

Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets

Tom Kocmi Cohere	Ekaterina Artemova Toloka AI	Eleftherios Avramidis DFKI	Rachel Bawden Inria, Paris, France
Ondřej Bojar Charles University	Konstantin Dranch CustomMT	Anton Dvorkovich Dubformer	Sergey Dukanov Dubformer
Mark Fishel University of Tartu	Markus Freitag Google	Thamme Gowda Microsoft	Roman Grundkiewicz Microsoft
Barry Haddow University of Edinburgh	Marzena Karpinska Microsoft	Philipp Koehn Johns Hopkins University	Howard Lakoungna Gates Foundation
Jessica M. Lundin Gates Foundation	Christof Monz University of Amsterdam	Kenton Murray JHU	Masaaki Nagata NTT
Stefano Perrella Sapienza University of Rome	Lorenzo Proietti Sapienza University of Rome	Martin Popel Charles University	
Maja Popović DCU & IU	Parker Riley Google	Mariya Shmatova Toloka AI	
Steinþór Steingrímsson The Árni Magnússon Institute	Lisa Yankovskaya University of Tartu	Vilém Zouhar ETH Zurich	

Abstract

This paper presents the results of the General Machine Translation Task organized as part of the 2025 Conference on Machine Translation (WMT). Participants were invited to build systems for any of the 30 language pairs. For half of these pairs, we conducted a human evaluation on test sets spanning four to five different domains. We evaluated 60 systems in total: 36 submitted by participants and 24 consisting of translations we collected from large language models (LLMs) and popular online translation providers. This year, we focused on creating challenging test sets by developing a difficulty sampling technique and using more complex source data. We evaluated system outputs with professional annotators using the Error Span Annotation (ESA) protocol, except for two language pairs, for which we used Multidimensional Quality Metrics (MQM) instead. We continued the trend of increasingly shifting towards document-level translation, providing the source texts as whole documents containing multiple paragraphs.

1 Introduction

The Tenth Conference on Machine Translation (WMT25)¹ was held in conjunction with the 2025

Conference on Empirical Methods in Natural Language Processing (EMNLP 2025). This 20th iteration of the conference builds on previous editions (Koehn and Monz, 2006; Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015, 2016, 2017, 2018; Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022, 2023, 2024a)² and hosted ten shared tasks on various aspects of machine translation (MT). This paper describes one of these tasks: the General Machine Translation shared task.

The goal of the General Machine Translation shared task is to explore the translation capabilities of current systems across a broad range of languages. Determining how best to test general MT performance is a research question in itself. Numerous phenomena could be evaluated, the most important of which include:

- **different domains** (news, medicine, IT, patents, legal, social, gaming, etc.),
- **styles of text** (formal or spoken language, fiction, technical reports, etc.),
- **non-standard user-generated content** (errors, code-switching, abbreviations, etc.),
- **source modalities** (text, speech, image).

¹www2.statmt.org/wmt25/

²WMT was organized as a workshop for ten years (2006-2015) before becoming a conference, making this the 20th event overall.

Since individually evaluating all of these phenomena is infeasible, we focus on a selection of domains: news, social/user-generated content, speech, literary, and educational texts. These domains were chosen to cover diverse styles while remaining broadly accessible; thus allowing for evaluation by human annotators without in-domain expertise. However, due to limited access to monolingual data, the set of evaluated domains is not identical across all language pairs.

Our evaluation includes the following 16 language pairs, with those new to this year’s task marked by *(new)*:

Czech→Ukrainian,
Czech→German,
Japanese→Simplified Chinese,
English→Egyptian Arabic (*new*),
English→Bhojpuri (*new*),
English→Simplified Chinese,
English→Czech,
English→Estonian (*new*),
English→Icelandic,
English→Italian (*new*),
English→Japanese,
English→Korean (*new*),
English→Maasai, Kenya (*new*),
English→Russian,
English→Serbian, Cyrillic script (*new*),
English→Ukrainian.

A new multilingual subtrack, subject only to automatic evaluation, also introduced 15 additional language pairs:

English→Bengali,
English→German,
English→Greek,
English→Hindi,
English→Indonesian,
English→Kannada,
English→Lithuanian,
English→Marathi,
English→Persian,
English→Romanian,
English→Serbian, Latin script,
English→Swedish,
English→Thai,
English→Turkish,
English→Vietnamese.

Furthermore, this year’s task is also distinguished by several key choices:

- **Non-textual modalities:** In addition to textual data, we also incorporated audio and image

sources. For the speech domain, participants received audio files with their automatic speech recognition (ASR) transcriptions. For the social domain, screenshots of posts were provided. Participants had the option to use the original audio directly, rather than relying on the provided ASR text.

- **Difficulty sampling:** We used the difficulty sampling method to select more challenging documents, hence increasing the overall difficulty of the test set.
- **Evaluation:** For most languages, we use the Error Span Annotation protocol (ESA; [Kocmi et al., 2024b](#)) which combines aspects of DA ([Graham et al., 2013](#)) and MQM ([Lommel et al., 2014a](#)). For English→Korean and Japanese→Chinese, we use the MQM annotation schema instead.
- **Document-level test set:** Each source text is an entire document (e.g., a news article or social media thread),³ which is then segmented while preserving paragraph boundaries. This allows us to evaluate translations within their full document context, while giving participants the flexibility to choose their translation strategy: processing the entire document at once, or splitting it by segments or paragraphs.⁴
- **Training corpora:** We prepared a list of recommended training corpora, adding document-level information and COMETKIWI22 ([Rei et al., 2022](#)) scores for most data sets.

Finally, as in previous years, this year’s shared task included several test suites that focused on a range of **translation challenges**, described in Section 8.

All submissions to the General MT Task, along with sources, references, and human judgments, are available in the dedicated GitHub repository.⁵

This paper is organized as follows. We first describe the data used in the shared task, detailing the collection and preparation of our test sets (**Section 2**) and outlining the permitted training data for the constrained track (**Section 3**). Next, we introduce the participating systems, including the large language model baselines added by the organizers (**Section 4**). We then explain our evaluation methodology, covering both automatic metrics (**Section 5**) and human evaluation protocols (**Section 6**). Finally, we present the official results

³In some cases, an initial section of a document was used rather than the full text.

⁴No sentence-level segmentation is provided.

⁵github.com/wmt-conference/wmt25-general-mt

(Section 7), describe the test suites (Section 8), and offer concluding remarks (Section 9).

Findings of the General MT Task. We make the following observations:

- ★ **The number of participants continues to grow:** Participation increased again this year, with a total of **36 submissions**. The majority of these participants used LLMs in their systems, most commonly by fine-tuning (Section 4).
- ★ **Automatic scores are biased:** Although Shy-hunyuan-MT placed first for all but one language pair in the automatic rankings, human evaluation revealed its performance was considerably lower than that of the top-rated systems (Section 7.2).
- ★ **Human translations are not always in the winning cluster:** Human references are in the winning cluster for only six out of 15 language pairs (Section 7.2).
- ★ **Constrained models challenge the performance of LLMs:** The top-performing constrained system was Shy-hunyuan-MT, which placed in the winning cluster for 11 language pairs within its category followed by Algharb placed in winning cluster for 6 language pairs. The best system overall, was Gemini 2.5 Pro, which was in top cluster in 14 language pairs (Section 7.2).
- ★ **Speech domain was the most challenging:** The speech domain texts were most challenging to translate (likely due to ASR errors) while literary texts were the easiest (Section 7.2).
- ★ **SOTA systems still struggle with robustness:** Analysis from six specialized test suites reveals that state-of-the-art (SOTA) systems still struggle with robustness to non-standard input, linguistic complexity, domain-specific terminology and gender choice/agreement in particular language pairs. This is despite notable improvements from advanced LLMs in areas like inclusivity and performance in certain specialized domains (Section 8).

2 Test Data

In this section, we describe the test data collection process (Section 2.1) and the creation of human reference translations (Section 2.2). This year, we introduced a new difficulty-based sub-sampling of source texts procedure (Sections 2.1.1 and 6). Motivated by the ever-increasing capabilities of modern MT systems, this step is designed to make our test

sets more challenging by selecting source documents that are estimated to be more difficult to translate.

2.1 Collecting test data

Collecting source data. As in previous years, the test sets consist of unseen translations created specifically for the shared task and released publicly as translation benchmarks. We collected public domain or open-licensed source data from a range of domains, focusing on the most recent data available to minimize potential overlap with the pre-training and fine-tuning data of the systems under evaluation. Importantly, for all language pairs, the source texts were originally written in the source language and subsequently translated into the target languages by human translators. This approach is crucial to avoid “translationese” in the source texts, which can negatively affect evaluation accuracy (Toral et al., 2018; Freitag et al., 2019; Läubli et al., 2020; Graham et al., 2020).

Domain and language coverage. We collected data from six domains (news, literary, speech, social, educational, and dialogue) and three source languages (Czech, English, and Japanese). However, not all domains cover all three source languages. Detailed statistics on our test sets, including the language coverage for each domain, are provided in Table 1.

2.1.1 Difficulty sampling

To identify documents that are particularly challenging for modern MT systems, we adopted the translation difficulty estimation introduced in Proietti et al. (2025). Specifically, we estimated the difficulty of the collected source documents using their best-performing estimator, `sentinel-src-25`.⁶ This model is an improved version of `sentinel-src`, a regression model, based on XLM-Roberta Large (Conneau et al., 2020), which was trained to predict translation quality using only the source text (Perrella et al., 2024).

For each document, we estimated the difficulty of each paragraph and then averaged the results to derive a document-level difficulty score. We then retained the most difficult documents for inclusion in our test sets. To ensure this process did not introduce ill-formed or garbled sources, we also

⁶huggingface.co/collections/Prosho/translation-difficulty-estimators-6816665c008e1d22426eb6c4

Language pair	News	Literary	Speech	Social	Education	Dialogue
#words						
English→*	8,917	9,921	9,007	8,925	-	9,297
Japanese→Chinese	10,427	12,121	10,655	10,589	-	-
Czech→*	8,643	-	8,898	9,311	9,022	7,660
#segments (#words/segment)						
English→*	94 (94.86)	85 (116.72)	62 (145.27)	91 (98.08)	-	52 (178.79)
Japanese→Chinese	93 (112.12)	74 (163.80)	59 (180.59)	108 (98.05)	-	-
Czech→*	132 (65.48)	-	86 (103.47)	99 (94.05)	83 (108.70)	52 (147.31)
#documents (#segments/document)						
English→*	14 (6.71)	2 (42.50)	62 (1.00)	9 (10.11)	-	26 (2.00)
Japanese→Chinese	32 (2.91)	2 (37.00)	59 (1.00)	13 (8.31)	-	-
Czech→*	38 (3.47)	-	86 (1.00)	48 (2.06)	56 (1.48)	26 (2.00)
#sentences (#sentences/segment)						
English→*	364 (3.87)	873 (10.27)	605 (9.76)	572 (6.29)	-	837 (16.10)
Japanese→Chinese	363 (3.90)	305 (4.12)	712 (12.07)	343 (3.18)	-	-
Czech→*	497 (3.77)	-	556 (6.47)	805 (8.13)	838 (10.10)	1017 (19.56)
Type-token ratio of source texts						
English→*	0.41	0.30	0.24	0.36	-	0.14
Japanese→Chinese	0.27	0.17	0.17	0.24	-	-
Czech→*	0.52	-	0.42	0.44	0.51	0.23

Table 1: Per-domain statistics of our test sets, calculated on the source side. To compute word counts, we used simple whitespace tokenization for Czech and English and the [MeCab](#) morphological analyzer for Japanese. Sentence segmentation was carried out using the language-specific Spacy models for English and Japanese, and the multilingual Spacy model for Czech ([Honnibal and Montani, 2017](#)), given that a language-specific one is not available.

manually validated the selected texts and discarded any problematic ones.

We applied this difficulty-based sub-sampling only to the news, speech, and social domains, as the amount of source data available for the other domains was insufficient for the procedure.

2.1.2 Test sets statistics

Source text segmentation. To balance the evaluation across domains and source languages, we aimed for a consistent size for each test set, that is, approximately 9,000 words distributed across 60 to 100 segments (see Table 1).⁷ We also aimed to keep the average segment length around 100 words, whenever possible. This design enables micro-averaging of results without any single category disproportionately influencing the final scores.

These choices were motivated by the aim to keep approximately the same number of segments per domain, which is important for balancing the do-

main for manual evaluation. The number of segments per domain is therefore more balanced than in last year’s test sets, as can be shown in Table 1. The composition of the texts included remains domain-specific due to the differing natures of the texts. For example, in the literary domain texts are longer and therefore only 2 documents were used for each source language. These were split into a large number of segments (42.5 per document on average for English→*). In contrast, the speech domain, where 59-86 documents were used depending on the source language, with each document forming its own segment. The longest segments are in the speech domain, while the shortest are in the news domain (respectively 145.27 and 94.86 words per segment on average for English→*).

Language-specific adjustments. For practical reasons, we were not always able to meet the objective of having a minimum of 100 words per segment. Most notably, the average segment length was longer in the speech domain for English (145.27 words) and Japanese (180.59 tokens), and in the dialogue domain for English (178.8 words)

⁷A segment contains one or more paragraphs, where each paragraph is defined by a line break in the source document. We then separate the segments from each other using double line breaks.

and Czech (147.3 words).

Finally, for Japanese, based on the 1-to-2 ratio derived from our observations of previous WMT Japanese-English test sets, we targeted approximately 18,000 characters per domain. We adopted a character-based metric, because the lack of spaces in Japanese complicates word segmentation. This approach also aligns with industry standards for translation pricing.

In the following paragraphs, we provide other information regarding the data collection process, but specific to each domain.

 **News domain.** For this domain, the data was prepared similarly to previous years (Kocmi et al., 2024a). We collected news articles published in February 2025 on online news sites and extracted the text while preserving the paragraph boundaries. This year we specifically aimed to extract articles that were more challenging to translate. For English, we limited the news crawl to opinion pieces on the basis that they tend to use a more complex, literary style than the straightforward event reporting. For Czech and Japanese, we extracted a larger pool of news covering the entire month to follow up with down-sampling.

 **Social domain.** The social domain data was sourced using the Mastodon Social API.⁸ Mastodon is a federated social network that is compatible with the W3C standard ActivityPub (Webber et al., 2018). Users publish short-form content known as “toots”, with the possibility of replying to other toots to form threads. We decided to use the original server, `mastodon.social` because of its large community and publicly available toots.

We collected data in March and April of 2025, using the reported language ID label to target the source languages of interest.

Given the document-level nature of the task, our aim was to collect threads comprising multiple toots. Our sourcing therefore involved regularly scraping random toots from the previous hour but also identifying and scraping any missing toots that made up threads only partially sourced (identified using the ‘in_reply_to_id’ attribute of already sourced toots). To avoid spam and uninformative toots, we removed empty toots, texts that appeared several times (probable spam), texts from accounts that produced a large number of toots overall (we set this to 100) and from accounts we heuristically

identified as being news outlets or bots (containing the keywords ‘bot’, ‘news’, ‘weather’, ‘sports’, ‘feeds’ or ‘press’ in their handle, as well as a few known media accounts). We created threads from the individual toots and manually selected threads of interest from threads of minimum 2 and maximum 100 toots. Our selection was based on having a diverse range of topics and targeting those characteristic of non-standard user-generated content.

The selected documents contain either a whole toot or a line of text within a toot (depending on whether the toot contained newlines, i.e. there is no distinction between newlines indicating a boundary between two toots and a newline within a toot). Each segment can therefore contain one or several sentences, depending on the original composition of the toots.

A new aspect of the task this year was the inclusion of screenshots capturing entire conversations. This decision was influenced by requests from human translators, researcher efforts to expand last year’s test set (Deutsch et al., 2025), the evolving nature of social media, and an overall interest in exploring the impact of multimodal translation. To accomplish this, we focused on Mastodon conversations that included an image. While the image is not needed for translation, our objective was to provide visual context for translators if needed. Furthermore, even without an image, non-standard domains like social media often convey meaning through layout features such as whitespace, text positioning, and other non-textual elements that may be lost without visual reference. After filtering, we had 481 conversations containing an image for final selection. See an example of the screenshots in Figure 1.

Due to insufficient data for Japanese in the social domain, we adjusted the size of the news test set to compensate, targeting approximately 24,000 characters.

Similarly, we were unable to obtain a sufficient amount of Czech Mastodon data. We therefore decided to use personal communication (usually between a Czech and Ukrainian speaker) from Charles Translator⁹ for the Czech Social domain. This data is similar to the domain called “Personal” last year. The texts were collected with users’ opt-in consent, filtered and pseudonymized in the same way as in the last three years (Kocmi et al., 2022). Each document is one conversation with one user.

⁸mastodon.social/api/v1/timelines/public

⁹translator.cuni.cz (Popel et al., 2024)

The lines reflect the formatting provided by the users. Segment boundaries (empty lines) were added based on the content (trying not to split a paragraph or sentences that are closely related), so as a result the average number of words per segments is close to 100.

📖 Literary Domain. For the English source texts, we selected an amateur-written story from the Archive of Our Own¹⁰ based on several criteria: its recency (published April 2025), its narrative quality, and the absence of explicit sexual or harmful content. The first two chapters, comprising approximately 9.9k words, were treated as two documents. The documents were then segmented so that each segment contained at least 100 words and all paragraph boundaries were maintained.

For the Japanese source texts, we selected two short stories¹¹ from Aozora Bunko, a public-domain digital archive of Japanese literature.¹² Selection was guided by three criteria: recent publication on the platform, the use of modern orthography (*shinjitai*), and a prose style accessible to contemporary readers. In line with the methodology for the English texts, both stories were segmented into passages of similar length ensuring that all paragraph-level boundaries were preserved. For the test set, we used all six chapters of the first story (*Bishōjo Ichiban-nori*) and the first four chapters of the second story (*Omokage*).

🗣️ Speech domain. The speech corpus was compiled from Creative Commons–licensed YouTube videos. For each language, we collected approximately 2,700 videos retrieved with 200 distinct queries spanning documentaries, instructional content, lectures, interviews, news, lifestyle vlogs, sports, and performing arts.

From each video, a segment was randomly sampled, with a minimum duration of 30 seconds and a maximum of 50 seconds, containing at least 30% speech. This constraint was introduced to balance the amount of context required for ASR and translation with the number of available documents.

Transcription of the segments was carried out using the *Whisper-large-v3* model (Radford et al., 2023). For English and Czech, punctuation was expected to be provided directly by Whisper. In

cases where the model hallucinated or returned text without punctuation, *Whisper-large-v2* was used instead. For Japanese, Whisper almost never returned punctuation. Therefore, an additional punctuation restoration model was applied.¹³

After the sampling procedure, 90% of the documents were discarded, and subsequent manual filtering retained approximately one third of the remaining examples.

While the shared task participants had access only to the original videos and automatic transcripts, the reference translations from Czech to Ukrainian and German were prepared with an initial manual correction of the ASR errors using the videos as a first step. As a result, the Ukrainian and German translations are expected to be more accurate in cases where the original transcript was ambiguous.

🎓 Education domain. The Educational domain includes selected exercises from an online portal *Škola s nadhledem*¹⁴ for elementary-school students from various subjects (chemistry, geography, Czech language, etc.). Some segments are not full sentences but short phrases. The reference translations into Ukrainian and German for this domain were created by professional translators within the EdUKate project. Last year, each page of an exercise was treated as a separate document, while this year, each exercise (with all its pages) was compiled into a single document. To meet the target of 9k words per document and 100 words per segment, we excluded documents with less than 90 words. Longer documents were split into multiple segments along page boundaries to ensure that no segment is longer than 200 words.

💬 Dialogue domain. The Dialogue domain texts originates in the MultiWOZ2.4 dataset (Ye et al., 2022)¹⁵ simulated dialogues between a user and a mock dialogue system (Wizard-of-Oz setup) responding to the user’s requests in multiple domains. Each document contains two parts: a description of what should be achieved in a dialogue with the agent (e.g. find a restaurant in Cambridge or finding and booking accommodation), and the dialogue of the user and the agent itself. As such, the sentences in this dataset are rather simple for translation, but the point lies in cross-sentence co-

¹⁰archiveofourown.org

¹¹*Bishōjo Ichiban-nori* (“Pretty Girl, First to Arrive”) from 1938 and *Omokage* (“Reminiscence”) from 1942 by Yamamoto Shūgorō.

¹²aozora.gr.jp

¹³huggingface.co/1-800-BAD-CODE/punct_cap_seg_47_language

¹⁴skolasnadhledem.cz

¹⁵github.com/smartyfh/MultiWOZ2.4

herence and primarily in gender and politeness preservation (or biases), which was promoted in the dataset through our translation process: We selected a subset of the MultiWOZ2.4 test set and professionally translated it from the original English to Czech. English typically keeps the gender of the parties unexpressed and conveys little or no markers of politeness. The first official translation to Czech was thus heavily biased towards the masculine gender and formal politeness level, both of which are explicit in Czech. We requested the translators to add more variance in this regard, i.e. pretend that one or both of the parties are female, and vary the politeness level (formal vs. informal). This varied Czech should not be treated as a canonical reference, but we used it as an interesting *source* for Czech→German translation because German is similarly explicit in gender and politeness as Czech. The participating systems in Czech→German translation thus have to demonstrate their ability to preserve these features (rather than losing them, for example, in implicit or explicit pivoting through English).

2.2 Human References

The test sets were translated by professional translation agencies according to the brief in Appendix B. Since each language pair was sponsored by a different partner, multiple translation agencies contributed, which may account for some variability of the final translations across languages.

Automatic quality assessment of human translations. The quality of human references is critical for reference-based metrics (Freitag et al., 2023). However, obtaining high-quality translations is challenging even with professional translators. This challenge was particularly salient this year, as our difficulty sampling approach (Section 2.1.1) intentionally selected hard-to-translate source texts. Therefore, following WMT24, we report scores from automatic evaluation methods to assess the quality of the collected human references. For this evaluation, we employ the GEMBA-ESA (Kocmi and Federmann, 2023a) as an LLM-as-a-Judge method, using two independent judges: GPT-4.1¹⁶ and Command A (Team, 2025). Our full automatic evaluation approach is detailed in Section 5.

¹⁶openai.com/index/gpt-4-1/

Discussion on the quality of human references.

