25: Statistical significance testing of the COMET score difference for each system pair for the uk→en.

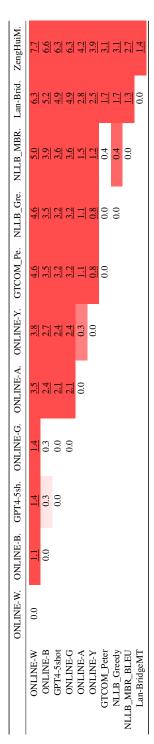


Table 26: Statistical significance testing of the COMET score difference for each system pair for the en→uk.

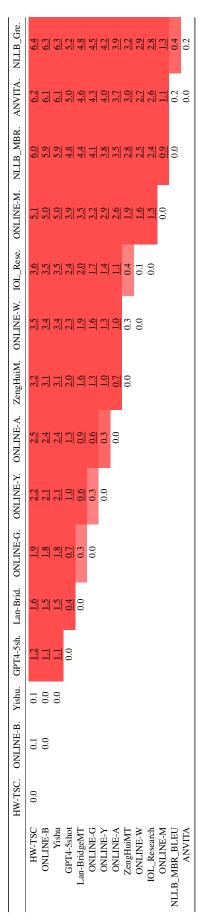


Table 27: Statistical significance testing of the COMET score difference for each system pair for the zh→en.

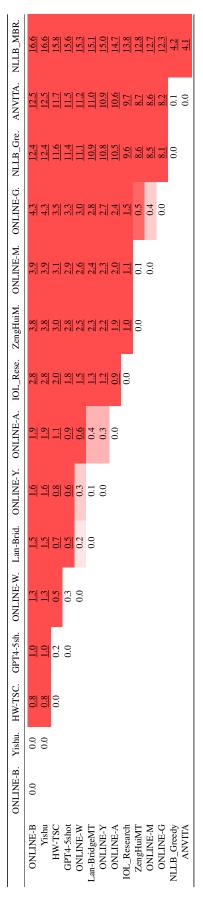


Table 28: Statistical significance testing of the COMET score difference for each system pair for the en→zh.

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.