RoCS-MT v2 at WMT 2025: Robust Challenge Set for Machine Translation

Rachel Bawden **Benoît Sagot**
Inria, Paris, France
firstname.lastname@inria.fr

Abstract

RoCS-MT (Robust Challenge Set for Machine Translation) was initially proposed at the test suites track of WMT 2023. Designed to challenge MT systems' translation performance on user-generated content (UGC), it contains examples sourced from English Reddit, with manually normalised versions, aligned labelled annotation spans and reference translations in five languages. In this article, we describe version 2 of RoCS-MT in the context of the 2025 WMT test suites track. This new version contains several improvements on the initial version including (i) minor corrections of normalisation, (ii) corrections to reference translations and addition of alternative references to accommodate for different possible genders (e.g. of speakers) and (iii) a redesign and re-annotation of normalisation spans for further analysis of different non-standard UGC phenomena. We describe these changes and provide results and preliminary analysis of the MT submissions to the 2025 general translation task.

1 Introduction

Large language models (LLMs) have truly arrived in the field of Machine Translation (MT); their performance often rivals that of dedicated MT systems across various domains (Kocmi et al., 2023; Xu et al., 2024; Cui et al., 2025). However, while they have opened up new possibilities for translation, enabling fine-grained control of output formats, styles and formalities, they are also characterised by new types of errors that were less present with dedicated models, such as translating in the wrong language, inappropriate copying of the source text, generation of outputs that are not translations of the source text, etc. Evaluation continues to be a highly important aspect of the field, and as MT technologies progress, so does the way in which we evaluate. The WMT shared tasks are a good example of this, with an evolution a few years ago to general translation instead of news translation (Kocmi et al., 2022), in order to challenge systems to translate a wider range of domains. One of the selected domains is social media content, known for covering a wide range of topics and containing non-standard language typical of user-generated content (UGC) (Foster, 2010; Seddah et al., 2012; Eisenstein, 2013; Baldwin and Li, 2015; van der Goot et al., 2018).

The translation of UGC has been a topic for a number of years (Belinkov and Bisk, 2018; Michel and Neubig, 2018; Vaibhav et al., 2019; Park et al., 2020; Nishimwe et al., 2024; Peters and Martins, 2025). In particular, a shared task was organised on the matter at WMT in 2023 (Kocmi et al., 2023), designed to target non-standard language from Reddit forums. Several parallel test sets of UGC texts exist (Ling et al., 2013; Vicente et al., 2016; Sluyter-Gäthje et al., 2018; Michel and Neubig, 2018; Rosales Núñez et al., 2019; Mubarak et al., 2020; Fujii et al., 2020; McNamee and Duh, 2022), for a range of languages, although they differ according to the language pairs covered and the degree of non-standardness present.

In the 2023 edition of the test suite track at WMT, we proposed the RoCS-MT test suite (Robust Challenge Set for MT) (Bawden and Sagot, 2023), designed to contain particularly challenging sentences with respect to their non-standard nature (e.g. with spelling mistakes, use of acronyms, marks of expressiveness, devowelling, contractions, etc.). Sourced from English Reddit, we aimed to cover a range of these phenomena in the texts selected, which we manually segmented into sentences, normalised into standard English (according to guidelines that normalised as reasonably possible whilst preserving fluency and meaning) and then translated by professional translators into French, German, Czech, Ukrainian and Russian.

For this 2025 edition of WMT (Kocmi et al., 2025), we resubmit RoCS-MT in an improved version (v2), after (i) some minor corrections to the

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 387–402
November 8–9, 2025. ©2025 Association for Computational Linguistics

source-side normalisations, (ii) corrections to the existing references and addition of multiple references to account for different genders and (iii) improvements to the annotations of the non-standard phenomena for additional analysis. We release this version in Huggingface's Datasets (Lhoest et al., 2021).[1] In this paper we describe those changes and provide results and analysis for the WMT2025 MT systems, with a major difference being that all systems this year were applied at the document level (at the level of Reddit post text in out case). We compare different segmentations of the texts and the performance of systems when applied to the original and normalised source texts.

2 The Test Suite

Composition The main composition of the challenge set is described in the article presenting the first version (Bawden and Sagot, 2023). The English source texts are taken from Reddit (all varieties of English including some non-native language, although we avoided code-switching). Candidate posts were identified using keyword searches on the Reddit API and chunks of text were manually selected from within those posts. The selected texts were manually segmented into sentences (non-trivial since many texts did not contain standard punctuation and sometimes contained newlines within sentences) and manually normalised. The normalisation guidelines we drew up aimed to balance (i) normalising as much as possible and at the same time (ii) rendering the output natural and realistic and (iii) not over-normalising to avoid losing the original text's style. For example, we did not use normalised variants that could be spontaneously and naturally used (e.g. we kept *lol* instead of *laughing out loud*. Finally, translations of the normalised texts were professionally produced in five languages: French, German, Czech, Ukrainian and Russian. Although not all these language directions are represented in this year's shared task, these references remain relevant for future use of the challenge set for these five languages.

Changes with respect to the First Version Several changes were carried out in this second version of the test suite, namely:

- Minor corrections to the normalisations of the source-side texts; we corrected a few wrong normalisations and typos and reverted some hypercorrections to make sure that certain grammatical variations due to dialectal differences were conserved (i.e. not over-normalising)
- Corrections to some references after a manual check, including making sure that emojis and emoticons were always included in the references (this was not the case previously, notably for the Russian translations). For certain target languages, we also complete the reference translations with alternations for multiple possible genders where appropriate (namely where the gender is underspecified given the available context of the Reddit post).
- Re-annotation of the normalisation spans according to a new annotation scheme that is organised hierarchically and structures the types differently.

The new annotation types can be found below. Detailed descriptions including examples can be found in Appendix A.

- **Punctuation, typographic conventions, symbols, etc.**: punct:diff, punct:norm, caps, slash_to_or, slash_to_and, slash_distribution, word_to_symbol, symbol_placement
- **Spacing**: spacing, spacing:camelcase
- **Phonetically similar spellings (including imitation of speech)**: phon, phon:char, phon:digits, phon:cute, phon:hesitate, phon:sound, phon:interjection
- **Other spelling variations (ergographic, expressiveness)**: elongation, devowelling, contraction, truncation, acronym, abbreviation
- **Spelling mistakes**: spell, spell:charswap
- **Misc**: digit_letter_sim, letter_to_digit, suffix
- **Added and dropped words**: word_drop, word_drop:pronoun, word_drop:det, word_add, word_add:det, symbol_drop, symbol_add
- **Grammar**: inflection, grammar, grammar:v, grammar:v:inflect
- **Lexical changes**: lex_choice, surrounding_emphasis, emoticon, censored

[1] https://huggingface.co/datasets/rbawden/RoCS-MT-v2

3 WMT 2025 submissions

There were 56 submissions to the 2025 general task (including variants of the same systems and systems run by the general MT task organisers) that translated the test suite. A range of architectures were used, with a majority using LLMs. In a bid to encourage document-level translation, one of the important factors to be taken into account in MT evaluation (Läubli et al., 2018), this year's general MT task focused on document-level MT, where documents were typically paragraphs of text. For the test suites, individual segments as provided in the test suite were concatenated to form source documents to be translated in one go. RoCS-MT was provided in two different formats: (i) manual segmentation and (ii) segmentation purely based on newlines within posts.

The language directions of the 2025 shared task overlap somewhat with the 2023 language directions (e.g. English to Czech, Chinese, Japanese, Ukrainian and Russian), although not all target languages for which we have reference translations are present (e.g. French and German), and many of the language directions are new and therefore do not have references (e.g. Arabic, Bhojpuri, Estonian, Icelandic, Italian, Korean, Massai, Serbian). We therefore choose in our initial results and analysis (see Section 4) to concentrate on quality estimation (QE) (i.e. without using the references for automatic analysis). We use two different metrics to calculate the scores (at the document level) used for the rankings: (i) CometKiwi (Rei et al., 2022),[2] and (ii) MetricX (Juraska et al., 2024).[3] This is to avoid bias towards a single metric, especially as many systems optimise for a particular QE metric. We acknowledge however that (i) these metrics may well have issues handling certain languages, particularly those not explicitly included in the training data of the underlying models, and (ii) the scores may well favour models that optimise for QE in general, even if the same QE model is not used.

4 Results and Analysis

In this section, we present results for the submitted systems along with several brief analyses. In order to compute rankings, we took into account the fact that not all systems took part in every language pair, and adopted the ranking algorithm commonly used for nations in the Olympic Games, which addresses a comparable situation. We observed that if systems are ranked based on their absolute scores for each language pair, a few systems are consistently ranked first, meaning that most systems never achieve a rank of 1 or 2, even if they have relatively high overall scores. The result of this is that an overall weaker system that has low scores in most languages but happens to be ranked highly for a single language pair can be ranked higher than a system that has high but not the best scores across all language pairs. We therefore choose to apply the "gold first" algorithm on quartile-based rankings instead of raw rankings; for a given language pair, all systems within the top 10% of scores are assigned to rank 1 and treated as such by the "gold first" algorithm, the next 10% to rank 2, and so on. In other words, systems are then ordered according to the number of language pairs in which they achieve rank 1 (i.e. belong to the top 10%). In the event of a tie, the number of rank 2 placements (top 10–20%) is considered, followed, if necessary, by the number of rank 3 placements (top 20–30%), and so on.

We applied this approach to get rankings from each of the two metrics used. These rankings are based on the original inputs (i.e. before normalisation) that have been manually segmented into sentences. We then computed overall rankings based on the average of the two rankings. Table 1 displays both metric-specific rankings and the overall ranking, the number of first-, second- and third-ranks for each system and each metric and the absolute gap between each system's CometKiwi-based and MetricX-based ranks, to get an idea of how consistent they are and the confidence we can place in the resulting rankings. Appendix B provides raw CometKiwi and MetricX scores per language pair, first comparing scores on original and normalised inputs, then comparing scores on manual and newline segmentations applied to original inputs.

Results highlight a small group of systems that dominate performance. Yolu occupies the top position with nine first-place results across all twelve language pairs on both metrics, followed closely by Shy-hunyuan-MT and CommandA-WMT. Laniqo and SalamandraTA follow, despite the fact that Laniqo participated in only seven pairs. Among the organiser-run systems, GPT-4.1 ranks 7th, closely followed by ONLINE-B and both TowerPlus models. Several larger LLM-based baselines, such as Claude-4 (18th) and both Gemini models (also

[2] WMT22-COMETKIWI-DA model. Higher is better.

[3] METRICX-24-HYBRID-XL-V2P6 model. Lower is better.

