An Automatic Evaluation of the WMT22 General Machine Translation Task

Benjamin Marie
nlp@benjaminmarie.com

Abstract
This report presents an automatic evaluation of the general machine translation task of the Seventh Conference on Machine Translation (WMT22). It evaluates a total of 185 systems for 21 translation directions including high-resource to low-resource language pairs and from closely related to distant languages. This large-scale automatic evaluation highlights some of the current limits of state-of-the-art machine translation systems. It also shows how automatic metrics, namely chrF, BLEU, and COMET, can complement themselves to mitigate their own limits in terms of interpretability and accuracy.

1 Introduction
In its 2022 edition, the Conference on Machine Translation (WMT22)\(^1\) organized a new “general” machine translation shared task. The participants in this task were asked to prepare machine translation (MT) systems for translating texts from potentially any domains and most likely “news, social, conversational, and ecommerce.”

For this task, WMT22 proposed a large variety of language pairs: Czech↔English (cs-en), Czech↔Ukrainian (cs-uk), German↔English (de-en), French↔German (fr-de), English↔Croatian (en-hr), English↔Japanese (en-ja), English↔Livonian (en-liv), English↔Russian (en-ru), Russian↔Yakut (ru-sah), English↔Ukrainian (en-uk), and English↔Chinese (en-zh). The organizers classified all these language pairs as in Table 1 according to the relatedness of the languages and the quantity of resources available for training an MT system. In total, there are 11 language pairs and 21 translation directions.

In this paper, we present an automatic evaluation\(^2\) of the 185 systems submitted\(^3\) to the general machine translation task. We used the translations officially released by WMT22.\(^4\) We analyze the results with different metrics and methods. Our main observations are as follows:

- Low-resource distant language pairs remain extremely challenging. The best systems for en↔liv and ru↔sah are very far from reaching the reference. ja↔en and en↔uk also remain challenging despite being medium-resource. Among the high-resource language pairs, the systems built for en↔ru struggled the most to approximate the reference translation.

- A BLEU score difference higher than 0.9 is always statistically significant in this task.

- A naive combination of all the evaluated systems using minimum Bayes risk (MBR) decoding with COMET can improve the BLEU, chrF and COMET scores.

- Only a few systems seem to have exploited document-level features or a context larger than one segment.

- Absolute COMET scores are meaningless and must be completed with another metric score to get an estimate of the system’s translation quality.

---

\(^1\)https://www.statmt.org/wmt22/

\(^2\)This is not an official paper of WMT22. WMT22 organizers will perform a human evaluation.

\(^3\)It includes the “Online-#” systems added by the organizers.

\(^4\)https://github.com/wmt-conference/wmt22-news-systems
The paper is organized as follows. In each section, we present an overview of the main findings with the numerous associated tables given in the appendix. In Section 2, we score and rank the systems with three different metrics. These rankings are completed by a statistical significance testing on all possible pairs of systems in Section 3. Section 4 presents the results of a combination of all the submissions using MBR and an oracle guided by COMET. Section 5 presents various statistics of the submitted systems to better highlight how they stand out. For a more in-depth evaluation with COMET, we evaluate all the systems with 5 different COMET models in Section 6. Finally, Section 7 concludes this evaluation.

2 Ranking with Metrics

For this evaluation, we will rely on 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 “wmt20-comet-da.”

Note that chrF and BLEU are merely used for diagnostic purposes and to answer the question: How far are we from reaching particular reference translations? On the other hand, we used COMET mainly to produce more accurate rankings that would better correlate with the human evaluation.

chrF and BLEU scores are computed with SacreBLEU (Post, 2018). If there are two reference translations available, both are used to compute the chrF and BLEU scores, while we only used the first available reference to compute COMET scores. The systems are ranked given their scores but a rank is assigned only to the systems that have been declared “constrained” by their authors, i.e., systems that only used the data provided by the organizers. The rankings and scores are given in Section A, from Table 2 to 22. In these tables, systems with a rank “n/a” are systems that are not constrained.

Having two references, we obtain absolute BLEU scores rarely seen in the machine translation research literature with, for instance, 60.9 BLEU points for JDExploreAcademy for cs→en. Even higher BLEU scores are observed for en→zh due to the use of smaller tokens that makes the 4-gram matching a much easier task. Such high scores may be misleading about the translation quality. Absolute BLEU scores do not inform us on the translation quality itself, but they do inform us that these systems produce many 4-grams that are in the reference translations.

While chrF and BLEU directly indicate how well the translation matches the references with a score between 0 and 100 points, COMET scores are not bounded between 0 and 100 making the interpretation of absolute scores extremely challenging. For instance, at the extremes, AMU obtains 104.9 COMET points for uk→cs and AIST obtains -152.7 COMET points for liv→en. BLEU and chrF absolute scores can be used for diagnostic purposes and answer basic questions: How close are we from the reference with a given tokenization? Is the system likely generating text in the target languages? etc. COMET cannot, but is much more reliable for ranking systems as demonstrated in previous work (Rei et al., 2020).

 BLEU and chrF yields the same rankings for only 5 translation directions: uk→cs, de→fr, sah→ru, and ru→sah. Nonetheless, BLEU and chrF yields ranking that are very similar and find the same best

---

5https://github.com/mjpost/sacrebleu
6We did not use the reference “stud” created by students.
systems for 20 among the 21 translation directions. On the other hand, for 11 among the 21 language pairs, COMET finds a best system that is not among the best systems found by BLEU and chrF. For some system pairs, the difference between BLEU and COMET can be very significant (more on this in Section 3). For instance for cs → uk, in terms of score differences, the difference between Online-B and AMU is of 3.6 BLEU points and -5.1 COMET points.

Surprisingly, for some translation directions, constrained systems outperform systems that are not constrained. According to COMET, this is the case for cs → uk, uk → cs, de → en, ja → en, and en → ja.

For some other directions, online systems seem to be better by a large margin. For instance, for de → fr, Online-W is better than the best constrained system by 18.3 BLEU points.

3 Statistical Significance Testing

We would like to measure how reliable is the conclusion that a system is better than another one according to some metric. In other words, we would like to test whether the difference between systems’ metric scores is statistically significant. There are several tools and techniques to perform statistical significance testing. For this evaluation, we chose the most commonly used: paired bootstrap resampling (Koehn, 2004). We used the implementation in SacreBLEU with its default parameters for chrF and BLEU. For COMET, we used the default parameters of “comet-compare.” Results are reported in Section B, from Table 23 to 85. In these tables, the difference between the metric scores is reported for each system pair. The background color is darker for more significant differences (lower p-value) and the score difference is underline if the p-value is below 0.05.

A first interesting observation we have is that a difference in BLEU higher than 0.9 points (cs → uk) is always significant (p-value < 0.05). In chrF, the largest difference that is not significant is 0.6 points (en → zh), while it reaches 2.6 points (liv → en) for COMET.7

The significance of the difference between systems can be very different depending on the metric. Overall, there is no clear pattern. For some translation directions, one metric will find a cluster of similar systems while another metric will find that they are all significantly different. This is well-illustrated with en → ru where BLEU finds a cluster of similar systems while COMET finds that they are almost all significantly different.

The three metrics only agree on a system that is significantly better than all the others for a few translation directions: cs → en (Online-W), fr → de (Online-W), en → liv (TAL-SJTU), sah → ru (Online-G), and en → uk (Online-B).

It will be interesting to observe how well the clusters of systems found with statistical significance testing match the clusters found with WMT22 human evaluation.

4 Naive MBR and Oracle Combinations

As shown by the tables presented in Section 3, most of the evaluated translations are significantly different according to COMET. One way to exploit this diversity is to combine the translations to generate a better translation. There are many methods to combine machine translation outputs. In this section, we only present two different methods for combination:

- Minimum Bayes Risk (MBR): Naively use all the evaluated translations and performs MBR decoding using COMET.

- Oracle: Select the best segments among all the evaluated translations according to their COMET score. This is an oracle experiment since we use the COMET scores computed with a reference translation. This experiment is mainly to get an idea of what would be the best translation obtained from the combination of all the systems.

The results of these combinations are reported in Section C, from Table 86 to 96. These MBR and oracle combinations rely on COMET but we also report on the chrF and BLEU scores of the resulting translations. “Baseline” denotes the score of the best single system for each metric.

7Note that this would highly vary depending on the model used with COMET.
As we observe in the tables, the MBR combination improves on all the three metrics for cs→uk, uk→cs (with 1.9 BLEU points of improvement), de→en, en→de, and liv→en.

However, for some translation directions, the MBR combination largely fails in improving the metric scores and may even decrease the COMET scores compared to the best single system. For instance, the MBR combination leads to lower COMET scores for cs→en, de→fr, ru→sah, ssa→ru, and uk→en. This is presumably due to the presence of several translations of a very different and lower quality compared to the best single translation.

On the other hand, the oracle combination improves the scores on all metrics for all translation directions. This shows that BLEU and chrF tend to agree, mostly, that a segment with a higher COMET score is better.

5 Sentence Matching and Normalization

In this section, we look at various statistics to try to identify some salient characteristics of the evaluated systems. Statistics are presented in Section E from Table 108 to 118.

Exact Match We counted how many segments in the system translation are identical to their corresponding segments in the reference translation.

Self Mismatch We counted the number of non-unique source segments translated differently by the system and the reference translation. It leads to interesting observations. Most systems always identically translate the same source segments, but there are some notable exceptions. For instance, Online-W generated different translations for the same segment for almost all its translation directions. For 47 of the 122 non-unique source segments of the en→ja, Online-W generates a different translation. Among the other submissions to this translation direction, only LanguageX (1) and NAIST-NICT-TIT (22) generated different translations for the same source segments. In contrast, the human translator(s?) produced 63 different translations for the 122 non-unique segments. This is the highest number among all the translation directions. The case of Japanese is particularly interesting here because in this language the context can frequently and radically lead to different translations for the same text. Getting different translations for the same segment is a clue that the system either takes into account document-level information and/or use sampling strategies with some randomness at inference time. CUNI-DocTransformer is another interesting example for cs→en since it translates differently as many non-unique source segments (7) as the reference translation, probably by exploiting document-level features as suggested by the name of the system. For 10 translation directions, the reference translation translates all the non-unique segments identically.

Sentence Normalization We also normalized the translation outputs to observe how BLEU and COMET are sensitive to changes in the punctuation marks and encoding issues. It can also highlight whether a system relied on some special post-processing to increase the metric scores. For normalization, we used the following sequence of Moses (Koehn et al., 2007) scripts:

```
tokenizer/replace-unicode-punctuation.perl | tokenizer/normalize-punctuation.perl -l <target_language> | /tokenizer/remove-non-printing-char.perl
```

As expected, COMET is almost insensitive to this normalization. On the other hand, this normalization has a stronger impact on the BLEU scores, but it can greatly vary from one system to another. For instance for en→cs, it has no effect on JDEpoloAcademy while the score of Online-Y drops by 1.4 BLEU points. For de→fr, the normalization increases the BLEU score of Online-A by 4.9 point and becomes better than Online-W for which the normalization has no effect on BLEU. Nonetheless, Online-W remains around 10 COMET points better than Online-A.

6 Evaluation With Different COMET Models

In Section 2, we observed that COMET scores can be very high (above 100) or very low (below -100) making their interpretation very difficult. In this section, in an attempt to obtain scores easier to interpret
we reevaluate all the translations with 5 different COMET models:

- wmt20-comet-da: The same model used in Section 2 trained with direct assessment (DA).
- wmt21-comet-da: A more recent model trained with DA on more data.
- wmt21-comet-mqm: A model trained with MQM human judgments.
- wmt21-comet-qe-mqm: A reference-less model trained with MQM human judgments.
- wmt21-comet-qe-da: A reference-less model trained with DA.

The results using these 5 models are presented in Section D from Table 97 to 107.

We can observe that wmt20-comet-da scores are actually quite different from all the other models. While the maximum score obtained by a system with wmt20-comet-da is 104.9 (uk→cs), the scores obtained with the other 4 models never exceed 15.9 for all translation directions. The absolute scores remain difficult to interpret with the 4 new models. More particularly, with wmt21-comet-da for ja→en, the best system is scored at 1.1 while for zh→en, wmt21-comet-da scores are negative for all the systems. We can conclude that absolute COMET scores are not informative whatever model we use. Negative COMET scores can be assigned to an excellent machine translation.

Nonetheless, the rankings obtained by all 5 models are very similar albeit the reference-less models that tend to produce more different rankings.

7 Conclusion

In this paper, we presented a large-scale evaluation of 185 systems for the 21 translation directions of 11 language pairs. The variety of scenarios, going from high-resource to low-resource language pairs, for closely related or distant language pairs, draws an interesting picture of the state of machine translation. According to three different automatic metrics, significant progress remains to be done to get close to a reference translation, especially for the most difficult scenario involving distant low-resource language pairs.

For only 5 translation directions, WMT22 systems submitted are significantly better than online systems according to COMET. This suggests that the online systems for these translation directions used less training data.

We highlight that while COMET is a state-of-the-art metric for ranking systems, its score are uninformative, or misleading for the readers that are unaware of COMET, as even systems that looks reasonably good may only achieve a negative COMET score. We believe that this peculiarity\(^\text{10}\) is a major obstacle for its wide adoption in MT research.

We also highlighted that most submitted systems still do not exploit document-level features or context larger than one segment.

It will be interesting to observe how the human evaluation conducted by WMT22 will align with these findings.

Acknowledgments

We would like to thank the WMT organizers for releasing the translations and Tom Kocmi for providing preliminary results as well as insightful comments and suggestions on the first draft of this report.

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.

\(^{10}\)The large majority of NLP evaluation metrics only yield positive scores.
Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. 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.
### Table 2: Scores for the cs→en translation task: chrF (nrefs:2|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:2|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System               | Rank | chrF  | System               | Rank | chrF  | System               | Rank | chrF  |
|----------------------|------|-------|----------------------|------|-------|----------------------|------|-------|
| Online-W             | n/a  | 74.5  | Online-W             | n/a  | 68.7  | Online-W             | n/a  | 77.4  |
| JDExploreAcademy     | 1    | 75.0  | JDExploreAcademy     | 1    | 60.9  | JDExploreAcademy     | 1    | 74.8  |
| Lan-Bridge           | n/a  | 7.6   | Lan-Bridge           | n/a  | 60.8  | Lan-Bridge           | n/a  | 71.9  |
| Online-B             | n/a  | 74.5  | Online-B             | n/a  | 60.6  | Online-B             | n/a  | 72.8  |
| Online-A             | n/a  | 74.1  | Online-A             | n/a  | 59.4  | Online-A             | n/a  | 73.8  |
| CUNI-DocTransformer  | 2    | 73.0  | CUNI-DocTransformer  | 2    | 57.9  | CUNI-DocTransformer  | 2    | 70.6  |
| CUNI-Transformer     | 3    | 72.4  | CUNI-Transformer     | 3    | 57.6  | CUNI-Transformer     | 3    | 69.8  |
| Online-G             | n/a  | 63.2  | Online-G             | n/a  | 54.3  | Online-G             | n/a  | 61.2  |
| SHOPLINE-PL          | 4    | 70.1  | SHOPLINE-PL          | 4    | 53.5  | SHOPLINE-PL          | 4    | 58.7  |
| Online-Y             | n/a  | 68.8  | Online-Y             | n/a  | 51.0  | Online-Y             | n/a  | 19.3  |
| ALMAnaCH-Inria       | 5    | 58.3  | ALMAnaCH-Inria       | 5    | 36.8  | ALMAnaCH-Inria       | 5    | 58.7  |

### Table 3: Scores for the en→cs translation task: chrF (nrefs:2|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:2|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System               | Rank | chrF  | System               | Rank | chrF  | System               | Rank | chrF  |
|----------------------|------|-------|----------------------|------|-------|----------------------|------|-------|
| Online-B             | n/a  | 71.3  | Online-B             | n/a  | 56.5  | Online-B             | n/a  | 97.9  |
| Lan-Bridge           | 1    | 70.4  | Lan-Bridge           | 1    | 53.4  | Lan-Bridge           | 1    | 97.5  |
| Online-W             | n/a  | 70.4  | Online-W             | n/a  | 53.1  | Online-W             | n/a  | 97.3  |
| Online-A             | 2    | 67.5  | Online-A             | 2    | 49.4  | Online-A             | 2    | 95.3  |
| JDExploreAcademy     | 3    | 67.2  | JDExploreAcademy     | 3    | 47.8  | JDExploreAcademy     | 3    | 94.7  |
| Online-G             | n/a  | 65.3  | Online-G             | n/a  | 46.2  | Online-G             | n/a  | 93.5  |
| CUNI-Bergamot        | 3    | 66.0  | CUNI-Bergamot        | 3    | 45.6  | CUNI-Bergamot        | 3    | 91.7  |
| Online-Y             | n/a  | 64.6  | Online-Y             | n/a  | 44.8  | Online-Y             | n/a  | 86.7  |
| CUNI-Transformer     | 4    | 64.5  | CUNI-Transformer     | 4    | 44.7  | CUNI-Transformer     | 4    | 83.7  |
| AMU                  | 5    | 64.2  | AMU                  | 5    | 38.9  | AMU                  | 5    | 82.3  |

### A Ranking Tables

From Table 2 to 22.