Table 2 shows the average GEMBA-ESA scores for the human reference translations, broken down by language pair and domain. The two language pairs with the lowest average GEMBA-ESA scores are English→Russian and English→Icelandic. For Russian, this aligns with its human evaluation results, i.e., human translations end up in the third cluster. For Icelandic, however, the pattern diverges: its human reference is the only item in the first cluster, outperforming the best MT system (Gemini 2.5 Pro) by a margin of 20 points (ESA). Given this discrepancy, and the fact that GPT-4.1 largely disagrees with Command A’s lower score for the Icelandic reference, it is possible that Command A is systematically underrating this particular translation.¹⁷ This hypothesis is further supported by Command A’s training data as out of all target languages included in the evaluation, Command A was not optimized to support Icelandic, Estonian, and Serbian (Team, 2025). Consistent with this, both Estonian and Serbian show notable negative difference values (Command A < GPT-4.1): -12.36 and -7.83, respectively. The main counterexample to this trend is English→Japanese; although Japanese is supported by Command A, it also shows a notable negative difference (-8.91) and, like Icelandic, its human reference also ranks alone in the first cluster. Across the remaining language pairs, Command A and GPT-4.1 yield similar absolute scores, with a mean difference of -4.90 points.

Finally, the English→Bhojpuri and English→Maasai pairs were excluded from this QE evaluation, as metric reliability has not been established for these low-resource languages (Section 5; Falcão et al., 2024; Singh et al., 2024; Wang et al., 2024; Sindhuja et al., 2025).

2.3 Test Suites

In addition to the test sets of the regular domains, the test sets given to the system participants were augmented with several *test suites*, which are custom-made test sets focusing on particular aspects of MT translation. The test suites were contributed and evaluated by test suite providers as part of a decentralized sub-task, detailed in Section 8. Across all language pairs of the shared task, test suites contributed 72,449 source test segments

¹⁷Command A scored the Icelandic human reference translation 15.04 points lower than GPT-4.1.

	Lit.	News	Social	Speech	Dial.	Edu	Avg.	Hum.
CS-DE	—	75.9	74.4	72.0	91.3	74.5	77.6	2
CS-UK	—	78.4	79.9	72.3	—	77.2	77.0	2
EN-AR	69.2	66.0	79.0	77.7	—	—	73.0	1
EN-CS	79.0	74.9	78.1	77.1	89.2	—	79.7	3
EN-ET	82.9	75.1	79.2	69.6	—	—	76.7	1
EN-IS	78.1	70.2	75.5	67.9	—	—	72.9	1
EN-JA	83.1	76.7	84.1	80.0	—	—	81.0	1
EN-KO	85.2	82.4	83.9	83.0	—	—	83.6	1
EN-RU	74.3	66.3	75.1	62.9	—	—	69.7	3
EN-SR	84.0	73.3	81.9	76.4	—	—	78.9	4
EN-UK	85.0	82.4	85.4	83.3	—	—	84.0	2
EN-ZH	79.2	68.8	78.0	72.4	—	—	74.6	2
JA-ZH	79.9	82.7	86.6	72.6	—	—	80.5	1

Table 2: GEMBA-ESA scores for human references. Each domain cell is the arithmetic mean of Command A and GPT-4.1; the Avg. column reports the macro-average across available domains. The last column is the human cluster assigned using the ESA protocol.

(detailed numbers can be found in Table 14).

3 Training Data

Similar to the previous years, we provide a selection of parallel and monolingual corpora for model training. The provenance and statistics of the selected parallel datasets are provided in the appendix in Table 18 and Table 19. Specifically, our parallel data selection include large multilingual corpora such as Europarl-v10 (Koehn, 2005), Paracrawl-v9 (Bañón et al., 2020), CommonCrawl, NewsCommentary-v18.1, WikiTitles-v3, WikiMatrix (Schwenk et al., 2021), TildeCorpus (Rozis and Skadiņš, 2017), OPUS (Tiedemann, 2012), CCAIghed (El-Kishky et al., 2020), UN Parallel Corpus (Ziemski et al., 2016), and language-specific corpora such as YandexCorpus,¹⁸ ELRC EU Acts, JParaCrawl (Morishita et al., 2020), Japanese-English Subtitle Corpus (Pryzant et al., 2018), KFTT (Neubig, 2011), TED (Cettolo et al., 2012), and back-translated news.

Links for downloading these datasets were provided on the task web page. We have automated the data preparation pipeline using MT-DATA (Gowda et al., 2021).¹⁹ This year’s monolingual data include the following: News Crawl, News Discussions, News Commentary, CommonCrawl, Europarl-v10 (Koehn, 2005), Extended CommonCrawl (Conneau et al., 2020), Leipzig Corpora (Goldhahn et al., 2012), UberText and Legal Ukrainian.

¹⁸github.com/mashashma/WMT2022-data

¹⁹statmt.org/wmt25/mtdata

Our automated dataset preparation pipeline made a best-effort attempt to preserve document identifiers whenever available in the source. In addition, we have precomputed and shared a quality estimation (QE) metric scores on training data to facilitate data filtering. Specifically, we shared COMETKiwi22 (Rei et al., 2022) annotations for all parallel segments in the training corpora. These were computed using the fast and efficient PYMARIAN framework (Gowda et al., 2024), which enabled QE scoring at scale.

4 System Submissions

This year, we received 96 submissions from 36 participating teams. While the number of submissions is slightly lower than last year’s 105, the number of participating teams increased by roughly a third.

In line with previous years, we included translations from three public MT services, anonymized as ONLINE- $\{B, G, W\}$. We also added contrastive translations from 20 LLMs—including commercial products like GPT-4.1 and open weights models like Llama3.1—and one encoder-decoder system (NLLB). This brought the total to 60 participants.

All participating systems are listed in Table 3. A more detailed description of each submitted system is included in Appendix C, as provided by the authors at the submission time. Section 4.1 discusses the general trends in chosen architectures and training strategies. Section 4.2 presents details on LLM benchmark usage in the task. Section 4.3 outlines two tracks, constrained or unconstrained, to which participants could submit outputs. Section 4.4 describes the submission system platform.

4.1 Architectures and Strategies

Each participating team was asked to submit a form detailing their approach along with an optional description paper. This section discusses the submissions with a summary of strategic details, such as the approach or base model, is provided in Table 3.

Architectures. This year generative language models (denoted as LLM in the table) were the predominant approach, used by all but one external submission. Still, six external team and one organizer submission mentioned using the encoder-decoder architecture, and several others used hybrid approaches.

Base models. The most popular pretrained language models were Qwen (7 submissions), Eu-

Team	Approach	Model	Size	Data	Training	Post-proc	Best rank
Algharb	LLM	Qwen 3	14B	—	SFT, CO	MBR, QE	1-1 (en-zh)
AMI	LLM	Llama 3.2	3B	B, S	CPT, SFT, CO	MBR, QE, rules	11-11 (en-is)
COILD-BHO	LLM	Llama 2	7B	B	CPT, SFT, CO	APE	13-15 (en-bho)
CommandA-WMT	LLM	CommandA	111B	E, S	CO	MBR, reason	5-6 (en-cs)
CUNI-Doc Transformer	enc-dec	from scratch	<1B	B	CPT	checkpoint avg	—
CUNI-EdUKate-v1	LLM	EuroLLM it	9B	B, E, S	SFT, CO, Ada	—	—
CUNI-SFT	LLM	EuroLLM	9B	B, E	SFT	—	16-17 (en-sr)
CUNI-Transformer	enc-dec	from scratch	<1B	B, S	CPT	checkpoint avg	—
DLUT_GTCOM	LLM, enc-dec	Gemma 3	27B	B, E, S	CPT, SFT	prompt	10-15 (en-sr)
HYT	LLM	Hunyuan-TurboS	—	B, E, S	CPT, RL	prompt	—
GemTrans	LLM	Gemma 3	27B	S	SFT, RL	APE	1-4 (en-it)
In2x	LLM	Qwen 2.5	72B	—	CPT	MBR, ensemble	9-16 (jp-zh)
NTTSU	LLM	Qwen 3	14B	B, S	CPT, SFT, CO	MBR	14-15 (jp-zh)
SalamandraTA	LLM	Salamandra	2B, 7B	B, E	CPT, SFT	MBR, TRR	11-15 (en-sr)
SH	LLM	DeepSeek-R1-Distill-Qwen-Japanese	14B	—	SFT, CO	MBR, APE	—
Shy-hunyuan-MT	LLM	Hunyuan	7B	B, E, S	SFT, CO	CO prompt	1-3 (en-ko)
Systran	LLM	EuroLLM	9B	B, E	CPT, SFT	MBR, QE, ensemble, prompt	13-16 (en-jp)
TranssionMT	enc-dec, LLM	NLLB, EuroLLM	1B, 9B	B, E, S	CPT, SFT	ensemble	9-13 (en-mas)
UvA-MT	LLM	Gemma 3	12B	S	SFT	—	5-10 (en-zh)
Wenyii	LLM	Qwen 3	14B	B, E, S	SFT, CO	MBR, QE, ensemble, APE	1-3 (en-uk)
Yandex	LLM	YandexGPT	—	B, E, S	CPT, SFT, CO	—	8-10 (en-ru)
Yolu	LLM	Qwen 3	14B	B, E, S	CPT, SFT, CO	MBR, APE, prompt	7-8 (en-et)
CUNI-MH-v2	LLM	EuroLLM it	9B	E, S	Ada, CO	—	16-16 (en-cs)
KYUoM	enc-dec	NLLB	600M	—	Ada	QE	—
Lanigo	LLM	EuroLLM it	90B	S	Ada/CO	MBR, QE, rules	12-17 (en-et)
CGFOKUS	LLM	Qwen 3	235B	—	—	prompt	—
Erlendur	LLM, hybrid	Claude 3.5 Sonnet	—	E	—	APE, prompt	3-4 (en-is)
IR-MultiagentMT	LLM	GPT-4o-mini	—	E	—	prompt	—
IRB-MT	LLM MAS	Gemma 3 it	12B	—	—	reason, prompt	7-7 (en-arz)
Kaze-MT	LLM	Qwen 2.5	72B	—	—	QE, ensemble	—
KIKIS	LLM	Plamo-2-translate	18B	B	—	reason, ensemble, prompt	13-16 (en-jp)
RuZH-Eole	LLM + Estimator	TowerPlus	9B	—	—	QE	17-18 (en-zh)
SRPOL	LLM, hybrid	EuroLLM, NLLB	9B, 3B	—	—	QE, ensemble	12-15 (en-et)

bb88, ctpc_nlp, TranssionTranslate: no information provided, no paper submitted

SYSTEMS ADDED BY THE ORGANIZERS: all LLMs, except NLLB (enc-dec):

Model	Size	Best rank	Model	Size	Best rank	Model	Size	Best rank
AyaExpanse	8B	4-6 (en-mas)	Gemini 2.5 Pro	—	1-1 (many)	NLLB	3.3B	8-10 (en-bho)
AyaExpanse	32B	7-13 (en-mas)	Gemma 3	12B	9-13 (en-mas)	Qwen 2.5	7B	9-13 (en-mas)
Claude4	—	2-4 (cs-de)	Gemma 3	27B	6-10 (cs-uk)	Qwen 3	235B	6-11 (en-zh)
CommandA	111B	3-3 (en-arz)	GPT-4.1	—	1-3 (cs-uk)	TowerPlus	9B	6-6 (en-is)
CommandR	7B	11-14 (en-arz)	Llama 3.1	8B	9-13 (en-mas)	TowerPlus	72B	8-10 (en-is)
DeepSeek V3	671B	2-6 (cs-de)	Llama-4-Maverick	400B	4-5 (en-mas)	ONLINE-B	—	3-4 (en-sr)
EuroLLM	9B	14-16 (en-it)	Mistral	7B	—	ONLINE-G	—	—
EuroLLM	22B	13-17 (en-et)	Mistral-Medium	—	2-5 (en-jp)	ONLINE-W	—	—

Table 3: Submissions to the General MT shared task, including the externally contributed submissions as well as the systems added by the organizers. Row coloring shows unconstrained-track (dark gray) and constrained-track (white) submissions. Entries are ordered lexicographically, with first the submissions that modified the foundation models somehow (training, tuning, etc), then submissions that created adapters without modifying the models and finally the submissions that used models as is. The last column shows the best rank achieved by the submission, as defined in the official results (see Section 7.4) and the translation direction where the rank was achieved.

Abbreviations: **LLM** (decoder-only language model), **enc-dec** (encoder-decoder), **B** (basic data preproc), **E** (elaborate data preproc), **S** (synthetic data), **CPT** (continued pre-training), **SFT** (supervised fine-tuning), **CO** (contrastive optimization/preference tuning), **Ada** (adapters), **MBR** (Minimum Bayes Risk decoding), **QE** (quality estimation), **rules** (rule-based post-processing/regular expressions), **reason** (reasoning in LLMs), **prompt** (prompting), **APE** (automatic post-editing), and **TRR** (translation re-ranking).

roLLM (6 submissions), and Gemma (4 submissions). Twenty-three submissions modified their base model via continued pre-training, supervised fine-tuning, preference optimizations (CPO/DPO), or reinforcement learning (RL). Three submissions trained adapters without changing the model, and eight used prompting without any model training.

Data preparation. For data preparation, 17 submissions reported using basic filtering (e.g., OPUS cleaner or empirical steps), while 15 reported more elaborate techniques (e.g., filtering with quality estimation by utilizing LLM-as-a-judge or CometKiwi). Synthetic data generation, such as back-translation, was reported by 16 teams.

Post-editing. Finally, for post-editing, 16 submissions reported using LLMs (prompt engineering) for automatic post-editing and/or reasoning. Eleven teams reported using Minimum Bayes Risk (MBR) decoding, and nine mentioned using quality estimation separately.

4.2 LLM Benchmark

LLMs have become popular tools for machine translation (Ataman et al., 2025; Chatterji et al., 2025). Following last year, we provide a systematic and unbiased evaluation of the most popular language models on our blind test sets.

Evaluated models. When deciding which LLMs to evaluate, we selected the best performing constrained and best performing unconstrained model from each popular LLM family. In addition, we collected three popular translation provider services as in previous years. The final list of systems is presented in Table 5.

Prompting LLMs for translation. We designed a unified script to collect translations from all LLMs in an identical setup. We used a zero-shot, instruction-following approach, translating full documents at once. To ensure deterministic outputs, we set the temperature to zero. If a model failed to translate a full document while preserving paragraph breaks, we segmented it into paragraphs and translated each one separately.²⁰ This generic setup may disadvantage systems tuned for specific MT instructions, such as TowerLLM or EuroLLM; these are marked with [M].

²⁰The code for collecting translations is available at: github.com/wmt-conference/wmt-collect-translations

Language model	Doc-lvl	Tokens (in/out)	Cost (\$)
Gemini-2.5-Pro [†]	95.1%	2.1 / 16.4 M	250.8
Claude-4	67.5%	2.6 / 3.5 M	60.4
CommandA	70.6%	2.3 / 3.0 M	35.3
GPT-4.1	97.1%	2.0 / 3.5 M	31.7
DeepSeek-V3	57.2%	2.5 / 2.8 M	6.6
AyaExpand-32B	54.9%	2.5 / 3.0 M	5.7
AyaExpand-8B	43.2%	2.6 / 2.8 M	5.5
Mistral-Medium [‡]	54.1%	2.2 / 2.0 M	5.0
Qwen3-235B	65.6%	2.5 / 3.4 M	2.5
Llama-4-Maverick	70.4%	2.3 / 2.2 M	2.5
Qwen2.5-7B	35.7%	3.2 / 3.5 M	2.0
Mistral-7B	35.1%	1.8 / 3.0 M	1.2
Llama-3.1-8B	27.1%	3.3 / 3.1 M	1.2
CommandR7B	49.3%	2.7 / 3.9 M	0.7

Table 4: Ratio of document-level translated data. Token counts are in millions. [†]Gemini-2.5-Pro used reasoning, increasing cost. [‡]Mistral-Medium did not translate four language pairs. Pricing for open-weight models is estimated via together.ai.

Supported languages. We collected translations for all language directions and tried to collect information about which languages are supported and which are not by looking into the original technical reports to see which languages are mentioned.

As shown in Table 4 in column *Doc-lvl*, one of the key limitations of current LLMs is failure to translate a document at once. This is caused by their window size or a failure to keep paragraph breaks.

API inference and cost. We collect all translations via the respective service APIs during the submission period. Table 4 shows the number of input and output tokens as determined by each model’s tokenizers. The estimated costs shown are for the main test set and do not include the test suites. While we disabled the "reasoning" mode for Qwen3-235B to prevent collection errors, we did not disable it for Gemini 2.5 Pro, which significantly increased its translation cost.

4.3 Constrained and Unconstrained Tracks

To promote fair comparison and encourage replicability, the WMT25 General MT Task distinguished between two tracks: *Constrained* and *Unconstrained*. These tracks differ in terms of model size, licensing, and reproducibility requirements.

Constrained Track: Systems in this track must adhere to the following criteria:

- Use only models and training data that are publicly available under open-source licenses per-

mitting unrestricted non-commercial use (e.g., Apache, MIT).

- The final model must not exceed 20 billion parameters. Larger models may be used during intermediate stages (e.g., for distillation), but the submitted outputs must be produced by a final model that complies with the size limit.
- Model weights must be released under an open-source license before the camera-ready deadline to ensure replicability.

This track is designed to foster transparency and reproducibility, allowing other research groups to replicate and build upon submitted systems.

Unconstrained Track: This track imposes no restrictions on model size, training data, or licensing. It includes systems built with proprietary tools, closed-source models, or undisclosed training pipelines, such as commercial LLMs (e.g., GPT-based systems). While participation in this track is open, systems are expected to provide as much detail as possible about their setup to support inter-pretability.

Although this year, we did not restrict training data for either track, we curated a set of corpora that covers the majority of publicly available resources to support participants in building competitive systems.²¹

To assist participants in the constrained track, we provided a non-exhaustive list of suggested models under the 20B parameter limit, including: textual models: Aya Expanse 8B, Aya 101 (13B), Cohere R 7B, LLaMA 7B and 13B, Qwen 2.5 7B, Mistral 7B and 8B, EuroLLM, NLLB; and multimodal models: Whisper, Seamless M4T. The complete list of systems is presented in Table 5.

Systems were evaluated within their respective tracks: constrained systems were compared only against other constrained systems, while unconstrained systems were evaluated in a broader context that includes all submissions.

4.4 System Submission Platform

We used the open-source OCELoT platform²² to collect system submissions again this year. Given that not all submissions could be included in the human evaluation due to resource constraints, we did not require pre-verification of participating teams. This allowed broader participation and flexibility in the submission process.

²¹www2.statmt.org/wmt25/mtdata

²²github.com/AppraiseDev/OCELoT

	Model	# Params	Open?
Constrained systems	AyaExpanse-8B	8B	✓
	CommandR7B	7B	✓
	EuroLLM-9B	9B	✓
	Gemma-3-12B	12B	✓
	Llama-3.1-8B	8B	✓
	Mistral-7B	7.3B	✓
	NLLB (NLLB-200)	3.3B	✓
	Qwen2.5-7B	7.6B	✓
	TowerPlus-9B	9B	✓
	AyaExpanse-32B	32B	✓
Unconstrained systems	Claude-4	—	✗
	CommandA	111B	✓
	DeepSeek-V3	671B (37B act.)	✓
	EuroLLM-22B	22B (preview)	✓
	Gemma-3-27B	27B	✓
	Gemini-2.5-Pro	—	✗
	GPT-4.1	—	✗
	Llama-4-Maverick	—	✓
	Mistral-Medium	—	✗
	ONLINE-B	—	✗
	ONLINE-G	—	✗
	ONLINE-W	—	✗
	Qwen3-235B	235B (22B act.)	✓
	TowerPlus-72B	72B	✓

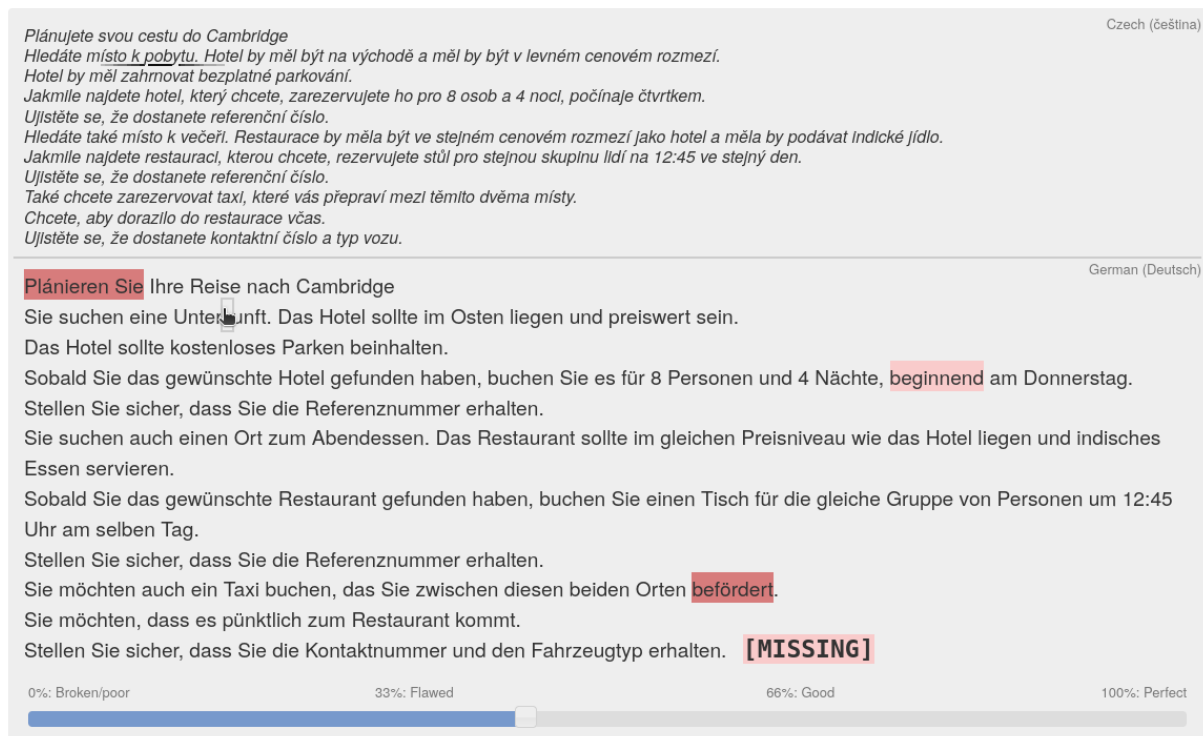
Table 5: List of constrained and unconstrained systems with parameter count. Open-weight models were marked with a cmark (✓).