System	#lp	CometKiwi		MetricX		Overall rank	Δrank
		Rank	"Medals"	Rank	"Medals"		
Yolu	12	1	9, 1, 0	3	9, 1, 1	1	2
Shy-hunyuan-MT	12	4	6, 4, 2	1	11, 0, 1	2	3
CommandA-WMT	12	2	7, 3, 1	4	8, 2, 0	3	2
Laniqo$^{\diamond}$	7	5	6, 1, 0	5	4, 3, 0	4	0
SalamandraTA	11	3	7, 2, 1	8	1, 0, 1	5	5
GemTrans	12	12	1, 2, 1	2	10, 1, 0	6	10
UvA-MT	12	7	2, 4, 4	16	0, 5, 3	7	9
*GPT-4.1	12	17	0, 4, 1	6	1, 8, 1	7	11
*ONLINE-B	11	8	2, 1, 4	20	0, 2, 2	9	12
*TowerPlus 9B	12	9	1, 5, 2	21	0, 2, 0	10	12
*TowerPlus 72B	12	11	1, 2, 3	22	0, 2, 0	11	11
SRPOL	7	6	3, 4, 0	28	0, 1, 0	12	22
IR-MultiagentMT	12	23	0, 1, 1	17	0, 4, 4	13	6
TranssionTranslate	12	10	1, 2, 4	33	0, 0, 2	14	23
*CommandA	12	18	0, 2, 4	26	0, 1, 1	15	8
NNTSU	1	32	0, 0, 1	12	1, 0, 0	15	20
Erlendur	1	32	0, 0, 1	12	1, 0, 0	15	20
In2x	1	15	1, 0, 0	30	0, 1, 0	18	15
*Claude4	12	21	0, 1, 2	24	0, 1, 2	18	3
*DeepSeek V3	12	22	0, 1, 2	23	0, 1, 7	18	1
Algharb$^{\diamond}$	12	26	0, 1, 0	19	0, 2, 3	18	7
*Gemini 2.5 Pro	12	38	0, 0, 0	7	1, 1, 7	18	31
*Gemma 3 27B	12	28	0, 0, 1	18	0, 3, 4	23	10
*AyaExpanse 32B	12	13	1, 0, 2	35	0, 0, 1	24	22
*AyaExpanse 8B	12	20	0, 2, 0	29	0, 1, 0	25	9
KIKIS	1	39	0, 0, 0	12	1, 0, 0	26	27
*EuroLLM 22B	12	19	0, 2, 0	36	0, 0, 1	27	17
*Gemma 3 12B	12	25	0, 1, 0	32	0, 0, 2	28	7
Yandex	1	46	0, 0, 0	12	1, 0, 0	29	34
Systran$^{\diamond}$	1	15	1, 0, 0	44	0, 0, 0	30	29
*Llama 3.1 8B	12	14	1, 0, 0	47	0, 0, 0	31	33
Kaze-MT$^{\diamond}$	12	52	0, 0, 0	9	1, 0, 0	31	43
KYUoM$^{\diamond}$	12	52	0, 0, 0	10	1, 0, 0	33	42
ctpc_nlp	12	52	0, 0, 0	11	1, 0, 0	34	41
Wenyiil$^{\diamond}$	12	30	0, 0, 1	34	0, 0, 2	35	4
*Mistral-Medium	9	40	0, 0, 0	25	0, 1, 1	36	15
*CommandR	12	24	0, 1, 1	43	0, 0, 0	37	19
*Qwen3 235B	12	41	0, 0, 0	27	0, 1, 1	38	14
*ONLINE-W	8	27	0, 1, 0	42	0, 0, 0	39	15
AMI$^{\diamond}$	1	32	0, 0, 1	37	0, 0, 1	39	5
*EuroLLM 9B	12	29	0, 0, 1	41	0, 0, 0	41	12
IRB-MT	12	42	0, 0, 0	31	0, 0, 3	42	11
*Llama-4-Maverick	12	37	0, 0, 0	38	0, 0, 0	43	1
CUNI-MH-v2	1	32	0, 0, 1	46	0, 0, 0	44	14
bb88	1	32	0, 0, 1	49	0, 0, 0	45	17
*NLLB	12	44	0, 0, 0	39	0, 0, 0	46	5
*Mistral 7B	12	31	0, 0, 1	53	0, 0, 0	47	22
DLUT_GTCOM	2	45	0, 0, 0	40	0, 0, 0	48	5
CUNI-SFT	3	48	0, 0, 0	45	0, 0, 0	49	3
TranssionMT	8	43	0, 0, 0	51	0, 0, 0	50	8
*Qwen 2.5	12	47	0, 0, 0	48	0, 0, 0	51	1
CGFOKUS	1	51	0, 0, 0	49	0, 0, 0	52	2
*ONLINE-G	10	49	0, 0, 0	52	0, 0, 0	53	3
SH	1	50	0, 0, 0	54	0, 0, 0	54	4
CUNI-DocTransformer	1	55	0, 0, 0	55	0, 0, 0	55	0
COILD-BHO	1	55	0, 0, 0	55	0, 0, 0	55	0

Table 1: Main ranking table. For each system and for both the CometKiwi and MetricX metrics, we provide their rank according to the metric, computed using the "gold first" scoring algorithm (Rank), the number of language pairs for which the system ranked first, second and third ("Medals"). For each system we also provide the number of language pairs it participated in (#lp), an overall rank based on the average between the CometKiwi- and the MetricX-based ranks, and the difference between the two original ranks, which shows for each system how consistent the two metrics are. Systems run by the task organisers are marked with an asterisk, while systems fine-tuned to optimise a Qe metric, such as CometKiwi and MetricX, are indicated with a $^{\diamond}$.

18th and 23rd, respectively), appear in the middle of the table, while Llama-4-Maverick, Mistral 7B, Qwen 2.5 and ONLINE-G fall into the lower ranks. NLLB, a widely used dedicated MT model, ranks only 46th, with consistently low ranks in both metric-specific rankings (44th and 39th). Conversely, several single-pair submissions are quite successful, achieving first place according to one of the metrics, and sometimes second or third according to the other one (NNTSU, Erelendur and In2x reach a joint overall 15th place). Overall, the results indicate that dedicated MT systems continue to outperform many general-purpose LLMs when evaluated using CometKiwi and MetricX.

These results are somewhat surprising, particularly the relatively low performance of several large LLMs—Qwen3 235B ranks only 38th overall, and GPT-4.1 is ranked 17th according to the CometKiwi-based ranking—and popular reference models such as NLLB, despite generally being evaluated quite highly in MT tasks, whether by automatic metrics or human evaluation. Three main factors may account for this outcome. First, strong performance on edited data does not necessarily translate into equally strong performance on non-standard data; success on the former is not a guarantee of robustness. Secondly, automatic evaluation metrics—especially CometKiwi, on which our ranking is based—may perform poorly when applied to translations of non-standard text. Thirdly, several systems used QE metrics for optimisation and may therefore have gained higher rankings than their actual quality would otherwise justify.[4] We leave these questions open for future investigation, and invite the reader to take our results and conclusions with a pinch of salt.

Another surprising observation is that several systems are positioned very differently in our two metric-specific rankings. The most extreme case is KazeMT, ranked 52nd using CometKiwi but 9th using MetricX. Another example of a large Δrank is Gemini 2.5 Pro, ranked 38th using CometKiwi but 7th using MetricX. Several single-pair submissions also display large discrepancies, such as Yandex and Systran, which both achieve first place according to one metric (MetricX for Yandex and CometKiwi for Systran), but do not perform as well according to the other metric. Such discrepancies could be explained, at least in some cases, in the way these models were trained or fine-tuned, for instance by optimising for QE, as mentioned above.

4.1 Original versus Normalised Texts

We first compare the impact of translating the original inputs (containing non-standard language) against the normalised inputs (both with manual segmentation). The full results are given in Appendix B (Tables 3 and 4). The scores for original texts are in general lower for CometKiwi and higher for MetricX than for the normalised ones. This is somewhat unsurprising for several reasons: (i) it is expected that more standard texts are easier to translate, as the majority of the texts that the models were trained on was standard, and the UGC texts are characterised by high levels of variation, (ii) metric scores are likely to penalise translations that are less standard. Concerning (i), there is some indication that is going on. For example, the difference between translations from normalised and original texts is very large for the lowest-resource language directions, at least for CometKiwi (English to Icelandic and to Maasai), showing that the models are struggling more with the non-standard texts. Concerning (ii), some further investigation is necessary here to ascertain whether the difference in scores are a property of the metrics themselves or whether they translate into real differences in translation quality. Our observations in Section 4.3 indicate that there is more going on than these basic scores and that we should not trust the metric scores alone.

4.2 Manual Segmentation versus Newline Segmentation

We then compare the impact of the text segmentation by looking at the scores based on inputs with manual segmentation and those separated on newlines (both with raw inputs). Results are shown in the same appendix section (Tables 5 and 6). The differences between the two segmentation types appears less than the differences previously observed between original and normalised inputs. In reality, given that the posts were given as complete documents, the segmentation has less of an impact than if the systems had been translating on the sentence level, as was the case for most systems in previous years.

4.3 A First Qualitative discussion

Table 2 shows the results of all systems (apart from those whose output was obviously the result of a

[4] We indicate systems that self-declare as using QE in some way with a $^\diamond$.

System	Text	CometKiwi	MetricX	tgt lang	elong tr
Source	*"politics used to be moooorrreeee poollliitte we neeeeed a return to The Discouuurrrse"*				
Qwen 2.5	politika byla dávno většinou politická, potřebujeme návrat k The Discouuurrrse.	45.4	17.4	+	−−−
Llama 3.1 8B	politika byla kdysi mnohem menší, potřebujeme návrat k diskuzi.	46.2	15.9	+++	−−−
ctpc_nlp	„politics used to be moooorrreeee poolllitte we neeeeed a return to The Discouuurrrse.“	46.3	17.4	−−−	−−−
Mistral 7B	politika byla dávno více politická, potřebujeme se vrátil do The Discouuurrrse.	47.7	16.9	+	−−−
GPT-4.1	„politika býýýývala mnohem slušnějšššííí, potřebujeme návrat k diskuuuziii.“	49.6	14.6	+++	++
CUNI-SFT	politika bývala moooorrreeee poollliitte, potřebujeme návrat k diskusi.	49.6	16.3	−	−−−
Gemma 3 27B	„politika bývala mooooc víc slušná, potřebujeme návrat k Diissskuuzi.“	51.3	10.9	+++	++
EuroLLM 9B	„politika bývala moooorrreeee poollliitte, potřebujeme návrat k Diskouuurrrse.“	51.4	18.0	−−	+
CUNI-DocTransformer	„politika bývala moooorrreeee poollliitte my jsme potřebovali návrat do The Discourrrse.“	51.8	17.9	−−	−−−
Gemini 2.5 Pro	„politika bejvávala slu-šně-ě-ě-jšíííí, po-tře-bu-je-meee se vrátit k Diskurzuuuu.“	52.8	11.5	++	+!
ONLINE-W	politics used to be moooorrreeee poollliitte we neeeeed a return to The Discouuurrrse.	52.8	17.7	−−−	−−−
Wenyiil	„politika bejváááaála kdysi slušněěěě, potřebíííijem návrat k Diskouuuurrrsu“	53.6	10.7	++(!)	+++
CUNI-MH-v2	politika bývala moooorrreeee poollliitte, potřebujeme návrat k Diskurzu.	53.6	18.4	−	−−−
Algharb	„politika bejvááááála kdysi slušněěěěějšššššíí, musíme se vráááátit k Diiiskuuurrrzu“	53.7	10.0	+++(!)	+++
AyaExpanse-32B	politika bývala víceeee čistááá, potřebujeme se vrátit k Diiskusi.	54.5	14.0	+++	++
SRPOL	„politika bývala moooorrreeee poollliitte, potřebujeme návrat k diskuusii.“	54.6	16.2	−	+
GemTrans	„politika bývala kdysi mnohem civilnější a potřebujeme návrat k seriózní debatě.“	54.7	4.6	+++	−−−
Tower Plus 72B	„politika bývala více politická, potřebujeme návrat k Diskurzu.“	55.1	10.8	+++	−−−
IRB-MT	politika bývala dřív mnohem uhlazenější, potřebujeme návrat k slušné konverzaci.	55.1	7.8	+++	−−−
Qwen3 235B	„dřív byla politika víc vstřícná, potřebujeme se vrátit ke civilizované debatě.“	55.3	9.6	+++	−−−
Gemma 3 12B	politika bývala dřív mnohem, mnohem politčtější, potřebujeme návrat k Diskuuzi.	55.8	10.8	+++	−−−
CommandA	„politika bývala moooc poolliittická, potřebujeme návrat k Diissccouurrrse.“	56.2	12.2	−	+
Yolu	politika dřív bývala mnohem... mnohem... polemickyjší a my potřebujeme návrat k společné debatě.	56.3	9.0	+++	+(!)
Shy-hunyuan-MT	„politika dřív byla mnohem slušnější, potřebujeme návrat k racionálnímu diskurzu“	57.3	7.6	+++	−−−
Claude-4	politika bývala vííííc sluuušnááá, potřebujeme se vrátit k Diskuuuurzu.	57.5	12.1	+++	+
DeepSeek-V3	„politika bývala slušnějšíííí, musíme se vrátit k diskurzuuuu“	57.5	9.8	+++	+
IR-MultiagentMT	politika bývala mnohem více otevřená, potřebujeme se vrátit k diskurzu.	57.7	10.1	+++	−−−
Mistral-Medium	politika bývala kdysiiiii moooooc civilizovanější, my potřebujeme návrat k Diskuuuuuursu.	58.0	9.0	+++	++
Laniqo	„politika se stala blbá a potřebujeme návrat ke klasickému politickému diskurzu.“	58.3	9.0	+++	−−−
UvA-MT	politika bývala mnohem politčtější, potřebujeme návrat k Diskuuzi.	58.5	8.9	+++	+
TowerPlus 9B	„politika bývala moooooc lepší, potřebujeme návrat k debatám.“	58.7	9.7	+++	+
Llama-4-Maverick	politika bývala mnohem uhlazenější, potřebujeme návrat k onomu Diskurzu.	58.7	10.1	+++	−−−
AyaExpanse 8B	politika byla kdysi zábavnější, potřebujeme se vrátit k diskusi.	60.1	12.1	+++	−−−
CommandA-WMT	„politika bývala víc politická, potřebujeme se vrátit k diskurzu“	61.9	6.8	+++	−−−
SalamandraTA	„politika bývala mnohem menší, takže jsme potřebovali návrat k The Discouuurrrse“	62.5	11.8	++	−−−
TranssionTranslate	„politika bývala mooooooočná, že potřebujeme návrat k The Discouurrse.“	64.8	21.6	++	+
ONLINE-B	„politika bývala mooooooodně dobrá, potřebujeme návrat k The Discouurrse.“	66.7	11.0	++	+
NLLB	Politika bývala moooorrreeee poollliitte, potřebujeme návrat do The Discouuurrrse	68.0	15.3	−−	−−−