### Table 4: Scores for the cs→uk translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System               | Rank | chrF  | System               | Rank | chrF  | System               | Rank | chrF  |
|----------------------|------|-------|----------------------|------|-------|----------------------|------|-------|
| Online-B             | n/a  | 64.0  | Online-B             | n/a  | 38.3  | AMU                  | 1    | 99.4  |
| Lan-Bridge           | n/a  | 64.0  | Lan-Bridge           | n/a  | 38.3  | Online-B             | n/a  | 94.3  |
| HuaweiTSC            | 1    | 62.6  | HuaweiTSC            | 1    | 36.0  | Lan-Bridge           | n/a  | 94.3  |
| Online-A             | n/a  | 62.4  | Online-A             | n/a  | 35.9  | CharlesTranslator    | n/a  | 94.8  |
| CUNI-JL-JH           | 2    | 61.6  | CUNI-JL-JH           | 2    | 35.0  | HuaweiTSC            | 2    | 90.7  |
| CUNI-Transformer     | 3    | 61.6  | CUNI-Transformer     | 3    | 34.8  | CUNI-JL-JH           | 3    | 90.1  |
| AMU                  | 4    | 61.5  | AMU                  | 4    | 34.7  | Online-A             | n/a  | 88.4  |
| CharlesTranslator    | n/a  | 61.3  | CharlesTranslator    | n/a  | 33.9  | Online-G             | n/a  | 87.9  |
| Online-G             | n/a  | 60.8  | Online-G             | n/a  | 32.5  | CUNI-Transformer     | 4    | 87.4  |
| Online-Y             | n/a  | 59.6  | Online-Y             | n/a  | 32.1  | Online-Y             | n/a  | 78.4  |
| ALMAnaCH-Inria       | 5    | 54.5  | ALMAnaCH-Inria       | 5    | 26.8  | ALMAnaCH-Inria       | 5    | 61.4  |
### Table 5: Scores for the uk→cs translation task:

| System                  | Rank | chrF  |
|-------------------------|------|-------|
| AMU                     | 1    | 60.7  |
| Lan-Bridge              | n/a  | 60.4  |
| Online-B               | n/a  | 60.3  |
| HuaweiTSC              | 2    | 59.6  |
| CharlesDownloader      | n/a  | 59.0  |
| CUNI-Transformer       | 3    | 59.0  |
| CUNI-JL-JH             | 4    | 58.7  |
| Online-A               | n/a  | 57.5  |
| Online-G               | n/a  | 56.3  |
| Online-Y               | n/a  | 55.3  |
| ALMANaCH-Inria         | 5    | 50.7  |

| System                  | Rank | BLEU |
|-------------------------|------|------|
| AMU                     | 1    | 37.0 |
| Lan-Bridge              | n/a  | 36.5 |
| Online-B               | n/a  | 36.4 |
| HuaweiTSC              | 2    | 36.0 |
| CharlesDownloader      | n/a  | 35.9 |
| CUNI-Transformer       | 3    | 35.8 |
| CUNI-JL-JH             | 4    | 35.1 |
| Online-A               | n/a  | 33.3 |
| Online-G               | n/a  | 31.5 |
| Online-Y               | n/a  | 29.6 |
| ALMANaCH-Inria         | 5    | 25.3 |

| System                  | Rank | COMET |
|-------------------------|------|-------|
| AMU                     | 1    | 104.9 |
| Lan-Bridge              | n/a  | 96.5 |
| Online-B               | n/a  | 94.6 |
| HuaweiTSC              | 2    | 91.5 |
| CharlesDownloader      | n/a  | 90.2 |
| CUNI-Transformer       | 3    | 89.1 |
| CUNI-JL-JH             | 4    | 88.5 |
| Online-A               | n/a  | 85.4 |
| Online-G               | n/a  | 84.2 |
| Online-Y               | n/a  | 78.7 |
| ALMANaCH-Inria         | 5    | 62.5 |

### Table 6: Scores for the de→en translation task:

| System                  | Rank | chrF  |
|-------------------------|------|-------|
| Online-A               | n/a  | 66.4  |
| Online-G               | n/a  | 66.3  |
| Lan-Bridge             | n/a  | 66.2  |
| JDExploreAcademy       | 1    | 65.8  |
| Online-B               | n/a  | 65.7  |
| Online-Y               | n/a  | 65.7  |
| PROMT                  | n/a  | 65.7  |
| LT22                   | 2    | 58.4  |

| System                  | Rank | BLEU |
|-------------------------|------|------|
| Online-A               | n/a  | 50.2  |
| Online-G               | n/a  | 49.7  |
| Lan-Bridge             | n/a  | 49.7  |
| JDExploreAcademy       | 1    | 49.1  |
| Online-Y               | n/a  | 49.3  |
| PROMT                  | n/a  | 49.2  |
| LT22                   | 2    | 40.3  |

| System                  | Rank | COMET |
|-------------------------|------|-------|
| Online-A               | n/a  | 58.0  |
| Online-G               | n/a  | 57.0  |
| Lan-Bridge             | n/a  | 56.6  |
| JDExploreAcademy       | 1    | 54.7  |
| Online-Y               | n/a  | 54.5  |
| Online-A               | n/a  | 54.4  |
| PROMT                  | n/a  | 51.8  |
| LT22                   | 2    | 25.6  |

### Table 7: Scores for the en→de translation task:

| System                  | Rank | chrF  |
|-------------------------|------|-------|
| Online-B               | n/a  | 69.7  |
| Online-Y               | n/a  | 69.1  |
| Online-A               | n/a  | 68.8  |
| Online-W               | n/a  | 68.7  |
| Online-G               | n/a  | 68.7  |
| PROMT                  | n/a  | 67.5  |
| OpenNMT               | n/a  | 66.7  |
| LT22                   | 1    | 54.4  |

| System                  | Rank | BLEU |
|-------------------------|------|------|
| Online-B               | n/a  | 52.3  |
| Online-Y               | n/a  | 50.5  |
| Online-A               | n/a  | 50.1  |
| Online-W               | n/a  | 49.6  |
| Online-G               | n/a  | 49.4  |
| PROMT                  | n/a  | 49.0  |
| OpenNMT               | n/a  | 48.9  |
| LT22                   | 1    | 28.3  |

| System                  | Rank | COMET |
|-------------------------|------|-------|
| Online-B               | n/a  | 65.6  |
| Online-Y               | n/a  | 63.3  |
| Online-A               | n/a  | 61.6  |
| Online-W               | n/a  | 60.6  |
| Online-G               | n/a  | 60.2  |
| OpenNMT               | n/a  | 57.3  |
| PROMT                  | n/a  | 55.9  |

### Table 8: Scores for the de→fr translation task:

| System                  | Rank | chrF  |
|-------------------------|------|-------|
| Online-W               | n/a  | 81.2  |
| Online-B               | n/a  | 65.2  |
| eTranslation           | 1    | 68.4  |
| Online-Y               | n/a  | 67.5  |
| Online-A               | n/a  | 67.2  |
| Online-G               | n/a  | 66.5  |
| Lan-Bridge             | n/a  | 65.0  |

| System                  | Rank | BLEU |
|-------------------------|------|------|
| Online-W               | n/a  | 64.8  |
| Online-B               | n/a  | 46.6  |
| eTranslation           | 1    | 46.5  |
| Online-Y               | n/a  | 45.0  |
| Online-A               | n/a  | 44.4  |
| Lan-Bridge             | n/a  | 41.8  |
| Online-G               | n/a  | 41.1  |
| eTranslation           | 1    | 55.5  |

| System                  | Rank | COMET |
|-------------------------|------|-------|
| Online-W               | n/a  | 77.9  |
| Online-B               | n/a  | 63.8  |
| Online-Y               | n/a  | 61.6  |
| Online-A               | n/a  | 59.3  |
| eTranslation           | 1    | 55.5  |

| System                  | Rank | BLEU |
|-------------------------|------|------|
| Online-W               | n/a  | 81.2  |
| Online-B               | n/a  | 65.2  |
| eTranslation           | 1    | 68.4  |
| Online-Y               | n/a  | 67.5  |
| Online-A               | n/a  | 67.2  |
| Online-G               | n/a  | 66.5  |
| Lan-Bridge             | n/a  | 65.0  |

| System                  | Rank | COMET |
|-------------------------|------|-------|
| Online-W               | n/a  | 77.9  |
| Online-B               | n/a  | 63.8  |
| Online-Y               | n/a  | 61.6  |
| Online-A               | n/a  | 59.3  |
| eTranslation           | 1    | 55.5  |

### Table 9: Scores for the fr→de translation task:

| System                  | Rank | chrF  |
|-------------------------|------|-------|
| Online-W               | n/a  | 74.6  |
| Online-Y               | n/a  | 66.8  |
| Online-A               | n/a  | 65.5  |
| Online-G               | n/a  | 64.5  |
| LT22                   | 1    | 54.4  |

| System                  | Rank | BLEU |
|-------------------------|------|------|
| Online-B               | n/a  | 58.4  |
| Online-Y               | n/a  | 46.2  |
| Online-A               | n/a  | 43.6  |
| Online-G               | n/a  | 41.3  |
| LT22                   | 1    | 28.3  |

| System                  | Rank | COMET |
|-------------------------|------|-------|
| Online-W               | n/a  | 70.5  |
| Online-Y               | n/a  | 63.7  |
| Online-A               | n/a  | 57.8  |
| Online-G               | n/a  | 52.3  |
| LT22                   | 1    | 10.5  |
Table 10: Scores for the en→hr translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System         | Rank | chrF  | System         | Rank | chrF  | System         | Rank | chrF  |
|----------------|------|-------|----------------|------|-------|----------------|------|-------|
| Online-B       | n/a  | 58.5  | Lan-Bridge     | n/a  | 58.5  | HuaweiTSC      | 1    | 56.8  |
| Online-A       | n/a  | 56.5  | SRPOL          | 2    | 56.3  |
| NiuTrans       | 3    | 56.3  | HuaweiTSC      | n/a  | 56.5  |
| Online-Y       | n/a  | 54.3  | Online-G       | n/a  | 53.2  |
| Online-G       | n/a  | 53.2  |

Table 11: Scores for the ja→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System         | Rank | chrF  | System         | Rank | chrF  | System         | Rank | chrF  |
|----------------|------|-------|----------------|------|-------|----------------|------|-------|
| NT5            | 1    | 36.8  | JDEExploreAcademy | 2  | 36.1  | LanguageX      | n/a  | 36.1  |
| JDEExploreAcademy | 2  | 36.1  | NT5            | n/a  | 51.3  |
| LanguageX      | n/a  | 36.1  | Online-B       | n/a  | 35.5  |
| DLUT           | 3    | 36.1  | Online-W       | 3    | 35.2  |
| Online-W       | 3    | 35.2  | LanguageX      | n/a  | 27.2  |
| Online-Y       | n/a  | 34.4  | Online-A       | n/a  | 34.1  |
| Lan-Bridge     | n/a  | 34.1  | NAIJ-NICT-TIT | 4    | 48.3  |
| Online-A       | n/a  | 34.1  | AISP-SJTU      | 5    | 48.0  |
| NAIJ-NICT-TIT | 4    | 48.3  | Online-Y       | n/a  | 48.2  |
| AISP-SJTU      | 5    | 48.0  | Online-A       | n/a  | 48.6  |
| Online-Y       | n/a  | 48.2  | NAIJ-NICT-TIT | 5    | 43.9  |
| KXB            | 6    | 43.4  | Online-G       | n/a  | 48.7  |
| Online-G       | n/a  | 48.7  |

Table 12: Scores for the en→ja translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System         | Rank | chrF  | System         | Rank | chrF  | System         | Rank | chrF  |
|----------------|------|-------|----------------|------|-------|----------------|------|-------|
| TartuNLP       | n/a  | 53.5  | TAL-SJTU       | n/a  | 53.2  |
| TAL-SJTU       | n/a  | 53.2  | HuaweiTSC      | n/a  | 48.4  |
| Liv4ever       | n/a  | 46.7  | Liv4ever       | n/a  | 26.0  |
| NiuTrans       | 1    | 35.6  | NiuTrans       | 1    | 13.0  |
| NiuTrans       | 1    | 35.6  | NiuTrans       | 1    | 13.0  |

Table 13: Scores for the liv→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), COMET (wmt20-comet-da).

| System         | Rank | chrF  | System         | Rank | chrF  | System         | Rank | chrF  |
|----------------|------|-------|----------------|------|-------|----------------|------|-------|
| TAL-SJTU       | n/a  | 43.8  | TAL-SJTU       | n/a  | 39.6  |
| Liv4ever       | n/a  | 39.6  | TAL-SJTU       | n/a  | 39.6  |
| Liv4ever       | n/a  | 39.6  | Liv4ever       | n/a  | 39.6  |
| NiuTrans       | n/a  | 39.6  | Liv4ever       | n/a  | 39.6  |
| HuaweiTSC      | n/a  | 39.6  | HuaweiTSC      | n/a  | 39.6  |
| NiuTrans       | 1    | 30.5  | NiuTrans       | 1    | 12.3  |

Table 14: Scores for the liv→vr translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), COMET (wmt20-comet-da).

| System         | Rank | chrF  | System         | Rank | chrF  | System         | Rank | chrF  |
|----------------|------|-------|----------------|------|-------|----------------|------|-------|
| TAL-SJTU       | n/a  | 43.8  | TAL-SJTU       | n/a  | 39.6  |
| Liv4ever       | n/a  | 39.6  | TAL-SJTU       | n/a  | 39.6  |
| Liv4ever       | n/a  | 39.6  | Liv4ever       | n/a  | 39.6  |
| NiuTrans       | n/a  | 39.6  | NiuTrans       | 1    | 12.3  |
| NiuTrans       | 1    | 30.5  | NiuTrans       | 1    | 12.3  |
### Table 15: Scores for the ru→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System              | Rank | chrF | BLEU  |
|---------------------|------|------|-------|
| Online-G            | n/a  | 68.9 | 46.7  |
| Lan-Bridge          | n/a  | 68.5 | 45.2  |
| HuaweiTSC          | 2    | 68.5 | 45.1  |
| Online-B            | n/a  | 68.3 | 45.0  |
| Online-A            | n/a  | 68.3 | 43.9  |
| Online-Y            | n/a  | 68.2 | 43.8  |
| SRPOL               | 3    | 67.2 | 43.6  |
| Online-W            | n/a  | 66.3 | 42.6  |
| ALMAAnaCH-Inria     | 4    | 57.9 | 30.3  |

### Table 16: Scores for the en→ru translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System              | Rank | chrF | BLEU  |
|---------------------|------|------|-------|
| Online-G            | n/a  | 59.5 | 34.9  |
| Lan-Bridge          | n/a  | 59.0 | 33.2  |
| HuaweiTSC          | 1    | 58.4 | 32.8  |
| Online-W            | n/a  | 58.3 | 32.8  |
| Online-Y            | n/a  | 58.3 | 32.6  |
| Online-A            | n/a  | 58.1 | 32.4  |
| SRPOL               | 2    | 56.6 | 32.4  |
| eTranslation        | 4    | 55.8 | 29.8  |

### Table 17: Scores for the sah→ru translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System              | Rank | chrF | BLEU  |
|---------------------|------|------|-------|
| Lan-Bridge          | 1    | 56.6 | 28.3  |
| Online-G            | n/a  | 55.5 | 29.6  |
| Lan-Bridge          | 1    | 28.3 | 7.1   |

### Table 18: Scores for the ru→sah translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System              | Rank | chrF | BLEU  |
|---------------------|------|------|-------|
| Lan-Bridge          | 1    | 67.3 | 44.6  |
| Online-B            | n/a  | 67.2 | 44.4  |
| Online-G            | n/a  | 66.0 | 43.2  |
| Online-A            | n/a  | 65.2 | 42.3  |
| PROMT               | n/a  | 64.7 | 42.1  |
| Online-Y            | n/a  | 64.6 | 41.8  |
| ARC-NKUA           | n/a  | 64.6 | 41.9  |
| HuaweiTSC          | 1    | 63.9 | 41.6  |
| ALMAAnaCH-Inria     | 2    | 55.6 | 30.0  |

### Table 19: Scores for the uk→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System              | Rank | chrF | BLEU  |
|---------------------|------|------|-------|
| Lan-Bridge          | n/a  | 67.3 | 44.6  |
| Online-B            | n/a  | 67.2 | 44.4  |
| Online-G            | n/a  | 66.0 | 43.2  |
| Online-A            | n/a  | 65.2 | 42.3  |
| PROMT               | n/a  | 64.7 | 42.1  |
| Online-Y            | n/a  | 64.6 | 41.8  |
| ARC-NKUA           | n/a  | 64.6 | 41.9  |
| HuaweiTSC          | 1    | 63.9 | 41.6  |
| ALMAAnaCH-Inria     | 2    | 55.6 | 30.0  |
| System     | Rank | chrF |
|------------|------|------|
| Online-B   | n/a  | 59.3 |
| Lan-Bridge | n/a  | 58.8 |
| Online-G   | n/a  | 57.2 |
| Online-A   | n/a  | 56.0 |
| Online-Y   | n/a  | 54.9 |
| eTranslation | 1    | 54.8 |
| HuaweiTSC  | 2    | 54.8 |
| ARC-NKUA   | n/a  | 54.0 |

| System     | Rank | BLEU |
|------------|------|------|
| Online-B   | n/a  | 32.5 |
| Lan-Bridge | n/a  | 29.5 |
| Online-A   | n/a  | 28.0 |
| Online-G   | n/a  | 27.2 |
| Online-Y   | n/a  | 26.9 |
| eTranslation | 1    | 26.5 |
| HuaweiTSC  | 1    | 26.2 |
| ARC-NKUA   | n/a  | 25.2 |

| System | Rank | COMET |
|--------|------|-------|
| Online-B   | n/a  | 73.3 |
| Lan-Bridge | n/a  | 69.9 |
| Online-A   | n/a  | 68.9 |
| Online-G   | n/a  | 64.9 |
| Online-Y   | n/a  | 54.6 |
| HuaweiTSC  | 1    | 54.4 |
| ARC-NKUA   | n/a  | 49.2 |

Table 20: Scores for the en→uk translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System     | Rank | chrF |
|------------|------|------|
| JDExploreAcademy | 1    | 62.4 |
| LanguageX  | n/a  | 61.7 |
| AISP-SITU  | 2    | 60.5 |
| HuaweiTSC  | n/a  | 59.6 |
| Online-B   | n/a  | 59.4 |
| Lan-Bridge | n/a  | 59.3 |
| Online-Y   | n/a  | 58.6 |
| Online-A   | n/a  | 58.4 |
| NiuTrans   | 3    | 57.2 |
| Online-W   | n/a  | 55.6 |

| System     | Rank | BLEU |
|------------|------|------|
| JDExploreAcademy | 1    | 37.9 |
| LanguageX  | n/a  | 36.4 |
| AISP-SITU  | 2    | 34.4 |
| HuaweiTSC  | n/a  | 34.3 |
| Online-B   | n/a  | 34.1 |
| Lan-Bridge | n/a  | 33.3 |
| Online-Y   | n/a  | 31.6 |
| Online-A   | n/a  | 31.1 |
| NiuTrans   | 3    | 30.5 |
| Online-W   | n/a  | 29.0 |

| System     | Rank | COMET |
|------------|------|-------|
| JDExploreAcademy | 1    | 45.7 |
| LanguageX  | n/a  | 45.0 |
| AISP-SITU  | 2    | 41.7 |
| HuaweiTSC  | n/a  | 40.9 |
| Online-B   | n/a  | 35.3 |
| Online-W   | n/a  | 31.7 |
| NiuTrans   | 3    | 31.4 |
| DLUT       | 4    | 30.7 |

Table 21: Scores for the zh→en translation task: chrF (nrefs:2|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:2|case:mixed|eff:no|tok:zh|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).