Participants were asked to submit their systems in a single JSONL file, covering all language pairs, and to use a provided verification script to ensure that the submission adheres to strict formatting and completeness requirements. The verification process included the following checks:

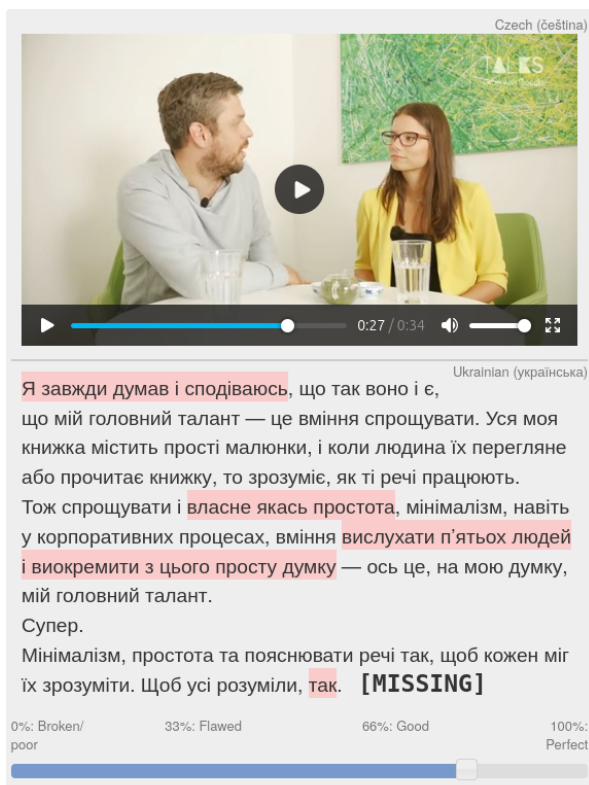
- **Completeness of translations:** while participants could submit outputs for only a subset of languages, each submitted language had to be fully translated.
- **Format verification:** ensured that all documents from the testset were translated with correct paragraph boundaries preserved.
- **Inclusion of testsuite translations:** verified that testsuite segments were not omitted because of their length or other reasons.

To avoid premature publication of rankings based on automatic metrics, all submissions were displayed anonymously on the leaderboards during the submission period.

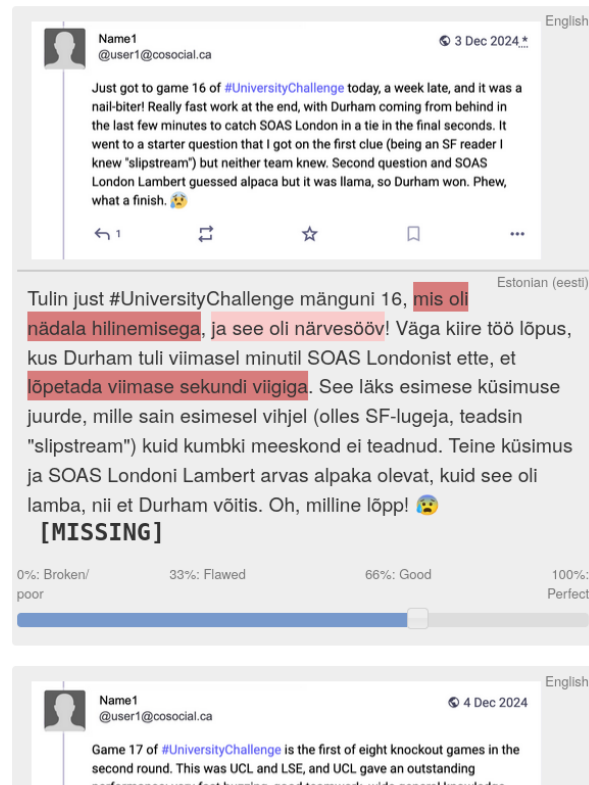
Teams had one week after the submission deadline to select a single primary submission, specify the track for that submission, and submit an abstract for their system description paper. These steps were mandatory for a system to be included in the human evaluation campaign.



(a) Screenshot of Czech→German annotations in the dialogue domain. Mouseover on the target side shows coarse alignment on the source side.



(b) Screenshot of Czech→Ukrainian annotations in the speech domain. The video can be paused and replayed.



(c) Screenshot of English→Estonian annotations in the social domain. The image partially shows the next segment which provides context.

Figure 1: Three screenshots of ESA (Kocmi et al., 2024b) annotations. ESA shows multiple segments within a document at once as well as video or image sources. After marking the individual error spans, the annotator assigns the final segment score from 0 to 100.

Language pairs	Annotators' profile	Tool
English→Chinese,Japanese	Microsoft annotators: bilingual target-language native speakers, professional translators or linguists, experienced in machine translation evaluation.	Appraise ESA
Czech→Ukrainian,German English→Czech	ÚFAL Charles University annotators: linguists, annotators, researchers, and students who were native speakers of one of the languages and high proficiency in the other.	Appraise ESA
Ukrainian,Russian English→Arabic,Serbian Maasai,Bhojpuri	Toloka AI paid expert crowd: Bilingual native target-language speakers who were high-performing on the platform.	Appraise ESA
English→Estonian	Professional translators from Luisa Language Solutions.	Appraise ESA
English→Icelandic	The Árni Magnússon Institute for Icelandic Studies annotators: bilingual target-language native speakers, paid translators with 3–25 years of experience in Icelandic↔English translation.	Appraise ESA
English→Italian	Cohere annotators: professional in-house employees experienced with annotations of general LLM outputs, bilingual speakers native in Italian.	Appraise ESA
English→Korean Japanese→Chinese	Professional translators from Venga Global.	Anthea MQM

Table 6: Annotators’ profiles and annotation tools for each language pair in the human evaluation. All annotators were paid a fair wage in their respective countries.

5 Automatic Evaluation

As in the last year, a high number of submissions²³ made full comprehensive manual evaluation infeasible. We therefore employed automatic metrics to select systems for human evaluation via a procedure we call AUTORANK. This year’s procedure improves upon WMT24 by incorporating a broader set of metrics and a revised aggregation method.

Metrics. For most language pairs (see “low-resource exception” below) the AUTORANK is a combination of three distinct families of evaluation methods:

- **LLM-as-a-Judge (reference-less).** We use GEMBA-ESA (Kocmi and Federmann, 2023b) with two independent judges: GPT-4.1²⁴ and Command A (Team, 2025), both in a reference-less setting.
- **Trained Reference-based Metrics.** Two supervised metrics trained to approximate human quality judgments with references: MetricX-24-Hybrid-XL²⁵ (Juraska et al., 2024) and XCOMET-XL²⁶ (Guerreiro et al., 2024).

²³We received submissions from 36 unique teams. A total of 43 teams initially registered, but 7 later withdrew or were disqualified.

²⁴openai.com/index/gpt-4-1/

²⁵huggingface.co/google/metricx-24-hybrid-xl-v2p6

²⁶huggingface.co/Unbabel/XCOMET-XL

- **Trained Quality Estimation (QE).** The reference-less QE metric CometKiwi-XL²⁷ (Rei et al., 2023), which is also trained to mimic human judgments.

This combination of reference-based and reference-less (or QE) methods is designed to balance their complementary failure modes. Reference-based metrics typically achieve a higher correlation with human judgments when high-quality references are available, while reference-less methods reduce susceptibility to reference bias when references are suboptimal (Freitag et al., 2023). We also account for known issues with specific metrics. To mitigate a common QE pitfall, i.e., being fooled by fluent output in the wrong language, the GEMBA-ESA prompt explicitly specifies the target language. However, while GEMBA-ESA is intended to reduce bias toward systems that use re-ranking, we note that some participants incorporated it directly as a reward model.

System-level scores. The system-level score for each language pair is the average of its paragraph-level (segment-level) scores from each metric across the testset. In particular, paragraphs constitute the input units for all the metrics. We make one exception for language pairs without human

²⁷huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

references by excluding CometKiwi-XL from the AUTORANK computation. This avoids redundancy, as the other hybrid metrics (MetricX-24-Hybrid-XL and XCOMET-XL) can also run in a reference-less (QE) mode to provide the necessary QE signal.

Low-resource exception. For the two lowest-resource languages in the testset, i.e., Bhojpuri and Maasai, we rely solely on chrF++ (Popović, 2017), computed with sacrebleu²⁸ (Post, 2018). This approach was chosen because the reliability of our main metrics is unestablished for these languages (Falcão et al., 2024; Singh et al., 2024; Wang et al., 2024; Sindhujan et al., 2025), whereas human references required for chrF++²⁹ were available. Moreover, our cross-metric correlation study—based on Pearson correlations of paragraph-level scores across all systems within each language pair—shows that Bhojpuri, Maasai, and Marathi have the weakest inter-metric agreement (Kocmi et al., 2025b). This observation further supports our use of chrF++ for Bhojpuri and Maasai. For Marathi, reference translations are not available, so its evaluation necessarily relies on QE metrics.

From system-level scores to AUTORANK. To combine the metrics into a single score, we first normalize them using median-interpercentile scaling to address differences in scale and reduce the influence of low-performing outliers. We then compute the average using equal weights. Finally, we linearly rescale the results to the range from 1 to N systems. A detailed description is provided below:

Let S be the set of submitted systems for a given language pair, $|S| = N$, and let M be the set of automatic metrics used for that language pair (for Bhojpuri and Maasai, $|M| = 1$). For each metric $m \in M$ and system $s \in S$, we compute a system-level score $x_s^{(m)}$ as the average of that metric over all available test segments. To combine scores across metrics, we first map them to a common scale; however, classical min-max normalization is highly sensitive to outliers. In fact, anchoring the scale at the single worst and best system allows an extremely low-scoring outlier to set the lower bound and compress the remaining scores into a narrow band near the top, obscuring meaningful differences among competitive systems. To down-

weight extremes without discarding any system, we apply a median-interpercentile scaling to each metric m :

$$\tilde{x}^{(m)} = \text{median} \{x_s^{(m)} \mid s \in S\}, \quad (1a)$$

$$D^{(m)} = \max \left(\varepsilon, Q_{100}^{(m)} - Q_{25}^{(m)} \right), \quad (1b)$$

$$z_s^{(m)} = \frac{x_s^{(m)} - \tilde{x}^{(m)}}{D^{(m)}}. \quad (1c)$$

Where $\varepsilon > 0$ and $Q_p^{(m)}$ denotes the p -th percentile of $\{x_s^{(m)} : s \in S\}$. Importantly, Eq. (1) is continuous and monotonic: it keeps all systems and preserves their order within each metric. Then, for each system, we average the robust-scaled values across metrics:

$$\bar{z}_s = \frac{1}{|M|} \sum_{m \in M} z_s^{(m)}. \quad (2)$$

Averaging after robust scaling yields a single comparable score that preserves the magnitude of performance differences between systems (in standardized units) while preventing any single metric’s outliers from dominating. Finally, for readability and to follow the WMT convention from last year (lower is better in AUTORANK, i.e., 1 is best and N worst), we apply a final linear mapping to the set $\{\bar{z}_s\}_{s \in S}$. Specifically, within $\{\bar{z}_s\}_{s \in S}$ the system with the highest average score is assigned 1, the system with the lowest average score is assigned N , and all remaining systems are placed linearly between these two endpoints. This remapping is applied only once—after the cross-metric aggregation—so it preserves the ordering and relative spacing between systems while retaining the outlier mitigation provided by the robust scaling. We refer to the resulting value as AUTORANK in the various tables.

Selecting systems for human evaluation. Following the procedure established in the preliminary report (Kocmi et al., 2025b), we use AUTORANK to select the subset of systems that undergo manual assessment. The target size is 18 systems per language pair, although this number can be higher in certain cases. Selection proceeds in two steps:

1. **Prioritizing constrained systems.** We first select the top-8 performing constrained systems according to AUTORANK.
2. **Filling to target.** We then add the best remaining systems—constrained or unconstrained in

²⁸github.com/mjpost/sacrebleu

²⁹SacreBLEU signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.5.1.

order of AUTORANK—until the language pair reaches the target number of systems.

Systems not selected for human evaluation keep their AUTORANK ranking as the official result for that language pair (Kocmi et al., 2025b).

5.1 Identification of Error Cases and Impact on Automatic Metrics

Generating text in wrong language. A problem that is particularly common with LLMs used for MT is generating text in the wrong language, either because of copying the source or getting the wrong target language (Bawden and Yvon, 2023; Zhang et al., 2023). To diagnose this issue, we ran language identification on the outputs of the systems using `fasttext` (Joulin et al., 2017b,a), considering a system-language pair problematic if the target language is incorrectly detected in more than 10% of examples.

The most problematic target languages detected by `fasttext` were Serbian, both in Latin script (13 systems) and Cyrillic script (9 systems), Kannada (6 systems), and Marathi (6 systems). Many of these outputs were incorrectly detected as closely related languages: Croatian, Serbo-Croatian and Bosnian for Serbian (particularly when in Latin script), and Hindi for Marathi. In the case of Serbian, these misclassifications appear to be mainly a consequence of the language identification tool’s bias: when the same outputs are transcribed into Cyrillic, Serbian is correctly detected. For Marathi, which shares a script with Hindi, the issue is more substantive, with outputs containing mixed or predominantly Hindi content. Kannada outputs, written in a different script, are visually telling: many contain Devanagari script (used for Hindi) or a mix of Devanagari and Kannada characters.

Copying the source text. Another commonly observed issue is generating output in the source language, particularly directly copying the source text. In our case, this mostly corresponds to English source texts, which is unsurprising, given that English was by far the most common source language. Aside from clear failure cases (e.g. NLLB and Gemma-3-12B outputs indicating “FAILED”), many of output language errors, particularly from CommandR7B and EuroLLM-9B (other models to a lesser extent), come from copying large portions of the source text. In some cases, this copied content is mixed with text in the correct target language.

Impact of incorrect language and source copying on automatic evaluation. Copying source text instead of translating it can pose challenges for automatic evaluation tools, particularly those that reward semantic similarity without taking into account the intended target language. In AUTORANK, this applies to the CometKiwi-XL metric, which calculates scores solely based on the source text and does not account for the correctness of the target language. To estimate the impact of source copying on this QE metric, we run a controlled experiment comparing predicted scores for (i) reference texts, (ii) source texts and (iii) different degrees of mixing between source and reference texts to simulate different levels of copying.³⁰ The results in Table 7 show that, in this setup, source copying does not artificially inflate scores as might be expected. Scores for source-only outputs are well below those for the reference, and the greater the proportion of copied source text in the reference, the lower the score. Interestingly, partial copying leads to approximately the same score degradation as only partial translation, i.e., when the same copied portion is empty.

The metrics generally show positive correlation with each other. However, for a significant number of paragraphs, metrics diverge in their estimation of translation quality. To better understand whether specific patterns exist related to source copying or incorrect language generation, we sought to find instances where metric scores diverged for individual paragraphs. For each paragraph (specific to a language direction), we ranked the system scores and assigned each to a decile (1–10) based on their relative performance. Repeating this process for each metric, we defined the spread for a given paragraph as the difference between the highest and lowest decile across metrics. We observed a large number of divergent instances, even with wide spreads (see Table 8), indicating that metrics behave very differently in certain scenarios. While some metrics are clearly more similar to each other (e.g., LLM-based metrics and the two COMET-based metrics), the dissimilarities in rankings are present between all metrics.

Refer to Figure 2 for how often each metric appears as the score in the lowest or highest decile

³⁰For each paragraph we took the first 25%, 50% or 75% of whitespace separated tokens from the source text and concatenated them with the last 75%, 50% or 25% of the reference text. We also tested using only the initial parts of the source text without concatenating the reference texts.

Hypothesis	Score
Source	0.23
Ref.	0.55
Source (†25%)	0.27
Source (†50%)	0.26
Source (†75%)	0.26
Source (†25%) · Ref. (†75%)	0.46
Source (†50%) · Ref. (†50%)	0.38
Source (†75%) · Ref. (†25%)	0.32
Ref. (†25%)	0.32
Ref. (†50%)	0.38
Ref. (†75%)	0.45
Ref. (†25%)	0.33
Ref. (†50%)	0.41
Ref. (†75%)	0.46

Table 7: CometKiwi-XL scores for different hypothesis (compared against the source). · indicates concatenation, and percentages indicate a percentage of the total text tokens taken either consecutively from the start of the text (†) or the end (†).

relative to others when spreads are greater than 5. No consistent pattern emerges: some metrics appear to reward source copying in certain examples, but not in others, and the overall correlation with the amount of copying and the metric scores is often small. For example, CometKiwi-XL is negatively correlated with the amount of copying (as measured by the 4-gram precision count calculated by sBLEU) and XCOMET-XL shows the strongest positive correlation at 0.061 (Pearson coefficient). Similarly, correlations between scores and correct target language decision are also low.

Some further investigation is necessary in following years to better understand the impact of errors on the different metrics, particularly if they are to be used for automatic ranking. Notably, preliminary experiments indicate that the different metrics used are not necessarily well aligned in terms of scores of individual paragraphs, and a more in-depth study could help to understand discrepancies.

	0	2	4	6	8
#paragraphs	370k	320k	206k	89k	22k
%paragraphs	100	86.35	55.73	23.97	5.83

Table 8: Number of paragraphs for which the decile spread across metrics is equal to or higher than thresholds 0–9.

	CometKiwi-XL	GEMBA-ESA-CMDA	GEMBA-ESA-GPT4.1	MetricX-24-Hybrid-XL	XCOMET-XL
CometKiwi-XL	0	7.1k	6.1k	5.7k	7.3k
GEMBA-ESA-CMDA	9.1k	0	4.9k	9.4k	12.4k
GEMBA-ESA-GPT4.1	7.0k	4.7k	0	7.0k	9.6k
MetricX-24-Hybrid-XL	5.0k	7.0k	5.2k	0	7.0k
XCOMET-XL	3.7k	9.7k	7.6k	5.3k	0

Figure 2: Disagreement matrix showing the number of paragraph–system instances where metric i assigned the minimum decile and metric j assigned the maximum decile. Only cases where the spread between the highest and lowest decile across metrics was greater than 5 are included.

6 Human Evaluation

The human evaluation is done primarily using Error Span Annotation (ESA; Kocmi et al., 2024b). For English→Korean and Japanese→Chinese we rely on the Multidimensional Quality Metrics (MQM; Lommel et al., 2014a).

The ESA Protocol. The annotators (professional translators but not experts in MQM/ESA-style annotations) were asked to mark each error as well as its severity, “Minor” or “Major”. In addition, the annotators were also asked to assign a score from 0 to 100, similar to Direct Assessment (DA), to the whole annotation segments (usually a paragraph). However, the ESA score should be more robust than DA alone because the annotators are primed by the highlighted errors at the time of the scoring.

The ESA interface. The interface is shown in Figure 1 with annotator instructions in Appendix A. At the start of annotation, each annotator was exposed to an interactive tutorial where they were asked to interact with the system. The source for the speech domain is a video which is shown in a native HTML video player. The output of the ESA annotation is a list of errors and their severity (minor or major) and the final score from 0 to 100 for each segment.

Task setup. The whole annotation was split into “tasks” where each task had a balanced number of words to make it approximately 1 hour long. Each task is done by a single annotator and contains segments from a single domain but contains output from multiple systems. In contrast to previous years, we do not include a quality control check due

to their annotation costs and low reliability (Zouhar et al., 2025a). Instead, we include “control tasks” for each language, which is the same task that each participating annotator has to fill out exactly once. Because these control tasks are fixed, this allows us to model annotator bias and reliability. Finally, each segment is annotated exactly twice, which can be used to estimate inter-annotator agreement and is especially useful for the metrics shared task (Lavie et al., 2025). See list of changes in contrast to previous version in Appendix A.

Diversity sampling. From the whole testset which all systems translated, we select a subset to human-evaluate. Specifically, we select a 50% of the original data which contains sources that lead to the most diverse translations (as measured by average pairwise ChrF). This ensures that we do not spend the evaluation budget on segments that have very similar translations, which contribute less to the final system ranking (Zouhar et al., 2025b).

The MQM protocol. MQM (Multidimensional Quality Metrics; Lommel et al. (2014b); Freitag et al. (2021)) is the translation evaluation framework that ESA is based on. Professional translators annotate error spans, assigning to each a severity (Major or Minor) and a category from a two-level error hierarchy (e.g. Accuracy/Mistranslation or Fluency/Grammar). Instead of then asking annotators to assign a numeric score to each segment’s translation, scores are automatically calculated by applying a severity- and category-dependent weighting scheme to each error and summing the results.

MQM interface. MQM ratings were collected using the open-source Anthea³¹ framework. Similarly to the ESA annotations, for the speech domain the video was used as the source side of the evaluation.

Task setup. Due to concerns with rater fatigue, steps were taken to limit the expected time for any individual MQM rating task. To this end, documents were truncated at paragraph boundaries to include no more than 12 source sentences; if the first paragraph contained more than 12 sentences, the document was skipped. For the literary domain in particular, to avoid truncating the vast majority of segments in the very-long documents, the text

³¹github.com/google-research/google-research/tree/master/anthea

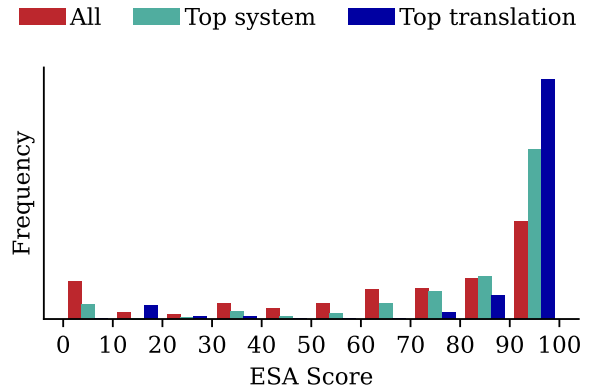


Figure 3: Distribution of final human segment-level scores (ESA) of all systems, top-system in each language, and top translation for each segment. The buckets have a width of 10 (corresponding to xticks). Results include all languages.

was instead split into chunks of one or more paragraphs up to the 12-sentence limit, and each chunk was treated as a separate document for the purposes of conducting the evaluation.