Table 2: Example of character repetition linked to a mark of expressivity for en–cs (same source text as in (Bawden and Sagot, 2023) to illustrate 2023 en–de results). For each system we provide the CometKiwi score (multiplied by 100; higher is better) and the MetricX score (lower is better) for the corresponding document, as scores were computed at the document level. Systems are ordered by increasing CometKiwi score. The two last columns provide a manual assessment of how much of the input sentence was translated into Czech—or at least not kept in English—(“tgt lang”) and of how well the elongation phenomenon was transferred to the output sentence (“elong tr”; non-translated tokens are ignored). Systems whose outputs obviously result from an error are not included.

bug) on the example already used in (Bawden and Sagot, 2023) to illustrate the behaviour of MT systems in the presence of several instances of the elongation phenomenon, by which one or more characters are repeated to express emphasis. A first glance at the results shows that there is not necessarily a convincing correlation between perceived translation quality and the automatic evaluation provided by the CometKiwi metric, whereas MetricX results look slightly more correlated. Looking more closely at the translations, two main observations can be made:

- Firstly, a number of systems tend to keep unchanged original English tokens that have undergone elongation, and sometimes even the whole input. The fact that the two last tokens are capitalised in the input sentence makes it even more difficult for most systems to actually try to translate them.
- Secondly, not all systems attempt to transfer the elongation phenomenon into their output. Some seem to (try to) produce standard Czech rather than preserving the non standard expressivity mark. Some even try to render the same expressivity using another non standard phenomenon.

To better understand what is at play here, we decided to manually annotate these translations for two features: how much of the input sentence was (tentatively) translated into Czech, and how much of the elongation phenomenon was transferred into the output sentence (ignoring tokens kept in their original English form). Comparing these annotations with system types is interesting. Although a single example is in no way sufficient to allow for any generalisations, it seems that generic LLMs are more liable to preserving elongation and, more generally, to produce better translations, whereas dedicated MT models seem to produce more standard outputs and/or not to translate significant parts of the input. Interestingly, this is not reflected in the CometKiwi scores, but it is more visible in the MetricX scores. For instance, the best CometKiwi-scored translation contains two segments that are still in English, a situation that invariably leads to bad (high) MetricX scores. However, CometKiwi ad MetricX seem consistently bad at penalising the absence of elongations in the produced translation. The best CometKiwi-scored translation does not contain any elongation in genuinely Czech tokens, and the best Metric-X-scored translation, which is perfect Czech, does not include any elongation whatsoever. On the contrary, the output of GPT-4.1 is good in both regards—it is entirely in Czech and does contain elongations—, and is a good translation, but it is scored poorly by both CometKiwi and MetricX. This shows that modern metrics such as CometKiwi and MetricX might not be reliable when it comes to assessing translation quality of non-standard content. We leave a more quantitative and systematic exploration of these questions and their implications for MT evaluation in general to future work.

Although the test suite this year presents new annotations for the non-standard phenomena present in the test suite that are more consistent and interesting for analysis, we also leave the analysis on a per-phenomenon basis to future work, in which we will go into more detail and length.

5 Conclusion

We have presented a new version (v2) of the RoCS-MT challenge set, first presented at the WMT 2023 test suites shared task track. This 2025 edition has several improvements, with minor corrections to source texts, some corrections to references and improved categorisation of non-standard phenomena. We describe these changes and also use the challenge set to compare systems submitted to this year's shared task, comparing translation from the original UGC inputs and their manually normalised versions. A major difference with previous years of the shared task is a switch to document-level MT, so whole chunks of posts were submitted to systems for translation. We nevertheless compare two different segmentation types (to see if initially manually segmenting into sentences and then concatenating the sentences with newlines could help translation) and discuss preliminary insights into the shortcomings of popular metrics such as CometKiwi and MetricX when applied on non-standard text MT.

Acknowledgments

This work was partly funded by both authors' chairs in the PRAIRIE institute, now PRAIRIE-PSAI, funded by the French national agency ANR, respectively as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and as part of the "France 2030" strategy under the reference ANR-23-IACL-0008.

References

Tyler Baldwin and Yunyao Li. 2015. An in-depth analysis of the effect of text normalization in social media. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 420–429, Denver, Colorado. Association for Computational Linguistics.

Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216, Singapore. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Jennifer Foster. 2010. "cba to check the spelling": Investigating Parser Performance on Discussion Forum Posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384, Los Angeles, California. Association for Computational Linguistics.

Ryo Fujii, Masato Mita, Kaori Abe, Kazuaki Hanawa, Makoto Morishita, Jun Suzuki, and Kentaro Inui.

2020. PheMT: A phenomenon-wise dataset for machine translation robustness on user-generated contents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5929–5943, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186, Sofia, Bulgaria. Association for Computational Linguistics.

Paul McNamee and Kevin Duh. 2022. The multilingual microblog translation corpus: Improving and evaluating translation of user-generated text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 910–918, Marseille, France. European Language Resources Association.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Hamdy Mubarak, Sabit Hassan, and Ahmed Abdelali. 2020. Constructing a bilingual corpus of parallel tweets. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 14–21, Marseille, France. European Language Resources Association.

Lydia Nishimwe, Benoît Sagot, and Rachel Bawden. 2024. Making sentence embeddings robust to user-generated content. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10984–10998, Torino, Italia. ELRA and ICCL.

Jungsoo Park, Mujeen Sung, Jinhyuk Lee, and Jaewoo Kang. 2020. Adversarial subword regularization for robust neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1945–1953, Online. Association for Computational Linguistics.

Ben Peters and Andre Martins. 2025. Did translation models get more robust without anyone Even noticing? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2445–2458, Vienna, Austria. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2019. Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 2–14, Turku, Finland. Linköping University Electronic Press.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a Treebank of Noisy User Generated Content. In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India. The COLING 2012 Organizing Committee.

Henny Sluyter-Gäthje, Pintu Lohar, Haithem Afli, and Andy Way. 2018. FooTweets: A bilingual parallel corpus of world cup tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Rob van der Goot, Rik van Noord, and Gertjan van Noord. 2018. A taxonomy for in-depth evaluation of normalization for user generated content. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 684–688, Miyazaki, Japan. European Language Resources Association (ELRA).

Iñaki San Vicente, Iñaki Alegría, Cristina España-Bonet, Pablo Gamallo, Hugo Gonçalo Oliveira, Eva Martínez Garcia, Antonio Toral, Arkaitz Zubiaga, and Nora Aranberri. 2016. TweetMT: A parallel microblog corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2936–2941, Portorož, Slovenia. European Language Resources Association (ELRA).

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. In *Proceedings of the Forty-First International Conference on Machine Learning*.

A Normalisation Span Classification

The original and normalised texts were aligned and different non-standard phenomena classified. Below is the list of normalisation categories with examples.

Punctuation, typographic conventions, symbols, etc.

- **punct:diff**: extra punctuation is included or necessary punctuation is removed (e.g. missing final punctuation, missing apostrophes, commas, etc.).
  e.g. im→I'm
- **punct:norm**: punctuation to be normalised according to certain conventions (e.g. same apostrophes and quotes).
  e.g. that's→that's
- **caps**: capitalisation differs from what is considered standard (e.g. lowercase initial characters, all uppercase, etc.).
  e.g. IM SO HAPPY→I'm so happy
- **slash_to_or**: a slash is used, where in normalised speech an "or" would be used to represent a list of items. This applies to the whole list, including where etc. is included. Not that this does not include cases where the items are alternatives in the discourse
  e.g. cat/exhaust/etc→cat or exhaust, etc.
  e.g. truth/dare→truth or dare
  e.g. [counter-example] AW WELL MY DOG/CHILD IS VERY FRIENDLY SO LET ME APPROACH→aw, well my dog/child is very friendly, so let me approach
- **slash_to_and**: a slash is used, where in normalised speech an "and" would be used. This applies to the whole list, including where etc. is included.
  e.g. work/paint→work and paint
- **slash_distribution**: the use of a slash to separate two items where the slash does not separate two complete items (i.e. part of one element is distributed to both items thanks to the slash). An example makes this easier to understand:
  e.g. just disrespects any / everyone→just disrespects any / everyone
- word_to_symbol: the use of a symbol to represent a word
  e.g. +→and, &→and
  e.g. →around
  e.g. $$$→money

- **symbol_placement**: non-standard placement of a symbol with respects to English norms.
e.g. 100$→$100

Spacing

- **spacing:** missing or added spacing in the original text
e.g. aswell→as well
e.g. over thinking→overthinking

- **spacing:camelcase:** the use camelcase (capital letters at the beginning of words) instead of using spaces
e.g. surroundedUs→surrounded us
e.g. sawThat→saw that

Phonetically similar spellings (including imitation of speech)