| System     | Rank | chrF |
|------------|------|------|
| Online-B   | n/a  | 70.3 |
| HuaweiTSC  | n/a  | 61.0 |
| LanguageX  | n/a  | 59.5 |
| Online-A   | n/a  | 58.1 |
| Manifold   | 1    | 57.7 |
| JDExploreAcademy | 2    | 55.8 |
| AISP-SITU  | 3    | 55.4 |
| Lan-Bridge | n/a  | 53.7 |
| DLUT       | 4    | 53.1 |
| Online-G   | n/a  | 52.4 |
| Online-W   | n/a  | 51.4 |

| System     | Rank | BLEU |
|------------|------|------|
| Online-B   | n/a  | 80.0 |
| HuaweiTSC  | n/a  | 73.3 |
| LanguageX  | n/a  | 69.4 |
| Online-A   | n/a  | 68.6 |
| JDExploreAcademy | 2    | 68.4 |
| AISP-SITU  | 3    | 67.6 |
| Lan-Bridge | n/a  | 66.8 |
| DLUT       | 4    | 63.9 |
| Online-G   | n/a  | 63.2 |
| Online-W   | n/a  | 62.0 |

| System     | Rank | COMET |
|------------|------|-------|
| LanguageX  | n/a  | 63.9 |
| Online-B   | n/a  | 61.8 |
| Online-W   | n/a  | 61.0 |
| Manifold   | 2    | 60.2 |
| AISP-SITU  | 3    | 56.6 |
| DLUT       | 4    | 52.2 |
| Online-G   | n/a  | 51.3 |

Table 22: Scores for the en→zh translation task: chrF (nrefs:2|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.1.0), BLEU (nrefs:2|case:mixed|eff:no|tok:zh|smooth:exp|version:2.1.0), COMET (wmt20-comet-da).
Table 23: Statistical significance testing of the chrF score difference for each system pair for the cs→en.

|       | JDE | Lan | CUN | SHO | ALM |
|-------|-----|-----|-----|-----|-----|
| Online-W | 0.0 | 4.3 | 5.1 | 5.2 | 5.6 |
| JDEExploreAcademy | 0.0 | 0.4 | 0.5 | 0.6 | 0.7 |
| Lan-Bridge | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 |
| Online-B | 0.0 | 0.5 | 0.6 | 0.7 | 0.8 |
| Online-A | 0.0 | 0.8 | 0.9 | 1.0 | 1.1 |
| CUNI-DocTransformer | 0.0 | 0.6 | 0.7 | 0.8 | 0.9 |
| CUNI-Transformer | 0.0 | 0.2 | 0.3 | 0.4 | 0.5 |
| Online-G | 0.0 | 0.3 | 0.4 | 0.5 | 0.6 |
| SHOPLINE-PL | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 |
| Online-Y | 0.0 | 0.2 | 0.3 | 0.4 | 0.5 |
| ALMAnaCH-Inria | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 |

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

|       | JDE | Lan | CUN | SHO | ALM |
|-------|-----|-----|-----|-----|-----|
| Online-W | 0.0 | 2.8 | 3.0 | 3.2 | 3.4 |
| JDEExploreAcademy | 0.0 | 0.2 | 0.3 | 0.4 | 0.5 |
| Lan-Bridge | 0.0 | 0.4 | 0.5 | 0.6 | 0.7 |
| Online-B | 0.0 | 0.7 | 0.8 | 0.9 | 1.0 |
| Online-A | 0.0 | 1.0 | 1.1 | 1.2 | 1.3 |
| CUNI-DocTransformer | 0.0 | 1.3 | 1.4 | 1.5 | 1.6 |
| CUNI-Transformer | 0.0 | 1.6 | 1.7 | 1.8 | 1.9 |
| Online-G | 0.0 | 2.0 | 2.1 | 2.2 | 2.3 |
| SHOPLINE-PL | 0.0 | 2.4 | 2.5 | 2.6 | 2.7 |
| Online-Y | 0.0 | 2.8 | 2.9 | 3.0 | 3.1 |
| ALMAnaCH-Inria | 0.0 | 3.2 | 3.3 | 3.4 | 3.5 |

B Statistical Significance Testing Tables

From Table 23 to 85.

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

|       | JDE | Lan | CUN | SHO | ALM |
|-------|-----|-----|-----|-----|-----|
| Online-W | 0.0 | 2.7 | 3.0 | 3.2 | 3.4 |
| JDEExploreAcademy | 0.0 | 0.2 | 0.3 | 0.4 | 0.5 |
| Lan-Bridge | 0.0 | 0.4 | 0.5 | 0.6 | 0.7 |
| Online-B | 0.0 | 0.7 | 0.8 | 0.9 | 1.0 |
| Online-A | 0.0 | 1.0 | 1.1 | 1.2 | 1.3 |
| CUNI-DocTransformer | 0.0 | 1.3 | 1.4 | 1.5 | 1.6 |
| CUNI-Transformer | 0.0 | 1.6 | 1.7 | 1.8 | 1.9 |
| Online-G | 0.0 | 2.0 | 2.1 | 2.2 | 2.3 |
| SHOPLINE-PL | 0.0 | 2.4 | 2.5 | 2.6 | 2.7 |
| Online-Y | 0.0 | 2.8 | 2.9 | 3.0 | 3.1 |
| ALMAnaCH-Inria | 0.0 | 3.2 | 3.3 | 3.4 | 3.5 |
| System                     | chrF Score Difference | System                     | chrF Score Difference | System                     | chrF Score Difference | System                     | chrF Score Difference | System                     | chrF Score Difference | System                     | chrF Score Difference |
|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|
| Online-B                  |                       | Online-W                  |                       | JDExploreAcademy          |                       | CUNI-DocTransformer       |                       | Online-G                  |                       | Online-Y                  |                       |
| 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       |
| Lan-Bridge                | 3.1                   | CUNI-Bergamot             | 0.4                   | Online-A                  | 0.0                   | CUNI-Transformer          | 0.0                   | Online-G                  | 0.0                   | Online-Y                  | 0.0                   |
| 0.0                       |                       | 0.4                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       |

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

| System                     | BLEU Score Difference | System                     | BLEU Score Difference | System                     | BLEU Score Difference | System                     | BLEU Score Difference | System                     | BLEU Score Difference | System                     | BLEU Score Difference |
|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|
| Online-B                  |                       | Online-W                  |                       | JDExploreAcademy          |                       | CUNI-DocTransformer       |                       | Online-G                  |                       | Online-Y                  |                       |
| 0.0                       |                       | 0.4                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       |
| Lan-Bridge                | 3.1                   | CUNI-Bergamot             | 0.4                   | Online-A                  | 0.0                   | CUNI-Transformer          | 0.0                   | Online-G                  | 0.0                   | Online-Y                  | 0.0                   |
| 0.0                       |                       | 0.4                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       |

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

| System                     | COMET Score Difference | System                     | COMET Score Difference | System                     | COMET Score Difference | System                     | COMET Score Difference | System                     | COMET Score Difference | System                     | COMET Score Difference |
|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|
| Online-B                  |                       | Online-W                  |                       | JDExploreAcademy          |                       | CUNI-DocTransformer       |                       | Online-G                  |                       | Online-Y                  |                       |
| 0.0                       |                       | 0.4                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       |
| Lan-Bridge                | 3.1                   | CUNI-Bergamot             | 0.4                   | Online-A                  | 0.0                   | CUNI-Transformer          | 0.0                   | Online-G                  | 0.0                   | Online-Y                  | 0.0                   |
| 0.0                       |                       | 0.4                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       |

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

| System                     | chrF Score Difference | System                     | chrF Score Difference | System                     | chrF Score Difference | System                     | chrF Score Difference | System                     | chrF Score Difference | System                     | chrF Score Difference |
|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|
| Online-B                  |                       | Online-W                  |                       | JDExploreAcademy          |                       | CUNI-DocTransformer       |                       | Online-G                  |                       | Online-Y                  |                       |
| 0.0                       |                       | 0.4                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       |
| Lan-Bridge                | 3.1                   | CUNI-Bergamot             | 0.4                   | Online-A                  | 0.0                   | CUNI-Transformer          | 0.0                   | Online-G                  | 0.0                   | Online-Y                  | 0.0                   |
| 0.0                       |                       | 0.4                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       |

Table 29: Statistical significance testing of the chrF score difference for each system pair for the cs→uk.

| System                     | BLEU Score Difference | System                     | BLEU Score Difference | System                     | BLEU Score Difference | System                     | BLEU Score Difference | System                     | BLEU Score Difference | System                     | BLEU Score Difference |
|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|---------------------------|-----------------------|
| Online-B                  |                       | Online-W                  |                       | JDExploreAcademy          |                       | CUNI-DocTransformer       |                       | Online-G                  |                       | Online-Y                  |                       |
| 0.0                       |                       | 0.4                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       |
| Lan-Bridge                | 3.1                   | CUNI-Bergamot             | 0.4                   | Online-A                  | 0.0                   | CUNI-Transformer          | 0.0                   | Online-G                  | 0.0                   | Online-Y                  | 0.0                   |
| 0.0                       |                       | 0.4                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       | 0.0                       |                       |

Table 30: Statistical significance testing of the BLEU score difference for each system pair for the cs→uk.
### Table 31: Statistical significance testing of the COMET score difference for each system pair for the cs→uk.

|        | AMU | Ond. | Lan. | Cha | Hua. | CUN. | Ond. | Ond. | CUN. | Ond. | ALM. |
|--------|-----|------|------|-----|------|------|------|------|------|------|------|
| AMU    | 0.0 | 5.1  | 7.6  | 8.6 | 8.7  | 9.3  | 11.0 | 11.5 | 12.0 | 21.0 | 38.0 |
| Online-B | 0.0 | 0.75 | 7.1  | 7.4 | 4.7  | 5.9  | 6.4  | 6.9  | 7.9  | 12.9 | 17.9 |
| Lan-Bridge | 0.0 | 1.0  | 1.1  | 1.2 | 1.3  | 1.4  | 1.5  | 1.6  | 1.7  | 1.8  | 1.9  |
| CharlesTranslator | 0.0 | 0.1  | 0.7  | 2.4 | 3.9  | 3.4  | 13.4 | 17.4 | 17.7 | 17.7 | 17.7 |
| HuaweiTSC | 0.0 | 0.6  | 1.7  | 1.6 | 2.3  | 2.8  | 2.3  | 3.3  | 3.7  | 5.9  | 9.6  |
| CUNI-JL-JH | 0.0 | 1.7  | 2.3  | 2.3 | 3.7  | 4.3  | 17.3 | 17.7 | 17.7 | 17.7 | 17.7 |
| Online-G | 0.0 | 0.5  | 1.0  | 10.0 | 17.0 | 20.6 |
| Online-A | 0.0 | 9.0  | 26.0 |
| Online-Y | 0.0 | 17.0 |
| ALMAnaCH-Inria | 0.0 | 0.0 |

### Table 32: Statistical significance testing of the chrF score difference for each system pair for the uk→cs.

|        | AMU | Lan. | Ond. | Cha | Hua. | CUN. | Ond. | Ond. | CUN. | Ond. | ALM. |
|--------|-----|------|------|-----|------|------|------|------|------|------|------|
| AMU    | 0.0 | 0.3  | 0.4  | 0.8 | 1.7  | 1.7  | 2.0  | 2.0  | 2.2  | 2.3  | 10.0 |
| Lan-Bridge | 0.0 | 0.1  | 1.4  | 1.7 | 2.9  | 3.7  | 4.7  | 5.7  | 7.7  | 9.7  |
| Online-B | 0.0 | 0.7  | 1.7  | 1.6 | 2.3  | 3.7  | 3.7  | 4.9  | 4.9  | 9.6  |
| CharlesTranslator | 0.0 | 0.1  | 0.6  | 0.9 | 1.7  | 2.2  | 2.7  | 2.7  | 5.7  | 5.7  |
| HuaweiTSC | 0.0 | 0.3  | 1.7  | 1.7 | 3.7  | 4.7  | 4.7  | 7.7  | 7.7  | 7.7  |
| CUNI-JL-JH | 0.0 | 0.3  | 1.7  | 1.7 | 3.7  | 4.7  | 4.7  | 7.7  | 7.7  | 7.7  |
| Online-A | 0.0 | 0.3  | 1.7  | 1.7 | 3.7  | 4.7  | 4.7  | 7.7  | 7.7  | 7.7  |
| Online-G | 0.0 | 0.3  | 1.7  | 1.7 | 3.7  | 4.7  | 4.7  | 7.7  | 7.7  | 7.7  |
| Online-Y | 0.0 | 0.3  | 1.7  | 1.7 | 3.7  | 4.7  | 4.7  | 7.7  | 7.7  | 7.7  |
| ALMAnaCH-Inria | 0.0 | 0.0 |

### Table 33: Statistical significance testing of the BLEU score difference for each system pair for the uk→cs.

|        | AMU | Ond. | Lan. | Cha | CUN. | Ond. | Ond. | CUN. | Ond. | ALM. |
|--------|-----|------|------|-----|------|------|------|------|------|------|
| AMU    | 0.0 | 8.4  | 10.3 | 13.4 | 14.7 | 15.8 | 16.4 | 16.4 | 19.5 | 20.7 | 26.2 |
| Online-B | 0.0 | 1.2  | 1.3  | 1.4  | 1.5  | 1.6  | 1.7  | 1.7  | 1.9  | 1.9  | 1.9  |
| Lan-Bridge | 0.0 | 4.1  | 4.4  | 5.3  | 5.3  | 5.3  | 5.1  | 5.1  | 9.3  | 10.4 | 15.9 |
| HuaweiTSC | 0.0 | 1.3  | 2.4  | 3.0  | 3.6  | 4.1  | 4.1  | 4.1  | 15.9 | 15.9 | 15.9 |
| CharlesTranslator | 0.0 | 1.3  | 1.4  | 1.5  | 1.5  | 1.5  | 1.5  | 1.5  | 15.9 | 15.9 | 15.9 |
| CUNI-JL-JH | 0.0 | 0.6  | 1.7  | 3.7  | 3.7  | 3.7  | 3.7  | 3.7  | 3.7  | 3.7  | 3.7  |
| CUNI-Transformer | 0.0 | 3.1  | 3.3  | 3.8  | 3.8  | 3.8  | 3.8  | 3.8  | 3.8  | 3.8  | 3.8  |
| Online-A | 0.0 | 1.2  | 1.2  | 1.2  | 1.2  | 1.2  | 1.2  | 1.2  | 1.2  | 1.2  | 1.2  |
| Online-G | 0.0 | 3.1  | 3.1  | 3.1  | 3.1  | 3.1  | 3.1  | 3.1  | 3.1  | 3.1  | 3.1  |
| Online-Y | 0.0 | 16.2 |
| ALMAnaCH-Inria | 0.0 | 0.0 |

### Table 34: Statistical significance testing of the COMET score difference for each system pair for the uk→cs.
Table 35: Statistical significance testing of the chrF score difference for each system pair for the de→en.

|       | Onl. | JDE. | Lan. | PRO. | Ope. | LT2. |
|-------|------|------|------|------|------|------|
| Online-A | 0.0  | 0.0  | 0.2  | 0.6  | 0.7  | 0.7  |
| Lan-Bridge | 0.0  | 0.0  | 0.1  | 0.5  | 0.6  | 0.7  |
| JDEExploreAcademy | 0.0  | 0.1  | 0.1  | 0.2  | 0.3  | 0.3  |
| Online-B | 0.0  | 0.0  | 0.1  | 0.5  | 0.6  | 0.7  |
| Online-Y | 0.0  | 0.0  | 0.0  | 0.2  | 0.3  | 0.3  |
| PROMT | 0.0  | 0.0  | 0.2  | 0.3  | 0.3  | 0.3  |
| Online-W | 0.0  | 0.0  | 0.1  | 0.5  | 0.6  | 0.7  |
| LT22 | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |

Table 36: Statistical significance testing of the BLEU score difference for each system pair for the de→en.

|       | JDE. | Lan. | Onl. | PRO. | Ope. | LT2. |
|-------|------|------|------|------|------|------|
| JDEExploreAcademy | 0.0  | 0.0  | 0.4  | 0.5  | 0.5  | 0.6  |
| Online-B | 0.0  | 0.0  | 0.1  | 0.2  | 0.3  | 0.3  |
| Online-G | 0.0  | 0.0  | 0.0  | 0.1  | 0.2  | 0.2  |
| Online-Y | 0.0  | 0.0  | 0.1  | 0.2  | 0.3  | 0.3  |
| Online-A | 0.0  | 0.0  | 0.1  | 0.2  | 0.3  | 0.3  |
| Online-W | 0.0  | 0.0  | 0.1  | 0.2  | 0.3  | 0.3  |
| PROMT | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| LT22 | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |

Table 37: Statistical significance testing of the COMET score difference for each system pair for the de→en.

|       | Onl. | JDE. | Lan. | PRO. | Ope. |
|-------|------|------|------|------|------|
| Online-B | 0.0  | 0.0  | 0.1  | 0.1  | 0.1  |
| JDEExploreAcademy | 0.0  | 0.0  | 0.1  | 0.1  | 0.1  |
| Lan-Bridge | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| Online-A | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| Online-W | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| Online-G | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| PROMT | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| OpenNMT | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |

Table 38: Statistical significance testing of the chrF score difference for each system pair for the en→de.

|       | Onl. | JDE. | Lan. | PRO. | Ope. |
|-------|------|------|------|------|------|
| Online-B | 0.0  | 0.0  | 0.1  | 0.1  | 0.1  |
| JDEExploreAcademy | 0.0  | 0.0  | 0.1  | 0.1  | 0.1  |
| Online-G | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| Online-Y | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| Online-A | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| Online-W | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| PROMT | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| OpenNMT | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |

Table 39: Statistical significance testing of the BLEU score difference for each system pair for the en→de.
|                      | Online-W | Online-Y | Online-B | Online-A | Online-G | Lan-Bridge | OpenNMT | PROMT |
|----------------------|----------|----------|----------|----------|----------|------------|----------|-------|
| JDExploreAcademy     | 2.3      | 4.4      | 5.3      | 4.4      | 6.8      | 8.3        | 9.7      |       |
| Online-W             | 0.0      | 0.4      | 0.8      | 0.8      | 1.3      | 1.3        | 1.3      |       |
| Online-Y             | 0.0      | 0.5      | 0.9      | 0.9      | 1.2      | 1.2        | 1.2      |       |
| Online-B             | 0.0      | 0.0      | 0.4      | 0.4      | 0.7      | 0.7        | 0.7      |       |
| Online-A             | 0.0      | 0.4      | 1.8      | 1.8      | 2.4      | 2.4        | 2.4      |       |
| Online-G             | 0.0      | 0.0      | 1.4      | 1.4      | 1.7      | 1.7        | 1.7      |       |
| Lan-Bridge           | 0.0      | 1.5      | 3.9      | 3.9      | 5.0      | 5.0        | 5.0      | 6.4   |
| OpenNMT             | 0.0      | 1.4      | 5.0      | 5.0      | 6.0      | 6.0        | 6.0      | 6.4   |
| PROMT                | 0.0      | 1.4      | 5.0      | 5.0      | 6.0      | 6.0        | 6.0      | 6.4   |