6.1 Human Evaluation Analysis

Score distribution. We analyze the score distribution from three perspectives: (1) scores across all systems, (2) scores of the top system in each language pair, and (3) score of top translation for each source segment. While Figure 3 shows that a near-perfect translation (a score of 90-100) is achievable for most source segments by at least one system, the performance of any single system is more modest. Even the top-performing system often fails to achieve a score above 90, a trend that is naturally more pronounced for lower-ranked systems.

Cost and reliability. We also analyze the annotation process itself, focusing on the trade-off between cost and reliability (Kocmi et al., 2024b; Zouhar et al., 2025a). Table 9 provides an overview of annotation volume, time (as a proxy for cost), and inter-annotator agreement (pairwise Pearson). Our analysis finds that while annotation speed and the number of annotated errors vary considerably, neither is predictive of inter-annotator agreement.

Domain and language difficulty. In Table 10, we present the score of the top-performing system for each language and domain. Focusing on the top-performing system mitigates the effect of low-performing outliers. While absolute scores are not directly comparable across languages (due to dif-

	Annotations	Ann./Sys.	Time/Seg.	Time/Word	Minor/Major	Annotators	IAA
English→Czech	8480	424	77.2s	0.9s	2.2/0.9	17	0.41
English→Masai	7030	370	56.9s	0.7s	0.2/1.2	4	0.17
English→Serbian (Cyrilic)	7828	412	88.5s	0.9s	2.0/1.6	4	0.57
English→Japanese	7182	378	89.2s	1.0s	1.0/0.2	40	0.28
English→Ukrainian	7562	398	23.0s	0.3s	0.8/0.3	2	0.64
English→Estonian	7220	380	90.1s	1.0s	2.7/1.1	8	0.53
English→Arabic (Egyptian)	7562	398	40.3s	0.5s	1.3/0.4	3	0.96
Czech→Ukrainian	8550	450	36.1s	0.5s	0.9/0.7	7	0.35
Czech→German	9702	462	81.2s	1.1s	2.0/1.8	12	0.52
English→Russian	7600	400	68.3s	1.5s	1.8/0.9	6	0.36
English→Bhojpuri	7182	378	96.8s	1.3s	1.7/2.0	4	0.85
English→Italian	7740	430	100.7s	1.0s	1.4/0.6	7	0.51
English→Chinese	7524	396	96.2s	1.5s	1.6/0.4	39	0.23
English→Icelandic	7676	404	62.0s	0.7s	3.2/1.7	5	0.69

Table 9: Overview of collected human evaluation data: Number of annotations, Number of annotations per system, Time per segment, Time per source word, Minor and major errors per segment, Number of annotators, Inter-annotator agreement (pairwise Pearson correlation on control subset that is annotated by all).

	Literary	Speech	Social	News	Avg.
En.→Czech	96.1	87.7	86.5	89.9	90.1
En.→Maasai	17.5	8.7	10.9	8.8	11.5
En.→Ukrainian	95.2	87.2	90.2	91.2	90.9
En.→Bhojpuri	98.8	88.3	95.1	95.9	94.5
En.→Japanese	98.4	86.3	92.1	88.9	91.4
En.→Icelandic	87.4	87.3	86.8	88.4	87.5
En.→Estonian	96.8	71.0	88.9	83.0	84.9
En.→Chinese	98.3	83.7	92.4	87.7	90.5
En.→Russian	94.8	77.0	83.7	83.6	84.8
En.→Arabic (Egy.)	88.4	80.8	84.3	74.7	82.0
En.→Serbian (Cyr.)	98.6	89.5	96.0	93.5	94.4
En.→Italian	87.0	71.9	83.4	86.1	82.1
Cz.→Ukrainian		89.4	92.9	94.7	92.3
Cz.→German		90.4	95.7	88.8	91.6

Table 10: Human evaluation scores for the top-performing system per language, by domain.

ferent sets of systems, annotators, and sources), we again observe a consistent pattern (Kocmi et al., 2024a): the speech domain receives the lowest scores, suggesting it is the most challenging, likely due to its reliance on ASR-generated text. The social and news domains follow in difficulty. Surprisingly, the literary domain achieved the highest scores.³²

Most language pairs show similar difficulty patterns, with English→Maasai as a notable outlier. Overall, English→Arabic (Egyptian) and English→Italian proved to be the most challenging.

³²This could be due to two factors: first, its source texts were not subject to difficulty sampling due to limited number of stories. Second, an evaluation that asks annotators to mark errors may award high scores to translations that are technically error-free, even if they do not fully capture stylistic qualities such as the author’s voice or the reader’s enjoyment (Carpuat et al., 2025).

7 Official Ranking Results

We now describe how we compute the final ranking and then discuss the final results and potential issues. The ranking is presented in tabular form in Section 7.4.

7.1 Human Ranking Computation

We calculate three different scores: the human ESA or MQM score, rank, and the cluster. The human ESA or MQM scores are the micro-average of the segment-level scores. This disregards any domain balancing, though we show per-domain results in Appendix D. For the statistical analysis and clustering, we use the Wilcoxon signed-rank test, a paired non-parametric test (Wilcoxon, 1945), with $p < 0.05$. The rank ranges differ from last year’s implementation. Systems are sorted by their average human score and for a system in row i we define its rank range $\langle i_{\downarrow}, i_{\uparrow} \rangle$ as follows: $i_{\downarrow} := \max\{j | j < i, \text{significant}(i, j)\} + 1$ and $i_{\uparrow} := \min\{j | j > i, \text{significant}(j, i)\} - 1$. In words, the ranks expand from i until a system that is statistically distinguishable is encountered. Lastly, the clusters are the maximal partition of systems such that ranks of systems from one cluster do not overlap with ranks of systems in another cluster.

7.2 Human Evaluation Discussion

In this section, we discuss the results of the human evaluation presented in Section 7.3 (constrained systems only) and in Section 7.4 (all systems).

7.3 Ranking of Constrained Systems

English→Chinese		
Rank	System	Human
1-1	Algharb	88.4
2-2	Shy-hunyuan-MT	88.2
3-3	Human	82.1
4-5	SRPOL	77.7
4-6	IRB-MT	76.5
5-6	RuZh	75.7
7-7	Laniquo	70.5

English→Ukrainian		
Rank	System	Human
1-1	Algharb	90.0
2-2	Shy-hunyuan-MT	88.4
3-3	Human	87.3
4-4	TowerPlus-9B[M]	84.2
5-5	IRB-MT	82.9
6-7	SRPOL	79.9
6-7	Laniquo	79.8

English→Italian		
Rank	System	Human
1-1	Shy-hunyuan-MT	78.7
2-3	TowerPlus-9B[M]	61.2
2-3	IRB-MT	60.3
4-6	SalamandraTA	57.5
4-6	AyaExpanse-8B	57.0
4-6	EuroLLM-9B[M]	56.6
7-8	Gemma-3-12B	53.6
7-8	Laniquo	53.4

English→Czech		
Rank	System	Human
1-1	Shy-hunyuan-MT	87.1
2-2	Human	84.5
3-3	Algharb	76.7
4-4	CUNI-MH-v2	71.0
5-6	SRPOL	67.5
5-7	Laniquo	66.1
6-7	TowerPlus-9B[M]	65.8
8-8	SalamandraTA	60.3

English→Arabic (Egyptian)		
Rank	System	Human
1-1	Human	78.5
2-2	IRB-MT	51.9
3-5	CommandR7B	3.7
3-6	Algharb	3.2
3-6	Shy-hunyuan-MT	3.2
4-6	AyaExpanse-8B	2.0
7-7	SRPOL	0.9

English→Russian		
Rank	System	Human
1-1	Shy-hunyuan-MT	80.2
2-2	Algharb	73.3
3-3	Human	70.5
4-4	IRB-MT	65.4
5-7	RuZh	57.9
5-7	SRPOL	56.9
5-7	Laniquo	56.2

English→Estonian		
Rank	System	Human
1-1	Human	83.1
2-3	Algharb	70.4
2-3	Shy-hunyuan-MT	70.3
4-5	SRPOL	49.4
4-6	Laniquo	48.6
5-6	SalamandraTA	46.7
7-7	IRB-MT	32.4

English→Bhojpuri		
Rank	System	Human
1-2	Human	92.6
1-2	Algharb	91.1
3-3	NLLB	75.6
4-4	COILD-BHO	68.7
5-5	IRB-MT	59.6
6-6	SalamandraTA	35.7
7-7	Shy-hunyuan-MT	1.7

Czech→German		
Rank	System	Human
1-1	Shy-hunyuan-MT	87.2
2-2	Human	82.8
3-4	Algharb	80.9
3-4	TowerPlus-9B[M]	79.8
5-7	CUNI-MH-v2	77.2
5-7	Gemma-3-12B	76.8
5-7	SRPOL	76.7
8-9	IRB-MT	71.7
8-9	Laniquo	70.0

Czech→Ukrainian		
Rank	System	Human
1-1	Shy-hunyuan-MT	91.8
2-2	Human	90.1
3-4	TowerPlus-9B[M]	85.3
3-5	Algharb	84.1
4-6	Laniquo	83.4
5-6	IRB-MT	82.7
7-7	SRPOL	80.8

Japanese→Chinese		
Rank	System	Human
1-1	Human	-3.5
2-3	Algharb	-5.8
2-3	Shy-hunyuan-MT	-6.1
4-5	NTTSU	-11.3
4-5	TowerPlus-9B[M]	-13.3
6-6	IRB-MT	-13.9
7-7	Laniquo	-18.3

English→Serbian (Cyrilic)		
Rank	System	Human
1-1	Shy-hunyuan-MT	92.2
2-2	Human	88.7
3-3	IRB-MT	77.6
4-5	SalamandraTA	75.5
4-5	Gemma-3-12B	74.8
6-7	CUNI-SFT	60.9
6-7	Llama-3.1-8B	58.4
8-8	NLLB	53.5
9-9	EuroLLM-9B[M]	41.8

English→Icelandic		
Rank	System	Human
1-1	Human	87.5
2-2	Shy-hunyuan-MT	63.2
3-3	TowerPlus-9B[M]	57.4
4-4	AMI	39.9
5-5	SalamandraTA	31.3
6-6	NLLB	24.1
7-7	IRB-MT	20.7
8-8	Gemma-3-12B	16.5
9-9	Llama-3.1-8B	10.5

English→Korean		
Rank	System	Human
1-2	Human	-1.9
1-2	Shy-hunyuan-MT	-2.5
3-4	Algharb	-4.4
3-5	IRB-MT	-5.6
4-5	Gemma-3-12B	-5.9
6-6	TowerPlus-9B[M]	-7.2
7-7	Laniquo	-9.1

English→Japanese		
Rank	System	Human
1-1	Human	89.2
2-2	Algharb	85.7
3-3	Shy-hunyuan-MT	79.9
4-5	KIKIS	76.2
4-5	Systran	75.6
6-6	NTTSU	73.3
7-7	Laniquo	67.8

English→Masai		
Rank	System	Human
1-1	Human	9.6
2-3	AyaExpanse-8B	6.0
2-3	Shy-hunyuan-MT	4.8
4-6	Llama-3.1-8B	3.0
4-6	Gemma-3-12B	3.0
4-6	Qwen2.5-7B	2.8
7-9	CommandR7B	1.6
7-9	TowerPlus-9B[M]	0.8
7-9	EuroLLM-9B[M]	0.7

Human evaluation shifts overall ranking. Overall, the best-performing model in the human evaluation is Gemini 2.5 Pro (see Figure 5).³³ It places in the top cluster for 14 of the 16 evaluated language pairs and is on par with or surpasses human translation in 10 of those pairs.³⁴

This result differs notably from the AUTORANK results (Kocmi et al., 2025b), where Shy-hunyan-MT ranked first for all but one language pair (English→Bhojpuri). The discrepancy may be due to Shy-hunyan-MT’s use of GRPO with XCOMET-XXL and GEMBA with DeepSeek V3 as its training signals.³⁵ However, despite its lower ranking in the human evaluation, Shy-hunyan-MT remains the best-performing *constrained system*, winning in this category for 11 language pairs. The second best constrained system is Algharb, which ranks first in six language pairs.

Finally, human translations, often treated as a “gold standard,” place in the top cluster for only six of the 15 language pairs where they were available. This finding highlights the inherent difficulty of translation, though it could also reflect the stylistic or lexical preferences of the annotators.

7.4 Official Ranking Results Tables

Results tables legend

The human score is the micro-average of human judgments across all domains and double annotations (single annotations for MQM language pairs). AutoRank is calculated from automatic metrics as per (Kocmi et al., 2025b). Significance testing is done using a [Wilcoxon signed rank test](#) with a p -value threshold of 5%. The rank range for the i th model begins as $\langle i, i \rangle$ and is expanded in both directions until a significant difference is found. Clusters are formed such that their constituent rank ranges do not overlap.

Systems are either constrained (white), or unconstrained (gray). Systems that do not officially support the language pair are marked with  and those where language support cannot be verified are marked with . The [M] suffix marks systems (submitted by the WMT organizers) that were trained/tuned with specific MT instructions, but prompted without these specific instructions (using a generic setup, same for all LLMs, see Section 4.2), which could disadvantage these systems. See Appendix D for a per-domain breakdown of system performances.

³³The translations with Gemini 2.5 Pro were collected with the “thinking” mode enabled, and it is unclear how much this contributed to its overall performance.

³⁴Note that while human translations for WMT are prepared by professional translators, they are not necessarily free of errors or undergo the same level of post-editing as human translations used in high-stake scenarios (e.g., sworn translation) or published translations (e.g., literary translation).

³⁵Similar shifts, i.e., a higher automatic rank and lower placement after human evaluation, are visible for other systems, some of which also use metrics as a signal in training or reranking (see Figure 4).

English→Arabic (Egyptian)			
Rank	System	Human	AutoRank
1-1	Human	78.5	
2-2	GPT-4.1	77.0	6.7
3-3	CommandA	74.0	8.6
4-4	Gemini-2.5-Pro	60.6	5.8
5-6	DeepSeek-V3?	56.8	7.0
5-6	Claude-4	55.7	7.8
7-7	IRB-MT	51.9	11.1
8-9	Mistral-Medium	36.0	7.7
8-9	CommandA-WMT	34.6	4.1
10-10	UvA-MT	29.0	4.2
11-14	CommandR7B	3.7	11.6
11-14	GemTrans	3.7	3.5
11-16	Algharb	3.2	2.7
11-16	Shy-hunyan-MT	3.2	1.0
13-16	AyaExpanse-8B	2.0	9.9
12-16	ONLINE-B	1.7	6.5
17-19	Yolu	1.4	5.5
15-18	Wenyiil	1.4	2.5
19-19	SRPOL 	0.9	8.1
20-39	19 systems not human-evaluated		...

English→Estonian			
Rank	System	Human	AutoRank
1-1	Human	83.1	
2-2	Gemini-2.5-Pro	78.8	2.5
3-4	Wenyiil	72.6	2.6
3-4	GPT-4.1	72.2	3.0
5-6	Algharb	70.4	3.9
5-6	Shy-hunyan-MT	70.3	1.0
7-8	ONLINE-B	60.2	6.0
7-8	Yolu	59.5	3.8
9-9	TranssionTranslate?	57.1	7.3
10-11	Claude-4?	53.0	6.5
10-12	GemTrans	51.7	5.1
11-14	CommandA-WMT 	50.1	6.1
12-15	SRPOL	49.4	5.7
12-17	Laniko	48.6	5.2
13-17	EuroLLM-22B-pre.[M]	47.2	8.1
14-18	SalamandraTA	46.7	6.3
14-18	UvA-MT	46.4	5.9
16-18	Gemma-3-27B	45.9	7.6
19-19	IRB-MT	32.4	11.4
20-40	20 systems not human-evaluated		...

English→Bhojpuri			
Rank	System	Human	AutoRank
1-1	Gemini-2.5-Pro	94.9	1.0
2-3	Human	92.6	
2-3	Algharb	91.1	2.8
4-4	Wenyiil	90.9	2.5
5-6	Claude-4?	83.2	4.5
5-6	GPT-4.1?	82.8	5.5
7-8	TranssionTranslate?	79.5	4.3
7-10	DeepSeek-V3?	77.3	5.1
8-10	Llama-4-Maverick	76.4	6.5
8-10	NLLB	75.6	6.6
11-12	CommandA 	72.6	6.5
11-12	Yolu	72.4	5.7
13-14	TranssionMT	70.1	6.2
13-15	COILD-BHO	68.7	8.9
14-15	ONLINE-B	67.2	4.1
16-16	IRB-MT	59.6	11.4
17-17	Gemma-3-27B?	56.0	8.3
18-18	SalamandraTA	35.7	12.1
19-19	Shy-hunyan-MT	1.7	11.5
20-37	17 systems not human-evaluated		...

English→Masai			
Rank	System	Human	AutoRank
1-1	Gemini-2.5-Pro?	9.8	6.1
2-2	Human	9.6	
3-3	Claude-4?	7.7	2.6
4-6	AyaExpanse-8B	6.0	8.2
4-5	Llama-4-Maverick	5.2	3.2
6-6	Shy-hunyuan-MT	4.8	1.0
7-13	AyaExpanse-32B	3.1	7.1
4-8	DeepSeek-V3?	3.0	6.2
9-13	Llama-3.1-8B	3.0	8.1
9-13	Gemma-3-12B?	3.0	8.8
9-13	Qwen2.5-7B?	2.8	8.6
9-13	Qwen3-235B	2.7	3.0
9-13	TranssionMT	2.5	5.9
14-18	CommandR7B	1.6	4.3
14-18	CommandA-WMT	1.5	6.4
14-16	CommandA	1.3	7.9
17-18	TowerPlus-9B[M]	0.8	5.3
17-18	EuroLLM-9B[M]	0.7	8.2
19-19	EuroLLM-22B-pre.[M]	0.5	8.2
20-29	9 systems not human-evaluated		...

English→Russian			
Rank	System	Human	AutoRank
1-1	Gemini-2.5-Pro	83.4	4.4
2-2	Shy-hunyuan-MT	80.2	1.0
3-5	Wenyiil	78.2	4.8
3-5	GPT-4.1	76.2	5.4
3-5	Claude-4	75.9	8.7
6-9	DeepSeek-V3?	73.6	5.7
5-8	Algharb	73.3	5.2
6-9	CommandA-WMT	73.2	4.2
8-10	Yandex	72.0	4.5
9-11	Human	70.5	
10-12	UvA-MT	69.1	4.5
11-14	Qwen3-235B	67.6	8.8
12-15	IRB-MT	65.4	10.1
12-15	Yolu	64.5	6.9
13-16	GemTrans	62.5	5.1
15-16	Gemma-3-27B	61.7	8.9
17-19	RuZh?	57.9	9.6
17-19	SRPOL	56.9	10.6
17-19	Laniquo	56.2	8.8
20-42	22 systems not human-evaluated		...

English→Ukrainian			
Rank	System	Human	AutoRank
1-3	Gemini-2.5-Pro	90.3	3.3
1-3	Algharb	90.0	4.2
1-3	Wenyiil	89.5	3.5
4-5	Shy-hunyuan-MT	88.4	1.0
4-5	GemTrans	88.2	4.6
6-7	GPT-4.1	87.9	3.5
5-8	Human	87.3	
7-9	UvA-MT	86.4	4.4
8-13	CommandA-WMT	86.3	3.9
9-13	Llama-4-Maverick	86.2	8.8
9-13	DeepSeek-V3?	85.8	5.0
9-14	Claude-4?	85.6	7.0
9-13	Yolu	85.4	6.0
14-16	Mistral-Medium?	84.5	6.0
14-16	TowerPlus-9B[M]	84.2	8.8
14-16	CommandA	84.0	7.4
17-17	IRB-MT	82.9	8.2
18-19	SRPOL	79.9	8.4
18-19	Laniquo	79.8	7.7
20-44	24 systems not human-evaluated		...

English→Italian			
Rank	System	Human	AutoRank
1-4	Gemini-2.5-Pro	79.4	4.4
1-4	GemTrans	79.4	5.2
1-4	GPT-4.1	79.0	4.5
1-4	Shy-hunyuan-MT	78.7	1.0
5-7	CommandA-WMT	75.5	2.6
5-8	Mistral-Medium?	73.8	7.1
5-10	CommandA	73.2	8.4
6-10	Claude-4	72.1	8.4
7-10	UvA-MT	71.8	5.3
7-10	DeepSeek-V3?	71.7	6.1
11-11	Qwen3-235B	67.0	7.2
12-13	TowerPlus-9B[M]	61.2	11.3
12-13	IRB-MT	60.3	10.2
14-16	SalamandraTA	57.5	10.3
14-16	AyaExpanse-8B	57.0	14.9
14-16	EuroLLM-9B[M]	56.6	15.2
17-18	Gemma-3-12B	53.6	15.5
17-18	Laniquo	53.4	7.6
19-34	15 systems not human-evaluated		...

English→Icelandic			
Rank	System	Human	AutoRank
1-1	Human	87.5	
2-2	Gemini-2.5-Pro	77.6	1.8
3-4	Erlendur	68.3	2.2
3-4	GPT-4.1	68.0	1.9
5-5	Shy-hunyuan-MT	63.2	1.0
6-6	TowerPlus-9B[M]	57.4	3.9
7-7	ONLINE-B	51.8	4.4
8-10	Claude-4?	47.8	5.2
8-10	TowerPlus-72B[M]	46.3	5.7
8-10	TranssionTranslate?	46.2	5.8
11-11	AMI	39.9	7.4
12-12	GemTrans	34.8	7.0
13-14	SalamandraTA	31.3	8.6
13-15	UvA-MT	30.6	6.8
14-15	CommandA-WMT	29.0	6.8
16-16	NLLB	24.1	15.2
17-17	IRB-MT	20.7	11.9
18-18	Gemma-3-12B	16.5	13.8
19-19	Llama-3.1-8B	10.5	24.9
20-35	15 systems not human-evaluated		...