- **phon:** the word uses a variant of spelling based on the phonetic similarity of the sequence of characters. This also includes the use of individual letters to represent words or syllables because of an equivalence in their pronunciation (u→you, b→be, c→see).
e.g. saturday sesh→Saturday session (also a case of truncation)
e.g. sup→What's up (also truncation)
e.g. bcos -> because
e.g. n→and
e.g. tho→though
e.g. speakin→speaking

- **phon:char:** a character is used in the place of a word or syllable because of its phonetic similarity with the word or syllable
e.g. b→be
e.g. c→see
e.g. u→you

- **phon:digits:** a digit is used in the place of a word or part of a word.
e.g. m8→mate
e.g. 2→to
e.g. as1 that will play→as one that will play (in a context where 1 could be incorrect, otherwise this should not be normalised)

- **phon:cute:** the spelling of a word to indicate "cute" or babyish pronunciation, e.g. using 'w' to replace an initial letter
e.g. wecommended→recommended

- **phon:hesitate:** words that are written in a way to imitate hesitation
e.g. terribl-. . . .yyy→terribly
e.g. Y-y-yyy-yes→Yes

- **phon:sound:** the case of words that are used to indicate a sound (very rare)
e.g. bRRrrRRrrRRrr→brr rrr rrr

- **phon:interjection:** interjections that are normalised to single (and more standard) variations
e.g. bla→blah
e.g. URGHHH→ugh
e.g. Nawh→no

Other spelling variations (ergographic, expressiveness)

- **elongation:** characters are repeated, usually as a a mark of expressiveness.
e.g. *meeeeelltiiiinggg*→melting
e.g. sooo→so

- **devowelling:** a word with the vowels removed (initial vowels are often kept however). This can often result in double characters being reduced to single ones (messages→msgs) In this category are also words where part of the word has been devowelled.
e.g. wt→what, ovr→over, ppl→people
e.g. askd→asked (initial vowel kept)
e.g. travllr→traveller

- **contraction:** when several words are contracted into a single one. This has some overlap with the characteristics of phonetic distance, in that it is due to the pronunciation of the words that the contraction occurs.
e.g. gonna→going to
e.g. innit→isn't it?

- **truncation:** a word is shortened, either at the end (traditional truncation) or sometimes at the beginning, often by removing a syllable or a suffix. Note the difference with acronymisation, which involves keeping initial characters.
e.g. sesh→session
e.g. cuz→because, till→until
e.g. ofc→of course -> CHANGED, NOW ACRONYM
e.g. w→with -> CHANGED, NOW ACRONYM

- **acronym:** a word or sequence of word is represented as an acronym, i.e. the initial characters of the word (or syllables) are retained and the others are elided.
  e.g. RN→right now
  e.g. gf→girlfriend
  e.g. never mind→nvm
  e.g. w→with
  We also include in this category words that are partially acronymised (i.e. where one syllable is represented by its initial but the rest is not).is acronymised but the rest is not.
  e.g. oline→offensive line
  e.g. gmeet -> Google meet
  e.g. bday→Birthday
  e.g. ofc→of course
  Note that sometimes slashes are included in the acronym
  e.g. b/c→because,
  e.g. w/o→without.

- **abbreviation:** abbreviations for units of measurement and other standard cases
  e.g. ft→feet, 2k→2000, hrs→hours
  e.g. Ex→for example

Spelling mistakes (distinguished from spelling variation identified as being intentional)

- **spell:** the word contains a spelling error that is not clearly intentional (covered by the other phenomena such as truncation, devowelling, etc.) and not covered by the other more specific categories.

- **spell:charswap:** the characters in the word are present but not in the right order (most often consecutive characters being swapped)
  e.g. nobel→noble
  e.g. furhter→futher

Misc

- **digit_letter_sim:** Very rare, but where a digit is used in the place of a letter due to the typographic similarity (see in 3ver→ever).

- **letter_to_digit:** Very rare, but where a digit is used in place of a letter not because of their typographic similarity, but because as a sort of tautology (seen in 1nce→Once).

- **suffix:** the addition of a suffix to a word, either as a diminutive or other
  e.g. lolsky→lol
  e.g. meanie→mean
  e.g. doggy→dog

Added and dropped words

- **word_drop:** a word is not present in the original text and is present in the normalised version
  e.g. It also confusing. . . →It's also confusing
  e.g. u wanna see?→Do you want to see?

- **word_drop:pronoun:** the original text omits a pronoun (often the case of subject pronouns at the beginning of sentences) that is included in the normalised version.
  e.g. Was gunna try distortion. . . →I was going to try distortion. . .

- **word_drop:det:** the original text omits an article (e.g. the or a) that is included in the normalised version.
  e.g. Pretty creative way. . . →A pretty create way. . . word_add: a word is present in the original text and is removed in the normalised version
  e.g. . . . in ten days ago→. . . ten days ago
  e.g. also for uses of word "like" as a filler

- **word_add:det:** the original text includes an article where the normalised version removes it
  e.g. . . . adds an 12kg of salt→. . . adds 12kg of salt symbol_drop: the original text omits a symbol that is included in the normalised version.
  e.g. 32c→32°C

- **symbol_add:** the original text includes a symbol that is removed in the normalised version.
  . . . no issue w being over 12+ ft. . . →. . . no issue with being over 12 feet. . .

Grammar

- **inflection:** a word is not correctly inflected (e.g. with respect to number, tense, etc.)
  e.g. . . . wondering what ppl thought are→. . . wondering what people's thoughts are

- **grammar:** inflection-related errors
  e.g. . . . wondering what ppl thought are→. . . wondering what people's thoughts are
  e.g. if your good→if you're good

- **grammar:v:**
- **grammar:v:inflect**

Lexical changes

- **lex_choice:** a use of a non-standard lexical choice, including dialectisms (e.g. cannae, ain't), malapropisms (e.g. genually), foreign words and generally wrong choices of words (e.g. wrong part of speech, wrong semantic choice of words, lacking punctuation, use of an antonym by accident, etc.)
  e.g. I am confusion→I am confused
  e.g. genually→genuinely
  e.g. pish→piss
  e.g. ain't→aren't
  e.g. cannae→cannot
  e.g. y'all→everyone/all/all your (depending on the context)
  e.g. sans guac→without guacamole

- **surrounding_emphasis:** emphasis added to certain words typographically (removed in the normalised variants).
  e.g. *without*→without
  e.g. ∼find∼→find

- emoticon: emoticon that is a variant on the common emoticons :-), :-D, :-(, :-/ and >:-)
  e.g. :-/////→:-/
  e.g. (:→:-)
  e.g. :)̂→:-)

- censored: the word contains symbols in an effort to censor the word
  e.g. upv*te→upvote
  e.g. s**t→shit, sh**→shit

B Raw Automatic Scores per Language pair

The raw scores (calculated at the document level) can be found in this appendix section. In each of the tables, the systems are ordered by the ranking across all languages for that particular metric (as described in Section 4). Note that higher CometKiwi scores are better and lower MetricX scores are better.

Tables 3 and 6 provide the CometKiwi and Metricx scores respectively for manually segmented texts, with a comparison of original and normalised input texts.

Tables 5 and 6 provide the CometKiwi and Metricx scores respectively for original inputs, with a comparison of manually segmented and newline-segmented texts.