**Table 40:** Statistical significance testing of the COMET score difference for each system pair for the en→de.

|                      | Online-B | Online-Y | Online-W | Online-A | Online-G | LT2 |
|----------------------|----------|----------|----------|----------|----------|-----|
| Online-B             | 0.0      | 7.5      | 9.1      | 10.1     | 12.0     | 20.2|
| Online-Y             | 0.0      | 1.0      | 1.0      | 2.8      | 11.1     |     |
| Online-W             | 0.0      | 0.0      | 1.3      | 1.3      | 10.1     |     |
| Online-A             | 0.0      | 0.0      | 1.2      | 1.2      | 3.8      |     |
| Online-G             | 0.0      | 0.0      | 0.7      | 0.7      | 3.8      |     |
| LT2                  | 0.0      | 1.5      | 3.9      | 3.9      | 5.0      | 5.0 |

**Table 41:** Statistical significance testing of the chrF score difference for each system pair for the de→fr.

|                      | Online-B | Online-Y | Online-W | Online-A | Online-G | LT2 |
|----------------------|----------|----------|----------|----------|----------|-----|
| Online-B             | 0.0      | 12.2     | 14.8     | 17.0     | 19.4     | 30.0|
| Online-Y             | 0.0      | 2.6      | 4.8      | 7.2      | 11.1     |     |
| Online-W             | 0.0      | 0.0      | 7.2      | 7.2      | 17.3     |     |
| Online-A             | 0.0      | 0.0      | 7.2      | 7.2      | 17.3     |     |
| Online-G             | 0.0      | 0.0      | 0.7      | 0.7      | 3.8      |     |
| LT2                  | 0.0      | 1.5      | 3.9      | 3.9      | 5.0      | 5.0 |

**Table 42:** Statistical significance testing of the BLEU score difference for each system pair for the de→fr.

|                      | Online-W | Online-Y | Online-B | Online-A | Online-G | Lan-Bridge | OpenNMT | PROMT |
|----------------------|----------|----------|----------|----------|----------|------------|----------|-------|
| Online-W             | 0.0      | 12.5     | 12.8     | 13.7     | 14.0     | 15.2       | 16.2     |       |
| Online-Y             | 0.0      | 0.3      | 1.2      | 1.5      | 2.7      | 3.7        | 3.7      |       |
| Online-B             | 0.0      | 0.0      | 0.3      | 0.3      | 1.5      | 1.5        | 1.5      |       |
| Online-A             | 0.0      | 0.0      | 1.2      | 1.2      | 3.8      | 3.8        | 3.8      |       |
| Online-G             | 0.0      | 0.0      | 0.3      | 0.3      | 1.5      | 1.5        | 1.5      |       |
| Lan-Bridge           | 0.0      | 0.0      | 1.5      | 1.5      | 3.8      | 3.8        | 3.8      | 6.4   |
| OpenNMT             | 0.0      | 1.4      | 5.0      | 5.0      | 6.0      | 6.0        | 6.0      | 6.4   |
| PROMT                | 0.0      | 1.4      | 5.0      | 5.0      | 6.0      | 6.0        | 6.0      | 6.4   |

**Table 43:** Statistical significance testing of the COMET score difference for each system pair for the de→fr.

|                      | Online-W | Online-Y | Online-B | Online-A | Online-G | Lan-Bridge | OpenNMT | PROMT |
|----------------------|----------|----------|----------|----------|----------|------------|----------|-------|
| Online-W             | 0.0      | 18.2     | 18.5     | 19.8     | 20.4     | 23.0       | 23.8     |       |
| Online-Y             | 0.0      | 0.0      | 1.6      | 1.6      | 3.3      | 3.3        | 3.3      |       |
| Online-B             | 0.0      | 1.4      | 3.7      | 3.7      | 4.7      | 4.7        | 4.7      |       |
| Online-A             | 0.0      | 0.6      | 1.2      | 1.2      | 3.7      | 3.7        | 3.7      |       |
| Online-G             | 0.0      | 0.0      | 1.4      | 1.4      | 3.7      | 3.7        | 3.7      |       |
| Lan-Bridge           | 0.0      | 0.0      | 1.4      | 1.4      | 3.7      | 3.7        | 3.7      | 6.4   |
| OpenNMT             | 0.0      | 1.4      | 5.0      | 5.0      | 6.0      | 6.0        | 6.0      | 6.4   |
| PROMT                | 0.0      | 1.4      | 5.0      | 5.0      | 6.0      | 6.0        | 6.0      | 6.4   |

**Table 44:** Statistical significance testing of the chrF score difference for each system pair for the fr→de.

|                      | Online-W | Online-Y | Online-B | Online-A | Online-G | Lan-Bridge | OpenNMT | PROMT |
|----------------------|----------|----------|----------|----------|----------|------------|----------|-------|
| Online-W             | 0.0      | 18.2     | 18.5     | 19.8     | 20.4     | 23.0       | 23.8     |       |
| Online-Y             | 0.0      | 0.0      | 1.6      | 1.6      | 3.3      | 3.3        | 3.3      |       |
| Online-B             | 0.0      | 1.4      | 3.7      | 3.7      | 4.7      | 4.7        | 4.7      |       |
| Online-A             | 0.0      | 0.6      | 1.2      | 1.2      | 3.7      | 3.7        | 3.7      |       |
| Online-G             | 0.0      | 0.0      | 1.4      | 1.4      | 3.7      | 3.7        | 3.7      |       |
| Lan-Bridge           | 0.0      | 0.0      | 1.4      | 1.4      | 3.7      | 3.7        | 3.7      | 6.4   |
| OpenNMT             | 0.0      | 1.4      | 5.0      | 5.0      | 6.0      | 6.0        | 6.0      | 6.4   |
| PROMT                | 0.0      | 1.4      | 5.0      | 5.0      | 6.0      | 6.0        | 6.0      | 6.4   |

**Table 45:** Statistical significance testing of the BLEU score difference for each system pair for the fr→de.
Table 46: Statistical significance testing of the COMET score difference for each system pair for the fr→de.

|         | Online-W | Online-B | Online-Y | Online-A | eTranslation | Lan-Bridge | NiuTrans | Online-G |
|---------|----------|----------|----------|----------|--------------|------------|----------|----------|
| 0.0     | 14.1     | 0.0      | 0.0      | 0.0      | 0.0          | 3.8        | 0.0      | 0.0      |
| 16.3    | 2.2      | 0.0      | 0.0      | 0.0      | 0.0          | 3.8        | 0.0      | 0.0      |
| 18.6    | 2.4      | 0.0      | 0.0      | 0.0      | 0.0          | 3.8        | 0.0      | 0.0      |
| 22.4    | 2.2      | 0.0      | 0.0      | 0.0      | 0.0          | 3.8        | 0.0      | 0.0      |
| 26.7    | 4.2      | 0.0      | 0.0      | 0.0      | 0.0          | 3.8        | 0.0      | 0.0      |
| 29.6    | 5.3      | 0.0      | 0.0      | 0.0      | 0.0          | 3.8        | 0.0      | 0.0      |

Table 47: Statistical significance testing of the chrF score difference for each system pair for the en→hr.

|         | Online-B | Lan-Bridge | NiuTrans | Online-A | SRPOL | Online-Y | Online-G |
|---------|----------|------------|----------|----------|-------|----------|----------|
| 0.0     | 0.0      | 0.1        | 0.0      | 0.0      | 0.0   | 0.0      | 0.0      |
| 0.1     | 1.7      | 2.0        | 2.2      | 2.2      | 4.2   | 3.1      | 3.7      |
| 2.0     | 3.7      | 7.7        | 7.7      | 7.7      | 7.7   | 7.7      | 7.7      |
| 2.2     | 7.7      | 7.7        | 7.7      | 7.7      | 7.7   | 7.7      | 7.7      |
| 2.2     | 7.7      | 7.7        | 7.7      | 7.7      | 7.7   | 7.7      | 7.7      |
| 4.2     | 3.1      | 3.1        | 3.1      | 3.1      | 3.1   | 3.1      | 3.1      |
| 5.3     | 3.7      | 3.7        | 3.7      | 3.7      | 3.7   | 3.7      | 3.7      |

Table 48: Statistical significance testing of the BLEU score difference for each system pair for the en→hr.

|         | Online-B | Lan-Bridge | SRPOL | HuaweiTSC | NiuTrans | Online-A | Online-Y | Online-G |
|---------|----------|------------|-------|-----------|----------|----------|----------|----------|
| 0.0     | 0.0      | 0.8        | 0.0   | 0.0       | 0.0      | 0.0      | 0.0      | 0.0      |
| 0.8     | 10.9     | 10.4       | 10.9  | 15.3      | 15.3     | 15.3     | 15.3     | 15.3     |
| 10.9    | 10.4     | 10.9       | 10.9  | 15.3      | 15.3     | 15.3     | 15.3     | 15.3     |
| 12.7    | 14.8     | 16.1       | 23.6  | 23.6      | 23.6     | 23.6     | 23.6     | 23.6     |
| 14.8    | 16.1     | 23.6       | 23.6  | 23.6      | 23.6     | 23.6     | 23.6     | 23.6     |
| 16.1    | 23.6     | 23.6       | 23.6  | 23.6      | 23.6     | 23.6     | 23.6     | 23.6     |
| 23.6    | 23.6     | 23.6       | 23.6  | 23.6      | 23.6     | 23.6     | 23.6     | 23.6     |

Table 49: Statistical significance testing of the COMET score difference for each system pair for the en→hr.

|         | Online-W | NTS | JDE | DLU | Online-W | Online-B | LanguageX | Lan-Bridge | Online-A | NAI | Online-A | AIS | Online-KYB | AIS |
|---------|----------|-----|-----|-----|----------|----------|-----------|------------|----------|------|----------|-----|-------------|-----|
| 0.0     | 1.5      | 1.8 | 1.9 | 1.9 | 2.6      | 2.9      | 2.9        | 3.3        | 3.5      | 3.4 | 3.3       | 3.5 | 3.0         | 3.0 |
| 0.4     | 0.0      | 0.2 | 0.4 | 0.4 | 0.0      | 0.0      | 0.0        | 0.0        | 0.0      | 0.0 | 0.0       | 0.0 | 0.0         | 0.0 |
| 0.4     | 0.0      | 0.1 | 0.1 | 0.1 | 0.0      | 0.0      | 0.0        | 0.0        | 0.0      | 0.0 | 0.0       | 0.0 | 0.0         | 0.0 |
| 0.4     | 0.0      | 0.1 | 0.1 | 0.1 | 0.0      | 0.0      | 0.0        | 0.0        | 0.0      | 0.0 | 0.0       | 0.0 | 0.0         | 0.0 |
| 0.4     | 0.0      | 0.1 | 0.1 | 0.1 | 0.0      | 0.0      | 0.0        | 0.0        | 0.0      | 0.0 | 0.0       | 0.0 | 0.0         | 0.0 |

Table 50: Statistical significance testing of the chrF score difference for each system pair for the ja→en.
Table 51: Statistical significance testing of the BLEU score difference for each system pair for the ja→en.

| NT5 | Onl | JDE | DLU | Onl | Onl | Lan | NAI | Lan | AIS | Onl | Onl | KYB | AIS |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     |     |     |     |     |     |     |     |     |     |     |     |     |     |
| Online-W | 0.0 | 0.9 | 1.5 | 2.4 | 4.8 | 8.7 | 9.1 | 9.2 | 9.7 | 10.1 | 11.9 | 19.8 | 24.2 | 194.8 |
| JDExploreAcademy | 0.0 | 0.6 | 1.5 | 5.9 | 7.6 | 7.8 | 8.3 | 8.7 | 9.2 | 10.0 | 11.9 | 19.7 | 24.2 | 193.6 |
| Online-B | 0.0 | 0.6 | 0.9 | 5.9 | 7.6 | 7.8 | 8.3 | 8.7 | 9.2 | 10.0 | 11.9 | 19.7 | 24.2 | 193.6 |
| NAI-S-NCT-TIT | 0.0 | 0.4 | 0.9 | 5.7 | 7.6 | 7.8 | 8.3 | 8.7 | 9.2 | 10.0 | 11.9 | 19.7 | 24.2 | 193.6 |
| LanguageX | 0.0 | 0.4 | 0.9 | 5.7 | 7.6 | 7.8 | 8.3 | 8.7 | 9.2 | 10.0 | 11.9 | 19.7 | 24.2 | 193.6 |
| Online-Y | 0.0 | 1.3 | 2.2 | 5.7 | 7.6 | 7.8 | 8.3 | 8.7 | 9.2 | 10.0 | 11.9 | 19.7 | 24.2 | 193.6 |
| Online-G | 0.0 | 1.3 | 2.2 | 5.7 | 7.6 | 7.8 | 8.3 | 8.7 | 9.2 | 10.0 | 11.9 | 19.7 | 24.2 | 193.6 |
| KYB | 0.0 | 1.3 | 2.2 | 5.7 | 7.6 | 7.8 | 8.3 | 8.7 | 9.2 | 10.0 | 11.9 | 19.7 | 24.2 | 193.6 |
| AIST | 0.0 | 1.3 | 2.2 | 5.7 | 7.6 | 7.8 | 8.3 | 8.7 | 9.2 | 10.0 | 11.9 | 19.7 | 24.2 | 193.6 |

Table 52: Statistical significance testing of the COMET score difference for each system pair for the ja→en.

| NT5 | JDE | Lan | DLU | Onl | Lan | NAI | Lan | AIS | Onl | KYB | AIS |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     |     |     |     |     |     |     |     |     |     |     |     |
| NT5 | 0.0 | 0.7 | 0.8 | 0.8 | 1.3 | 1.6 | 2.4 | 2.7 | 2.8 | 3.0 | 3.1 | 8.3 | 8.9 |
| JDExploreAcademy | 0.0 | 0.1 | 0.1 | 0.6 | 0.9 | 1.2 | 2.0 | 2.7 | 3.0 | 3.2 | 3.3 | 8.3 | 8.9 |
| LanguageX | 0.0 | 0.1 | 0.1 | 0.6 | 0.9 | 1.2 | 2.0 | 2.7 | 3.0 | 3.2 | 3.3 | 8.3 | 8.9 |
| DLUT | 0.0 | 0.1 | 0.1 | 0.6 | 0.9 | 1.2 | 2.0 | 2.7 | 3.0 | 3.2 | 3.3 | 8.3 | 8.9 |
| Online-B | 0.0 | 0.1 | 0.1 | 0.6 | 0.9 | 1.2 | 2.0 | 2.7 | 3.0 | 3.2 | 3.3 | 8.3 | 8.9 |
| Online-Y | 0.0 | 0.1 | 0.1 | 0.6 | 0.9 | 1.2 | 2.0 | 2.7 | 3.0 | 3.2 | 3.3 | 8.3 | 8.9 |
| Online-G | 0.0 | 0.1 | 0.1 | 0.6 | 0.9 | 1.2 | 2.0 | 2.7 | 3.0 | 3.2 | 3.3 | 8.3 | 8.9 |
| KYB | 0.0 | 0.1 | 0.1 | 0.6 | 0.9 | 1.2 | 2.0 | 2.7 | 3.0 | 3.2 | 3.3 | 8.3 | 8.9 |
| AIST | 0.0 | 0.1 | 0.1 | 0.6 | 0.9 | 1.2 | 2.0 | 2.7 | 3.0 | 3.2 | 3.3 | 8.3 | 8.9 |

Table 53: Statistical significance testing of the chrF score difference for each system pair for the en→ja.

| NT5 | Lan | DLU | Onl | Lan | AIS | Lan | NAI | Lan | AIS | NT5 | Lan | AIS |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     |     |     |     |     |     |     |     |     |     |     |     |     |
| NT5 | 0.0 | 0.4 | 0.5 | 0.5 | 0.6 | 1.1 | 2.4 | 2.6 | 2.8 | 3.0 | 3.1 | 8.3 | 8.9 |
| LanguageX | 0.0 | 0.4 | 0.5 | 0.6 | 1.1 | 2.4 | 2.6 | 2.8 | 3.0 | 3.1 | 8.3 | 8.9 |
| DLUT | 0.0 | 0.4 | 0.5 | 0.6 | 1.1 | 2.4 | 2.6 | 2.8 | 3.0 | 3.1 | 8.3 | 8.9 |
| Online-W | 0.0 | 0.4 | 0.5 | 0.6 | 1.1 | 2.4 | 2.6 | 2.8 | 3.0 | 3.1 | 8.3 | 8.9 |
| JDExploreAcademy | 0.0 | 0.4 | 0.5 | 0.6 | 1.1 | 2.4 | 2.6 | 2.8 | 3.0 | 3.1 | 8.3 | 8.9 |
| Online-B | 0.0 | 0.4 | 0.5 | 0.6 | 1.1 | 2.4 | 2.6 | 2.8 | 3.0 | 3.1 | 8.3 | 8.9 |
| Online-Y | 0.0 | 0.4 | 0.5 | 0.6 | 1.1 | 2.4 | 2.6 | 2.8 | 3.0 | 3.1 | 8.3 | 8.9 |
| Online-G | 0.0 | 0.4 | 0.5 | 0.6 | 1.1 | 2.4 | 2.6 | 2.8 | 3.0 | 3.1 | 8.3 | 8.9 |
| KYB | 0.0 | 0.4 | 0.5 | 0.6 | 1.1 | 2.4 | 2.6 | 2.8 | 3.0 | 3.1 | 8.3 | 8.9 |

Table 54: Statistical significance testing of the BLEU score difference for each system pair for the en→ja.
Table 55: Statistical significance testing of the COMET score difference for each system pair for the en→ja.