English→Serbian (Cyrilic)			
Rank	System	Human	AutoRank
1-1	Gemini-2.5-Pro	94.2	3.0
2-3	GPT-4.1	92.5	3.4
2-4	Shy-hunyuan-MT	92.2	1.0
3-4	ONLINE-B	90.6	6.1
5-5	Claude-4?	90.0	6.8
6-6	Human	88.7	
7-7	TranssionTranslate?	85.1	8.0
8-9	GemTrans	81.5	4.6
8-9	DeepSeek-V3?	78.7	8.6
10-11	IRB-MT	77.6	9.9
10-15	DLUT_GTCOM	77.2	9.3
11-14	CommandA-WMT	76.5	7.0
10-15	UvA-MT	76.2	5.8
11-15	SalamandraTA	75.5	8.8
13-15	Gemma-3-12B	74.8	12.1
16-17	CUNI-SFT	60.9	13.5
16-17	Llama-3.1-8B	58.4	19.4
18-18	NLLB	53.5	19.8
19-19	EuroLLM-9B[M]	41.8	22.3
20-34	14 systems not human-evaluated		...

Rank	System	Czech→German	Human	AutoRank
1-1	Gemini-2.5-Pro		90.7	2.5
2-4	GPT-4.1		89.5	2.4
2-4	Claude-4		88.8	4.8
2-6	DeepSeek-V3?		88.1	3.5
4-7	Shy-hunyuan-MT		87.2	1.0
4-8	Mistral-Medium		87.0	4.2
5-7	CommandA		86.8	4.8
8-8	CommandA-WMT		85.6	2.1
9-12	Human		82.8	
9-13	GemTrans		82.6	6.3
9-13	Gemma-3-27B		82.0	7.2
9-13	Wenyiil		82.0	10.9
10-15	Algharb		80.9	13.2
13-15	TowerPlus-9B[M]		79.8	10.3
13-15	UvA-MT		79.5	7.0
16-19	CUNI-MH-v2		77.2	14.2
16-18	Gemma-3-12B		76.8	11.5
16-18	SRPOL		76.7	11.0
19-19	Yolu		75.3	9.3
20-21	IRB-MT		71.7	12.4
20-21	Laniquo		70.0	10.3
22-42	20 systems not human-evaluated			...

Rank	System	English→Czech	Human	AutoRank
1-1	Gemini-2.5-Pro		88.7	3.4
2-2	Shy-hunyuan-MT		87.1	1.0
3-4	DeepSeek-V3?		85.1	5.1
3-4	Human		84.5	
5-6	CommandA-WMT		82.6	3.6
5-6	Wenyiil		82.4	4.5
7-9	GPT-4.1		80.8	4.0
7-9	Mistral-Medium?		80.4	7.1
7-10	Claude-4?		79.6	9.0
9-11	UvA-MT		78.6	6.5
10-14	Algharb		76.7	6.4
11-14	CommandA		76.4	8.8
11-15	Yolu		75.6	6.3
11-15	Gemma-3-27B		75.6	9.2
13-15	GemTrans		73.2	5.1
16-16	CUNI-MH-v2		71.0	12.1
17-18	SRPOL		67.5	8.7
17-19	Laniquo		66.1	8.8
18-19	TowerPlus-9B[M]		65.8	11.6
20-20	SalamandraTA		60.3	10.5
21-44	23 systems not human-evaluated			...

Rank	System	English→Chinese	Human	AutoRank
1-1	Algharb		88.4	4.2
2-4	Shy-hunyuan-MT		88.2	1.0
2-5	Claude-4		86.9	7.2
2-5	Wenyiil		86.3	4.0
3-6	DeepSeek-V3		85.0	7.3
5-10	GemTrans		84.4	5.0
6-11	Qwen3-235B		84.0	4.9
5-10	GPT-4.1		84.0	4.7
6-11	Gemini-2.5-Pro		83.8	4.0
5-10	UvA-MT		83.4	6.4
11-13	Human		82.1	
11-15	CommandA-WMT		81.3	5.7
11-15	Llama-4-Maverick		80.7	8.1
12-16	Mistral-Medium?		79.9	5.0
12-16	Yolu		79.0	4.9
14-17	SRPOL		77.7	10.5
16-18	IRB-MT		76.5	9.5
17-18	RuZh?		75.7	10.6
19-19	Laniquo		70.5	9.3
20-40	20 systems not human-evaluated			...

Rank	System	English→Japanese	Human	AutoRank
1-1	Human		89.2	
2-4	Gemini-2.5-Pro		85.8	2.5
2-6	Algharb		85.7	3.3
2-5	Mistral-Medium?		84.8	5.5
3-6	Wenyiil		84.4	3.0
5-6	GPT-4.1		83.7	2.9
7-7	CommandA-WMT		82.2	3.7
8-12	Shy-hunyuan-MT		79.9	1.0
8-13	DeepSeek-V3?		79.3	4.7
8-13	Claude-4		79.3	5.8
8-13	UvA-MT		79.3	6.5
8-14	ONLINE-B		78.0	6.3
9-16	In2x?		77.8	2.3
12-16	GemTrans		76.2	5.6
13-16	KIKIS		76.2	3.2
13-16	Systran		75.6	7.5
17-18	NTTSU		73.3	8.1
17-18	Yolu		72.6	6.1
19-19	Laniquo		67.8	9.5
20-45	25 systems not human-evaluated			...

Rank	System	Czech→Ukrainian	Human	AutoRank
1-2	Gemini-2.5-Pro		92.9	1.1
1-3	GPT-4.1		92.1	1.3
2-3	Shy-hunyuan-MT		91.8	1.0
4-8	GemTrans		90.2	4.4
4-6	Human		90.1	
4-10	Mistral-Medium?		89.4	4.2
6-10	Claude-4?		89.1	3.7
4-10	DeepSeek-V3?		89.0	3.2
6-10	CommandA-WMT		88.7	1.3
6-10	Gemma-3-27B		88.6	5.0
11-12	CommandA		86.4	4.6
11-13	Wenyiil		85.7	5.4
12-15	TowerPlus-9B[M]		85.3	7.9
13-16	Algharb		84.1	7.2
13-17	UvA-MT		83.5	5.1
14-17	Laniquo		83.4	7.7
15-17	IRB-MT		82.7	9.1
18-19	SRPOL		80.8	7.8
18-19	Yolu		80.1	6.0
20-44	24 systems not human-evaluated			...

Rank	System	Japanese→Chinese	Human	AutoRank
1-1	Human		-3.5	
2-2	Gemini-2.5-Pro		-4.4	3.3
3-6	Algharb		-5.8	4.3
3-7	Claude-4		-5.9	6.4
3-7	Shy-hunyuan-MT		-6.1	1.0
3-7	GPT-4.1		-6.2	4.5
4-7	Wenyiil		-6.9	4.5
8-10	CommandA-WMT		-7.7	5.2
8-10	DeepSeek-V3		-8.1	6.5
8-13	Kaze-MT		-8.6	3.9
10-13	Mistral-Medium		-10.0	6.6
10-13	In2x		-10.0	3.0
10-13	Qwen3-235B		-10.9	7.6
14-15	GemTrans		-10.9	6.6
14-15	NTTSU		-11.3	5.9
16-17	Yolu		-12.6	7.1
16-17	TowerPlus-9B[M]		-13.3	11.5
18-18	IRB-MT		-13.9	12.4
19-19	Laniquo		-18.3	11.3
20-42	22 systems not human-evaluated			...

English→Korean			
Rank	System	Human	AutoRank
1-3	Human	-1.9	
1-3	Shy-hunyuan-MT	-2.5	1.0
1-3	Gemini-2.5-Pro	-2.7	2.5
4-6	GPT-4.1	-3.3	2.9
4-7	Claude-4	-3.4	4.4
4-7	DeepSeek-V3 	-3.8	5.1
5-10	GemTrans	-4.1	5.0
7-12	CommandA-WMT	-4.3	2.9
5-12	Wenyiil	-4.3	3.0
5-12	Algharb	-4.4	3.1
8-15	Mistral-Medium 	-4.7	6.1
7-15	CommandA	-4.7	6.0
11-16	UvA-MT	-5.2	4.3
11-16	Qwen3-235B	-5.5	6.5
11-16	IRB-MT	-5.6	8.6
13-16	Gemma-3-12B	-5.9	9.2
17-18	TowerPlus-9B[M]	-7.2	10.1
17-18	Yolu	-7.3	7.0
19-19	Laniko	-9.1	9.2
20-37	17 systems not human-evaluated		...

While best system scores high many stay in the middle. Although the best system almost always achieves a score of 90 or higher, no system (including human references) achieves a perfect score of 100 (see Table 10 and Appendix D). The spread of scores is also quite large: mid-tier models often score 60 or below, and the worst-performing systems can score as low as 0, depending on the language pair and domain (see Figures 6 and 7).

Translation into low-resource languages remains a challenge. The translation quality for Maasai, a low-resource language, is largely unusable. We observe that for this language, systems often produce large portions of their output in Swahili, and the few acceptable spans in Maasai are frequently overlooked by annotators.³⁶ Additional errors arise from translations that fail to capture the meaning and context of the source text. While human translations into Maasai also score low (9.6 ESA scores), we were able to independently confirm that these are generally understandable by native speakers and often sound natural. However, they may exhibit extensive code-mixing of English and Swahili (Figures 6 and 7). We speculate that our Maasai annotators, when evaluating a large portion of poor-quality system outputs, failed to notice the one translation that was mostly reasonable.

A similar issue occurs with Egyptian Arabic, where systems tend to output Modern Standard Arabic (MSA). This type of error would likely be overlooked by quality estimation metrics, which

³⁶While many Maasai people speak Swahili, Maasai (or Maa) is a distinct language from a different language family.

typically do not incorporate target language labels into their pipeline (though they could potentially be trained to do so).

7.5 Additional Analysis of English→Serbian translations

In addition to the official human evaluation, an analysis of English→Serbian translations was carried out by an MT researcher with experience in human translation. One part of the analysis deals with the two scripts (Latin and Cyrillic), and another one with translation quality taking into account both errors as well as exceptionally good idiomatically translated parts named “rewards”.

Scripts. Due to historical and cultural reasons, the Serbian language is bi-alphabetical, using both Latin and Cyrillic scripts. Serbian Cyrillic alphabet is a highly phonetic alphabet with a one-to-one correspondence between letters and sounds. Serbian Latin alphabet is almost perfectly compatible with Cyrillic with a one-to-one correspondence, except for the three digraphs each representing one sound (Љ ↔ /lj/, Њ ↔ /nj/, Ћ ↔ /dž/). Serbian speakers switch easily between the two scripts without much thinking. The choice of the script is partly random but also influenced by the context, medium, or even ideological preferences. However, the scripts should not be mixed within a single text with a few exceptions: URLs or foreign brand names which should be in Latin even in Cyrillic texts. The automatic conversion of Serbian Cyrillic into Latin script is easier than the other way round. The primary reason is the one-to-one character mapping from Cyrillic to Latin, while converting from Latin to Cyrillic introduces ambiguity due to the three previously mentioned digraphs.

Training data for generative NLP including MT are available in both scripts, and it might be challenging to ensure that the outputs are written in the desired script. Therefore, the WMT translations were checked in this aspect, by measuring the percentage of words written in another script.

The results are presented in Table 11. In Cyrillic translations there is always a small percent of words written in Latin, because of URLs, named entities or similar. Overall, there is more mixing in Cyrillic translations: found in more systems and also to the larger extent. The probable reason is that there is more available data in Latin script. Another observation is that some systems use only one script (e.g. ONLINE-B and ONLINE-G only Cyrillic,

% Cyrillic in Latin		% Latin in Cyrillic	
ONLINE-G	98.9	CUNI-SFT	100.0
ONLINE-B	98.0	Llama-3.1-8B	99.7
		UvA-MT	95.7
Mistral-7B	12.9		
Gemma-3-12B	11.2	AyaExpanse-8B	85.8
TowerPlus-72B	4.9	CommandR7B	79.2
Gemma-3-27B	3.4	Qwen2.5-7B	76.9
CommandR7B	3.2	TowerPlus-9B	69.6
IRB-MT	2.8	EuroLLM-9B	66.3
Qwen3-235B	1.9	EuroLLM-22B	66.0
Qwen2.5-7B	1.7	IRB-MT	25.8
TowerPlus-9B	0.4	Gemma-3-12B	22.4
AyaExpanse-8B	0.3		
Yolu	0.0	Mistral-7B	8.3
Wenyiil	0.0	TowerPlus-72B	3.6
UvA-MT	0.0	Shy	2.9
TranssionTranslate	0.0	IR-MultiagentMT	2.7
TranssionMT	0.0	GPT-4.1	2.4
Shy	0.0	Gemma-3-27B	2.2
SalamandraTA	0.0	ONLINE-B	2.1
NLLB	/	Llama-4-Maverick	2.1
Llama-4-Maverick	0.0	DeepSeek-V3	2.1
Llama-3.1-8B	0.0	AyaExpanse-32B	2.0
IR-MultiagentMT	0.0	Gemini-2.5-Pro	1.9
GPT-4.1	0.0	DLUT_GTCOM	1.9
GemTrans	0.0	TranssionTranslate	1.7
Gemini-2.5-Pro	0.0	Claude-4	1.7
EuroLLM-9B	0.0	CommandA	1.6
EuroLLM-22B	0.0	Qwen3-235B	1.4
DLUT_GTCOM	/	ONLINE-G	1.4
DeepSeek-V3	0.0	NLLB	1.4
CUNI-SFT	0.0	SalamandraTA	1.3
CommandA	0.0	GemTrans	1.1
Claude-4	0.0	Yolu	/
AyaExpanse-32B	0.0	Wenyiil	/
Algharb	0.0	TranssionMT	/
		Algharb	/

(a) % Unexpected Cyrillic script detected in a Latin-script translation.

(b) % Unexpected Latin script detected in a Cyrillic-script translation

Table 11: Comparison of script intrusions across Latin and Cyrillic translations.

CUNI-SFT and Llama-3.1-8B only in Latin, UvA-MT almost all in Latin).

We further explored qualitative differences between the outputs in different scripts. The first step was automatic: Cyrillic outputs were converted into Latin and then compared by word bi-gram overlap (F-score). For some systems, there was almost no difference, but for the majority there were notable differences (for example, around 75% overlap score for Gemini and GPT, around 60% for Shy, see the full table can be seen in Appendix Table 20). However, qualitative manual inspection did not identify any major or systematic differences regarding translation quality or types of errors.

Errors and rewards. We next looked into the annotated error spans similarly to Popovic (2021).

However, due to discrepancies in error span annotations it was difficult to determine the nature of the annotated spans. For example, a number of non-existing or obviously incorrect words were not marked at all, and overall scores were not lowered at all or only slightly, while there were completely correct passages which are marked as errors. Furthermore, a number of segments seemed to be heavily penalized only for using the Latin script: all words there were marked as errors without looking at actual errors, and the scores were lowered but inconsistently, ranging from 90 to 10. The affected systems were CUNI-SFT, UvA-MT, Llama-3.1-8B and EuroLLM-9B. After qualitative inspection, it seemed that the translation quality of Llama-3.1-8B and EuroLLM-9B was indeed low, while CUNI-SFT and UvA-MT might be underestimated.

For these reasons, a full additional ESA annotation has been carried out on a small selected set: last ten documents from each of the domains, so 40 documents in total. The following nine translations were included: the best ranked systems in the official evaluation (Gemini, GPT, Shy, ONLINE-B, Claude) and the human translation, the two low-ranked systems which were potentially over-penalised for the script (UvA and CUNI), as well as one system from the middle cluster (GemTrans).

For each translated documents, first all errors were marked, and then overall scores were assigned (the same process as in the official evaluation, although no distinction between major and minor errors was made). During the evaluation, the annotator observed that, apart from error spans, there were passages translated exceptionally well, namely idiomatically: diverging from the source, but fully keeping the meaning while sounding completely natural in the target language. These spans were then marked as “rewards”.

The results from the additional span annotation are aggregated as word-level error rates and reward rates, namely number of words in the marked spans divided by total number of words (length). Also, a total score is presented, where error spans are subtracted from the length and reward spans are added. In addition, the official scores and error rates are extracted for the 40 selected documents and presented for comparison.

The results can be seen in Table 12. The largest percentage of idiomatic translations can be found in the middle-ranked GemTrans translation (5.67%), followed by human translation (4.17%). Gemini

	official e.		additional evaluation			
	score	%err.	score	%err.	%rew.	total
Gemini	89.3	5.3	87.4	5.82	3.61	97.8
GPT	93.8	2.8	82.8	8.29	0.95	92.6
Shy	86.4	5.7	85.8	7.02	3.52	96.5
ONLINE-B	84.0	7.7	70.7	13.3	0.42	87.1
Claude	85.9	9.1	63.2	15.2	0.29	85.1
Human	86.8	19.2	86.9	7.84	4.19	96.4
GemTrans	84.2	6.6	81.1	9.60	5.67	96.1
UvA	81.7	76.6	64.8	14.7	1.18	86.4
CUNI	54.2	81.2	41.4	27.2	0.18	73.0

Table 12: Results of the additional analysis on the selected set of translations for English→Serbian together with the scores from the official evaluation.

official evaluation		additional evaluation			
score	%err.	score	%err.	%rew.	total
GPT	GPT	Gemini	Gemini	GemTrans	Gemini
Gemini	Gemini	Human	Shy	Human	Shy
Human	Shy	Shy	Human	Gemini	Human
Shy	GemTrans	GPT	GPT	Shy	GemTrans
Claude	ONLINE-B	GemTrans	GemTrans	UvA	GPT
GemTrans	Claude	ONLINE-B	ONLINE-B	GPT	ONLINE-B
ONLINE-B	Human	UvA	UvA	ONLINE-B	UvA
UvA	UvA	Claude	Claude	Claude	Claude
CUNI	CUNI	CUNI	CUNI	CUNI	CUNI

Table 13: Rankings of the selected translations for English→Serbian according to different scores.

and Shy have around 3.5% idiomatically translated words, while all other systems have around 1% or less.

Table 13 presents rankings of the systems according to different scores: official overall score and error rate, additional overall score and error rate, as well as reward rate and total span-based score. It can be noted that Gemini, Shy, GPT and Human are almost always on the top, while CUNI is always the last. Furthermore, Gemini clearly surpasses human translation in terms of both scores and error rates in both evaluations.

Claude and ONLINE-B might be over-estimated in the official evaluation, since in the additional one they obtained notably lower scores, higher error rates, and almost no rewards, being comparable to possibly under-estimated UvA-MT. Also, human translation might be under-estimated in the official evaluation, but it is clearly worse than Gemini and comparable with Shy and GPT.

Furthermore, the middle-ranked GemTrans system turned out to be very interesting, since it generated a notable amount of idiomatic translations (5.67%), even more than humans (4.19%), but also exhibits relatively high number of errors (9.6%), many of them being morphological/agreement issues which were typical for statistical systems.

According to idiomatic translations, GemTrans and human are ranked the best, followed by the two overall top-ranked Gemini and Shy. And according to the total rate, taking into account both errors and rewards, the best three translations are Gemini, Shy and Human, followed by GemTrans and GPT.

It might be worth noting that there were 15 translations (originating from 10 source segments) without any errors: 7 were translated by Gemini, 5 by

Shy, and 3 by human translators. Three of them are from the literary domain, one from speech and 11 from the social domain. Of those, 11 also have idiomatic translations: 7 by Gemini, 2 by Shy, and 2 by human translators.

As for different domains, there are no notable differences regarding idiomatic translations, although there are slightly more in news and literature (3.2% and 2.2%) than in speech and social (2.0% and 1.5%).

8 Test Suites Sub-task: “Help us break LLMs vol. 2”

For a second year in a row, we have invited the community to submit test suites in the sub-task under the call “help us break LLM”. The aim is again to demonstrate evaluation methods that can expose weaknesses in LLMs which cannot be detected using standard evaluation methods. With more LLMs participating this year, and the technology advancing quickly, this call remains particularly relevant.

8.1 Setup of the Sub-task

Each test suite is a customised extension of the standard test sets that focuses on specific aspects of the machine translation (MT) output. Evaluation of the MT output takes place in a decentralised manner. Test suite providers were invited to submit their customised test sets following the setup introduced at the Third Conference on Machine Translation (Bojar et al., 2018). For this purpose, each test suite provider submitted a source-side test set, which the organisers of the General MT Shared Task then appended to their standard test sets. After generating the corresponding system outputs, the organizers returned them to the respec-

tive providers, who then conducted the evaluation according to their own methodological approach. Detailed results and analyses for each test suite are presented in separate description papers, while a consolidated summary is provided below.

8.2 Submissions

Six test suites are participating this year, covering a wide range of translation phenomena, domains, and language pairs. An overview of the test suites can be seen in Table 14. Descriptions of each submission, along with their main findings, can be found below.

DFKI (Manakhimova et al., 2025). This test suite offers a fine-grained linguistically motivated analysis of the shared task MT outputs for English–Russian, based on 465 manually devised test items, which cover 55 phenomena in 13 categories. Extending their previous test suite submissions (e.g. Avramidis et al., 2020; Macketanz et al., 2021, 2022; Manakhimova et al., 2023, 2024), the submission of this year analyzes how English–Russian machine translation (MT) systems submitted to WMT25 perform on linguistically challenging translation tasks, similar to problems used in university translator training.

The findings show that in 2025, even top-performing MT systems still struggle with translation problems that require deep understanding and rephrasing, much like human novices do. The best systems exhibit marked improvements in handling such ‘extra-credit’ challenges, often producing more natural translations rather than producing word-for-word renditions. However, persistent structural and lexical problems remain: literal word order carry-overs, misused verb forms, and rigid phrase translations were common, mirroring errors typically seen in beginner translator assignments.

EAA-Terminology (Hauksdottir and Steingrims-son, 2025). The EEA terminology test suite is a novel evaluation set designed to assess the capabilities of machine translation (MT) systems in handling terminology found in the EEA Agreement. It is designed for English-to-Icelandic translations, but can be easily adapted for other languages. The test suite evaluates four subdomains of the terminology in EEA regulations: science, technology, finance, and society. The test suite consists of 300 text examples in the form of sentences in English, stored in a single text file, which is to be translated by the MT systems. The suite also contains a gold

standard translation meant for comparison, where each example has been translated as expected into Icelandic.

GENDER1PERSON (Popović and Lapshinova-Koltunski, 2025). The GENDER1PERSON test suite is designed for measuring gender bias in translating first-person singular forms from English into two Slavic languages, Russian and Serbian. The test suite consists of 1 000 Amazon product reviews, uniformly distributed over 10 different product categories. The bias is measured through a gender score ranging from -100 (all reviews are feminine) to 100 (all reviews are masculine).

The test suite shows that the majority of the systems participating in the WMT-2025 task for these two target languages prefer the masculine writer’s gender. There is no single system which is biased towards the feminine variant. Furthermore, for each language pair, there are seven systems which are considered balanced, having the gender scores between -10 and 10.