System	Rank	en-ar_EG		en-bho_IN		en-cs_CZ		en-et_EE		en-is_IS		en-ja_JP		en-ko_KR		en-mas_KE		en-ru_RU		en-sr_Latn_RS		en-uk_UA		en-zh_CN	
		norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig
Yolu	1	78.5	73.4	71.9	68.9	82.1	76.3	84.8	79.6	35.9	22.7	83.7	79.7	83.3	79.3	35.9	22.7	81.1	76.0	83.8	78.5	80.7	75.5	80.7	76.1
CommandA-WMT	2	71.4	65.6	75.5	72.8	80.9	75.1	83.0	77.8	75.0	70.0	82.7	78.5	82.5	78.1	65.5	62.6	80.0	74.7	80.6	75.0	80.0	75.3	79.6	74.7
SalamandraTA	3	72.1	66.8	60.7	57.4	81.9	76.0	84.7	78.9	77.6	72.2	82.1	78.4	81.1	77.1	–	–	81.1	75.5	83.7	78.3	80.5	75.1	80.1	75.4
Shy-hunyuan-MT	4	75.8	70.7	80.5	76.6	80.6	74.8	83.2	77.6	77.5	71.4	82.1	77.7	81.5	76.9	55.9	52.2	79.4	73.9	82.8	77.0	79.3	73.7	79.2	73.9
Laniqo	5	–	–	–	–	80.8	74.7	83.9	78.0	–	–	82.4	78.3	82.2	78.1	–	–	80.3	74.7	–	–	79.9	74.6	79.7	74.7
SRPOL	6	76.8	71.2	–	–	81.0	74.4	83.1	77.4	–	–	82.5	78.5	–	–	–	–	79.7	74.1	–	–	79.3	73.8	79.3	74.2
UvA-MT	7	71.6	66.4	73.0	68.2	79.9	73.8	81.3	75.7	71.9	67.9	82.3	78.2	82.1	77.6	37.3	37.1	79.5	73.8	81.9	76.7	78.5	73.5	79.2	74.4
*ONLINE-B	8	76.4	71.4	60.0	54.5	79.5	73.1	82.1	75.8	76.7	71.4	82.3	78.3	81.2	76.6	–	–	78.7	73.3	80.3	74.5	78.3	72.1	79.1	73.8
*TowerPlus 9B	9	48.9	46.0	79.1	75.8	79.6	72.9	62.3	57.9	76.2	70.8	81.7	77.5	81.6	77.2	43.6	41.3	79.0	73.7	63.8	60.2	78.3	72.8	78.9	73.6
TranssionTranslate	10	67.5	61.7	59.9	54.9	79.5	72.9	82.2	76.0	76.5	71.1	82.8	78.5	81.9	77.1	35.9	22.7	79.4	73.4	76.6	65.3	78.1	71.7	79.1	73.8
*TowerPlus 72B	11	66.4	60.4	79.7	76.2	78.8	72.7	72.6	67.4	75.5	70.4	81.6	77.1	81.4	77.0	44.0	38.7	79.1	73.4	73.6	68.8	78.2	72.5	79.0	74.1
GemTrans	12	75.7	70.3	80.4	76.9	78.9	72.3	81.3	74.8	72.7	67.5	80.7	76.2	80.4	75.7	36.5	34.3	78.3	72.4	81.2	75.1	78.0	72.7	78.0	73.1
*AyaExpanse 32B	13	66.0	60.5	65.2	62.8	78.2	72.7	54.1	50.9	50.9	47.8	81.6	77.0	81.3	76.4	46.5	43.2	78.2	73.1	72.9	68.3	77.6	72.4	77.9	72.6
*Llama 3.1 8B	14	57.2	52.0	66.3	62.1	72.9	66.7	67.2	62.6	56.4	53.4	75.7	72.2	76.8	72.0	50.9	48.8	74.8	69.2	73.6	68.1	73.3	68.2	76.4	70.9
Systran	15	–	–	–	–	–	–	–	–	–	–	84.3	81.0	–	–	–	–	–	–	–	–	–	–	–	–
In2x	15	–	–	–	–	–	–	–	–	–	–	82.8	78.7	–	–	–	–	–	–	–	–	–	–	–	–
*GPT-4.1	17	62.7	57.9	62.4	59.6	79.8	73.2	82.5	76.5	76.2	70.5	81.6	77.2	81.5	76.8	34.6	32.2	78.6	72.6	82.0	76.2	78.2	72.5	78.5	73.1
*CommandA	18	64.1	59.3	63.0	60.2	79.5	73.5	77.3	71.3	67.9	63.7	82.1	77.7	81.9	77.4	42.1	39.9	78.7	73.1	79.5	74.5	78.2	72.5	78.9	73.8
*EuroLLM 22B	19	72.1	66.0	77.2	73.6	79.0	72.5	81.9	76.1	45.7	43.1	81.3	76.7	81.2	75.9	39.7	36.7	78.5	73.0	79.3	74.2	77.7	72.1	78.3	73.1
*AyaExpanse 8B	20	74.6	69.7	74.7	71.1	77.7	71.4	39.3	37.5	39.6	37.4	81.0	76.7	80.9	76.2	39.9	38.2	77.6	72.0	60.7	56.8	77.4	71.9	77.6	72.5
*Claude4	21	64.8	59.4	63.1	59.5	79.7	72.8	81.6	74.9	75.9	69.4	82.2	77.5	81.7	76.9	39.8	37.6	78.7	72.3	81.1	75.3	78.2	72.2	78.6	73.0
*DeepSeek V3	22	64.6	58.9	65.1	62.5	78.9	72.4	81.5	75.3	74.5	68.8	81.2	76.4	80.8	75.7	38.5	35.4	77.8	72.2	81.0	75.4	77.0	71.7	76.5	70.5
IR-MultiagentMT	23	68.2	63.7	63.4	61.5	78.3	72.5	80.7	74.9	75.0	69.6	81.3	76.4	80.7	76.5	38.4	36.2	78.0	73.0	80.9	75.2	77.7	72.5	77.9	72.9
*CommandR	24	72.8	66.5	66.0	62.2	74.2	67.8	42.4	39.5	41.1	38.7	79.9	75.3	79.2	74.6	44.6	41.2	72.4	66.1	61.6	57.2	72.5	66.9	76.7	71.4
*Gemma 3 12B	25	62.3	57.5	59.3	57.2	77.9	71.5	78.9	72.7	69.2	64.3	80.3	75.5	79.5	74.8	43.9	42.1	77.7	72.6	79.6	74.0	77.4	72.4	77.6	72.5
Algharb	26	74.8	69.2	59.1	56.7	78.1	71.6	81.4	74.9	35.9	22.7	80.5	75.8	80.0	75.1	35.9	22.7	76.9	70.6	80.9	74.7	76.8	70.9	77.2	71.3
*ONLINE-W	27	74.6	68.7	–	–	79.1	68.9	80.5	68.0	–	–	79.6	73.8	81.4	75.4	–	–	79.0	72.5	–	–	78.1	72.1	78.3	73.0
*Gemma 3 27B	28	63.9	59.1	62.4	59.8	78.5	72.5	80.9	75.1	73.2	68.1	81.0	76.5	80.2	75.4	35.3	32.6	78.1	72.4	80.7	75.0	77.7	72.5	77.8	72.5
*EuroLLM 9B	29	70.3	64.7	69.9	65.4	78.7	72.4	81.2	75.0	43.2	41.5	80.3	75.9	80.3	75.6	32.5	28.5	77.8	72.3	77.6	70.6	77.2	71.6	77.4	72.0
Wenyiil	30	74.0	67.6	58.2	55.6	76.8	68.7	79.8	72.4	35.9	22.7	79.4	74.3	78.9	73.3	35.9	22.7	76.0	69.0	79.9	72.8	76.2	69.8	76.6	70.4
*Mistral 7B	31	48.6	45.0	55.6	52.8	64.7	59.4	41.5	39.1	42.5	40.4	71.2	67.9	71.5	67.6	41.2	40.1	71.2	66.2	68.9	64.6	70.1	65.4	71.2	65.9
CUNI-MH-v2	32	–	–	–	–	79.2	73.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
NNTSU	32	–	–	–	–	–	–	–	–	–	–	82.3	78.2	–	–	–	–	–	–	–	–	–	–	–	–
bb88	32	–	–	–	–	–	–	–	–	–	–	82.1	77.8	–	–	–	–	–	–	–	–	–	–	–	–
AMI	32	–	–	–	–	–	–	–	–	75.4	69.4	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Erlendur	32	–	–	–	–	–	–	–	–	75.0	69.1	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama-4-Maverick	37	62.8	57.7	60.2	57.3	78.5	71.8	81.3	74.4	72.9	67.0	80.4	75.8	80.1	75.4	37.2	34.8	78.4	72.6	80.6	74.7	78.0	72.3	78.1	72.4
*Gemini 2.5 Pro	38	58.9	54.8	59.4	56.5	77.8	70.7	80.7	74.4	74.6	68.8	80.1	75.1	79.3	74.0	37.6	35.2	76.3	69.7	80.4	74.3	76.1	70.2	76.4	70.2
KIKIS	39	–	–	–	–	–	–	–	–	–	–	81.9	77.3	–	–	–	–	–	–	–	–	–	–	–	–
*Mistral-Medium	40	65.6	60.3	63.3	60.6	79.2	72.5	79.9	73.3	72.2	66.9	81.6	76.7	81.2	76.2	–	–	–	–	–	–	78.0	72.0	78.3	72.8
*Qwen3 235B	41	64.7	59.0	62.8	60.4	77.7	71.3	76.2	69.7	68.4	63.5	81.2	76.9	80.4	76.0	38.9	36.4	78.3	72.2	78.6	72.5	76.8	71.2	78.6	73.0
IRB-MT	42	61.1	56.3	59.5	56.3	76.5	70.5	77.8	72.0	68.9	64.6	78.8	74.2	78.2	72.9	38.5	36.6	76.6	71.1	79.0	73.4	76.2	70.6	76.1	70.1
TranssionMT	43	67.5	60.6	59.4	57.4	68.9	58.2	71.9	63.0	–	–	–	–	–	–	38.8	36.4	71.9	64.1	78.5	71.3	72.0	65.3	–	–
*NLLB	44	70.0	63.3	62.0	58.4	77.2	68.9	79.0	71.6	68.8	63.5	74.0	68.1	77.6	71.4	23.9	22.5	76.2	69.1	23.9	22.5	75.2	67.6	64.6	59.6
DLUT_GTCOM	45	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	78.7	72.6	–	–	77.9	71.5	–	–
Yandex	46	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	78.2	72.6	–	–	–	–	–	–
*Qwen 2.5	47	54.7	50.2	61.3	58.9	65.5	59.4	49.8	46.8	42.9	40.4	76.9	72.5	71.8	67.2	36.7	33.8	72.1	67.0	61.3	56.8	63.6	59.5	76.8	71.2
CUNI-SFT	48	–	–	–	–	78.3	70.7	–	–	–	–	–	–	–	–	–	–	–	–	77.9	71.5	76.9	70.7	–	–
*ONLINE-G	49	62.9	53.5	–	–	73.2	59.9	74.4	61.6	68.6	59.7	71.9	61.2	68.2	57.3	–	–	78.8	72.2	76.8	70.7	77.1	71.0	71.9	62.9
SH	50	–	–	–	–	–	–	–	–	–	–	80.2	75.6	–	–	–	–	–	–	–	–	–	–	–	–
CGFOKUS	51	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	76.6	70.7	–	–
Kaze-MT	52	32.1	27.5	32.0	27.6	32.5	27.9	32.5	28.1	32.3	27.7	33.3	28.7	33.5	29.0	33.1	28.5	32.8	28.2	33.1	28.4	33.1	28.5	32.4	28.2
ctpc_nlp	52	32.1	27.5	32.0	27.6	62.4	53.5	32.5	28.1	32.3	27.7	33.3	28.7	33.5	29.0	33.1	28.5	32.8	28.2	33.1	28.4	33.1	28.5	32.4	28.2
KYUoM	52	32.1	27.5	32.0	27.6	32.5	27.9	32.5	28.1	32.3	27.7	33.3	28.7	33.5	29.0	33.1	28.5	32.8	28.2	33.1	28.4	63.8	54.6	32.4	28.2
CUNI-DocTransformer	55	–	–	–	–	76.3	58.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
COILD-BHO	55	–	–	48.6	47.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 3: CometKiwi (higher is better) scores across language directions for original and normalised inputs (manual segmentation in both instances). System are sorted by their CometKiwi-based ranking (over all language pairs). Systems run by the task organisers are marked with an asterisk.