| System A         | System B         | TAR | TAL  | Hua  | Liv  | Niu  |
|------------------|------------------|-----|------|------|------|------|
| JDExploreAcademy | NTS              | 0.0 | 1.1  |      |      |      |
| NTS              | LanguageX        | 0.0 | 1.1  | 1.6  | 1.6  | 2.1  |
| LanguageX        | Online-B         | 0.0 | 0.3  | 1.0  | 1.0  | 1.0  |
| Online-B         | DLUT             | 0.0 | 0.7  | 0.7  | 0.7  | 0.7  |
| DLUT             | Online-W         | 0.0 | 0.7  | 0.7  | 0.7  | 0.7  |
| Online-W         | Lan-Bridge       | 0.0 | 0.3  | 0.3  | 0.3  | 0.3  |
| Lan-Bridge       | Online-A         | 0.0 | 0.3  | 0.3  | 0.3  | 0.3  |
| Online-A         | NAIST-NICT-TIT   | 0.0 | 0.3  | 0.3  | 0.3  | 0.3  |
| NAIST-NICT-TIT   | AISP-SJTU        | 0.0 | 0.3  | 0.3  | 0.3  | 0.3  |
| AISP-SJTU        | KYB              | 0.0 | 0.3  | 0.3  | 0.3  | 0.3  |
| KYB              | Online-G         | 0.0 | 0.3  | 0.3  | 0.3  | 0.3  |

Table 56: Statistical significance testing of the chrF score difference for each system pair for the liv→en.

| System A         | System B         | TAR | TAL  | Hua  | Liv  | Niu  |
|------------------|------------------|-----|------|------|------|------|
| TartuNLP         | TAL-SJTU         | 0.0 | 0.4  | 0.4  | 0.4  | 0.4  |
| TAL-SJTU         | HuaweiTSC        | 0.0 | 0.4  | 0.4  | 0.4  | 0.4  |
| HuaweiTSC        | Liv4ever         | 0.0 | 0.4  | 0.4  | 0.4  | 0.4  |
| Liv4ever         | Niutrans         | 0.0 | 0.4  | 0.4  | 0.4  | 0.4  |

Table 57: Statistical significance testing of the BLEU score difference for each system pair for the liv→en.

| System A         | System B         | TAR | TAL  | Hua  | Liv  | Niu  |
|------------------|------------------|-----|------|------|------|------|
| TartuNLP         | TAL-SJTU         | 0.0 | 0.5  | 0.5  | 0.5  | 0.5  |
| TAL-SJTU         | HuaweiTSC        | 0.0 | 0.5  | 0.5  | 0.5  | 0.5  |
| HuaweiTSC        | Liv4ever         | 0.0 | 0.5  | 0.5  | 0.5  | 0.5  |
| Liv4ever         | Niutrans         | 0.0 | 0.5  | 0.5  | 0.5  | 0.5  |

Table 58: Statistical significance testing of the COMET score difference for each system pair for the liv→en.

| System A         | System B         | TAR | TAL  | Hua  | Liv  | Niu  |
|------------------|------------------|-----|------|------|------|------|
| TartuNLP         | TAL-SJTU         | 0.0 | 2.0  | 2.0  | 2.0  | 2.0  |
| TAL-SJTU         | HuaweiTSC        | 0.0 | 2.0  | 2.0  | 2.0  | 2.0  |
| HuaweiTSC        | Liv4ever         | 0.0 | 2.0  | 2.0  | 2.0  | 2.0  |
| Liv4ever         | Niutrans         | 0.0 | 2.0  | 2.0  | 2.0  | 2.0  |

Table 59: Statistical significance testing of the chrF score difference for each system pair for the en→liv.

| System A         | System B         | TAR | TAL  | Hua  | Liv  | Niu  |
|------------------|------------------|-----|------|------|------|------|
| TartuNLP         | TAL-SJTU         | 0.0 | 2.0  | 2.0  | 2.0  | 2.0  |
| TAL-SJTU         | HuaweiTSC        | 0.0 | 2.0  | 2.0  | 2.0  | 2.0  |
| HuaweiTSC        | Liv4ever         | 0.0 | 2.0  | 2.0  | 2.0  | 2.0  |
| Liv4ever         | Niutrans         | 0.0 | 2.0  | 2.0  | 2.0  | 2.0  |

Table 60: Statistical significance testing of the BLEU score difference for each system pair for the en→liv.
### Table 61: Statistical significance testing of the COMET score difference for each system pair for the en→liv.

|       | TAL  | TAT  | Hua. | Liv. | Niu. |
|-------|------|------|------|------|------|
| TAL-SJTU | 0.0  | 2.3  | 9.9  | 9.3  | 52.4 |
| TurnuSPL | 0.0  | 2.0  | 2.6  | 10.1 |
| HuaweiTSC | 0.0  | 0.6  | 3.0  | 4.9  |
| Livlever | 0.0  | 2.5  | 6.7  | 8.0  |
| NiuTrans | 0.0  | 0.0  | 0.0  | 0.0  |

### Table 62: Statistical significance testing of the chrF score difference for each system pair for the ru→en.

|       | Onl. | Lan. | JDE  | Onl. | Hua. | Onl. | Onl. | SRP | Onl. | ALM. |
|-------|------|------|------|------|------|------|------|-----|------|------|
| Online-G | 0.0  | 0.3  | 1.5  | 2.0  | 2.5  | 2.8  | 2.3  | 2.7 | 3.7  | 5.7  |
| Lan-Bridge | 0.0  | 0.1  | 0.2  | 0.3  | 0.3  | 1.3  | 2.7  | 3.0 | 4.0  | 12.3 |
| HuaweiTSC | 0.0  | 0.1  | 0.6  | 1.2  | 2.2  | 3.4  | 5.0  | 5.4 | 6.8  | 18.8 |
| Online-A | 0.0  | 0.1  | 0.2  | 0.3  | 0.3  | 1.3  | 2.7  | 3.0 | 4.0  | 12.3 |
| Online-Y | 0.0  | 0.1  | 0.2  | 0.3  | 0.3  | 1.3  | 2.7  | 3.0 | 4.0  | 12.3 |
| SRPOL   | 0.0  | 0.2  | 0.3  | 0.3  | 0.3  | 1.3  | 2.7  | 3.0 | 4.0  | 12.3 |
| Online-W | 0.0  | 0.3  | 0.6  | 0.9  | 2.2  | 3.4  | 5.0  | 5.4 | 6.8  | 18.8 |
| ALMAnaCH-Inria | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0 | 0.0  | 0.0  |

### Table 63: Statistical significance testing of the BLEU score difference for each system pair for the ru→en.

|       | Onl. | Lan. | JDE  | Onl. | Onl. | SRP | ALM. |
|-------|------|------|------|------|------|-----|------|
| Online-G | 0.0  | 0.3  | 0.8  | 1.3  | 1.8  | 2.3 | 3.0  |
| Lan-Bridge | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  |
| HuaweiTSC | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  |
| Online-A | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  |
| Online-Y | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  |
| SRPOL   | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  |
| Online-W | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  |
| ALMAnaCH-Inria | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0 | 0.0  |

### Table 64: Statistical significance testing of the COMET score difference for each system pair for the ru→en.

|       | Onl. | Lan. | JDE  | Onl. | Onl. | SRP | PRO | Hua. | eTr. |
|-------|------|------|------|------|------|-----|-----|-----|-----|
| Online-B | 0.0  | 0.3  | 0.8  | 1.3  | 1.8  | 2.3 | 3.0  | 3.6 | 3.9  |
| Lan-Bridge | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |
| HuaweiTSC | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |
| Online-A | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |
| SRPOL   | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |
| PROMT   | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |
| eTranslation | 0.0  | 0.3  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |

### Table 65: Statistical significance testing of the chrF score difference for each system pair for the en→ru.

|       | Onl. | Lan. | JDE  | Onl. | Onl. | SRP | PRO | Hua. | eTr. |
|-------|------|------|------|------|------|-----|-----|-----|-----|
| Online-B | 0.0  | 0.3  | 0.8  | 1.3  | 1.8  | 2.3 | 3.0  | 3.6 | 3.9  |
| Lan-Bridge | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |
| HuaweiTSC | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |
| Online-A | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |
| SRPOL   | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |
| PROMT   | 0.0  | 0.1  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |
| eTranslation | 0.0  | 0.3  | 0.6  | 0.9  | 1.2  | 1.8 | 2.3  | 2.7 | 3.0  |
|        | Onl. | Otl. | Otl. | JDE. | Lan. | Otl. | Hua. | PRO. | SRP. | eTr. |
|--------|------|------|------|------|------|------|------|------|------|------|
| Online-B | 0.0  | 1.7  | 1.9  | 2.1  | 2.2  | 2.3  | 2.5  | 4.2  | 4.4  | 4.3  | 5.1  |
| Online-Y | 0.0  | 0.2  | 0.4  | 0.5  | 0.6  | 0.8  | 2.3  | 2.7  | 2.9  | 3.7  | 3.9  |
| Online-A | 0.0  | 0.3  | 0.4  | 0.5  | 0.7  | 1.1  | 1.2  | 1.3  | 1.4  | 1.5  | 1.6  |
| Online-G | 0.0  | 0.1  | 0.2  | 0.4  | 0.7  | 1.1  | 2.7  | 2.9  | 3.1  | 3.5  | 3.7  |
| JDEExploreAcademy | 0.0  | 0.1  | 0.3  | 0.7  | 1.1  | 1.5  | 2.9  | 3.3  | 3.7  | 3.9  | 4.0  |
| Lan-Bridge | 0.0  | 0.2  | 0.4  | 0.7  | 1.1  | 1.5  | 2.9  | 3.3  | 3.7  | 3.9  | 4.0  |
| Online-W | 0.0  | 1.6  | 1.8  | 2.0  | 2.7  | 3.1  | 3.7  | 4.0  | 4.3  | 4.5  | 4.7  |
| HuaweiTSC | 0.0  | 0.2  | 0.4  | 0.7  | 1.1  | 1.5  | 2.9  | 3.3  | 3.7  | 3.9  | 4.0  |
| PROMT | 0.0  | 0.2  | 0.4  | 0.7  | 1.1  | 1.5  | 2.9  | 3.3  | 3.7  | 3.9  | 4.0  |
| SRPOL | 0.0  | 0.2  | 0.4  | 0.7  | 1.1  | 1.5  | 2.9  | 3.3  | 3.7  | 3.9  | 4.0  |
| eTranslation | 0.0  | 0.2  | 0.4  | 0.7  | 1.1  | 1.5  | 2.9  | 3.3  | 3.7  | 3.9  | 4.0  |

Table 66: Statistical significance testing of the BLEU score difference for each system pair for the en→ru.

|        | Onl. | Otl. | Otl. | JDE. | Lan. | Otl. | Hua. | PRO. | SRP. | eTr. |
|--------|------|------|------|------|------|------|------|------|------|------|
| Online-W | 0.0  | 1.9  | 2.2  | 2.3  | 2.5  | 2.7  | 2.9  | 3.3  | 3.5  | 3.7  | 3.9  |
| Online-G | 0.0  | 0.3  | 0.5  | 0.7  | 0.9  | 1.1  | 1.3  | 1.5  | 1.7  | 1.9  | 2.1  |
| Online-B | 0.0  | 1.7  | 2.0  | 2.3  | 2.5  | 2.8  | 3.1  | 3.4  | 3.6  | 3.9  | 4.1  |
| JDEExploreAcademy | 0.0  | 0.1  | 0.3  | 0.5  | 0.7  | 0.9  | 1.1  | 1.3  | 1.5  | 1.7  | 1.9  |
| Lan-Bridge | 0.0  | 0.2  | 0.4  | 0.6  | 0.8  | 1.0  | 1.2  | 1.4  | 1.6  | 1.8  | 2.0  |
| Online-A | 0.0  | 1.6  | 1.8  | 2.0  | 2.2  | 2.4  | 2.6  | 2.8  | 3.0  | 3.2  | 3.4  |
| PROMT | 0.0  | 0.2  | 0.4  | 0.6  | 0.8  | 1.0  | 1.2  | 1.4  | 1.6  | 1.8  | 2.0  |
| SRPOL | 0.0  | 0.2  | 0.4  | 0.6  | 0.8  | 1.0  | 1.2  | 1.4  | 1.6  | 1.8  | 2.0  |
| HuaweiTSC | 0.0  | 0.2  | 0.4  | 0.6  | 0.8  | 1.0  | 1.2  | 1.4  | 1.6  | 1.8  | 2.0  |
| eTranslation | 0.0  | 0.2  | 0.4  | 0.6  | 0.8  | 1.0  | 1.2  | 1.4  | 1.6  | 1.8  | 2.0  |

Table 67: Statistical significance testing of the COMET score difference for each system pair for the en→ru.

|        | Lan. | Otl. |
|--------|------|------|
| Lan-Bridge | 0.0  | 9.7  |
| Online-G | 0.0  | 0.0  |

Table 68: Statistical significance testing of the chrF score difference for each system pair for the ru→sah.

|        | Lan. | Otl. |
|--------|------|------|
| Lan-Bridge | 0.0  | 0.6  |
| Online-G | 0.0  | 0.0  |

Table 69: Statistical significance testing of the BLEU score difference for each system pair for the ru→sah.

|        | Otl. | Lan. |
|--------|------|------|
| Online-G | 0.0  | 25.9 |
| Lan-Bridge | 0.0  | 0.0  |

Table 70: Statistical significance testing of the COMET score difference for each system pair for the ru→sah.

|        | Otl. | Lan. |
|--------|------|------|
| Online-G | 0.0  | 27.2 |
| Lan-Bridge | 0.0  | 0.0  |

Table 71: Statistical significance testing of the chrF score difference for each system pair for the sah→ru.

|        | Otl. | Lan. |
|--------|------|------|
| Online-G | 0.0  | 22.6 |
| Lan-Bridge | 0.0  | 0.0  |

Table 72: Statistical significance testing of the BLEU score difference for each system pair for the sah→ru.
Table 73: Statistical significance testing of the COMET score difference for each system pair for the sah$\rightarrow$ru.

|        | Lan. | Onl. | Onl. | Onl. | PRO. | Onl. | ARC. | Hua. | ALM. |
|--------|------|------|------|------|------|------|------|------|------|
| Lan-Bridge | 0.0  | 0.1  | 1.3  | 2.1  | 2.6  | 2.8  | 2.7  | 3.4  | 11.3 |
| Online-B  | 0.0  | 0.7  | 1.2  | 2.0  | 2.3  | 2.7  | 2.6  | 3.3  | 11.6 |
| Online-G  | 0.0  | 0.7  | 0.5  | 0.7  | 0.7  | 0.5  | 0.7  | 0.8  | 0.7  |
| Online-A  | 0.0  | 0.1  | -0.1 | 0.1  | 0.2  | 0.3  | 0.3  | 0.4  | 0.4  |
| PROMT     | 0.0  | 0.2  | 0.0  | 0.2  | 0.3  | 0.3  | 0.3  | 0.4  | 0.4  |
| Online-Y  | 0.0  | 0.2  | 0.0  | 0.2  | 0.3  | 0.3  | 0.3  | 0.4  | 0.4  |
| ARC-NKUA  | 0.0  | 0.2  | 0.0  | 0.2  | 0.3  | 0.3  | 0.3  | 0.4  | 0.4  |
| HuaweiTSC | 0.0  | 0.2  | 0.0  | 0.2  | 0.3  | 0.3  | 0.3  | 0.4  | 0.4  |
| ALMAnaCH-Inria | 0.0  | 0.2  | 0.0  | 0.2  | 0.3  | 0.3  | 0.3  | 0.4  | 0.4  |

Table 74: Statistical significance testing of the chrF score difference for each system pair for the uk$\rightarrow$en.

|        | Lan. | Onl. | Onl. | Onl. | PRO. | Onl. | ARC. | Hua. | ALM. |
|--------|------|------|------|------|------|------|------|------|------|
| Lan-Bridge | 0.0  | 0.1  | 1.4  | 2.3  | 2.5  | 2.5  | 2.6  | 3.4  | 11.7 |
| Online-B  | 0.0  | 0.7  | 2.2  | 2.2  | 2.4  | 2.6  | 2.6  | 2.8  | 14.6 |
| Online-G  | 0.0  | 0.1  | 0.2  | 0.3  | 0.5  | 0.6  | 0.6  | 0.7  | 12.5 |
| Online-A  | 0.0  | 0.2  | 0.1  | 0.2  | 0.3  | 0.4  | 0.4  | 0.5  | 12.5 |
| PROMT     | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 12.5 |
| Online-Y  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 12.5 |
| ARC-NKUA  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 12.5 |
| HuaweiTSC | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 12.5 |
| ALMAnaCH-Inria | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 12.5 |

Table 75: Statistical significance testing of the BLEU score difference for each system pair for the uk$\rightarrow$en.

|        | Onl. | Lan. | Onl. | Onl. | Hua. | Onl. | PRO. | ARC. | ALM. |
|--------|------|------|------|------|------|------|------|------|------|
| Online-B  | 0.0  | 0.2  | 5.1  | 10.4 | 12.6 | 12.8 | 13.0 | 13.0 | 40.8 |
| Lan-Bridge | 0.0  | 0.2  | 4.9  | 10.2 | 12.5 | 12.8 | 13.0 | 13.0 | 40.6 |
| Online-G  | 0.0  | 0.7  | 7.7  | 7.7  | 7.7  | 7.7  | 7.7  | 7.8  | 35.3 |
| Online-A  | 0.0  | 0.7  | 7.7  | 7.7  | 7.7  | 7.7  | 7.7  | 7.8  | 35.3 |
| HuaweiTSC | 0.0  | 0.3  | 0.5  | 0.5  | 0.5  | 0.5  | 0.5  | 0.5  | 28.3 |
| Online-Y  | 0.0  | 0.2  | 0.2  | 0.2  | 0.2  | 0.2  | 0.2  | 0.2  | 28.3 |
| PROMT     | 0.0  | 0.2  | 0.2  | 0.2  | 0.2  | 0.2  | 0.2  | 0.2  | 28.3 |
| ARC-NKUA  | 0.0  | 0.2  | 0.2  | 0.2  | 0.2  | 0.2  | 0.2  | 0.2  | 28.3 |
| HuaweiTSC | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 27.8 |
| eTranslation | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 27.8 |

Table 76: Statistical significance testing of the COMET score difference for each system pair for the uk$\rightarrow$en.