Finally, the analysis of different products showed that the choice of the writer’s gender depends to a large extent on the product. Moreover, it is demonstrated that even the systems with overall balanced scores are actually biased, but in different ways for different product categories.

IITP-legal (Singh et al., 2025a). The study critically examines various Machine Translation systems, particularly focusing on Large Language Models, using the WMT25 Legal Domain Test Suite for translating English into Hindi. It utilizes a dataset of 5, sentences designed to capture the complexity of legal texts, based on word frequency ranges from 5 to 54. Each frequency range contains 100 sentences, collectively forming a corpus that spans from simple legal terms to intricate legal provisions. Six metrics were used to evaluate the performance of the system: BLEU, METEOR, TER, CHRF++, BERTScore and COMET. The findings reveal diverse capabilities and limitations of LLM architectures in handling complex legal texts. Notably, Gemini-2.5-Pro, Claude-4 and Llama-4-Maverick topped the performance charts, showcasing the potential of LLMs for nuanced translation. Despite these advances, the study identified areas for further research, especially in improving robustness, reliability, and explainability for use in critical legal contexts. The study also supports the WMT25 subtask focused on evaluat-

Test suite	Focus	Language pair	Segments
DFKI (Manakhimova et al., 2025)	linguistic phenomena	en→ru	5,553
EEA Terminology (Hauksdottir and Steingrímsson, 2025)	legal domain	en→is	256
GENDER1PERSON (Popović and Lapshinova-Koltunski, 2025)	gender choice and agreement	en→{ru, sr}	2,000
IITP-legal (Singh et al., 2025a)	legal domain	en→hi	5,000
SportsEval (Sigurdsson et al., 2025)	sports domain	en→is	300
RoCS-MT v2 (Bawden and Sagot, 2025)	non-standard user-generated texts	en→{ar, bho, cs, et, is, ja, ko, mas, ru, sr, uk, zh}	59,340

Table 14: Overview of the participating test suites.

ing weaknesses of large language models (LLMs). The dataset and related resources are publicly available.³⁷

RoCS-MT v2 (Bawden and Sagot, 2025). Robust Challenge Set for Machine Translation,³⁸ is designed to test MT systems’ ability to translate user-generated content with non-standard characteristics, such as spelling errors, devowelling, acronymisation, etc. The original English Reddit texts are associated with manual normalisations and translations in five languages (French, German, Czech, Ukrainian and Russian). This second version of the test suite presents several improvements over the previously published version (Bawden and Sagot, 2023), 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.

In previous editions of the shared task, we saw that non-standard UGC phenomena still posed problems for many models, although some of the larger, closer-sourced models handle them better. The behaviour of the systems varies greatly, with different handling of the translation of the phenomena, some systems producing more standardised outputs than others. In this edition, we saw that there was still a wide range in behaviour of systems, not all of which is accurately characterised by the automatic metric used in evaluation (COMET-Kiwi). We show some preliminary analysis showing that for elongation as a mark of expressivity (e.g. *mooorreeee* instead of *more*), some systems

are rewarded for copying the source text rather than translating the word, either in its standard or expressive form. This especially reveals issues in the current evaluation protocol for the test suite, and will require clarifying in future work.

SportsEval (Sigurdsson et al., 2025). Sports are a consistently popular domain in most news media. Although many sports attract extensive media attention and feature a rich, polysemous language, often shaped by active neologism and community-driven translations, the sports domain has received relatively little focus in MT research.

The SportsEval test suite was developed to examine the robustness of MT systems in translating sports-related texts from English into Icelandic. It covers five sports that are popular in Iceland: football (100 segments), basketball (100), chess (50), gymnastics (25), and golf (25), with a total of 300 segments. Each of these sports has a long-established domain-specific vocabulary in Icelandic, and mistranslations can easily render a text unintelligible. The segments range from single to multi-sentence passages, and most include multiple terms. While the majority are drawn from authentic usage examples, some have been adapted for brevity, and a small number are synthetic, created to illustrate specific terminological challenges.

In total, the test suite contains 971 term instances. Since some terms recur across segments, the design also enables an evaluation of translation consistency. The findings of our study indicate that current MT systems face considerable challenges in this domain.

8.3 Aggregated Results

In order to have a more general overview of the comparative system performance with regard to the test suites, we present the ranks produced by

³⁷ github.com/wmt25testsuite/wmt25

³⁸ github.com/rbawden/RoCS-MT,
huggingface.co/datasets/rbawden/RoCS-MT-v2.

	WMT25 human	EEA term	sports eval	RoCS MT-v2
Gemini-2.5-Pro	1	3	1	2
Erlendur	2	2	3	4
GPT-4.1	2	10	4	3
Shy-hunyuan-MT	4	6	2	1
TowerPlus-9B	5	8	7	6
ONLINE-B	6	5	6	5
TranssionTranslate	7	4	5	9
Claude-4	7	6	9	12
TowerPlus-72B	7	12	25	8
hybrid	–	8	8	–
AMI	10	16	11	11
GemTrans	11	21	16	7
SalamandraTA	12	11	17	13
UvA-MT	12	22	26	21
CommandA	13	26	23	30
ONLINE-G	–	1	9	31
NLLB	15	14	14	24
Gemma-3-27B	–	19	12	15
IRB-MT	16	24	19	23
DeepSeek-V3	–	15	20	14
Llama-4-Maverick	–	18	13	20
Gemma-3-12B	17	22	18	25
IR-MultiagentMT	–	12	31	10
CommandA-WMT	–	16	21	16
Llama-3.1-8B	18	27	22	32
Mistral-Medium	–	20	15	22
Qwen3-235B	–	25	24	29
AyaExpans-32B	–	28	27	33
EuroLLM-22B	–	29	30	34
CommandR-7B	–	29	31	36
Qwen2.5-7B	–	32	27	38
Mistral-7B	–	32	29	37
AyaExpans-8B	–	29	33	39
EuroLLM-9B	–	32	34	35

Table 15: Aggregated system ranking for English→Icelandic according to human evaluation and test suites.

test suites of the same language direction side by side, including the human ranks of the official WMT25 General MT test set (first column). In Tables 15, 16, 17 we present results for the language directions where we have more than one test suite. For visualisation purposes, the table rows are ordered primarily by the human ranks of the WMT25 General MT test set and then by the average of the rest of the test suites. It must be noted that this visualisation has to be taken with a grain of salt, as test suites employ different evaluation methods over different test sets of different sizes. Also, due to the different methods, the confidence intervals between the ranks have not been always taken into consideration.

One can see that there is quite some variety between the ranks of the WMT25 General MT test set and the test suites, with most obvious the ones of RoCS-MT-v2, indicating the non-standard nature of the data means that some systems which

	WMT25 human	dfki	gender 1person	RoCS MT-v2
Gemini-2.5-Pro	1	1	4	13
Shy-hunyuan-MT	2	5	7	1
GPT-4.1	3	5	16	8
Wenyiil	3	1	6	30
Claude-4	3	5	14	23
Algharb	5	1	1	18
DeepSeek-V3	6	5	17	10
CommandA-WMT	6	9	34	6
Yandex	8	1	3	4
UvA-MT	9	10	20	7
Qwen3-235B	10	–	22	17
Yolu	11	–	2	5
IRB-MT	11	–	12	12
GemTrans	12	–	9	2
hybrid	–	–	13	–
TowerPlus-9B	–	–	11	16
Gemma-3-27B	14	–	24	11
Gemma-3-12B	–	–	15	14
Laniko	16	–	8	3
SRPOL	16	–	32	24
RuZh	16	–	–	–
DLUT_GTCOM	–	–	19	15
SalamandraTA	–	–	10	25
ONLINE-G	–	–	5	33
IR-MultiagentMT	–	–	30	9
AyaExpans-32B	–	–	21	20
TowerPlus-72B	–	–	26	19
CommandA	–	–	25	21
ONLINE-W	–	–	18	31
EuroLLM-22B	–	–	23	26
TranssionTranslate	–	–	27	28
Llama-4-Maverick	–	–	33	22
AyaExpans-8B	–	–	29	27
ONLINE-B	–	–	28	29
Qwen2.5-7B	–	–	31	34
EuroLLM-9B	–	–	37	32
Llama-3.1-8B	–	–	36	35
CommandR	–	–	35	40
NLLB	–	–	38	39
Mistral-7B	–	–	39	41
TranssionMT	–	–	40	42

Table 16: Aggregated system ranking for English→Russian according to human evaluation and test suites.

otherwise perform good, are unusually poor. The gender1person testsuite also indicates that systems with high overall performance indicate a strong male bias towards the selection of the male gender. The linguistically motivated test suite by *dfki* also has quite some variability for the systems that are evaluated. Meanwhile, the two terminology test suites are a bit closer to the WMT25 General MT testset, albeit with a few exceptions.

8.4 Summary

The evaluation of multiple test suites across diverse language pairs and domains reveals persistent challenges for current MT systems. Fine-grained linguistic analysis for English–Russian indicates high

	WMT25 human	gender 1person	RoCS MT-v2
Gemini-2.5-Pro	1	1	10
Shy-hunyuan-MT	2	7	1
GPT-4.1	2	9	6
Yolu	–	4	2
ONLINE-B	3	3	26
Claude-4	5	10	15
Human	6	–	–
TranssionTranslate	7	35	34
GemTrans	8	5	3
Algharb	–	2	14
DeepSeek-V3	8	18	9
UvA-MT	10	17	5
IRB-MT	10	13	11
DLUT_GTCOM	10	–	–
SalamandraTA	11	28	4
CommandA-WMT	11	30	25
Wenyii	–	6	18
hybrid	–	12	–
Gemma-3-27B	–	16	8
Gemma-3-12B	13	14	12
EuroLLM-22B	–	11	17
CUNI-SFT	16	8	23
IR-MultiagentMT	–	25	7
Llama-3.1-8B	16	23	24
AyaExpanse-32B	–	15	21
NLLB	18	–	38
EuroLLM-9B	19	26	22
CommandA	–	24	16
Qwen3-235B	–	21	20
Llama-4-Maverick	–	29	13
TowerPlus-9B	–	19	29
AyaExpanse-8B	–	20	30
TowerPlus-72B	–	31	19
CommandR7B	–	22	33
Qwen2.5-7B	–	27	32
Mistral-7B	–	32	28
TranssionMT	–	34	27
ONLINE-G	–	33	31

Table 17: Aggregated system ranking for English→Serbian according to human evaluation and test suites.

error rates in semantic roles, domain-specific terminology, and proper names, although an increase in gender-inclusive renderings was observed compared to previous years. Domain-specific evaluations for English–Icelandic, including EEA terminology and sports-related texts, demonstrate substantial difficulties in maintaining terminological accuracy and consistency. Gender bias assessment for Russian and Serbian shows a systematic preference for masculine forms, with product-specific variation even among systems classified as balanced. Legal-domain evaluation for English–Hindi confirms the superior performance of advanced LLMs such as Gemini-2.5-Pro and Claude-4, while highlighting the need for improved robustness and explainability in critical applications. Finally, robustness testing on user-generated content un-

derscores ongoing weaknesses in handling non-standard linguistic phenomena, despite incremental progress in larger models.

9 Conclusions

The WMT 2025 General Machine Translation Task covered 30 language pairs, with human evaluation conducted on half of them across four to five domains. We prepared more challenging test set by utilizing novel difficulty sampling.

We evaluated 60 systems in total: 36 participant submissions and 24 systems collected from LLMs and popular online providers. Participation continued to grow compared to last year, and most teams utilize LLMs, often via fine-tuning.

We adopt ESA and MQM as the human evaluation protocols, which show the weak-points of models, especially in Egyptian Arabic dialect or in extremely low-resource language Maasai.

Automatic rankings did not always match human judgments: systems that topped automated metrics such as Shy-hunyuan-MT, did not consistently win under human evaluation, pointing to persistent metric bias in MBR and reinforcing that human evaluation should remain the final arbiter of translation quality.

Domain analyses showed speech as the most challenging (likely due to ASR noise), while literary was the easiest among those tested. Targeted test suites revealed remaining weaknesses in robustness to non-standard input, linguistic complexity, domain terminology, and gender choice/agreement, even as advanced LLMs improved inclusivity and performance in some specialized areas.

All source data, system outputs, and human judgments are released to support transparency, reproducibility, and further research on machine translation.

10 Limitations

We tested the general capabilities of MT systems. However, we have simplified this approach and only used three to five domains. Out of various possible modalities, we used audio and text.

Some models used pretrained metrics such as xComet or MetricX during their training, for example, using Minimum Bayes Risk or as a reward model. This significantly affected the automatic evaluation of such models giving them artificially higher scores. Furthermore, automatic metrics are limited, brittle, and biased (Karpinska et al., 2022;

Moghe et al., 2025), especially in novel domains (Zouhar et al., 2024a,b), which motivates them being superseded by human evaluation. Another potential problem may have been that test sets we use are paragraph-level; automatic metrics have usually been tested in a sentence-level scenario. Therefore, we strongly advise careful interpretation of automatic scores.

Although we use human judgments as the gold standard, giving us more reliable signal than automatic metrics, we should mention that human annotations are noisy (Wei and Jia, 2021) and their performance is affected by the quality of other evaluated systems (Mathur et al., 2020). Lastly, different annotators use different ranking strategies, which may have an effect on the system ranking.

11 Ethical Considerations

Inappropriate, controversial, and explicit content was filtered out prior to translation, keeping in mind the translators and not exposing them to such content or obliging them to translate it.

Human evaluation using Appraise for the collection of human judgements was fully anonymous. Automatically generated accounts associated with annotation tasks with single-sign-on URLs were distributed among pools of annotators and we do not store any personal information. We do store the mapping between which annotator (pseudonymized) annotated which segments. Annotators received standard professional translator’s or evaluator’s wage with respect to their countries.

Acknowledgments

This report would not have been possible without the partnership with Árni Magnússon Institute for Icelandic Studies, Charles University, Cohere, Custom.MT, Dubformer, Gates Foundation, Google, Institute of the Estonian Language, Microsoft, NTT, Toloka AI, University of Tartu, University of Tokyo. Furthermore, we are grateful to Toshiaki Nakazawa, Michael Karani and Youssef Nafea.

Vilém Zouhar gratefully acknowledges the support of the Google PhD Fellowship.

Barry Haddow’s participation was funded by UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee [grant numbers 10052546 and 10039436].

Rachel Bawden’s participation was funded by her chair position in the PRAIRIE institute funded

by the French national agency ANR under the project MaTOS - “ANR-22-CE23-0033-03” and as part of the “Investissements d’avenir” programme under the reference ANR-19-P3IA-0001.

Martin Popel’s participation was funded by TAČR grant EdUKate (TQ01000458).

Ondřej Bojar acknowledges the support by the grant CZ.02.01.01/00/23_020/0008518 (Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím). The reference translations and manual evaluations were also supported National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO.

This work has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

- Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. *Findings of the 2021 conference on machine translation (WMT21)*. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.
- Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. *Machine translation in the era of large language models: a survey of historical and emerging problems*. *Information*, 16(9).
- Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. *Fine-grained linguistic evaluation for state-of-the-art machine translation*. In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356. Association for Computational Linguistics.
- Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarriás, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. *ParaCrawl: Web-scale acquisition of parallel corpora*. In

- Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567. Association for Computational Linguistics.
- Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. [Findings of the 2020 conference on machine translation \(WMT20\)](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55. Association for Computational Linguistics.
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(WMT19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61. 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. Association for Computational Linguistics.
- Rachel Bawden and Benoît Sagot. 2025. RoCS-MT v2 at WMT 2025: Robust Challenge Set for Machine Translation. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Rachel Bawden and François Yvon. 2023. [Investigating the translation performance of a large multilingual language model: the case of BLOOM](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170. European Association for Machine Translation.
- Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. [Findings of the 2013 Workshop on Statistical Machine Translation](#). In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44. Association for Computational Linguistics.
- Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. [Findings of the 2014 workshop on statistical machine translation](#). In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58. Association for Computational Linguistics.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. [Findings of the 2017 conference on machine translation \(WMT17\)](#). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214. Association for Computational Linguistics.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névél, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. [Findings of the 2016 conference on machine translation](#). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198. Association for Computational Linguistics.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. [Findings of the 2015 workshop on statistical machine translation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46. Association for Computational Linguistics.
- Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. [Findings of the 2018 conference on machine translation \(WMT18\)](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303. Association for Computational Linguistics.
- Vicent Briva-Iglesias. 2025. [Are AI agents the new machine translation frontier? challenges and opportunities of single- and multi-agent systems for multilingual digital communication](#).
- Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. [\(meta-\) evaluation of machine translation](#). In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Association for Computational Linguistics.
- Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. [Further meta-evaluation of machine translation](#). In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106. Association for Computational Linguistics.
- Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. [Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation](#). In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53. Association for Computational Linguistics.
- Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. [Findings of the 2012 workshop on statistical machine translation](#). In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51. Association for Computational Linguistics.
- Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. [Findings of the 2009 Workshop on Statistical Machine Translation](#). In *Proceedings of the*

- Fourth Workshop on Statistical Machine Translation*, pages 1–28. Association for Computational Linguistics.
- Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. [Findings of the 2011 workshop on statistical machine translation](#). In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics.
- Marine Carpuat, Omri Asscher, Kalika Bali, Luisa Bentivogli, Frédéric Blain, Lynne Bowker, Monojit Choudhury, Hal Daumé III, Kevin Duh, Ge Gao, Alvin Grissom II, Marzena Karpinska, Elaine C. Khoong, William D. Lewis, André F. T. Martins, Mary Nurminen, Douglas W. Oard, Maja Popovic, Michel Simard, and François Yvon. 2025. [An interdisciplinary approach to human-centered machine translation](#).
- Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. [WIT3: Web inventory of transcribed and translated talks](#). In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268. European Association for Machine Translation.
- Aaron Chatterji, Thomas Cunningham, David Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. [How people use ChatGPT](#).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.
- Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. [WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284. Association for Computational Linguistics.
- Adam Dobrowolski, Paweł Przewłocki, Paweł Przybyś, Marcin Szymański, and Dawid Siwicki. 2025. [A* decoding for Machine Translation in LLMs - SRPOL participation in WMT2025](#). In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. [CCAligned: A massive collection of cross-lingual web-document pairs](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969. Association for Computational Linguistics.
- Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. [COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565. ELRA and ICCL.
- Mara Finkelstein, Geza Kovacs, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Markus Freitag, and David Vilar. 2025. [Google Translate’s Research Submission to WMT2025](#). In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Markus Freitag, Isaac Caswell, and Scott Roy. 2019. [APE at scale and its implications on MT evaluation biases](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44. Association for Computational Linguistics.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. [Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628. Association for Computational Linguistics.
- Javier Garcia Gilabert, Xixian Liao, Severino Da Dalt, Ella Bohman, Audrey Mash, Francesca De Luca Fornaciari, Irene Baucells, Joan Llop, Miguel Claramunt, Carlos Escolano, and Maite Melero. 2025. [From SALAMANDRA to SALAMANDRATA: BSC Submission for WMT25 General Machine Translation Shared Task](#). In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. [Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 759–765. European Language Resources Association (ELRA).
- Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vázquez-Reina, Montserrat Marimon, Oriol Pareras, Valle Ruiz-Fernández, and Marta Villegas. 2025. [Salamandra technical report](#).
- Thamme Gowda, Roman Grundkiewicz, Elijah Rippeth, Matt Post, and Marcin Junczys-Dowmunt. 2024. [PyMarian: Fast neural machine translation and evaluation in python](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 328–335. Association for Computational Linguistics.
- Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. [Many-to-English machine translation tools, data, and pretrained models](#). In *Proceedings of the 59th*

- Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316. Association for Computational Linguistics.
- Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. [Continuous measurement scales in human evaluation of machine translation](#). In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41. Association for Computational Linguistics.
- Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. [Statistical power and translationese in machine translation evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81. Association for Computational Linguistics.
- Cristian Grozea and Oleg Verbitsky. 2025. Evaluation of QWEN-3 for English to Ukrainian Translation. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Ivan Grubišić and Damir Korencic. 2025. IRB-MT at WMT25 Translation Task: A Simple Agentic System Using an Off-the-Shelf LLM. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xcomet: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Kamil Guttman, Zofia Rostek, Adrian Charkiewicz, Antoni Solarzski, Mikołaj Pokrywka, and Artur Nowakowski. 2025. Laniqo at WMT25 General Translation Task: Self-Improved and Retrieval-Augmented Translation. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Selma Dis Hauksdóttir and Steinthor Steingrímsson. 2025. Automated Evaluation for Terminology Translation related to the EEA Agreement. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Svanhvít Lilja Ingólfssdóttir, Petur Ragnarsson, Haukur Jónsson, Haukur Simonarson, Vilhjálmur Thorsteinsson, and Vésteinn Snæbjarnarson. 2023. [Byte-level grammatical error correction using synthetic and curated corpora](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7299–7316. Association for Computational Linguistics.
- Svanhvít Lilja Ingólfssdóttir, Haukur Páll Jónsson, Kári Steinn Aðalsteinsson, Róbert Fjölfnir Birkisson, Sveinbjörn Þórðarson, and Þorvaldur Páll Helgason. 2025. Miðeind at WMT25 General Machine Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Koichi Iwakawa, Keito Kudo, Subaru Kimura, Takumi Ito, and Jun Suzuki. 2025. KIKIS at WMT 2025 General Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Atli Jasonarson and Steinthor Steingrímsson. 2025. AMI at WMT25 General Translation Task: How Low Can We Go? Finetuning Lightweight Llama models for Low Resource Machine Translation. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Josef Jon, Miroslav Hrabal, Martin Popel, and Ondřej Bojar. 2025. CUNI at WMT25 General Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. 2017a. [Fast-text.zip: Compressing text classification models](#). In *International Conference on Learning Representations (ICLR)*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017b. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for 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. Association for Computational Linguistics.
- Nikolay Karpachev, Ekaterina Enikeeva, Dmitry Popov, Arsenii Bulgakov, Daniil Panteleev, Dmitrii Ulianov, Artem Kryukov, and Artem Mekhraliev. 2025. Yandex Submission to the WMT25 General Machine Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. [DEMETER: Diagnosing evaluation metrics for translation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561. Association for Computational Linguistics.
- Ahrii Kim. 2025a. Multi-agentMT: Deploying AI Agent in the WMT25 Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Ahrii Kim. 2025b. [RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165. Association for Computational Linguistics.
- Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh

- Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, Nithya Govindarajan, Wei-Yin Ko, Julia Kreutzer, Kelly Marchisio, Ahmet Üstün, Sebastian Vincent, and Ivan Zhang. 2025a. Command-A-Translate: Raising the Bar of Machine Translation with Difficulty Filtering. 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, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinhórfur Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46. 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, Maja Popović, Mariya Shmatova, and Jun Suzuki. 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. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakounga, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary ranking of WMT25 general machine translation systems.
- 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. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203. European Association for Machine Translation.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86.
- Philipp Koehn and Christof Monz, editors. 2006. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics.
- Alon Lavie, Greg Hanneman, Sweta Agrawal, Kanojia Diptesh, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhuja, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.
- Zheng Li. 2025. HYT at WMT25 General Machine Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014a. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172. European Association for Machine Translation.
- Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014b. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.
- Samuel Lübbli, Sheila Castilho, Graham Neubig, Rico Senrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–Machine parity in language translation. *Journal of Artificial Intelligence Research (JAIR)*, 67.
- Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947. European Language Resources Association.
- Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems

- for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073. Association for Computational Linguistics.
- Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. [Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT?](#) In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245. Association for Computational Linguistics.
- Shushen Manakhimova, Ekaterina Lapshinova-Koltunski, Maria Kunilovskaya, and Eleftherios Avramidis. 2025. Fine-Grained Evaluation of English-Russian MT in 2025: Linguistic Challenges Mirroring Human Translator Training. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2024. [Investigating the linguistic performance of large language models in machine translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 355–371. Association for Computational Linguistics.
- Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. [EuroLLM-9B: Technical report](#).
- Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. [Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997. Association for Computational Linguistics.
- Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. [Machine translation meta evaluation through translation accuracy challenge sets](#). *Computational Linguistics*, 51(1):73–137.
- Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. [JParaCrawl: A large scale web-based English-Japanese parallel corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3603–3609. European Language Resources Association.
- Graham Neubig. 2011. [The Kyoto free translation task](#).
- Lei Pang, Hanyi Mao, Qianxia Xiao, Chen Ruihan, Jingjun Zhang, Haixiao Liu, and Xiangyi Li. 2025. In2x at WMT25 Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. [Guardians of the machine translation meta-evaluation: Sentinel metrics fall in!](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244. Association for Computational Linguistics.
- Martin Popel, Jindřich Libovický, and Jindřich Helcl. 2022. [CUNI systems for the WMT 22 Czech-Ukrainian translation task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 352–357. Association for Computational Linguistics.
- Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. 2019. [English-Czech systems in WMT19: Document-level transformer](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 342–348. Association for Computational Linguistics.
- Martin Popel, Lucie Polakova, Michal Novák, Jindřich Helcl, Jindřich Libovický, Pavel Straňák, Tomas Krabac, Jaroslava Hlavacova, Mariia Anisimova, and Tereza Chlanova. 2024. [Charles translator: A machine translation system between Ukrainian and Czech](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3038–3045. ELRA and ICCL.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618. Association for Computational Linguistics.
- Maja Popovic. 2021. [On nature and causes of observed MT errors](#). In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 163–175. Association for Machine Translation in the Americas.
- Maja Popović and Ekaterina Lapshinova-Koltunski. 2025. GENDER1PERSON: Test Suite for estimating gender bias of first-person singular forms. In *Proceedings of the Tenth Conference on Machine Translation*, China. 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. Association for Computational Linguistics.
- Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. [Estimating machine translation difficulty](#).
- Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. [JESC: Japanese-English subtitle corpus](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. [Robust speech recognition via large-scale weak supervision](#). In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Ricardo Rei, Nuno M. Guerreiro, Josão Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. [Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task](#).

- In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848. Association for Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645. Association for Computational Linguistics.
- Roberts Rozis and Raivis Skadiņš. 2017. [Tilde MODEL - multilingual open data for EU languages](#). In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265. Association for Computational Linguistics.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. [WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361. Association for Computational Linguistics.
- Hayate Shiroma. 2025. SH at WMT25 General Machine Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Einar Sigurdsson, Magnús Már Magnússon, Atli Jasonarson, and Steinthor Steingrímsson. 2025. Up to Par? MT Systems Take a Shot at Sports Terminology. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Archchana Sindhuja, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. [When LLMs struggle: Reference-less translation evaluation for low-resource languages](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459. Association for Computational Linguistics.
- Anushka Singh, Ananya Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. [How good is zero-shot MT evaluation for low resource Indian languages?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649. Association for Computational Linguistics.
- Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2025a. Evaluation of LLM for English to Hindi Legal Domain Machine Translation Systems. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2025b. Instruction-Tuned English to Bhojpuri Neural Machine Translation Using Contrastive Preference Optimization. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Shaomu Tan. 2025. Kaze-MT at WMT25 Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Cohere Team. 2025. [Command a: An enterprise-ready large language model](#).
- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218. European Language Resources Association (ELRA).
- Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. [Attaining the unattainable? reassessing claims of human parity in neural machine translation](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123. Association for Computational Linguistics.
- Hao Wang, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao, Bo Zeng, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. Marco Large Translation Model at WMT2025: Transforming Translation Capability in LLMs via Quality-Aware Training and Decoding. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024. [Evaluating WMT 2024 metrics shared task submissions on AfriMTE \(the African challenge set\)](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 505–516. Association for Computational Linguistics.
- Christopher Lemmer Webber, Jessica Tallon, Erin Shepherd, Amy Guy, and Evan Prodromou. 2018. [Activitypub, w3c recommendation](#). Technical report, W3C.
- Johnny Wei and Robin Jia. 2021. [The statistical advantage of automatic NLG metrics at the system level](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6840–6854. Association for Computational Linguistics.
- Frank Wilcoxon. 1945. [Individual comparisons by ranking methods](#).
- Di Wu, Yan Meng, Maya Konstantinovna Nachesa, Seth Aycock, and Christof Monz. 2025. UvA-MT’s Participation in the WMT25 General Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Jiafeng Xiong and Yuting Zhao. 2025. KYUoM’s Submissions to the WMT 2025 General Machine Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. [MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation](#). In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360. Association for Computational Linguistics.

- Zhang Yin, Hiroyuki Deguchi, Haruto Azami, Guanyu Ouyang, Kosei Buma, Yingyi Fu, Katsuki Chousa, and Takehito Utsuro. 2025. NTTSU at WMT2025 General Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. [Prompting large language model for machine translation: A case study](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.
- Dakun Zhang, Yara Khater, Ramzi Rahli, Anna Rebollo, and Josep Crego. 2025. SYSTRAN @ WMT 2025 General Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Mao Zheng, Zheng Li, Yang Du, Bingxin Qu, and Mingyang Song. 2025. Shy-hunyuan-MT at WMT25 General Machine Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. [The United Nations parallel corpus v1.0](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534. European Language Resources Association (ELRA).
- Hao Zong, Chao Bei, Wentao Chen, Conghu Yuan, Huan Liu, and Degen Huang. 2025. DLUT and GTCOM’s Large Language Model Based Translation System for WMT25. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.
- Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024a. [Pitfalls and outlooks in using COMET](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288. Association for Computational Linguistics.
- Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024b. [Fine-tuned machine translation metrics struggle in unseen domains](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500. Association for Computational Linguistics.
- Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025a. [AI-assisted human evaluation of machine translation](#). 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 4936–4950. Association for Computational Linguistics.
- Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025b. [How to select datapoints for efficient human evaluation of NLG models?](#)

A Error Span Annotation Miscellaneous

The following instructions were shown in the annotation interface and could be accessed at any time.

Highlighting errors: *Select the part of translation where you have identified a translation error (drag or click start & end). Click on the highlight to change error severity (minor/major) or remove the highlight.*

Choose error severity:

- *Minor errors: Style, grammar, word choice could be better or more natural.*
- *Major errors:: The meaning is changed significantly and/or the part is really hard to understand.*

Tips:

- *Missing content: If something is missing, highlight the word [MISSING] to mark the error.*
- *Tip: Highlight the word or general area of the error (it doesn't need to be exact). Use multiple highlights for different errors.*
- *Tip: Pay particular attention to translation consistency between texts across the whole document.*
- *Tip: If the translation is in the wrong language, mark it fully and assign it 0*
- *Tip: If the translation contains additional text (e.g. "Here is the translation") or alternative secondary translation, mark it as a major error.*
- *Using external tools for annotations (chatbots, LLMs) is not allowed.*

Score the translation: *After marking errors, please use the slider and set an overall score based on meaning preservation and general quality:*

- *0: Broken/poor translation.*
- *33%: Flawed: significant issues*
- *66%: Good: insignificant issues with grammar, fluency, or consistency*
- *100%: Perfect: meaning and style aligned completely with the source*

Changes to ESA interface. We introduced the following changes to the ESA interface since the previous use in WMT 2024 (Kocmi et al., 2024a):

- Dropped the requirement for a task being strictly 100 segments.
- We added a crude character-based alignment (Figure 1, top).
- We include images on the source text (Figure 1, bottom right).
- We use a translated version of the tutorial for each language pair.
- Updated the annotation instructions and the annotation scale.
- Updated the interface style slightly (Figure 1).

B Translator Brief

The following instructions were given to human translators:

In this project we wish to translate data from several domains for use in the evaluation of Machine Translation (MT). The translations produced by you will be compared against the translations produced by a variety of different MT systems. They will be released to the research community to provide a benchmark, or “gold-standard” measure for translation quality. The translation therefore needs to be a high-quality rendering of the source text into the target language, as if it was originally written directly in the target language. However, there are some constraints imposed by the intended usage:

- *All translations must be “from scratch,” without post-editing from machine translation or usage of CAT tools. Post-editing machine translation would bias the evaluation, so we need to avoid it. We can detect post-editing and will reject translations that are post-edited.*
- *Translators must preserve paragraph boundaries, which are marked by empty lines in the source text, but they are free to adjust the number of sentences within each paragraph.*
- *Translators should avoid inserting parenthetical explanations into the translated text and obviously avoid losing any pieces of information from the source text. We will check the translations for quality and will reject translations that contain errors.*
- *If the original data contains errors, typos, or other problems, do not change the source sentences, instead try to prepare a correct translation as if the error wouldn't be in the source.*

- The data contain four domains (news, speech, social, literary), each folder containing one domain source and each domain needing a specific handling

The source files will be delivered as text files (sometimes known as “notepad” files), with paragraphs separated by an empty line. We need the translations to be returned in the same format. The translation file needs to have the same name as the original file.

Speech Domain. Your task is to translate the speech from provided video. We also provide you with automated transcription, which is not human edited and contains errors, thus should be used only as a guideline for translation from the video. Each file represents one segment of video. Videos correspond to different domains: they differ in formality, style, topics and number of speakers. The idea is to translate using the most similar language in the target language, matching as best as possible the characteristics of the source video.

Social Domain. The texts are from the social network Mastodon (similar to Twitter). Each file represents a thread or part of a thread from one or several users. Different posts within a thread are separated with empty line. Individual posts can also span several lines. The sentences have been selected so that they do not contain offensive or sensitive content (hate speech, taking-drugs, suicide, politically sensitive topics, etc.). However, profanities were kept as they were taken to be illustrative of the sociolect of online language. If however, you do not feel comfortable with translating something, please let us know.

The texts are particular in that they may contain spelling errors, slang, acronyms, marks of expressivity, etc. The idea is to translate using the most natural language in the target language, matching as best as possible the style and familiarity of the source text.

- Spelling mistakes should not be preserved in their translations, i.e. the translation should be spelt correctly. Introduce proper capitalization in translations.
- Marks of expressivity (e.g. asterisks *wow*, capitals letters WOW) should be conserved as best as possible. However, do not attempt to reproduce repeated characters (e.g. woowoow) in translation, as the choice as to which character to repeat is often arbitrary.
- There will be abbreviations and acronyms (e.g. btw -> by the way, fwiw -> for what it's worth). These do not need to be translated using abbreviation or acronyms unless an abbreviation/acronym is the best translation choice in the target language.
- Users (@user123) and URLs should be left as they are, i.e. not translated.
- Platform-specific elements such as hashtags should be translated as hashtags, but the content should be translated appropriately into the target language.
- Punctuation can be added if it necessary to avoid comprehension difficulties. Otherwise, we recommend following the punctuation of the source text.

Please always refer to the screenshots included alongside the source texts. These screenshots show the original context of the entire thread contained in the source text file and should be consulted during translation. Screenshot files have the same name as the corresponding source text files but with a .png image extension instead of .txt. The screenshots may also contain images attached in the thread that can provide further context for the translation.

Literary domain. The texts in this domain are stories written by aspiring writers. Each story should be translated as one coherent text, preserving characters' speech patterns and personalities consistently. Aim to maintain the original tone and register, retaining the emotional depth of the story. Dialogues should sound natural and follow the conventions of the target language.

C System Submission Summaries

This section lists all the submissions to the translation task and provides the authors' descriptions of their submission.

C.1 Algharb (Wang et al., 2025)

In this paper, we introduce a large language model system for translation, developed through a comprehensive training pipeline. Our submissions include translations from English to Chinese, Arabic, Czech, Japanese, Korean, Russian, Ukrainian, Serbian, Bhojpuri, and Estonian, as well as from Czech to German/Ukrainian and Japanese to Chinese. Our approach integrates Machine Translation-based Supervised Fine-Tuning with post-training reinforcement via Group Relative Policy Optimization (GRPO). For decoding, we employ a Minimum Bayes Risk (MBR) algorithm enhanced with a finetuned reranker. This combined strategy ensures the generation of robust and consistent high-quality translations across a diverse set of languages.

The model is available on Hugging Face: huggingface.co/AIDC-AI/Marco-MT-Algharb.

C.2 AMI (Jasonarson and Steingrímsson, 2025)

We present the submission of the Árni Magnússon Institute's team for the WMT25 General translation task. We focus on the English→Icelandic translation direction. We pre-train Llama 3.2 3B on 10B tokens of English and Icelandic texts and fine-tune on parallel corpora. Multiple translation hypotheses are produced first by the fine-tuned model, and then more hypotheses are added by that same model further tuned using contrastive preference optimization. The hypotheses are then post-processed using a grammar correction model and post-processing rules before the final translation is selected using minimum Bayes risk decoding. We found that while it is possible to generate translations of decent quality based on a lightweight model with simple approaches such as the ones we apply, our models are quite far behind the best participating systems and it would probably take somewhat larger models to reach competitive levels.

The model is available on Hugging Face: huggingface.co/arnastofnun/Llama-3.2-3B-wmt25-AMI-en-is.

C.3 CGFOKUS (Grozea and Verbitsky, 2025)

We report here the outcome of evaluating Qwen3 for the English to Ukrainian language pair of the general MT task of WMT 2025. In addition to the quantitative evaluation, a qualitative evaluation was performed, leveraging the cooperation with a native Ukrainian speaker - therefore we present an example-heavy analysis of the typical failures the LLMs still do when translating natural language, particularly into Ukrainian. We report also on the practicalities of using LLMs, such as on the difficulties of making them follow instruction, on ways to exploit the increased “smartness” of the reasoning models while simultaneously avoiding the reasoning part interfering wrongly with the chain of which the LLM is just one element.

C.4 COILD-BHO (Singh et al., 2025b)

This paper presents an English to Bhojpuri machine translation (MT) system developed for the WMT25 General MT Shared Task. Given the low-resource nature of Bhojpuri, we adopt a two-stage training pipeline: unsupervised pretraining followed by supervised fine-tuning. During pretraining, we use a 300,000-sentence corpus comprising 70% Bhojpuri monolingual data and 30% English data to establish language grounding. The fine-tuning stage utilizes 29,749 bilingual English to Bhojpuri sentence pairs (including training, validation, and test sets). To adapt the system to instruction-following scenarios, we apply a novel optimization strategy: Contrastive Preference Optimization (CPO). This technique enables the model to capture fine-grained translation preferences and maintain semantic fidelity in instruction-tuned settings. We evaluate our system across multiple metrics, demonstrating moderate performance in low-resource MT tasks, particularly in diverse domains such as literary, news, social, and speech.

The model is available at: drive.google.com/drive/folders/1ZzJ9ZlfqaT-fEo5umovNN4HWkZhOlqqC.

C.5 CommandA-WMT (Kocmi et al., 2025a)

We built our system on top of Command-A using a direct preference optimization with data preparation pipeline that emphasizes robust data quality control, primarily incorporating standard quality filtering along with a novel difficulty filtering component, which serves as the key innovation of our approach. The final translation is built through step-by-step reasoning, and we employ limited Minimum Bayes Risk decoding with a limited candidate pool size of 20, using MetricX-XL as the primary utility metric. For unsupported languages, we use a second model prepared identically but with an additional initial supervised fine-tuning step for the unsupported languages that Command-A model has not been trained on.

The model is available on Hugging Face: huggingface.co/CohereLabs/command-a-translate-08-2025

C.6 CUNI-EdUKate-v1 (Jon et al., 2025)

CUNI-EdUKate-v1 is an unconstrained system trained on educational domain data using LoRA, SFT, and Contrastive Preference Optimization. It is also fine-tuned from the EuroLLM-9B-Instruct model. It only supports the cs2uk language direction and, unlike CUNI-MH-v2, both training and inference were done at the sentence level.

C.7 CUNI-MH-v2 (Jon et al., 2025)

CUNI-MH-v2 is a constrained system trained on partially synthetic data sampled from the CzEng 2.0 dataset using LoRA and Contrastive Preference Optimization. We plan on releasing both the model and the filtered training data. It is fine-tuned from the EuroLLM-9B-Instruct model. We currently only support two language directions, en2cs and cs2de, and offer separate LoRA adapters for each. The translations were done on the paragraph level.

The models are available on Hugging Face: huggingface.co/hrabalm/CUNI-MH-v2-encs and huggingface.co/hrabalm/CUNI-MH-v2-csde.

C.8 CUNI-SFT (Jon et al., 2025)

This paper describes the joint effort of Phrase a.s. and CUNI/UFAL on the WMT25 Automated Translation Quality Evaluation Systems Shared Task. Both teams participated both in a collaborative and competitive manner, i.e. they each submitted a system of their own as well as a contrastive joint system ensemble. In Task 1, we show that such an ensembling—if chosen in a clever way—can lead to a performance boost. We present the analysis of various kinds of systems comprising both “traditional” NN-based approach, as well as different flavours of LLMs—off-the-shelf commercial models, their fine-tuned versions, but also in-house, custom-trained alternative models. In Tasks 2 and 3 we show Phrase’s approach to tackling the tasks via various GPT models: Error Span Annotation via the complete MQM solution using non-reasoning models (including fine-tuned versions) in Task 2, and using reasoning models in Task 3.

The model is available on Hugging Face: huggingface.co/ufal/wmt25-cuni-sft.

C.9 CUNI-Transformer and CUNI-DocTransformer (Popel et al., 2022, 2019)

CUNI-Transformer and CUNI-DocTransformer rely on standard NMT training with Block backtranslation and optionally document-level training.

The models are available at lindat.mff.cuni.cz/repository/items/b1cfdecf-fda3-4198-a537-e58a20ddea60 and lindat.mff.cuni.cz/repository/items/4b5d758f-ca9e-4ca6-8129-2331928ba950.

C.10 DLUT_GTCOM (Zong et al., 2025)

This paper presents the submission from Dalian University of Technology (DLUT) and Global Tone Communication Technology Co., Ltd. (GTCOM) to the WMT25 General Machine Translation Task. Amidst the paradigm shift from specialized encoder-decoder models to general-purpose Large Language Models (LLMs), this work conducts a systematic comparison of both approaches across five language pairs. For traditional Neural Machine Translation (NMT), we build strong baselines using deep Transformer architectures enhanced with data augmentation. For the LLM paradigm, we explore zero-shot performance

and two distinct supervised fine-tuning (SFT) strategies: direct translation and translation refinement. Our key findings reveal a significant discrepancy between lexical and semantic evaluation metrics: while strong NMT systems remain competitive in BLEU scores, fine-tuned LLMs demonstrate marked superiority in semantic fidelity as measured by COMET. Furthermore, we find that fine-tuning LLMs for direct translation is more effective than for refinement, suggesting that teaching the core task directly is preferable to correcting baseline outputs.

C.11 Erlendur (Ingólfssdóttir et al., 2025)

We present Miðeind’s system contribution for English-to-Icelandic translation. We participate in the Terminology Shared Task with the same system. Erlendur is a multilingual LLM-based translation system which employs a multi-stage pipeline approach, with enhancements especially for translations from English to Icelandic. We address translation quality and grammatical accuracy challenges in current LLMs through a hybrid prompt-based approach that can benefit lower-resource language pairs. In a preparatory step, the LLM analyzes the source text and extracts key terms for lookup in an English-Icelandic dictionary. Main results of the analysis and the retrieved dictionary results are then incorporated into the translation prompt. When provided with a custom glossary, the system identifies relevant terms from the glossary and incorporates them into the translation as well, to ensure consistency in terminology. For longer inputs, the system maintains translation consistency by providing contextual information from preceding text chunks. Lastly, Icelandic target texts are passed through our custom-developed seq2seq language correction model (Ingólfssdóttir et al., 2023), where grammatical errors are corrected. Using this hybrid method, Erlendur delivers high-quality translations, without fine-tuning.

C.12 HYT (Li, 2025)

This paper illustrates the submission system of the HYT team for the WMT25 General Machine Translation shared task. We submitted translations for all translation directions in the general machine translation task and test suites subtask. The ID of our submission in OCELoT system is 43, which can be categorized as being in the unconstrained track. The base model we use is Hunyuan-TurboS. Overall, we first performed continued pretraining(CPT) using open-source data to enhance the model’s multilingual capabilities. Then, we used DeepSeek-V3-03241 to synthesize a large amount of parallel data and performed Reinforcement learning on the CPT model. Finally, we used ensemble learning to further improve translation quality.

C.13 GemTrans (Finkelstein et al., 2025)

Large Language Models have shown impressive multilingual capabilities, where translation is one among many tasks. Google Translate’s submission to the 2025 WMT evaluation tries to research how these models behave when pushing their translation performance to the limit. Starting with the strong Gemma 3 model, we carry out supervised fine-tuning on high quality, synthetically generated parallel data. Afterwards we perform an additional Reinforcement Learning step, with reward models based on translation metrics to push the translation capabilities even further. Controlling the combination of reward models, including reference-based and quality estimation metrics, we found that the behaviour of the model could be tailored towards a more literal or more creative translation style. Our two submissions correspond to those two models. We chose the more creative system as our primary submission, targeting a human preference for better sounding, more naturally flowing text, although at the risk of losing on the accuracy of the translation. It is an open question to find the sweet spot between these two dimensions, which certainly will depend on the specific domain to handle and user preferences.