System	Rank	en-ar_EG		en-bho_IN		en-cs_CZ		en-et_EE		en-is_IS		en-ja_JP		en-ko_KR		en-mas_KE		en-ru_RU		en-sr_Latn_RS		en-uk_UA		en-zh_CN	
		norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig
Shy-hunyuan-MT	1	3.4	3.8	3.9	4.2	4.7	5.4	5.7	6.4	5.2	6.1	4.0	4.5	3.7	4.1	11.8	11.6	3.1	3.8	2.2	2.6	3.9	4.6	2.3	2.8
GemTrans	2	3.5	3.9	4.0	4.3	4.8	5.5	6.1	6.9	6.9	7.3	4.0	4.5	3.8	4.3	14.7	15.0	3.4	4.0	2.1	2.7	4.1	4.7	2.5	2.9
Yolu	3	3.6	4.2	5.5	5.8	5.0	5.9	5.7	6.7	9.3	11.4	4.2	4.6	4.0	4.5	9.3	11.4	3.7	4.6	2.1	2.7	4.4	5.3	2.7	3.2
CommandA-WMT	4	4.1	4.7	4.4	4.8	4.5	5.1	6.6	7.3	7.9	8.9	4.1	4.5	3.8	4.3	10.7	10.8	4.1	4.7	3.9	4.6	3.9	4.6	2.8	3.2
Laniqo	5	–	–	–	–	4.9	5.9	5.4	6.5	–	–	4.4	4.8	4.0	4.5	–	–	3.5	4.5	–	–	4.2	4.9	2.7	3.3
*GPT-4.1	6	5.5	6.0	6.4	6.6	5.6	6.6	6.6	7.4	6.1	7.0	4.4	4.8	4.2	4.7	15.0	15.0	4.3	5.1	2.6	3.2	4.9	5.7	2.9	3.4
*Gemini 2.5 Pro	7	6.2	6.5	6.2	6.5	5.9	6.8	6.7	7.6	6.0	6.9	4.4	4.9	4.5	5.0	15.7	16.0	4.6	5.4	3.0	3.3	5.2	6.1	3.0	3.5
SalamandraTA	8	7.6	8.6	9.7	10.3	5.7	7.2	6.7	8.4	6.9	8.4	6.0	6.9	6.0	6.9	–	–	4.7	6.2	2.4	3.0	5.0	6.3	3.6	4.4
Kaze-MT	9	8.9	9.3	8.4	8.8	9.4	9.7	8.7	9.2	8.6	9.0	8.7	9.0	8.8	9.1	9.3	9.7	9.1	9.5	8.4	8.7	9.2	9.5	8.8	9.1
KYUoM	10	8.9	9.3	8.4	8.8	9.4	9.7	8.7	9.2	8.6	9.0	8.7	9.0	8.8	9.1	9.3	9.7	9.1	9.5	8.4	8.7	12.5	15.3	8.8	9.1
ctpc_nlp	11	8.9	9.3	8.4	8.8	12.7	15.2	8.7	9.2	8.6	9.0	8.7	9.0	8.8	9.1	9.3	9.7	9.1	9.5	8.4	8.7	9.2	9.5	8.8	9.1
Yandex	12	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3.8	4.5	–	–	–	–	–	–
NNTSU	12	–	–	–	–	–	–	–	–	–	–	4.1	4.6	–	–	–	–	–	–	–	–	–	–	–	–
KIKIS	12	–	–	–	–	–	–	–	–	–	–	4.0	4.6	–	–	–	–	–	–	–	–	–	–	–	–
Erlendur	12	–	–	–	–	–	–	–	–	5.8	7.1	–	–	–	–	–	–	–	–	–	–	–	–	–	–
UvA-MT	16	5.0	5.6	6.5	7.1	5.9	6.6	8.2	8.9	9.4	9.9	4.7	5.2	4.3	4.8	13.0	13.0	4.2	5.0	2.6	3.1	4.9	5.5	3.2	3.5
IR-MultiagentMT	17	5.5	5.8	6.7	6.9	6.2	6.8	7.7	8.3	7.3	8.0	4.8	5.3	4.6	4.9	14.7	14.9	4.6	5.2	2.8	3.2	5.3	5.8	3.2	3.6
*Gemma 3 27B	18	5.5	5.9	6.7	6.8	6.1	6.7	7.7	8.6	8.2	8.8	4.6	5.0	4.4	5.0	15.4	15.7	4.4	5.2	2.7	3.2	5.2	6.0	3.1	3.5
Algharb	19	4.2	4.7	6.3	6.5	6.2	7.0	6.9	8.0	9.3	11.4	4.7	5.0	4.6	5.1	9.3	11.4	4.9	5.8	2.9	3.5	5.5	6.4	3.2	3.8
*ONLINE-B	20	4.0	4.9	7.2	9.8	5.9	7.3	6.9	8.3	6.1	7.2	4.3	4.9	4.4	5.0	–	–	5.6	6.7	4.0	4.9	5.3	6.6	3.0	3.9
*TowerPlus 9B	21	12.9	13.3	5.3	5.8	6.4	7.7	15.8	16.4	6.2	7.2	4.8	5.3	4.7	5.2	18.1	17.7	4.8	5.7	4.6	5.3	5.2	6.3	3.3	4.0
*TowerPlus 72B	22	7.1	7.7	5.3	5.8	6.6	7.8	13.0	13.5	6.5	7.6	4.6	5.3	4.7	5.3	18.3	18.1	4.8	5.8	3.7	4.2	5.6	6.6	3.4	4.1
*DeepSeek V3	23	5.8	6.4	5.8	6.1	6.0	7.0	7.7	8.5	7.6	8.5	4.6	5.0	4.4	5.0	15.3	15.5	4.5	5.2	2.7	3.3	5.3	6.0	3.0	3.5
*Claude4	24	5.7	6.2	5.9	6.4	6.3	7.6	7.9	9.1	7.1	8.3	4.5	5.1	4.4	4.9	15.3	15.5	4.9	6.0	2.9	3.6	5.5	6.5	3.0	3.7
*Mistral-Medium	25	6.1	6.6	6.6	7.0	6.3	7.2	9.1	10.1	9.7	10.2	4.5	5.1	4.5	5.0	–	–	–	–	–	–	5.4	6.2	2.8	3.4
*CommandA	26	5.6	6.1	6.6	7.2	6.1	7.0	11.0	11.9	12.3	12.7	4.7	5.1	4.4	4.9	15.9	15.7	5.0	5.9	3.0	3.6	5.5	6.4	3.3	3.7
*Qwen3 235B	27	7.4	7.9	7.6	8.0	6.7	7.9	10.6	11.9	11.3	12.4	4.7	5.3	4.5	4.9	15.8	16.1	4.8	5.8	3.5	4.3	5.8	6.7	2.9	3.5
SRPOL	28	4.5	5.4	–	–	5.6	7.1	7.0	8.5	–	–	4.9	5.6	–	–	–	–	5.0	6.1	–	–	5.3	6.4	3.3	4.0
*AyaExpanse 8B	29	4.8	5.3	6.9	7.4	7.0	7.7	23.9	23.9	23.2	23.2	5.0	5.4	4.7	5.2	21.6	21.2	5.9	6.5	5.5	6.1	6.0	6.7	3.7	4.0
In2x	30	–	–	–	–	–	–	–	–	–	–	4.2	4.7	–	–	–	–	–	–	–	–	–	–	–	–
IRB-MT	31	6.3	6.7	7.9	8.1	6.7	7.2	8.9	9.7	10.4	10.4	4.8	5.3	4.8	5.4	14.3	14.6	4.7	5.4	2.8	3.4	5.3	6.1	3.1	3.6
*Gemma 3 12B	32	6.5	6.9	8.1	8.3	6.6	7.5	9.3	10.0	10.9	11.3	5.2	5.7	5.1	5.6	14.1	14.1	5.0	5.6	2.9	3.4	5.3	6.1	3.4	3.9
TranssionTranslate	33	6.7	8.1	7.2	9.5	6.3	8.0	7.2	8.9	6.4	7.8	4.7	5.5	4.5	5.6	9.3	11.4	4.9	6.6	4.2	7.0	5.7	7.2	3.7	4.7
Wenyiil	34	4.5	5.5	6.7	7.1	7.0	8.2	8.0	9.5	9.3	11.4	5.1	5.7	5.1	5.9	9.3	11.4	5.5	6.8	3.1	3.8	6.0	7.1	3.4	4.1
*AyaExpanse 32B	35	5.7	6.1	8.1	8.1	6.4	7.0	20.8	20.6	20.3	20.4	4.8	5.1	4.6	5.0	19.3	19.1	5.2	5.8	3.8	4.3	5.6	6.2	3.4	3.9
*EuroLLM 22B	36	5.5	6.3	5.9	6.6	6.3	7.5	7.7	8.9	21.9	21.7	4.9	5.7	4.8	5.6	19.6	18.5	5.3	6.3	2.9	3.7	5.8	6.9	3.3	4.0
AMI	37	–	–	–	–	–	–	–	–	6.9	8.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama-4-Maverick	38	6.0	6.6	6.7	7.2	6.3	7.6	7.8	9.1	8.2	9.2	4.7	5.1	4.5	5.0	15.1	15.3	4.9	6.0	2.8	3.5	5.4	6.4	3.0	3.6
*NLLB	39	7.3	8.6	6.7	7.6	7.9	10.1	10.0	11.8	9.2	10.8	8.1	9.0	6.5	7.8	12.9	13.8	7.7	9.7	12.9	13.8	7.9	9.9	7.9	8.8
DLUT_GTCOM	40	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	4.8	5.7	–	–	5.9	7.3	–	–
*EuroLLM 9B	41	6.3	7.0	7.6	8.4	6.7	8.1	8.1	9.5	22.3	21.8	5.3	5.8	5.1	5.8	15.8	15.1	5.7	7.0	3.2	4.4	6.1	7.3	3.6	4.4
*ONLINE-W	42	5.1	6.4	–	–	6.6	10.0	9.1	13.5	–	–	5.3	6.2	4.5	6.3	–	–	5.1	7.0	–	–	5.5	7.0	3.7	4.4
*CommandR	43	5.4	6.2	10.1	10.5	8.6	9.8	23.3	23.2	21.6	21.9	5.4	6.0	5.5	6.0	20.1	19.6	8.5	9.8	6.2	6.7	8.7	9.7	3.8	4.5
Systran	44	–	–	–	–	–	–	–	–	–	–	4.5	5.2	–	–	–	–	–	–	–	–	–	–	–	–
CUNI-SFT	45	–	–	–	–	7.2	9.3	–	–	–	–	–	–	–	–	–	–	–	–	3.3	4.4	6.3	7.8	–	–
CUNI-MH-v2	46	–	–	–	–	6.3	7.8	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama 3.1 8B	47	10.3	10.7	8.9	9.2	9.3	10.4	15.5	15.7	17.8	17.8	6.4	6.7	6.4	6.9	16.6	16.4	7.7	8.7	4.0	4.6	8.3	9.1	3.8	4.4
*Qwen 2.5	48	11.0	11.5	11.7	11.8	11.9	12.8	22.0	22.0	22.8	22.8	6.8	7.4	7.5	7.9	20.6	20.0	7.6	8.6	5.8	6.7	12.0	12.7	3.6	4.2
CGFOKUS	49	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5.9	6.9	–	–
bb88	49	–	–	–	–	–	–	–	–	–	–	4.9	5.4	–	–	–	–	–	–	–	–	–	–	–	–
TranssionMT	51	6.8	8.9	7.2	8.5	12.3	16.0	14.2	17.0	–	–	–	–	–	–	15.9	16.7	10.0	12.3	3.5	5.0	9.4	11.0	–	–
*ONLINE-G	52	10.3	13.4	–	–	9.9	14.7	12.2	16.5	11.1	15.1	9.6	12.7	9.8	12.1	–	–	5.0	7.0	4.9	6.1	6.0	7.7	6.7	8.9
*Mistral 7B	53	14.9	15.3	14.1	14.1	12.9	13.5	23.6	23.7	22.6	22.6	8.5	8.8	8.2	8.5	22.1	22.2	9.5	10.5	4.6	5.3	10.0	10.5	5.4	6.0
SH	54	–	–	–	–	–	–	–	–	–	–	5.4	6.1	–	–	–	–	–	–	–	–	–	–	–	–
CUNI-DocTransformer	55	–	–	–	–	8.2	13.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
COILD-BHO	55	–	–	10.7	11.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 4: MetricX scores (lower is better) across language directions for original and normalised inputs (manual segmentation in both instances). System are sorted by their MetricX-based ranking (over all language pairs). Systems run by the task organisers are marked with an asterisk.