|        | Onl. | Lan. | Onl. | Onl. | Onl. | eTr. | Hua. | ARC. |
|--------|------|------|------|------|------|------|------|------|
| Online-B  | 0.0  | 0.2  | 2.2  | 3.3  | 3.4  | 3.5  | 3.5  | 5.2  |
| Lan-Bridge | 0.0  | 1.5  | 2.8  | 3.0  | 3.1  | 3.2  | 3.2  | 4.8  |
| Online-G  | 0.0  | 1.3  | 2.4  | 2.4  | 2.4  | 2.4  | 2.4  | 3.3  |
| Online-A  | 0.0  | 0.7  | 1.0  | 1.0  | 1.0  | 1.0  | 1.0  | 2.0  |
| eTranslation | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| HuaweiTSC | 0.0  | 0.3  | 0.5  | 0.5  | 0.5  | 0.5  | 0.5  | 2.8  |
| eTranslation | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |
| ARC-NKUA  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |

Table 77: Statistical significance testing of the chrF score difference for each system pair for the en$\rightarrow$uk.

|        | Onl. | Lan. | Onl. | Onl. | Onl. | Hua. | eTr. | ARC. |
|--------|------|------|------|------|------|------|------|------|
| Online-B  | 0.0  | 3.4  | 4.3  | 5.3  | 5.6  | 6.0  | 6.3  | 7.4  |
| Lan-Bridge | 0.0  | 3.4  | 4.3  | 5.3  | 5.6  | 6.0  | 6.3  | 7.4  |
| Online-G  | 0.0  | 3.4  | 4.3  | 5.3  | 5.6  | 6.0  | 6.3  | 7.4  |
| Online-A  | 0.0  | 0.3  | 0.7  | 1.0  | 1.0  | 1.0  | 1.0  | 2.0  |
| HuaweiTSC | 0.0  | 0.4  | 0.7  | 1.0  | 1.0  | 1.0  | 1.0  | 2.0  |
| eTranslation | 0.0  | 0.3  | 0.7  | 1.0  | 1.0  | 1.0  | 1.0  | 2.0  |
| ARC-NKUA  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |

Table 78: Statistical significance testing of the BLEU score difference for each system pair for the en$\rightarrow$uk.
Table 79: Statistical significance testing of the COMET score difference for each system pair for the en→uk.

|         | Orl. | Onl. | Lan. | eTr. | Hua. | Orl. | ARC. |
|---------|------|------|------|------|------|------|------|
| Online-B | 0.0  | 1.4  | 7.5  | 12.4 | 18.7 | 18.9 | 21.3 | 24.1 |
| Online-G | 0.0  | 0.4  | 9.0  | 13.3 | 15.5 | 17.3 | 19.7 | 20.7 |
| Lan-Bridge| 0.0  | 3.2  | 11.2 | 11.4 | 13.3 | 16.5 | 16.6 | 17.4 |
| eTranslation | 0.0  | 6.3  | 6.5  | 8.9  | 11.7 | 11.7 | 11.7 | 11.7 |
| HuaweiTSC | 0.0  | 0.2  | 7.6  | 5.4  | 5.4  | 5.4  | 5.4  | 5.4  |
| Online-Y | 0.0  | 3.4  | 7.5  | 7.5  | 7.5  | 7.5  | 7.5  | 7.5  |
| ARC-NKUA | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  | 0.0  |

Table 80: Statistical significance testing of the chrF score difference for each system pair for the zh→en.

|         | JDE  | Lan. | Onl. | AIS. | Hua. | Orl. | Lan. | Orl. | Niu | DLU | Orl. |
|---------|------|------|------|------|------|------|------|------|-----|-----|------|
| JDExploreAcademy | 0.0 | 0.2 | 1.5 | 2.8 | 3.0 | 3.3 | 1.8 | 6.4 | 5.2 | 6.1 | 6.8 |
| LanguageX | 0.0 | 0.8 | 1.7 | 2.7 | 3.6 | 3.0 | 3.2 | 4.3 | 3.7 | 3.7 | 3.7 |
| Online-G | 0.0 | 0.3 | 1.7 | 3.2 | 3.7 | 3.7 | 3.7 | 3.7 | 3.7 | 3.7 | 3.7 |
| AISP-SJTU | 0.0 | 1.0 | 1.5 | 1.9 | 2.1 | 2.1 | 3.7 | 3.7 | 3.7 | 3.7 | 3.7 |
| HuaweiTSC | 0.0 | 0.0 | 0.2 | 0.5 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |
| Online-B | 0.0 | 0.5 | 1.9 | 2.0 | 2.0 | 2.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 |
| Lan-Bridge | 0.0 | 0.7 | 1.9 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 |
| Online-Y | 0.0 | 0.2 | 2.7 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 |
| Online-A | 0.0 | 0.2 | 2.7 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 |
| NiuTrans | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| DLUT | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Online-W | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Table 81: Statistical significance testing of the BLEU score difference for each system pair for the zh→en.

|         | Orl. | JDE  | Lan. | Lan. | Hua. | Orl. | AIS. | Orl. | Orl. | Niu | DLU | Orl. |
|---------|------|------|------|------|------|------|------|------|------|-----|-----|------|
| JDExploreAcademy | 0.0 | 1.5 | 3.5 | 3.5 | 3.8 | 4.6 | 5.0 | 6.3 | 6.8 | 7.4 | 8.9 | 9.8 |
| LanguageX | 0.0 | 0.8 | 1.7 | 2.7 | 3.6 | 3.0 | 3.2 | 4.3 | 3.7 | 3.7 | 3.7 | 3.7 |
| Online-G | 0.0 | 0.3 | 1.7 | 3.2 | 3.7 | 3.7 | 3.7 | 3.7 | 3.7 | 3.7 | 3.7 | 3.7 |
| AISP-SJTU | 0.0 | 1.0 | 1.5 | 1.9 | 2.1 | 2.1 | 3.7 | 3.7 | 3.7 | 3.7 | 3.7 | 3.7 |
| HuaweiTSC | 0.0 | 0.0 | 0.2 | 0.5 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 | 0.9 |
| Online-B | 0.0 | 0.5 | 1.9 | 2.0 | 2.0 | 2.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 |
| Lan-Bridge | 0.0 | 0.7 | 1.9 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 |
| Online-Y | 0.0 | 0.2 | 2.7 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 |
| Online-A | 0.0 | 0.2 | 2.7 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 |
| NiuTrans | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| DLUT | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Online-W | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Table 82: Statistical significance testing of the COMET score difference for each system pair for the zh→en.
Table 83: Statistical significance testing of the chrF score difference for each system pair for the en→zh.

| System   | Online-B | HuaweiTSC | LanguageX | Online-A | Manifold | Online-Y | JDEExploreAcademy | AISP-SJTU | Lan-Bridge | DLUT | Online-G | Online-W |
|----------|----------|-----------|-----------|----------|----------|----------|-------------------|-----------|-----------|------|---------|---------|
|          | 0.0      | 9.3       | 10.8      | 12.3     | 12.7     | 13.4     | 14.5              | 15.0      | 16.7      | 17.2 | 17.9    | 18.8    |

Table 84: Statistical significance testing of the BLEU score difference for each system pair for the en→zh.

| System   | Online-B | HuaweiTSC | LanguageX | Online-A | Manifold | Online-Y | JDEExploreAcademy | AISP-SJTU | Lan-Bridge | DLUT | Online-G | Online-W |
|----------|----------|-----------|-----------|----------|----------|----------|-------------------|-----------|-----------|------|---------|---------|
|          | 0.0      | 6.7       | 7.1       | 10.5     | 11.4     | 11.6     | 11.7              | 12.4      | 13.4      | 16.1 | 16.7    | 18.0    |

Table 85: Statistical significance testing of the COMET score difference for each system pair for the en→zh.

| System   | LanguageX | JDEExploreAcademy | Lan-Bridge | Online-W | Online-Y | HuaweiTSC | AISP-SJTU | DLUT | Online-G | Online-W |
|----------|-----------|-------------------|-----------|---------|---------|-----------|-----------|------|---------|---------|
|          | 0.0       | 2.1               | 2.1       | 3.5     | 3.8     | 3.7       | 4.1              | 4.3     | 6.5      | 7.3    | 11.7    | 12.6    |

Table 85: Statistical significance testing of the COMET score difference for each system pair for the en→zh.
| System | COMET   | chrF   | BLEU   |
|--------|---------|--------|--------|
| Baseline | 77.5    | 79.3   | 64.2   |
| MBR     | 77.3    | 75.4   | 56.5   |
| Oracle  | 86.9    | 82.9   | 69.2   |

(a) cs → en

| System | COMET   | chrF   | BLEU   |
|--------|---------|--------|--------|
| Baseline | 97.9    | 69.0   | 48.2   |
| MBR     | 102.5   | 67.2   | 44.6   |
| Oracle  | 111.3   | 75.9   | 57.8   |

(b) en → cs

Table 86: MBR and oracle combination using COMET for cs-en.

| System | COMET   | chrF   | BLEU   |
|--------|---------|--------|--------|
| Baseline | 99.4    | 64.0   | 38.3   |
| MBR     | 104.3   | 64.9   | 38.8   |
| Oracle  | 110.8   | 70.8   | 46.5   |

(a) cs → uk

| System | COMET   | chrF   | BLEU   |
|--------|---------|--------|--------|
| Baseline | 104.9   | 60.7   | 37.0   |
| MBR     | 107.0   | 62.1   | 38.9   |
| Oracle  | 114.2   | 67.8   | 46.6   |

(b) uk → cs

Table 87: MBR and oracle combination using COMET for cs-uk.

Table 88: MBR and oracle combination using COMET for de-en.

C MBR and Oracle Combination Tables

From Table 86 to 96.
| System   | COMET  | chrF  | BLEU  | System   | COMET  | chrF  | BLEU  |
|----------|--------|-------|-------|----------|--------|-------|-------|
| Baseline | 70.5   | 74.6  | 58.4  | Baseline | 77.9   | 81.2  | 64.8  |
| MBR      | 70.8   | 69.0  | 49.1  | MBR      | 74.3   | 72.5  | 52.6  |
| Oracle   | 84.3   | 75.2  | 58.8  | Oracle   | 86.5   | 82.7  | 67.8  |

(a) de→fr

| System   | COMET  | chrF  | BLEU  | System   | COMET  | chrF  | BLEU  |
|----------|--------|-------|-------|----------|--------|-------|-------|
| Baseline | 80.4   | 58.5  | 31.5  | Baseline | 65.2   | 36.8  | 27.6  |
| MBR      | 84.6   | 58.2  | 31.0  | MBR      | 69.7   | 36.3  | 27.0  |
| Oracle   | 92.8   | 60.8  | 34.6  | Oracle   | 82.7   | 42.0  | 33.2  |

(b) fr→de

Table 89: MBR and oracle combination using COMET for de-fr.

| System   | COMET  | chrF  | BLEU  | System   | COMET  | chrF  | BLEU  |
|----------|--------|-------|-------|----------|--------|-------|-------|
| Baseline | 77.9   | 53.5  | 15.3  | Baseline | 65.2   | 70.0  | 46.7  |
| MBR      | 77.9   | 59.8  | 34.5  | MBR      | 68.0   | 70.0  | 46.8  |
| Oracle   | 87.6   | 64.4  | 40.1  | Oracle   | 76.4   | 73.6  | 53.1  |

(a) ja→en

| System   | COMET  | chrF  | BLEU  | System   | COMET  | chrF  | BLEU  |
|----------|--------|-------|-------|----------|--------|-------|-------|
| Baseline | -17.1  | 56.6  | 15.3  | Baseline | 31.1   | 55.5  | 29.6  |
| MBR      | -41.7  | 50.2  | 13.3  | MBR      | 20.1   | 51.8  | 26.1  |
| Oracle   | -5.6   | 60.6  | 22.9  | Oracle   | 32.9   | 55.5  | 29.7  |

(b) en→ja

Table 91: MBR and oracle combination using COMET for en-ja.

| System   | COMET  | chrF  | BLEU  | System   | COMET  | chrF  | BLEU  |
|----------|--------|-------|-------|----------|--------|-------|-------|
| Baseline | 62.6   | 67.3  | 44.6  | Baseline | 73.3   | 59.3  | 32.5  |
| MBR      | 62.0   | 67.4  | 45.1  | MBR      | 73.5   | 58.0  | 30.0  |
| Oracle   | 71.4   | 71.6  | 51.0  | Oracle   | 84.2   | 62.1  | 34.7  |

(a) liv→en

| System   | COMET  | chrF  | BLEU  | System   | COMET  | chrF  | BLEU  |
|----------|--------|-------|-------|----------|--------|-------|-------|
| Baseline | 73.3   | 59.3  | 32.5  | Baseline | 73.3   | 59.3  | 32.5  |
| MBR      | 73.5   | 58.0  | 30.0  | MBR      | 73.5   | 58.0  | 30.0  |
| Oracle   | 84.2   | 62.1  | 34.7  | Oracle   | 84.2   | 62.1  | 34.7  |

(b) en→uk

Table 95: MBR and oracle combination using COMET for en-uk.
| System   | COMET | chrF | BLEU |
|----------|-------|------|------|
| Baseline | 45.7  | 61.1 | 33.5 |
| MBR      | 52.9  | 60.5 | 31.8 |
| Oracle   | 64.8  | 63.9 | 36.2 |

(a) zh→en

| System   | COMET | chrF | BLEU |
|----------|-------|------|------|
| Baseline | 63.9  | 49.1 | 54.3 |
| MBR      | 67.3  | 46.7 | 51.3 |
| Oracle   | 77.7  | 53.6 | 57.8 |

(b) en→zh

Table 96: MBR and oracle combination using COMET for en-zh.
| System              | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|---------------------|----------------|----------------|-----------------|-------------------|------------------|
| Online-W            | 77.5 (1)       | 11.4 (1)       | 4.5 (1)         | 11.3 (1)          | 7.9 (3)          |
| JDExploreAcademy    | 74.8 (2)       | 10.1 (2)       | 4.5 (2)         | 11.3 (2)          | 8.3 (1)          |
| Online-B            | 71.9 (3)       | 9.0 (4)        | 4.4 (4)         | 11.2 (3)          | 8.0 (2)          |
| Lan-Bridge          | 71.8 (4)       | 9.0 (3)        | 4.4 (3)         | 11.1 (4)          | 7.9 (4)          |
| CUNI-DocTransformer | 70.6 (5)       | 8.4 (5)        | 4.3 (5)         | 11.0 (5)          | 7.2 (5)          |
| Online-A            | 69.8 (6)       | 8.2 (6)        | 4.3 (6)         | 11.0 (6)          | 7.0 (6)          |
| CUNI-Transformer    | 69.2 (7)       | 7.9 (7)        | 4.3 (7)         | 11.0 (7)          | 7.0 (7)          |
| Online-G            | 63.0 (8)       | 6.0 (8)        | 4.2 (8)         | 10.7 (8)          | 5.7 (9)          |
| SHOPLINE-PL         | 61.2 (9)       | 5.3 (9)        | 4.1 (9)         | 10.7 (9)          | 5.8 (8)          |
| Online-Y            | 58.7 (10)      | 4.3 (10)       | 4.1 (10)        | 10.6 (10)         | 5.3 (10)         |
| ALMANaCH-Inria      | 19.3 (11)      | -7.1 (11)      | 3.4 (11)        | 9.5 (11)          | -1.4 (11)        |

(a) cs → en

| System              | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|---------------------|----------------|----------------|-----------------|-------------------|------------------|
| Online-W            | 97.9 (1)       | 15.9 (2)       | 4.9 (2)         | 12.2 (1)          | 13.9 (1)         |
| Online-B            | 97.5 (2)       | 16.2 (1)       | 4.9 (1)         | 11.9 (4)          | 12.7 (4)         |
| CUNI-Bergamot       | 96.1 (3)       | 13.9 (5)       | 4.8 (3)         | 12.0 (3)          | 13.4 (2)         |
| JDExploreAcademy    | 95.3 (4)       | 15.1 (4)       | 4.8 (4)         | 12.1 (2)          | 13.3 (3)         |
| Lan-Bridge          | 94.7 (5)       | 15.4 (3)       | 4.8 (5)         | 11.8 (6)          | 12.5 (5)         |
| Online-A            | 92.2 (6)       | 13.5 (6)       | 4.7 (7)         | 11.7 (8)          | 11.9 (7)         |
| CUNI-Transformer    | 91.7 (7)       | 13.3 (7)       | 4.7 (6)         | 11.8 (5)          | 12.2 (6)         |
| Online-Y            | 83.7 (9)       | 10.7 (9)       | 4.5 (10)        | 11.3 (10)         | 9.7 (10)         |
| Online-G            | 82.3 (10)      | 10.5 (10)      | 4.5 (9)         | 11.3 (9)          | 9.7 (9)          |

(b) en → cs

Table 97: COMET scores (ranking) with different models for cs-en.

D Rankings With Different COMET Models

From Table 97 to 107.