System	Rank	en-ar_EG		en-bho_IN		en-cs_CZ		en-et_EE		en-is_IS		en-ja_JP		en-ko_KR		en-mas_KE		en-ru_RU		en-sr_Latn_RS		en-uk_UA		en-zh_CN	
		man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl
Yolu	1	73.4	72.7	68.9	66.2	76.3	75.6	79.6	78.7	22.7	23.0	79.7	78.8	79.3	78.2	22.7	23.0	76.0	75.2	78.5	78.2	75.5	74.6	76.1	75.8
CommandA-WMT	2	65.6	66.2	72.8	73.6	75.1	74.8	77.8	77.8	70.0	68.9	78.5	78.2	78.1	77.8	62.6	62.0	74.7	74.4	75.0	75.1	75.3	74.7	74.7	74.5
SalamandraTA	3	66.8	65.4	57.4	56.9	76.0	76.5	78.9	79.2	72.2	72.0	78.4	77.4	77.1	76.1	–	–	75.5	75.7	78.3	78.5	75.1	75.3	75.4	74.8
Shy-hunyuan-MT	4	70.7	70.2	76.6	76.6	74.8	74.4	77.6	77.2	71.4	71.8	77.7	77.7	76.9	76.8	52.2	50.5	73.9	73.6	77.0	76.8	73.7	73.8	73.9	73.9
Laniqo	5	–	–	–	–	74.7	73.8	78.0	77.4	–	–	78.3	77.2	78.1	76.9	–	–	74.7	73.8	–	–	74.6	73.8	74.7	73.9
SRPOL	6	71.2	71.6	–	–	74.4	74.5	77.4	77.2	–	–	78.5	78.4	–	–	–	–	74.1	74.2	–	–	73.8	73.8	74.2	74.2
UvA-MT	7	66.4	66.4	68.2	68.6	73.8	73.7	75.7	75.5	67.9	67.6	78.2	78.0	77.6	77.4	37.1	37.4	73.8	73.4	76.7	76.5	73.5	73.3	74.4	74.3
*ONLINE-B	8	71.4	71.1	54.5	54.3	73.1	72.8	75.8	75.6	71.4	71.2	78.3	78.2	76.6	76.4	–	–	73.3	72.9	74.5	74.4	72.1	72.0	73.8	73.6
*TowerPlus 9B	9	46.0	45.2	75.8	75.7	72.9	73.2	57.9	59.2	70.8	70.8	77.5	77.4	77.2	77.1	41.3	42.0	73.7	73.2	60.2	60.6	72.8	72.6	73.6	73.5
TranssionTranslate	10	61.7	60.2	54.9	54.6	72.9	72.9	76.0	75.7	71.1	71.1	78.5	78.3	77.1	76.9	22.7	23.0	73.4	73.3	65.3	64.8	71.7	71.8	73.8	73.5
*TowerPlus 72B	11	60.4	60.7	76.2	76.0	72.7	72.8	67.4	67.5	70.4	70.3	77.1	77.1	77.0	76.9	38.7	42.1	73.4	73.4	68.8	69.3	72.5	72.4	74.1	73.9
GemTrans	12	70.3	69.9	76.9	76.6	72.3	72.2	74.8	74.9	67.5	67.3	76.2	75.9	75.7	75.3	34.3	34.7	72.4	72.5	75.1	74.9	72.7	72.6	73.1	72.7
*AyaExpanse 32B	13	60.5	60.3	62.8	62.5	72.7	72.2	50.9	50.9	47.8	47.8	77.0	76.9	76.4	76.2	43.2	44.0	73.1	72.6	68.3	67.6	72.4	72.4	72.6	72.7
*Llama 3.1 8B	14	52.0	51.8	62.1	62.1	66.7	66.7	62.6	62.6	53.4	53.2	72.2	71.9	72.0	71.8	48.8	48.9	69.2	69.2	68.1	68.0	68.2	68.1	70.9	70.5
Systran	15	–	–	–	–	–	–	–	–	–	–	81.0	80.6	–	–	–	–	–	–	–	–	–	–	–	–
In2x	15	–	–	–	–	–	–	–	–	–	–	78.7	78.4	–	–	–	–	–	–	–	–	–	–	–	–
*GPT-4.1	17	57.9	57.8	59.6	59.3	73.2	73.1	76.5	76.3	70.5	70.5	77.2	77.0	76.8	76.5	32.2	32.4	72.6	72.4	76.2	76.1	72.5	72.4	73.1	73.0
*CommandA	18	59.3	59.1	60.2	60.2	73.5	73.4	71.3	71.0	63.7	63.9	77.7	77.8	77.4	77.2	39.9	40.3	73.1	73.0	74.5	74.4	72.5	72.4	73.8	73.6
*EuroLLM 22B	19	66.0	65.0	73.6	73.2	72.5	72.3	76.1	76.0	43.1	42.5	76.7	76.4	75.9	75.5	36.7	37.1	73.0	72.6	74.2	74.0	72.1	72.2	73.1	72.8
*AyaExpanse 8B	20	69.7	69.5	71.1	71.2	71.4	71.4	37.5	37.4	37.4	37.4	76.7	76.2	76.2	75.8	38.2	38.3	72.0	71.9	56.8	56.7	71.9	71.7	72.5	72.3
*Claude4	21	59.4	59.3	59.5	59.4	72.8	72.9	74.9	74.8	69.4	69.3	77.5	77.4	76.9	76.6	37.6	37.7	72.3	72.2	75.3	75.2	72.2	72.2	73.0	72.9
*DeepSeek V3	22	58.9	59.3	62.5	62.3	72.4	72.4	75.3	75.5	68.8	68.7	76.4	76.2	75.7	75.7	35.4	35.6	72.2	71.6	75.4	75.4	71.7	71.3	70.5	70.2
IR-MultiagentMT	23	63.7	63.4	61.5	61.7	72.5	72.4	74.9	74.9	69.6	69.2	76.4	76.6	76.5	76.6	36.2	36.2	73.0	72.2	75.2	75.5	72.5	72.6	72.9	72.8
*CommandR	24	66.5	66.7	62.2	62.7	67.8	67.8	39.5	40.4	38.7	38.7	75.3	75.2	74.6	73.8	41.2	41.6	66.1	65.9	57.2	57.0	66.9	66.9	71.4	70.9
*Gemma 3 12B	25	57.5	57.3	57.2	57.6	71.5	71.8	72.7	72.6	64.3	64.2	75.5	75.4	74.8	74.5	42.1	42.3	72.6	72.1	74.0	73.9	72.4	71.8	72.5	72.2
Algharb	26	69.2	69.0	56.7	56.2	71.6	71.2	74.9	74.9	22.7	23.0	75.8	75.4	75.1	74.6	22.7	23.0	70.6	70.3	74.7	74.4	70.9	70.7	71.3	71.1
*ONLINE-W	27	68.7	68.3	–	–	68.9	67.3	68.0	66.2	–	–	73.8	69.6	75.4	74.4	–	–	72.5	71.4	–	–	72.1	72.0	73.0	70.8
*Gemma 3 27B	28	59.1	59.2	59.8	59.4	72.5	72.3	75.1	75.1	68.1	68.0	76.5	76.1	75.4	75.3	32.6	32.7	72.4	72.3	75.0	74.9	72.5	72.2	72.5	72.5
*EuroLLM 9B	29	64.7	64.8	65.4	66.9	72.4	72.0	75.0	74.8	41.5	41.3	75.9	75.9	75.6	75.3	28.5	29.6	72.3	71.7	70.6	70.9	71.6	71.3	72.0	71.6
Wenyiil	30	67.6	68.3	55.6	55.6	68.7	69.2	72.4	73.2	22.7	23.0	74.3	74.6	73.3	73.7	22.7	23.0	69.0	69.4	72.8	73.5	69.8	69.6	70.4	70.5
*Mistral 7B	31	45.0	44.3	52.8	52.4	59.4	59.2	39.1	39.1	40.4	40.5	67.9	67.7	67.6	67.2	40.1	40.2	66.2	66.1	64.6	63.4	65.4	65.2	65.9	65.7
CUNI-MH-v2	32	–	–	–	–	73.0	72.5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
NNTSU	32	–	–	–	–	–	–	–	–	–	–	78.2	77.6	–	–	–	–	–	–	–	–	–	–	–	–
bb88	32	–	–	–	–	–	–	–	–	–	–	77.8	77.4	–	–	–	–	–	–	–	–	–	–	–	–
AMI	32	–	–	–	–	–	–	–	–	69.4	69.9	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Erlendur	32	–	–	–	–	–	–	–	–	69.1	69.1	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama-4-Maverick	37	57.7	57.5	57.3	57.3	71.8	71.6	74.4	74.7	67.0	66.6	75.8	75.5	75.4	75.3	34.8	34.5	72.6	72.4	74.7	74.5	72.3	72.2	72.4	72.0
*Gemini 2.5 Pro	38	54.8	54.5	56.5	56.6	70.7	70.6	74.4	74.5	68.8	68.7	75.1	75.1	74.0	73.6	35.2	35.1	69.7	69.9	74.3	74.2	70.2	69.9	70.2	70.1
KIKIS	39	–	–	–	–	–	–	–	–	–	–	77.3	77.1	–	–	–	–	–	–	–	–	–	–	–	–
*Mistral-Medium	40	60.3	59.8	60.6	60.5	72.5	72.4	73.3	73.1	66.9	66.8	76.7	76.7	76.2	75.8	–	–	–	–	–	–	72.0	72.1	72.8	72.7
*Qwen3 235B	41	59.0	59.6	60.4	60.2	71.3	71.3	69.7	69.9	63.5	63.3	76.9	76.7	76.0	75.5	36.4	36.6	72.2	72.1	72.5	72.8	71.2	71.2	73.0	73.0
IRB-MT	42	56.3	56.0	56.3	56.7	70.5	69.9	72.0	72.0	64.6	63.9	74.2	74.1	72.9	72.6	36.6	37.4	71.1	70.1	73.4	73.1	70.6	70.1	70.1	69.9
TranssionMT	43	60.6	60.3	57.4	57.1	58.2	57.5	63.0	62.0	–	–	–	–	–	–	36.4	35.6	64.1	63.0	71.3	71.0	65.3	64.5	–	–
*NLLB	44	63.3	60.7	58.4	58.5	68.9	63.6	71.6	66.5	63.5	61.3	68.1	64.6	71.4	67.5	22.5	23.8	69.1	64.2	22.5	23.8	67.6	64.4	59.6	57.4
DLUT_GTCOM	45	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	72.6	72.5	–	–	71.5	71.8	–	–
Yandex	46	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	72.6	72.3	–	–	–	–	–	–
*Qwen 2.5	47	50.2	51.0	58.9	58.5	59.4	59.2	46.8	47.4	40.4	40.4	72.5	72.3	67.2	67.2	33.8	34.5	67.0	66.8	56.8	57.0	59.5	59.5	71.2	70.7
CUNI-SFT	48	–	–	–	–	70.7	70.9	–	–	–	–	–	–	–	–	–	–	–	–	71.5	72.3	70.7	70.6	–	–
*ONLINE-G	49	53.5	53.1	–	–	59.9	59.4	61.6	61.6	59.7	58.8	61.2	60.0	57.3	56.1	–	–	72.2	71.8	70.7	70.6	71.0	70.6	62.9	61.6
SH	50	–	–	–	–	–	–	–	–	–	–	75.6	74.8	–	–	–	–	–	–	–	–	–	–	–	–
CGFOKUS	51	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	70.7	70.8	–	–
Kaze-MT	52	27.5	25.0	27.6	25.0	27.9	25.3	28.1	25.5	27.7	25.2	28.7	26.4	29.0	26.6	28.5	26.0	28.2	25.7	28.4	26.0	28.5	26.0	28.2	25.8
ctpc_nlp	52	27.5	25.0	27.6	25.0	53.5	55.1	28.1	25.5	27.7	25.2	28.7	26.4	29.0	26.6	28.5	26.0	28.2	25.7	28.4	26.0	28.5	26.0	28.2	25.8
KYUoM	52	27.5	25.0	27.6	25.0	27.9	25.3	28.1	25.5	27.7	25.2	28.7	26.4	29.0	26.6	28.5	26.0	28.2	25.7	28.4	26.0	54.6	55.3	28.2	25.8
CUNI-DocTransformer	55	–	–	–	–	58.6	57.4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
COILD-BHO	55	–	–	47.0	47.8	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 5: CometKiwi scores (higher is better) across languages directions for manual (man) segmentation versus newline (nl) segmentation on original outputs. System are sorted by their CometKiwi-based ranking (over all language pairs). Systems run by the task organisers are marked with an asterisk.