E Sentence Matching Statistic Tables

From Table 108 to 118.
Table 98: COMET scores (ranking) with different models for cs-uk.

| System          | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-----------------|----------------|----------------|-----------------|--------------------|------------------|
| AMU             | 99.4 (1)       | 14.1 (1)       | 4.6 (1)         | 12.5 (1)           | 14.1 (1)         |
| Online-B        | 94.3 (2)       | 13.6 (2)       | 4.5 (2)         | 12.0 (2)           | 11.6 (2)         |
| Lan-Bridge      | 91.8 (3)       | 12.7 (3)       | 4.4 (3)         | 11.9 (3)           | 11.0 (3)         |
| HuaweiTSC       | 90.8 (4)       | 11.8 (5)       | 4.4 (4)         | 11.7 (7)           | 10.7 (5)         |
| CharlesTranslator | 90.7 (5)     | 12.2 (4)       | 4.5 (5)         | 11.9 (4)           | 11.0 (4)         |
| CUNI-JL-JH      | 90.1 (6)       | 11.3 (7)       | 4.3 (6)         | 11.7 (8)           | 10.4 (7)         |
| Online-G        | 88.4 (7)       | 11.3 (8)       | 4.3 (7)         | 11.9 (5)           | 10.5 (6)         |
| Online-A        | 87.9 (8)       | 11.4 (6)       | 4.3 (8)         | 11.8 (6)           | 10.1 (9)         |
| CUNI-Transformer | 87.4 (9)      | 11.0 (9)       | 4.3 (9)         | 11.7 (9)           | 10.3 (8)         |
| Online-Y        | 78.4 (10)      | 8.4 (10)       | 4.1 (10)        | 11.3 (10)          | 7.9 (10)         |
| ALMAnaCH-Inria  | 61.4 (11)      | 3.4 (11)       | 3.8 (11)        | 10.8 (11)          | 4.8 (11)         |

Table 99: COMET scores (ranking) with different models for de-en.

| System          | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-----------------|----------------|----------------|-----------------|--------------------|------------------|
| AMU             | 104.9 (1)      | 15.7 (1)       | 5.0 (1)         | 12.9 (1)           | 13.8 (1)         |
| Online-B        | 96.5 (2)       | 13.8 (2)       | 4.8 (2)         | 12.5 (2)           | 11.4 (2)         |
| Lan-Bridge      | 94.6 (3)       | 13.3 (3)       | 4.7 (3)         | 12.4 (3)           | 11.0 (3)         |
| HuaweiTSC       | 91.5 (4)       | 12.0 (4)       | 4.6 (4)         | 12.2 (4)           | 10.6 (4)         |
| CharlesTranslator | 90.2 (5)      | 11.4 (5)       | 4.6 (5)         | 12.1 (5)           | 10.2 (5)         |
| CUNI-JL-JH      | 89.1 (6)       | 11.0 (7)       | 4.6 (6)         | 12.1 (6)           | 9.9 (7)          |
| CUNI-Transformer | 88.5 (7)      | 11.2 (6)       | 4.6 (7)         | 12.1 (7)           | 10.1 (6)         |
| Online-A        | 85.4 (8)       | 10.2 (8)       | 4.5 (8)         | 12.1 (8)           | 9.3 (8)          |
| Online-G        | 84.2 (9)       | 9.7 (9)        | 4.4 (9)         | 11.9 (9)           | 8.9 (9)          |
| Online-Y        | 78.7 (10)      | 8.0 (10)       | 4.3 (10)        | 11.8 (10)          | 7.9 (10)         |
| ALMAnaCH-Inria  | 62.5 (11)      | 3.1 (11)       | 4.0 (11)        | 11.2 (11)          | 4.8 (11)         |

(a) de→en

| System          | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-----------------|----------------|----------------|-----------------|--------------------|------------------|
| JDExploreAcademy | 58.0 (1)      | 4.6 (1)        | 4.1 (1)         | 11.6 (1)           | 9.8 (1)          |
| Online-B        | 57.0 (2)       | 4.2 (2)        | 4.1 (2)         | 11.5 (2)           | 9.3 (2)          |
| Lan-Bridge      | 56.6 (3)       | 4.1 (3)        | 4.1 (3)         | 11.5 (3)           | 9.2 (3)          |
| Online-G        | 55.3 (4)       | 3.7 (4)        | 4.1 (4)         | 11.4 (4)           | 8.9 (4)          |
| Online-Y        | 54.7 (5)       | 3.5 (7)        | 4.0 (5)         | 11.4 (5)           | 8.9 (5)          |
| Online-A        | 54.5 (6)       | 3.6 (5)        | 4.0 (6)         | 11.4 (6)           | 8.8 (7)          |
| Online-W        | 54.4 (7)       | 3.6 (6)        | 4.0 (7)         | 11.4 (7)           | 8.9 (6)          |
| PROMT           | 51.8 (8)       | 2.6 (8)        | 4.0 (8)         | 11.2 (8)           | 8.2 (8)          |
| LT22            | 25.6 (9)       | -4.7 (9)       | 3.5 (9)         | 10.6 (9)           | 3.3 (9)          |

(b) en→de

Table 99: COMET scores (ranking) with different models for de-en.
### Table 100: COMET scores (ranking) with different models for de-fr.

| System          | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-----------------|----------------|----------------|-----------------|-------------------|-----------------|
| Online-B        | 70.5 (1)       | 8.4 (1)        | 4.4 (1)         | 11.6 (2)          | 6.7 (2)         |
| Online-W        | 63.7 (2)       | 6.1 (2)        | 4.3 (2)         | 11.9 (1)          | 8.0 (1)         |
| Online-Y        | 57.8 (3)       | 4.3 (3)        | 4.1 (3)         | 11.6 (3)          | 6.3 (3)         |
| Online-A        | 52.3 (4)       | 3.4 (4)        | 4.1 (4)         | 11.5 (4)          | 5.4 (5)         |
| Online-G        | 44.9 (5)       | 1.1 (5)        | 3.9 (5)         | 11.2 (5)          | 4.9 (5)         |
| LT22            | 10.5 (6)       | -6.6 (6)       | 3.5 (6)         | 10.7 (6)          | 0.9 (6)         |

(a) de $\rightarrow$ fr

| System          | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-----------------|----------------|----------------|-----------------|-------------------|-----------------|
| Online-B        | 77.9 (1)       | 13.6 (1)       | 4.7 (1)         | 11.5 (1)          | 7.2 (1)         |
| Online-W        | 63.8 (2)       | 8.0 (2)        | 4.3 (2)         | 11.3 (2)          | 6.0 (2)         |
| Online-Y        | 61.6 (3)       | 7.1 (3)        | 4.3 (3)         | 11.3 (3)          | 5.8 (3)         |
| Online-A        | 59.3 (4)       | 6.7 (4)        | 4.2 (4)         | 11.2 (4)          | 5.5 (4)         |
| eTranslation    | 55.5 (5)       | 5.8 (5)        | 4.2 (5)         | 11.0 (5)          | 4.9 (5)         |
| Lan-Bridge      | 51.2 (6)       | 4.3 (6)        | 4.1 (6)         | 11.0 (6)          | 4.6 (6)         |
| Online-G        | 48.3 (7)       | 3.5 (7)        | 4.0 (7)         | 10.6 (7)          | 3.3 (7)         |

(b) fr $\rightarrow$ de

### Table 101: COMET scores (ranking) with different models for en-hr.

| System          | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-----------------|----------------|----------------|-----------------|-------------------|-----------------|
| NT5             | 42.1 (1)       | 1.1 (1)        | 3.7 (1)         | 10.8 (1)          | 5.9 (2)         |
| Online-W        | 41.2 (2)       | 1.1 (2)        | 3.7 (2)         | 10.8 (2)          | 5.6 (3)         |
| JDEexpAcademy   | 40.6 (3)       | 0.8 (3)        | 3.7 (3)         | 10.8 (3)          | 6.1 (1)         |
| Online-B        | 39.7 (4)       | 0.6 (4)        | 3.6 (4)         | 10.6 (4)          | 5.1 (4)         |
| DLUT            | 37.3 (5)       | -0.3 (5)       | 3.6 (5)         | 10.5 (5)          | 4.8 (5)         |
| NAIST-NICT-TIT  | 33.4 (6)       | -1.5 (6)       | 3.5 (6)         | 10.4 (6)          | 4.4 (7)         |
| LanguageX       | 33.0 (7)       | -1.5 (7)       | 3.5 (8)         | 10.2 (10)         | 3.6 (9)         |
| Online-A        | 32.9 (8)       | -1.6 (8)       | 3.5 (7)         | 10.3 (7)          | 4.2 (8)         |
| Online-Y        | 32.4 (9)       | -1.7 (9)       | 3.5 (9)         | 10.3 (8)          | 4.5 (6)         |
| Lan-Bridge      | 32.0 (10)      | -3.7 (10)      | 3.4 (10)        | 10.2 (11)         | 3.6 (10)        |
| AISP-SJTU       | 30.2 (11)      | -2.3 (11)      | 3.4 (11)        | 10.3 (9)          | 3.6 (11)        |
| Online-G        | 22.3 (12)      | -4.3 (12)      | 3.3 (12)        | 9.9 (13)          | 2.1 (12)        |
| KYB             | 17.4 (13)      | -5.5 (13)      | 3.2 (13)        | 10.0 (12)         | 1.9 (13)        |
| AIST            | -152.7 (14)    | -65.3 (14)     | -0.6 (14)       | 4.0 (14)          | -38.0 (14)      |

(a) en $\rightarrow$ hr

### Table 102: COMET scores (ranking) with different models for en-ja.

| System          | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-----------------|----------------|----------------|-----------------|-------------------|-----------------|
| JDEexpAcademy   | 65.2 (1)       | 10.4 (1)       | 4.4 (1)         | 12.1 (1)          | 13.3 (1)        |
| NT5             | 64.1 (2)       | 10.2 (2)       | 4.4 (2)         | 12.0 (2)          | 13.0 (2)        |
| LanguageX       | 62.1 (3)       | 9.4 (3)        | 4.4 (3)         | 11.9 (3)          | 12.2 (5)        |
| Online-W        | 60.8 (4)       | 9.4 (4)        | 4.4 (4)         | 11.9 (5)          | 12.5 (3)        |
| DLUT            | 60.5 (5)       | 9.1 (5)        | 4.3 (6)         | 11.9 (6)          | 12.4 (4)        |
| Online-W        | 59.8 (6)       | 9.1 (6)        | 4.4 (5)         | 12.0 (3)          | 12.2 (6)        |
| Online-Y        | 56.9 (7)       | 8.1 (7)        | 4.2 (8)         | 11.8 (7)          | 11.7 (8)        |
| Lan-Bridge      | 56.6 (8)       | 8.0 (8)        | 4.3 (7)         | 11.8 (8)          | 11.7 (9)        |
| Online-A        | 53.7 (9)       | 7.5 (9)        | 4.2 (9)         | 11.8 (9)          | 11.9 (7)        |
| NAIST-NICT-TIT  | 53.4 (10)      | 7.4 (10)       | 4.2 (10)        | 11.7 (10)         | 11.6 (10)       |
| AISP-SJTU       | 52.4 (11)      | 7.0 (11)       | 4.2 (11)        | 11.7 (11)         | 11.0 (11)       |
| KYB             | 31.8 (12)      | 1.5 (12)       | 3.8 (12)        | 11.2 (12)         | 7.8 (12)        |
| Online-G        | 25.0 (13)      | 0.5 (13)       | 3.7 (13)        | 11.0 (13)         | 6.7 (13)        |

(b) en $\rightarrow$ ja
Table 103: COMET scores (ranking) with different models for en-liv.

| System          | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-----------------|----------------|----------------|-----------------|--------------------|------------------|
| TAL-SJTU        | -29.5 (1)      | -15.5 (1)      | 0.7 (1)         | 2.9 (2)            | -40.0 (3)        |
| TartuNLP        | -36.8 (2)      | -17.1 (2)      | 0.5 (3)         | 2.9 (3)            | -40.2 (4)        |
| HuaweiTSC       | -38.8 (3)      | -17.7 (3)      | 0.4 (4)         | 2.8 (4)            | -41.7 (5)        |
| Liv4ever        | -39.4 (4)      | -18.4 (4)      | 0.6 (2)         | 3.2 (1)            | -37.6 (1)        |
| NiuTrans        | -81.9 (5)      | -31.3 (5)      | 0.1 (5)         | 2.8 (5)            | -38.4 (2)        |

Table 104: COMET scores (ranking) with different models for en-ru.

| System          | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-----------------|----------------|----------------|-----------------|--------------------|------------------|
| Online-G        | 65.2 (1)       | 8.2 (1)        | 4.4 (1)         | 11.3 (2)           | 8.1 (2)          |
| JDExploreAcademy| 64.9 (2)       | 8.1 (2)        | 4.4 (2)         | 11.5 (1)           | 8.6 (1)          |
| Online-Y        | 64.1 (3)       | 7.7 (3)        | 4.4 (3)         | 11.3 (3)           | 7.9 (3)          |
| Lan-Bridge      | 63.1 (4)       | 7.4 (4)        | 4.4 (4)         | 11.3 (4)           | 7.9 (4)          |
| Online-B        | 63.1 (5)       | 7.4 (5)        | 4.4 (5)         | 11.3 (5)           | 7.9 (5)          |
| Online-A        | 62.2 (6)       | 7.1 (6)        | 4.3 (6)         | 11.2 (7)           | 7.6 (7)          |
| Online-W        | 61.6 (7)       | 7.0 (7)        | 4.3 (7)         | 11.3 (6)           | 7.8 (6)          |
| HuaweiTSC       | 60.9 (8)       | 6.8 (8)        | 4.3 (8)         | 11.2 (8)           | 7.5 (8)          |
| SRPOL           | 59.5 (9)       | 6.3 (9)        | 4.3 (9)         | 11.2 (9)           | 7.4 (9)          |
| ALMAnach-Itali* | 26.9 (10)      | -4.0 (10)      | 3.6 (10)        | 10.1 (10)          | 1.4 (10)         |

Table 105: COMET scores (ranking) with different models for ru-sah.

| System          | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-----------------|----------------|----------------|-----------------|--------------------|------------------|
| Online-W        | 75.1 (1)       | 11.3 (1)       | 4.6 (1)         | 11.8 (1)           | 12.7 (1)         |
| Online-G        | 73.2 (2)       | 10.8 (2)       | 4.6 (2)         | 11.7 (2)           | 12.3 (2)         |
| Online-B        | 72.9 (3)       | 10.7 (3)       | 4.6 (3)         | 11.5 (3)           | 11.6 (3)         |
| Online-Y        | 69.8 (4)       | 9.6 (4)        | 4.5 (4)         | 11.5 (4)           | 11.1 (5)         |
| JDExploreAcademy| 69.7 (5)       | 9.6 (5)        | 4.5 (5)         | 11.5 (5)           | 11.4 (4)         |
| Lan-Bridge      | 67.4 (6)       | 9.1 (6)        | 4.4 (6)         | 11.3 (6)           | 10.5 (7)         |
| Online-A        | 67.4 (7)       | 8.6 (7)        | 4.4 (7)         | 11.3 (7)           | 10.6 (6)         |
| PROMT           | 60.4 (8)       | 6.9 (8)        | 4.3 (8)         | 11.2 (8)           | 9.8 (8)          |
| SRPOL           | 59.8 (9)       | 6.4 (9)        | 4.2 (9)         | 11.0 (10)          | 9.2 (10)         |
| HuaweiTSC       | 59.2 (10)      | 6.3 (10)       | 4.2 (10)        | 11.1 (9)           | 9.8 (8)          |
| eTranslation    | 58.0 (11)      | 6.0 (11)       | 4.2 (11)        | 11.0 (11)          | 9.0 (11)         |

Table 106: COMET scores (ranking) with different models for ru-sah.

| System          | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-----------------|----------------|----------------|-----------------|--------------------|------------------|
| Online-G        | -17.1 (1)      | -18.3 (1)      | 1.0 (1)         | 4.8 (1)            | -30.4 (1)        |
| Lan-Bridge      | -48.0 (2)      | -19.0 (2)      | 0.1 (2)         | 3.3 (2)            | -38.8 (2)        |

Table 107: COMET scores (ranking) with different models for ru-sah.
| System            | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-------------------|----------------|----------------|-----------------|--------------------|------------------|
| Online-B          | 62.6 (1)       | 7.4 (1)        | 4.4 (1)         | 11.5 (1)           | 7.0 (1)          |
| Lan-Bridge        | 62.4 (2)       | 7.4 (2)        | 4.4 (2)         | 11.5 (2)           | 6.9 (2)          |
| Online-G          | 57.5 (3)       | 5.8 (3)        | 4.3 (3)         | 11.3 (3)           | 6.0 (3)          |
| Online-A          | 52.2 (4)       | 4.1 (4)        | 4.2 (4)         | 11.1 (4)           | 5.1 (4)          |
| HuaweiTSC         | 50.1 (5)       | 3.4 (5)        | 4.2 (5)         | 11.1 (5)           | 5.0 (5)          |
| Online-Y          | 49.8 (6)       | 3.3 (7)        | 4.1 (6)         | 10.9 (8)           | 4.5 (8)          |
| PROMT             | 49.6 (7)       | 3.4 (6)        | 4.1 (7)         | 11.0 (6)           | 4.9 (6)          |
| ARC-NKUA          | 49.6 (8)       | 3.3 (8)        | 4.1 (8)         | 11.0 (7)           | 4.6 (7)          |
| ALMAnaCH-Inria    | 21.8 (9)       | -5.7 (9)       | 3.5 (9)         | 10.0 (9)           | -0.7 (9)         |

(a) uk → en

| System            | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-------------------|----------------|----------------|-----------------|--------------------|------------------|
| Online-B          | 73.3 (1)       | 10.4 (1)       | 4.5 (1)         | 11.6 (1)           | 11.9 (1)         |
| Online-G          | 69.9 (2)       | 9.2 (2)        | 4.4 (2)         | 11.6 (2)           | 11.7 (2)         |
| Lan-Bridge        | 65.8 (3)       | 8.5 (3)        | 3.3 (3)         | 11.3 (3)           | 10.4 (3)         |
| Online-A          | 60.9 (4)       | 6.8 (4)        | 4.2 (4)         | 11.2 (4)           | 10.0 (4)         |
| eTranslation      | 54.6 (5)       | 4.9 (5)        | 4.1 (5)         | 11.0 (5)           | 9.0 (5)          |
| HuaweiTSC         | 54.4 (6)       | 4.7 (6)        | 4.1 (6)         | 11.0 (6)           | 8.7 (6)          |
| Online-Y          | 52.0 (7)       | 4.0 (7)        | 4.0 (7)         | 10.8 (7)           | 7.9 (7)          |
| ARC-NKUA          | 49.2 (8)       | 3.4 (8)        | 4.0 (8)         | 10.7 (8)           | 7.7 (8)          |

(b) en → uk

Table 106: COMET scores (ranking) with different models for en-uk.

| System            | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-------------------|----------------|----------------|-----------------|--------------------|------------------|
| Online-G          | 45.7 (1)       | -0.0 (1)       | 3.3 (1)         | 9.3 (2)            | 4.6 (1)          |
| JDEExploreAcademy | 45.2 (2)       | -0.3 (2)       | 3.3 (2)         | 9.2 (3)            | 4.1 (5)          |
| LanguageX         | 45.0 (3)       | -0.3 (3)       | 3.3 (3)         | 9.2 (4)            | 4.0 (6)          |
| Lan-Bridge        | 43.1 (4)       | -0.5 (4)       | 3.3 (4)         | 9.4 (1)            | 4.3 (2)          |
| HuaweITSC         | 42.9 (5)       | -0.7 (5)       | 3.2 (5)         | 9.2 (5)            | 4.0 (7)          |
| Online-B          | 42.2 (6)       | -0.7 (6)       | 3.2 (6)         | 9.2 (6)            | 4.2 (4)          |
| AISP-SJTU         | 41.7 (7)       | -1.2 (7)       | 3.2 (7)         | 9.1 (8)            | 3.6 (8)          |
| Online-Y          | 40.9 (8)       | -1.5 (8)       | 3.2 (8)         | 9.2 (7)            | 4.3 (3)          |
| Online-A          | 35.3 (9)       | -2.8 (9)       | 3.1 (9)         | 8.9 (9)            | 3.1 (9)          |
| Online-W          | 31.7 (10)      | -3.3 (10)      | 3.0 (11)        | 8.8 (12)           | 2.6 (10)         |
| NiuTrans          | 31.4 (11)      | -3.5 (11)      | 3.1 (10)        | 8.9 (10)           | 2.5 (11)         |
| DLUT              | 30.7 (12)      | -4.0 (12)      | 3.0 (12)        | 8.9 (11)           | 2.4 (12)         |

(a) zh → en

| System            | wmt20-comet-da | wmt21-comet-da | wmt21-comet-mqm | wmt21-comet-qe-mqm | wmt21-comet-qe-da |
|-------------------|----------------|----------------|-----------------|--------------------|------------------|
| LanguageX         | 63.9 (1)       | 7.6 (1)        | 4.2 (1)         | 11.3 (3)           | 8.0 (5)          |
| JDEExploreAcademy | 61.8 (2)       | 6.8 (2)        | 4.2 (2)         | 11.3 (4)           | 8.1 (4)          |
| Online-G          | 61.8 (3)       | 6.8 (3)        | 4.2 (3)         | 11.3 (5)           | 8.0 (6)          |
| Lan-Bridge        | 61.4 (4)       | 6.5 (4)        | 4.2 (4)         | 11.4 (1)           | 8.2 (2)          |
| Online-W          | 61.0 (5)       | 6.2 (5)        | 4.1 (5)         | 11.3 (6)           | 8.2 (3)          |
| Online-Y          | 59.8 (7)       | 5.6 (8)        | 4.1 (7)         | 11.3 (7)           | 8.0 (7)          |
| HuaweITSC         | 59.6 (8)       | 6.0 (7)        | 4.1 (8)         | 11.3 (8)           | 7.8 (8)          |
| Online-A          | 57.4 (9)       | 4.9 (10)       | 4.0 (10)        | 11.2 (9)           | 7.5 (9)          |
| AISP-SJTU         | 56.6 (10)      | 5.1 (9)        | 4.1 (9)         | 11.2 (10)          | 7.2 (10)         |
| DLUT              | 52.2 (11)      | 3.7 (11)       | 4.0 (11)        | 11.2 (11)          | 7.0 (11)         |
| Online-G          | 51.3 (12)      | 3.0 (12)       | 3.9 (12)        | 10.9 (12)          | 5.7 (12)         |

(b) en → zh

Table 107: COMET scores (ranking) with different models for en-zh.
Table 108: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

| System                      | Exact Match | Self Mismatch | BLEU | COMET |
|-----------------------------|-------------|---------------|------|-------|
| Online-W                    | 145         | 68.78/66.6    | 72.5/5.2 |
| JDEExploreAcademy           | 98          | 60.96/69.7    | 74.8/7.4 |
| Lan-Bridge                  | 84          | 60.86/70.7    | 71.8/7.8 |
| Online-B                    | 84          | 60.60/65.0    | 71.9/7.1 |
| CUNI-DocTransformer         | 80          | 57.9/7/57.9    | 70.6/7.8 |
| Online-A                    | 83          | 59.49/49.4    | 69.8/8.9 |
| Online-G                    | 75          | 54.35/4/54.3   | 63.0/8.3 |
| SHOPLINE-PL                 | 55          | 53.53/34.1    | 61.2/6.1 |
| Online-Y                    | 57          | 51.65/51.5    | 58.7/5.7 |
| ALMAInCH-Iria               | 12          | 36.07/31.1    | 19.3/9.3 |

Table 109: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

| System                      | Exact Match | Self Mismatch | BLEU | COMET |
|-----------------------------|-------------|---------------|------|-------|
| AMU                         | 24          | 34.7/34.3     | 99.49/94.4 |
| Online-B                    | 80          | 38.3/37.6     | 94.30/4.2 |
| Lan-Bridge                  | 66          | 38.3/38.2     | 91.80/1.8 |
| CharlesTranslator           | 46          | 34.3/41.0     | 90.80/0.7 |
| Online-A                    | 51          | 36.0/35.7     | 90.7/0.0 |
| CUNI-JL-JH                  | 39          | 34.6/34.1     | 90.1/0.0 |
| Online-G                    | 29          | 32.5/32.5     | 88.4/8.3 |
| Online-A                    | 43          | 35.0/35.5     | 87.9/8.7 |
| CUNI-Transformer            | 42          | 35.0/34.8     | 87.9/8.7 |
| Online-Y                    | 40          | 32.1/31.9     | 78.4/7.8 |
| ALMAInCH-Iria               | 21          | 26.8/26.7     | 61.4/6.1 |

Table 109: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

| System                      | Exact Match | Self Mismatch | BLEU | COMET |
|-----------------------------|-------------|---------------|------|-------|
| AMU                         | 117         | 37.0/36.4     | 104.9/104.7 |
| Online-B                    | 127         | 36.4/35.7     | 96.59/6.3 |
| Lan-Bridge                  | 147         | 36.5/36.5     | 94.60/4.6 |
| HuaweiTSC                   | 132         | 36.0/35.7     | 91.59/1.4 |
| CharlesTranslator           | 123         | 35.9/35.5     | 90.20/0.0 |
| CUNI-JL-JH                  | 122         | 35.1/35.1     | 89.18/9.0 |
| CUNI-Transformer            | 126         | 35.8/35.8     | 88.5/8.4 |
| Online-A                    | 116         | 33.3/33.0     | 85.4/8.3 |
| Online-Y                    | 106         | 31.5/31.5     | 84.2/8.2 |
| ALMAInCH-Iria               | 88          | 29.6/29.1     | 78.7/7.8 |

Table 110: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

| System                      | Exact Match | Self Mismatch | BLEU | COMET |
|-----------------------------|-------------|---------------|------|-------|
| JDEExploreAcademy           | 98          | 49.3/49.3     | 96.3/95.9 |
| CUNI-Transformer            | 84          | 45.4/44.4     | 88.7/8.6 |
| Online-B                    | 107         | 53.4/53.4     | 95.3/95.3 |
| Lan-Bridge                  | 226         | 53.4/53.4     | 94.7/94.7 |
| Online-A                    | 105         | 49.1/49.1     | 92.2/92.2 |
| Online-G                    | 90          | 44.8/44.8     | 82.3/82.3 |
| Online-W                    | 137         | 53.1/53.1     | 97.8/97.8 |
| Online-Y                    | 99          | 44.7/44.7     | 83.7/83.8 |

Table 110: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

| System                      | Exact Match | Self Mismatch | BLEU | COMET |
|-----------------------------|-------------|---------------|------|-------|
| Online-W                    | 75          | 48.9/48.8     | 60.6/6.5 |
| JDEExploreAcademy           | 70          | 51.6/51.2     | 63.3/8.1 |
| Online-B                    | 74          | 52.3/52.1     | 62.1/8.2 |
| Lan-Bridge                  | 55          | 50.5/50.1     | 61.8/6.9 |
| Online-Y                    | 75          | 50.15/50.1    | 60.6/6.6 |
| Online-A                    | 47          | 50.6/50.6     | 64.9/6.9 |
| Lan-Bridge                  | 70          | 49.4/49.4     | 58.8/8.8 |
| OpenNMT                     | 86          | 49.4/49.4     | 57.1/8.1 |
| Online-G                    | 64          | 49.0/49.0     | 55.9/5.7 |

Table 111: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of non-unique source segments translated differently by the reference/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

| System                      | Exact Match | Self Mismatch | BLEU | COMET |
|-----------------------------|-------------|---------------|------|-------|
| Online-W                    | 315         | 64.9/64.9     | 77.0/77.0 |
| Online-B                    | 153         | 46.6/46.6     | 63.8/8.6 |
| Online-Y                    | 121         | 45.0/45.0     | 61.6/6.4 |
| Online-A                    | 133         | 44.4/44.4     | 59.3/9.3 |
| eTranslation                | 162         | 46.5/46.5     | 55.5/5.2 |
| Lan-Bridge                  | 107         | 41.8/41.8     | 51.2/12.2 |
| Online-G                    | 85          | 41.1/41.1     | 48.3/8.4 |

Table 111: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of non-unique source segments translated differently by the reference/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.
Table 112: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

| System          | Exact Match | Self Mismatch | BLEU  | COMET  |
|-----------------|-------------|---------------|-------|--------|
| JDExploreAcademy| 35          | 13/11/2008    | 22.7/2.9 | 33.4/3.4 |
| AIST            | 0           | 24/11/2008    | 10.0/0.1 | 152.7/152.6 |
| Online-W        | 45          | 14/11/2008    | 27.8/2.7 | 41.2/4.1  |
| JDEExploreAcademy| 75          | 01/15/2008    | 25.6/2.6 | 42.0/4.2  |
| Online-B        | 63          | 01/15/2008    | 24.7/2.4 | 39.7/3.9  |
| Online-A        | 39          | 01/15/2008    | 22.8/2.9 | 32.9/3.9  |
| LanguageX       | 36          | 01/15/2008    | 22.4/2.3 | 33.0/3.2  |
| Online-Y        | 32          | 01/15/2008    | 21.9/2.4 | 32.4/3.2  |
| Lan-Bridge      | 44          | 01/15/2008    | 22.8/2.7 | 32.0/3.1  |
| AISP-SJTU       | 36          | 01/15/2008    | 22.0/2.1 | 30.2/3.0  |
| Online-G        | 34          | 01/15/2008    | 19.7/2.1 | 22.3/2.5  |
| KYB             | 30          | 01/15/2008    | 18.1/2.2 | 17.4/1.7  |

Table 113: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

| System          | Exact Match | Self Mismatch | BLEU  | COMET  |
|-----------------|-------------|---------------|-------|--------|
| AISP-SJTU       | 13          | 22/63/122/2037| 30.4/2.8 | 32.9/3.2 |
| LanguageX       | 15          | 01/11/2008    | 22.0/2.1 | 24.2/2.7  |
| Online-A        | 11          | 01/11/2008    | 21.0/2.0 | 23.1/2.6  |
| Online-Y        | 10          | 01/11/2008    | 19.7/2.0 | 19.8/2.0  |
| Online-G        | 97          | 01/11/2008    | 12.4/2.8 | 25.0/20.2 |
| KYB             | 5           | 01/11/2008    | 33.2/3.7 | 31.8/6.6  |
| Online-G        | 6           | 01/11/2008    | 42.8/2.8 | 42.0/2.8  |

Table 114: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

| System          | Exact Match | Self Mismatch | BLEU  | COMET  |
|-----------------|-------------|---------------|-------|--------|
| TarteNLP        | 3           | 00/04/2020    | 29.9/2.9 | 5.7/5.7  |
| TAL-SJTU        | 4           | 00/04/2020    | 30.4/0.6 | 8.3/8.3  |
| HuaweiTSC       | 1           | 00/04/2020    | 23.2/1.3 | 44.0/4.4 |
| Liv4ever        | 3           | 00/04/2020    | 23.2/1.3 | 44.0/4.4 |
| NiuTrans        | 0           | 00/04/2020    | 0.0/0.0  | 48.0/48.0|

Table 115: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

| System          | Exact Match | Self Mismatch | BLEU  | COMET  |
|-----------------|-------------|---------------|-------|--------|
| Online-G        | 132         | 05/05/2016    | 46.7/4.6 | 65.2/6.5 |
| JDEExploreAcademy| 107         | 05/05/2016    | 45.1/4.3 | 64.9/4.9 |
| Online-Y        | 98          | 05/05/2016    | 43.8/4.7 | 64.1/6.1 |
| Lan-Bridge      | 98          | 05/05/2016    | 45.2/4.5 | 63.2/6.1 |
| Online-A        | 96          | 05/05/2016    | 43.9/3.8 | 62.2/6.2 |
| Online-W        | 71          | 05/05/2016    | 42.6/4.2 | 61.6/6.1 |
| HuaweiTSC       | 97          | 05/05/2016    | 45.1/4.1 | 60.8/6.9 |
| SRPOL           | 85          | 05/05/2016    | 43.8/4.3 | 59.3/5.9 |
| ALMAasCH-Itaria | 35          | 05/05/2016    | 30.3/0.3 | 26.9/2.6 |

(a) en→hr
(b) hr→en
(a) ja→en
(b) en→ja
(a) liv→en
(b) en→liv
(a) ru→en
(b) en→ru
### Table 116: Sentence Matching Statistics.

| System          | Exact Match | Self Mismatch | BLEU  | COMET  |
|-----------------|-------------|---------------|-------|--------|
| Online-G        | 7           | 0/0/14/1123   | 14.7/14.7 | -17.1/-17.2 |
| Lan-Bridge      | 0           | 0/0/14/1123   | 15.3/14.0 | -48.0/-48.9 |

(a) ru→sah

| System          | Exact Match | Self Mismatch | BLEU  | COMET  |
|-----------------|-------------|---------------|-------|--------|
| Online-G        | 32          | 0/15/1123     | 29.0/29.6 | 51.1/51.0 |
| Lan-Bridge      | 0           | 0/15/1123     | 7.0/7.1   | -75.9/-75.7 |

(b) sah→ru

Table 117: Sentence Matching Statistics.

| System          | Exact Match | Self Mismatch | BLEU  | COMET  |
|-----------------|-------------|---------------|-------|--------|
| Online-B        | 101         | 0/6/122/2037  | 32.5/31.6 | 73.3/73.1 |
| Online-G        | 36          | 2/6/122/2037  | 27.2/27.2 | 69.9/69.9 |
| Lan-Bridge      | 71          | 0/6/122/2037  | 29.5/29.5 | 65.8/65.7 |
| HuaweiTSC       | 45          | 0/6/122/2037  | 26.9/26.9 | 54.4/54.4 |
| eTranslation    | 53          | 0/6/122/2037  | 26.2/26.2 | 54.6/54.6 |
| HuaweiTSC       | 45          | 0/6/122/2037  | 26.2/26.2 | 54.6/54.6 |
| Online-Y        | 49          | 0/6/122/2037  | 26.9/26.9 | 52.0/52.0 |
| ARC-NKUA        | 50          | 0/6/122/2037  | 25.2/25.2 | 49.2/49.2 |
| Online-A        | 52          | 0/6/122/2037  | 25.8/25.8 | 49.8/49.8 |
| Online-W        | 51          | 0/6/122/2037  | 25.3/25.3 | 49.4/49.4 |
| NiuTrans        | 42          | 0/6/122/2037  | 24.5/24.5 | 48.9/48.9 |
| DLUT            | 40          | 0/6/122/2037  | 23.1/23.1 | 47.6/47.6 |

(a) uk→en

| System          | Exact Match | Self Mismatch | BLEU  | COMET  |
|-----------------|-------------|---------------|-------|--------|
| Online-G        | 44          | 0/3/122/2037  | 9.5/9.5   | -60.5/-60.5 |
| LanguageX       | 49          | 3/3/122/2037  | 4.1/4.1   | -54.1/-54.1 |
| Online-B        | 45          | 0/3/122/2037  | 4.1/4.1   | -54.1/-54.1 |
| Online-Y        | 44          | 0/3/122/2037  | 4.1/4.1   | -54.1/-54.1 |
| Online-A        | 43          | 0/3/122/2037  | 4.1/4.1   | -54.1/-54.1 |
| Online-W        | 42          | 0/3/122/2037  | 4.1/4.1   | -54.1/-54.1 |
| ARC-NKUA        | 41          | 0/3/122/2037  | 4.1/4.1   | -54.1/-54.1 |

(b) en→uk

### Table 118: Sentence Matching Statistics.

| System          | Exact Match | Self Mismatch | BLEU  | COMET  |
|-----------------|-------------|---------------|-------|--------|
| LanguageX       | 95          | 0/3/122/2037  | 24.9/4.9   | 63.9/60.6 |
| Online-B        | 65          | 0/3/122/2037  | 48.1/9.2   | 61.8/58.5 |
| LanguageX       | 59          | 0/3/122/2037  | 24.9/4.9   | 61.8/58.5 |
| HuaweiTSC       | 49          | 0/3/122/2037  | 23.6/23.6  | 51.4/51.4 |
| Online-W        | 32          | 0/3/122/2037  | 23.6/23.6  | 66.5/66.5 |
| Online-Y        | 44          | 0/3/122/2037  | 28.0/6.0   | 59.8/56.5 |
| HuaweiTSC       | 43          | 0/3/122/2037  | 25.3/25.3  | 59.8/56.5 |
| Online-A        | 50          | 0/3/122/2037  | 26.4/5.4   | 57.6/57.6 |
| AISP-SJTU       | 52          | 0/3/122/2037  | 25.7/5.7   | 56.6/56.6 |
| DLUT            | 35          | 0/3/122/2037  | 21.5/21.5  | 52.2/52.2 |
| Online-G        | 38          | 0/3/122/2037  | 16.5/16.5  | 51.3/51.3 |

(a) zh→en

| System          | Exact Match | Self Mismatch | BLEU  | COMET  |
|-----------------|-------------|---------------|-------|--------|
| LanguageX       | 49          | 0/3/122/2037  | 24.9/4.9   | 63.9/60.6 |
| LanguageX       | 59          | 0/3/122/2037  | 24.9/4.9   | 63.9/60.6 |
| Online-B        | 65          | 0/3/122/2037  | 48.1/9.2   | 61.8/58.5 |
| LanguageX       | 59          | 0/3/122/2037  | 24.9/4.9   | 61.8/58.5 |
| HuaweiTSC       | 49          | 0/3/122/2037  | 23.6/23.6  | 51.4/51.4 |
| Online-W        | 32          | 0/3/122/2037  | 23.6/23.6  | 66.5/66.5 |
| Online-Y        | 44          | 0/3/122/2037  | 28.0/6.0   | 59.8/56.5 |
| HuaweiTSC       | 43          | 0/3/122/2037  | 25.3/25.3  | 59.8/56.5 |
| Online-A        | 50          | 0/3/122/2037  | 26.4/5.4   | 57.6/57.6 |
| AISP-SJTU       | 52          | 0/3/122/2037  | 25.7/5.7   | 56.6/56.6 |
| DLUT            | 35          | 0/3/122/2037  | 21.5/21.5  | 52.2/52.2 |
| Online-G        | 38          | 0/3/122/2037  | 16.5/16.5  | 51.3/51.3 |

(b) en→zh

Table 116: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of non-unique source segments translated differently by the reference/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

Table 117: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of non-unique source segments translated differently by the reference/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.

Table 118: Sentence Matching Statistics. “Exact Match” counts how many segments in the translation are identical to the corresponding segments of the reference translation. “Self Mismatch” provides various statistics: the number of non-unique source segments translated differently by the system/the number of non-unique source segments translated differently by the reference/the number of duplicated source segments/the total number of source segments. “BLEU” and “COMET” give the scores computed on the original/normalized translations.