System	Rank	en-ar_EG		en-bho_IN		en-cs_CZ		en-et_EE		en-is_IS		en-ja_JP		en-ko_KR		en-mas_KE		en-ru_RU		en-sr_Latn_RS		en-uk_UA		en-zh_CN	
		man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl
Shy-hunyuan-MT	1	3.8	4.0	4.2	4.3	5.4	5.5	6.4	6.4	6.1	6.0	4.5	4.6	4.1	4.2	11.6	11.8	3.8	4.0	2.6	2.7	4.6	4.6	2.8	2.9
GemTrans	2	3.9	4.0	4.3	4.4	5.5	5.5	6.9	6.8	7.3	7.3	4.5	4.6	4.3	4.3	15.0	14.8	4.0	4.1	2.7	2.8	4.7	4.8	2.9	3.1
Yolu	3	4.2	4.2	5.8	6.2	5.9	5.9	6.7	6.8	11.4	11.4	4.6	4.7	4.5	4.5	11.4	11.4	4.6	4.5	2.7	2.8	5.3	5.2	3.2	3.1
CommandA-WMT	4	4.7	4.6	4.8	4.8	5.1	5.2	7.3	7.2	8.9	8.9	4.5	4.6	4.3	4.4	10.8	10.9	4.7	4.8	4.6	4.6	4.6	4.6	3.2	3.3
Laniqo	5	–	–	–	–	5.9	5.7	6.5	6.5	–	–	4.8	5.0	4.5	4.6	–	–	4.5	4.4	–	–	4.9	4.9	3.3	3.2
*GPT-4.1	6	6.0	6.1	6.6	6.8	6.6	6.6	7.4	7.4	7.0	7.0	4.8	4.9	4.7	4.8	15.0	14.8	5.1	5.2	3.2	3.2	5.7	5.7	3.4	3.5
*Gemini 2.5 Pro	7	6.5	6.5	6.5	6.5	6.8	6.8	7.6	7.6	6.9	6.9	4.9	4.9	5.0	5.1	16.0	15.8	5.4	5.5	3.3	3.5	6.1	6.1	3.5	3.5
SalamandraTA	8	8.6	9.0	10.3	10.4	7.2	6.9	8.4	8.1	8.4	8.3	6.9	7.1	6.9	7.1	–	–	6.2	6.1	3.0	3.0	6.3	6.1	4.4	4.5
Kaze-MT	9	9.3	10.6	8.8	10.0	9.7	11.0	9.2	10.5	9.0	10.2	9.0	10.2	9.1	10.4	9.7	11.0	9.5	10.7	8.7	10.1	9.5	10.8	9.1	10.3
KYUoM	10	9.3	10.6	8.8	10.0	9.7	11.0	9.2	10.5	9.0	10.2	9.0	10.2	9.1	10.4	9.7	11.0	9.5	10.7	8.7	10.1	15.3	15.4	9.1	10.3
ctpc_nlp	11	9.3	10.6	8.8	10.0	15.2	14.2	9.2	10.5	9.0	10.2	9.0	10.2	9.1	10.4	9.7	11.0	9.5	10.7	8.7	10.1	9.5	10.8	9.1	10.3
Yandex	12	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	4.5	4.6	–	–	–	–	–	–
NNTSU	12	–	–	–	–	–	–	–	–	–	–	4.6	4.6	–	–	–	–	–	–	–	–	–	–	–	–
KIKIS	12	–	–	–	–	–	–	–	–	–	–	4.6	4.7	–	–	–	–	–	–	–	–	–	–	–	–
Erlendur	12	–	–	–	–	–	–	–	–	7.1	7.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–
UvA-MT	16	5.6	5.5	7.1	7.1	6.6	6.6	8.9	8.8	9.9	10.0	5.2	5.2	4.8	4.9	13.0	13.0	5.0	5.0	3.1	3.1	5.5	5.6	3.5	3.6
IR-MultiagentMT	17	5.8	6.0	6.9	7.0	6.8	6.9	8.3	8.4	8.0	8.2	5.3	5.3	4.9	5.0	14.9	14.9	5.2	5.4	3.2	3.3	5.8	5.8	3.6	3.7
*Gemma 3 27B	18	5.9	5.9	6.8	7.0	6.7	6.8	8.6	8.5	8.8	8.9	5.0	5.1	5.0	5.1	15.7	15.8	5.2	5.3	3.2	3.3	6.0	6.1	3.5	3.6
Algharb	19	4.7	4.6	6.5	6.6	7.0	6.8	8.0	7.6	11.4	11.4	5.0	5.0	5.1	5.0	11.4	11.4	5.8	5.5	3.5	3.5	6.4	6.2	3.8	3.7
*ONLINE-B	20	4.9	5.0	9.8	9.8	7.3	7.4	8.3	8.4	7.2	7.2	4.9	5.0	5.0	5.1	–	–	6.7	6.8	4.9	5.0	6.6	6.7	3.9	4.1
*TowerPlus 9B	21	13.3	13.4	5.8	5.9	7.7	7.6	16.4	16.0	7.2	7.2	5.3	5.3	5.2	5.3	17.7	17.8	5.7	5.9	5.3	5.4	6.3	6.3	4.0	4.1
*TowerPlus 72B	22	7.7	7.8	5.8	5.8	7.8	7.7	13.5	13.3	7.6	7.5	5.3	5.3	5.3	5.3	18.1	17.9	5.8	5.9	4.2	4.2	6.6	6.6	4.1	4.1
*DeepSeek V3	23	6.4	6.3	6.1	6.2	7.0	6.9	8.5	8.5	8.5	8.5	5.0	5.1	5.0	5.1	15.5	15.4	5.2	5.4	3.3	3.3	6.0	6.0	3.5	3.6
*Claude4	24	6.2	6.2	6.4	6.4	7.6	7.4	9.1	9.0	8.3	8.3	5.1	5.2	4.9	5.0	15.5	15.4	6.0	6.0	3.6	3.6	6.5	6.5	3.7	3.7
*Mistral-Medium	25	6.6	6.7	7.0	6.9	7.2	7.1	10.1	10.1	10.2	10.2	5.1	5.1	5.0	5.0	–	–	–	–	–	–	6.2	6.2	3.4	3.4
*CommandA	26	6.1	6.2	7.2	7.1	7.0	6.9	11.9	11.8	12.7	12.5	5.1	5.1	4.9	4.9	15.7	15.7	5.9	5.9	3.6	3.7	6.4	6.4	3.7	3.8
*Qwen3 235B	27	7.9	7.9	8.0	8.1	7.9	7.7	11.9	11.7	12.4	12.1	5.3	5.3	4.9	5.0	16.1	16.1	5.8	5.9	4.3	4.2	6.7	6.7	3.5	3.5
SRPOL	28	5.4	5.3	–	–	7.1	6.9	8.5	8.2	–	–	5.6	5.5	–	–	–	–	6.1	6.0	–	–	6.4	6.2	4.0	4.0
*AyaExpanse 8B	29	5.3	5.3	7.4	7.4	7.7	7.8	23.9	23.7	23.2	23.2	5.4	5.6	5.2	5.2	21.2	20.9	6.5	6.5	6.1	6.1	6.7	6.7	4.0	4.2
In2x	30	–	–	–	–	–	–	–	–	–	–	4.7	4.7	–	–	–	–	–	–	–	–	–	–	–	–
IRB-MT	31	6.7	6.7	8.1	8.0	7.2	7.3	9.7	9.6	10.4	10.6	5.3	5.4	5.4	5.5	14.6	14.5	5.4	5.4	3.4	3.5	6.1	6.1	3.6	3.7
*Gemma 3 12B	32	6.9	7.0	8.3	8.3	7.5	7.4	10.0	10.0	11.3	11.3	5.7	5.6	5.6	5.7	14.1	14.1	5.6	5.7	3.4	3.6	6.1	6.3	3.9	4.0
TranssionTranslate	33	8.1	8.1	9.5	9.7	8.0	7.6	8.9	8.6	7.8	7.5	5.5	5.5	5.6	5.5	11.4	11.4	6.6	6.4	7.0	7.0	7.2	6.9	4.7	4.7
Wenyiil	34	5.5	5.0	7.1	7.1	8.2	7.7	9.5	8.6	11.4	11.4	5.7	5.4	5.9	5.5	11.4	11.4	6.8	6.0	3.8	3.6	7.1	6.7	4.1	3.8
*AyaExpanse 32B	35	6.1	6.1	8.1	8.1	7.0	7.1	20.6	20.9	20.4	20.5	5.1	5.2	5.0	5.0	19.1	19.0	5.8	5.8	4.3	4.3	6.2	6.2	3.9	3.9
*EuroLLM 22B	36	6.3	6.4	6.6	6.6	7.5	7.3	8.9	8.8	21.7	21.7	5.7	5.8	5.6	5.7	18.5	18.2	6.3	6.4	3.7	3.7	6.9	6.8	4.0	4.1
AMI	37	–	–	–	–	–	–	–	–	8.2	8.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama-4-Maverick	38	6.6	6.4	7.2	7.1	7.6	7.5	9.1	8.8	9.2	9.0	5.1	5.2	5.0	5.0	15.3	15.4	6.0	5.9	3.5	3.5	6.4	6.4	3.6	3.7
*NLLB	39	8.6	9.2	7.6	7.7	10.1	11.2	11.8	13.3	10.8	13.1	9.0	9.7	7.8	8.6	13.8	12.0	9.7	11.6	13.8	12.0	9.9	11.0	8.8	9.2
DLUT_GTCOM	40	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5.7	5.2	–	–	7.3	7.0	–	–
*EuroLLM 9B	41	7.0	7.1	8.4	8.0	8.1	8.1	9.5	9.5	21.8	21.7	5.8	5.9	5.8	5.9	15.1	14.2	7.0	7.1	4.4	4.4	7.3	7.3	4.4	4.4
*ONLINE-W	42	6.4	6.5	–	–	10.0	10.8	13.5	14.1	–	–	6.2	6.8	6.3	6.5	–	–	7.0	7.4	–	–	7.0	7.0	4.4	4.6
*CommandR	43	6.2	6.2	10.5	10.1	9.8	9.7	23.2	22.8	21.9	21.3	6.0	6.0	6.0	6.1	19.6	18.4	9.8	9.7	6.7	6.7	9.7	9.5	4.5	4.5
Systran	44	–	–	–	–	–	–	–	–	–	–	5.2	5.2	–	–	–	–	–	–	–	–	–	–	–	–
CUNI-SFT	45	–	–	–	–	9.3	8.7	–	–	–	–	–	–	–	–	–	–	–	–	4.4	4.3	7.8	7.5	–	–
CUNI-MH-v2	46	–	–	–	–	7.8	7.7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama 3.1 8B	47	10.7	10.6	9.2	9.0	10.4	10.3	15.7	15.6	17.8	17.6	6.7	6.6	6.9	7.0	16.4	15.8	8.7	8.6	4.6	4.5	9.1	8.8	4.4	4.4
*Qwen 2.5	48	11.5	11.2	11.8	11.9	12.8	12.8	22.0	21.7	22.8	22.6	7.4	7.5	7.9	7.8	20.0	19.5	8.6	8.4	6.7	6.6	12.7	12.5	4.2	4.3
CGFOKUS	49	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	6.9	6.8	–	–
bb88	49	–	–	–	–	–	–	–	–	–	–	5.4	5.2	–	–	–	–	–	–	–	–	–	–	–	–
TranssionMT	51	8.9	8.9	8.5	8.7	16.0	16.1	17.0	17.0	–	–	–	–	–	–	16.7	16.4	12.3	12.4	5.0	5.2	11.0	11.1	–	–
*ONLINE-G	52	13.4	13.5	–	–	14.7	14.8	16.5	16.5	15.1	15.5	12.7	12.9	12.1	12.2	–	–	7.0	7.1	6.1	6.1	7.7	7.8	8.9	9.2
*Mistral 7B	53	15.3	15.1	14.1	14.4	13.5	13.2	23.7	23.7	22.6	22.5	8.8	8.7	8.5	8.5	22.2	22.2	10.5	10.2	5.3	5.4	10.5	10.3	6.0	6.0
SH	54	–	–	–	–	–	–	–	–	–	–	6.1	6.1	–	–	–	–	–	–	–	–	–	–	–	–
CUNI-DocTransformer	55	–	–	–	–	13.6	14.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
COILD-BHO	55	–	–	11.0	11.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 6: MetricX scores (lower is better) across languages directions for manual (man) segmentation versus newline (nl) segmentation on original inputs. System are sorted by their MetricX-based ranking (over all language pairs). Systems run by the task organisers are marked with an asterisk.