C.14 In2x (Pang et al., 2025)

This paper presents the open-system submission by the In2x research team for the WMT25 General Machine Translation Shared Task. Our submission focuses on Japanese-related translation tasks, aiming to explore a generalizable paradigm for extending large language models (LLMs) to other languages. This paradigm encompasses aspects such as data construction methods and reward model design. The ultimate goal is to enable large language model systems to achieve exceptional performance in low-resource or less commonly spoken languages.

C.15 IR-MultiagentMT (Kim, 2025a)

We introduce our model, referred to as Multi-agentMT, for participation in the WMT 25 General Machine Translation Shared Task. This model operationalizes the notion of an AI Agent by employing a multi-agent workflow known as Prompt Chaining (Briva-Iglesias, 2025) alongside the automatic MQM (Multidimensional Quality Metrics) error annotation framework designated as RUBRIC-MQM (Kim, 2025b). Our primary submission is developed through the Translate-Postedit-Proofread paradigm, whereby the positions of the errors are clearly marked and enhanced throughout the process. Our study suggests that a semi-autonomous agent scheme in Machine Translation is viable with an older and smaller model in some language pairs, resulting in comparable results with 2.3x faster speed and only 2% of the budget.

C.16 IRB-MT (Grubišić and Korencić, 2025)

Large Language Models (LLMs) have been demonstrated to achieve state-of-art results on machine translation. LLM-based translation systems usually rely on model adaptation and fine-tuning, requiring datasets and compute. The goal of our team’s participation in the “General Machine Translation” and “Multilingual” tasks of WMT25 was to evaluate the translation effectiveness of a resource-efficient solution consisting of a smaller off-the-shelf LLM coupled with a self-refine agentic workflow. Our approach requires a high-quality multilingual LLM capable of instruction following. We select Gemma3-12B among several candidates using the pretrained translation metric MetricX-24-XL and a small development dataset. WMT25 automatic evaluations place our solution in the mid tier of all WMT25 systems, and also demonstrate that it can perform competitively for approximately 16% of language pairs.

C.17 Kaze-MT (Tan, 2025)

This paper describes the Kaze-MT submission to the WMT25 General Machine Translation task for the Japanese-Chinese track. The system relies on a minimalist Test-Time Scaling (TTS) pipeline composed of three stages: Sampling, Scoring, and Selection. In the sampling stage, we utilize zero-shot Qwen 2.5 models (72B and 14B) to generate 512 candidate translations under a fixed temperature schedule, encouraging diversity without compromising fluency. In the scoring stage, each candidate is evaluated using multiple quality estimation (QE) models, namely KIWI22, MetricX-24, and ReMedy-24. Finally, we select the final candidate based on rank aggregation across QE scores. Our approach requires no fine-tuning, in-context examples, or specialized decoding heuristics, and we participate in both constrained and unconstrained tracks. Preliminary results show competitive performance on automatic metrics, with final human evaluation results to be reported in the camera-ready version.

C.18 KIKIS (Iwakawa et al., 2025)

We participated in the constrained English–Japanese track of the WMT 2025 General Machine Translation Task. Our system collected the outputs produced by multiple subsystems, each of which consisted of LLM-based translation and reranking models configured differently (e.g., prompting strategies and context sizes), and reranked those outputs. Each subsystem generated multiple segment-level candidates and iteratively selected the most probable one to construct the document translation. We then reranked the document-level outputs from all subsystems to obtain the final translation. For reranking, we adopted a text-based LLM reranking approach with a reasoning model to take long contexts into account. Additionally, we built a bilingual dictionary on the fly from the parallel corpus to make the system more robust to rare words.

C.19 KYUoM (Xiong and Zhao, 2025)

This paper describes the KYUoM team’s submission system for the WMT 2025 general translation task. We focused on exploring the capabilities of inductive generalization from a multimodal domain to a text-based domain of machine translation. Our submission system consists of a two-stage adaptation process with multimodal domain learning in the first stage and textual domain adaptation in the second stage for the English to Ukrainian task in the unconstrained track. The main advance is using a GAT adapter to achieve two-stage continuous learning for cross-modal generalization.

C.20 Lanigo (Guttmann et al., 2025)

This work describes Lanigo’s submission to the constrained track of the WMT25 General MT Task. We participated in 11 translation directions. Our approach combines several techniques: fine-tuning the EuroLLM-9B-Instruct model using Contrastive Preference Optimization on a synthetic dataset, applying Retrieval-Augmented Translation with human-translated data, implementing Quality-Aware Decoding, and performing postprocessing of translations with a rule-based algorithm. We analyze the contribution of each method and report improvements at every stage of our pipeline.

The model is available on Hugging Face: huggingface.co/lanigo/WMT25-EuroLLM-9B-CPO.

C.21 NTTSU (Yin et al., 2025)

This paper presents the submission of NTTSU for the constrained track of the English–Japanese and Japanese–Chinese language directions at the WMT2025 general translation task. For each translation direction, we build translation models from a large language model by combining continual pretraining, supervised fine-tuning, and preference optimization based on the translation quality and adequacy. We finally generate translations via context-aware MBR decoding to maximize translation quality and document-level consistency.

The models are available on Hugging Face: huggingface.co/UtsuroLab/WMT25_En-Ja and huggingface.co/UtsuroLab/WMT25_Ja-Zh.

C.22 RuZH-Eole (no paper submission)

Eole NLP Submission uses Tower+ 9B model with an extra layer for quality estimation. It generates multiple hypotheses and rank them according to an internal score.

C.23 SalamandraTA (Gilabert et al., 2025)

In this paper, we present the SALAMANDRA^{TA} family of models, an improved iteration of SALAMANDRA LLMs (Gonzalez-Agirre et al., 2025) specifically trained to achieve strong performance in translation-related tasks for 38 European languages. SALAMANDRA^{TA} comes in two scales: 2B and 7B parameters. For both versions, we applied the same training recipe with a first step of continual pre-training on parallel data, and a second step of supervised fine-tuning on high-quality instructions. The BSC submission to the WMT25 General Machine Translation shared task is based on the 7B variant of SALAMANDRA^{TA}. We first adapted the model vocabulary to support the additional non-European languages included in the task. This was followed by a second phase of continual pre-training and supervised fine-tuning, carefully designed to optimize performance across all translation directions for this year’s shared task. For decoding, we employed two quality-aware strategies: Minimum Bayes Risk Decoding and Tuned Re-ranking using COMET and COMET-KIWI respectively.

We publicly release both the 2B and 7B versions of SALAMANDRA^{TA}, along with the newer SALAMANDRA^{TA}-V2 model, on Hugging Face: huggingface.co/LangTech-MT/salamandraTA-7b-instruct-WMT25.

C.24 SH (Shiroma, 2025)

We participated in the unconstrained track of the English-to-Japanese translation task at the WMT 2025 General Machine Translation Task. Our submission leverages several large language models, all of which are trained with supervised fine-tuning, and some further optimized via preference learning. To enhance translation quality, we introduce an automatic post-editing model and perform automatic post-editing. In addition, we select the best translation from multiple candidates using Minimum Bayes Risk (MBR) decoding with the use of COMET-22 and LaBSE-based cosine similarity as evaluation metrics.

C.25 Shy-hunyuan-MT (Zheng et al., 2025)

This paper presents our submission to the WMT25 shared task on machine translation, for which we propose Synergy-enhanced policy optimization, named Shy, a novel two-phase training framework that synergistically combines ensemble knowledge distillation with reinforcement learning optimization. In the first phase, we introduce a multi-stage training framework that harnesses the complementary strengths of multiple state-of-the-art large language models to generate diverse, high-quality translation

candidates. These candidates serve as pseudo-references to guide the supervised fine-tuning of our model, Hunyuan-7B, effectively distilling the collective knowledge of multiple expert systems into a single efficient model. In the second phase, we further refine the distilled model through Group Relative Policy Optimization, a reinforcement learning technique that employs a composite reward function. By calculating reward from multiple perspectives, our model ensures better alignment with human preferences and evaluation metrics. Extensive experiments across multiple language pairs demonstrate that our model **Shy-hunyuan-MT** yields substantial improvements in translation quality compared to baseline approaches. Notably, our framework achieves competitive performance with state-of-the-art systems while maintaining computational efficiency through knowledge distillation and strategic ensemble.

The model is available on Hugging Face: huggingface.co/collections/tencent/hunyuan-mt-68b42f76d473f82798882597.

C.26 SRPOL (Dobrowolski et al., 2025)

This work presents an innovative decoding approach utilizing the A* (A-star) algorithm, which generates a diverse and precise set of translation hypotheses. Subsequent reranking through the Noisy Channel Model Reranking and Quality Estimation selects the best among these diverse hypotheses, leading to a significant improvement in translation quality. This approach achieves up to a 0.5-point reduction in the MetricX-24 score and a 1.5-point increase in the COMET score. The A* algorithm can be applied to decoding in any LLMs or classic transformers. The experiment shows that by using freely available, open-source MT models, it is possible to achieve translation quality comparable to the best online translators and LLMs using only a PC under your desk.

C.27 Sysstran (Zhang et al., 2025)

We present an English-to-Japanese translation system built upon the EuroLLM-9B (Martins et al., 2025) model. The training process involves two main stages: continue pretraining (CPT) and supervised fine-tuning (SFT). After both stages, we further tuned the model using a development set to optimize performance. For training data, we employed both basic filtering techniques and high-quality filtering strategies to ensure data cleanliness. Additionally, we classify both the training data and development data into four different domains and we train and fine-tune with domain specific prompts during system training. Finally, we applied Minimum Bayes Risk (MBR) decoding and paragraph-level reranking for post-processing to enhance translation quality.

The models are available on Hugging Face: huggingface.co/collections/Sysstran/wmt25-en-ja-6867eed78ea21e28a282aaed.

C.28 TranssionMT (no paper submission)

The team employs the Transformer architecture and finetuning the translation of a specific language within the multilingual pretrained model to enhance its translation performance. They adopts various strategies, such as finetuning language model instructions, joint training of similar languages, integrated model decision-making, and non-English data mining, all aimed at improving the translation outcomes.

C.29 TranssionTranslate (no paper submission)

This paper presents our machine translation system developed for the WMT25 shared task. Our approach leverages state-of-the-art neural architectures, including transformer-based models with advanced pre-training and fine-tuning techniques. We focus on multilingual and domain-adaptive strategies to enhance translation quality across diverse language pairs. Key features include: (1) large-scale pretraining on parallel and monolingual corpora, (2) dynamic data filtering and domain adaptation, (3) ensemble and reranking methods to improve fluency and accuracy. We explore both supervised and zero-shot settings, particularly for low-resource languages. Our system demonstrates competitive performance on WMT25 evaluation benchmarks, achieving improvements in BLEU, TER, and human evaluation metrics. We analyze challenges such as rare word translation, syntactic divergence, and robustness to noisy inputs. The results highlight the effectiveness of our approach in balancing generalization and language-specific optimization. This work contributes insights into scalable and adaptive MT systems, with potential

applications in multilingual NLP tasks. Future directions include better handling of linguistic diversity and real-time adaptation.

C.30 UvA-MT (Wu et al., 2025)

The UvA-MT’s submission is competing in the unconstrained track across all 16 translation directions. Unusually, this year we use only the source side of the test set to generate synthetic data for LLM training, and translations are produced using pure beam search for submission. Overall, our approach can be seen as a special variant of data distillation, motivated by two key considerations: (1) perfect domain alignment, where the training and test domains are distributionally identical; and (2) the strong teacher model, GPT-4o-mini, offers high-quality outputs as both a reliable reference and a fallback in case of mere memorization. Interestingly, the outputs of the resulting model, trained on Gemma3-12B using Best-of-N (BoN) outputs from GPT-4o-mini, outperform the original BoN outputs in some high-resource languages across various metrics, including CometKiwi-XXL which is the very metric used for BoN selection. We attribute this to a successful model ensemble, where the student model (Gemma3-12B) retains the strengths of the teacher (GPT-4o-mini) while implicitly avoiding their flaws. Our experiments on other datasets, such as WMT24++, also confirm this observation.

C.31 Wenyil (no paper submission)

This paper introduces Wenyil, an advanced translation system based on a large language model (LLM). This multilingual model supports 13 language directions, and its superior performance is derived from a comprehensive training process that includes multi-stage supervised fine-tuning (SFT) for translation tasks and a two-stage post-training scheme. Furthermore, we propose a novel hybrid decoding strategy to overcome the limitations of standard decoding. This method integrates word alignment with an advanced Minimum Bayes Risk (MBR) re-ranking algorithm. This approach not only enhances translation stability but also ensures excellent accuracy across diverse linguistic contexts.

C.32 Yandex (Karpachev et al., 2025)

This paper describes Yandex’s submission to the WMT25 General Machine Translation task. We participate in the English-to-Russian translation direction and propose a purely LLM-based translation model. Our training procedure comprises a training pipeline of several stages built upon YandexGPT, an in-house general-purpose LLM. In particular, firstly, we employ continual pretraining (post-pretrain) for MT task for initial adaptation to multilinguality and translation. Subsequently, we use SFT on parallel document-level corpus in the form of P-Tuning. Following SFT, we propose a novel alignment scheme of two stages, the first one being a curriculum learning with difficulty schedule and a second one - training the model for tag preservation and error correction with human post-edits as training samples. Our model achieves results comparable to human reference translations on multiple domains.

C.33 Yolu (no paper submission)

This paper details Yolu’s submission for the WMT’25 General Machine Translation Task. Our work, situated within the constrained track, investigates the efficacy of Reinforcement Learning (RL) in enhancing machine translation. Our system is built upon the open-source Qwen3 model. We introduce a robust methodology for continuous performance improvement, which combines meticulous data cleaning with advanced data distillation techniques. This is complemented by a multi-stage optimization strategy, sequentially employing Continued Pre-Training (CPT), Supervised Fine-Tuning (SFT), Contrastive Preference Optimization (CPO), and a novel policy optimization algorithm, Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO). Furthermore, we integrate a Quality Estimation (QE) model to facilitate online QE distillation, thereby refining the model’s output during the decoding phase.

D Official Ranking Results (extends Section 7.4)

Results tables legend

The human score is the micro-average of human judgments across all domains and double annotations (single annotations for MQM language pairs). AutoRank is calculated from automatic metrics as per (Kocmi et al., 2025b). Significance testing is done using a [Wilcoxon signed rank test](#) with a p -value threshold of 5%. The rank range for the i th model begins as $\langle i, i \rangle$ and is expanded in both directions until a significant difference is found. Clusters are formed such that their constituent rank ranges do not overlap.

Systems are either constrained (white), or unconstrained (gray). Systems that do not officially support the language pair are marked with  and those where language support cannot be verified are marked with . The [M] suffix marks systems (submitted by the WMT organizers) that were trained/tuned with specific MT instructions, but prompted without these specific instructions (using a generic setup, same for all LLMs, see Section 4.2), which could disadvantage these systems.

English→Arabic (Egyptian)							
Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Human	78.5		83.8	74.7	79.1	77.3
2-2	GPT-4.1	77.0	6.7	80.4	74.7	77.3	76.2
3-3	CommandA	74.0	8.6	81.3	66.8	75.6	73.8
4-4	Gemini-2.5-Pro	60.6	5.8	88.4	0.9	84.3	80.8
5-6	DeepSeek-V3 	56.8	7.0	66.0	33.2	64.5	69.9
5-6	Claude-4	55.7	7.8	73.5	23.7	64.5	69.5
7-7	IRB-MT	51.9	11.1	62.2	20.0	68.2	61.6
8-9	Mistral-Medium	36.0	7.7	44.2	0.1	46.4	64.9
8-9	CommandA-WMT	34.6	4.1	37.0	30.4	18.1	66.8
10-10	UvA-MT	29.0	4.2	12.4	8.2	42.0	58.8
11-14	CommandR7B	3.7	11.6	0.0	0.6	3.2	13.9
11-14	GemTrans	3.7	3.5	0.0	0.1	1.6	17.6
11-16	Algharb	3.2	2.7	0.0	1.2	1.6	12.9
11-16	Shy-hunyuanyuan-MT	3.2	1.0	0.0	2.6	1.7	10.5
13-16	AyaExpanse-8B	2.0	9.9	0.0	0.0	1.7	8.2
12-16	ONLINE-B	1.7	6.5	0.0	0.6	1.8	5.0
17-19	Yolu	1.4	5.5	0.0	0.0	1.6	4.8
15-18	Wenyiil	1.4	2.5	0.0	0.6	1.7	3.5
19-19	SRPOL 	0.9	8.1	0.0	0.0	1.6	2.4
20-39	19 systems not human-evaluated		...				

English→Estonian							
Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Human	83.1		96.8	83.0	82.3	68.2
2-2	Gemini-2.5-Pro	78.8	2.5	72.3	78.1	88.9	71.0
3-4	Wenyiil	72.6	2.6	63.5	77.2	78.3	67.8
3-4	GPT-4.1	72.2	3.0	79.0	71.4	72.2	64.9
5-6	Algharb	70.4	3.9	51.9	77.0	79.7	68.0
5-6	Shy-hunyuanyuan-MT	70.3	1.0	71.3	73.8	69.3	65.2
7-8	ONLINE-B	60.2	6.0	80.3	52.6	58.8	49.7
7-8	Yolu	59.5	3.8	66.9	58.5	60.4	50.5
9-9	TranssionTranslate 	57.1	7.3	55.5	59.4	64.9	42.9
10-11	Claude-4 	53.0	6.5	51.0	53.5	58.7	45.2
10-12	GemTrans	51.7	5.1	38.6	51.0	58.3	57.6
11-14	CommandA-WMT 	50.1	6.1	53.7	48.7	52.0	45.2
12-15	SRPOL	49.4	5.7	40.4	53.8	54.1	46.2
12-17	Lanigo	48.6	5.2	50.1	53.7	45.8	43.5
13-17	EuroLLM-22B-pre.[M]	47.2	8.1	49.8	42.2	51.9	44.1
14-18	SalamandraTA	46.7	6.3	40.2	49.5	48.8	46.9
14-18	UvA-MT	46.4	5.9	55.0	38.5	45.4	49.7
16-18	Gemma-3-27B	45.9	7.6	32.6	51.7	46.9	51.4
19-19	IRB-MT	32.4	11.4	14.8	35.8	36.4	42.1
20-40	20 systems not human-evaluated		...				

English→Bhojpuri							
Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Gemini-2.5-Pro	94.9	1.0	98.8	95.9	95.1	88.3
2-3	Human	92.6		97.7	92.7	94.1	83.8
2-3	Algharb	91.1	2.8	89.1	95.8	92.1	84.5
4-4	Wenyiil	90.9	2.5	94.0	93.6	91.0	82.7
5-6	Claude-4 ?	83.2	4.5	90.7	88.7	73.7	81.2
5-6	GPT-4.1 ?	82.8	5.5	76.2	92.5	84.8	73.1
7-8	TranssionTranslate ?	79.5	4.3	88.6	82.1	72.0	76.9
7-10	DeepSeek-V3 ?	77.3	5.1	85.3	82.3	71.7	69.0
8-10	Llama-4-Maverick	76.4	6.5	71.9	78.1	76.8	78.6
8-10	NLLB	75.6	6.6	77.1	78.3	74.8	71.0
11-12	CommandA	72.6	6.5	84.6	73.0	67.8	65.2
11-12	Yolu	72.4	5.7	79.1	70.6	69.8	71.4
13-14	TranssionMT	70.1	6.2	53.7	76.0	74.4	74.2
13-15	COILD-BHO	68.7	8.9	75.9	83.3	71.6	32.7
14-15	ONLINE-B	67.2	4.1	62.0	70.5	62.6	76.0
16-16	IRB-MT	59.6	11.4	62.7	74.2	52.2	45.8
17-17	Gemma-3-27B ?	56.0	8.3	42.5	47.2	65.7	69.9
18-18	SalamandraTA	35.7	12.1	27.6	35.1	44.8	31.7
19-19	Shy-hunyuan-MT	1.7	11.5	0.0	3.0	1.7	1.9
20-37	17 systems not human-evaluated		...				

English→Masai							
Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Gemini-2.5-Pro ?	9.8	6.1	17.5	7.8	6.3	8.7
2-2	Human	9.6		16.5	4.6	10.9	7.9
3-3	Claude-4 ?	7.7	2.6	17.0	3.5	4.4	8.2
4-6	AyaExpanse-8B	6.0	8.2	13.8	2.8	5.4	2.6
4-5	Llama-4-Maverick	5.2	3.2	2.5	8.8	3.8	4.4
6-6	Shy-hunyuan-MT	4.8	1.0	12.7	2.9	1.7	2.7
7-13	AyaExpanse-32B	3.1	7.1	0.0	2.5	5.7	4.1
4-8	DeepSeek-V3 ?	3.0	6.2	1.7	4.0	1.8	4.8
9-13	Llama-3.1-8B	3.0	8.1	0.1	1.4	8.1	2.1
9-13	Gemma-3-12B ?	3.0	8.8	0.0	1.7	6.6	3.9
9-13	Qwen2.5-7B ?	2.8	8.6	0.1	2.6	5.1	3.1
9-13	Qwen3-235B	2.7	3.0	0.2	0.8	5.7	5.3
9-13	TranssionMT	2.5	5.9	0.9	1.6	3.2	5.4
14-18	CommandR7B	1.6	4.3	0.1	0.0	3.4	3.9
14-18	CommandA-WMT	1.5	6.4	0.0	3.9	0.0	1.5
14-16	CommandA	1.3	7.9	0.1	0.1	3.0	2.7
17-18	TowerPlus-9B[M]	0.8	5.3	0.0	1.2	0.1	2.1
17-18	EuroLLM-9B[M]	0.7	8.2	0.0	0.3	1.6	1.2
19-19	EuroLLM-22B-pre.[M]	0.5	8.2	0.0	1.1	0.0	0.6
20-29	9 systems not human-evaluated		...				

English→Russian							
Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Gemini-2.5-Pro	83.4	4.4	91.4	82.5	83.7	74.8
2-2	Shy-hunyuan-MT	80.2	1.0	89.8	83.4	78.5	67.5
3-5	Wenyiil	78.2	4.8	90.7	82.4	70.3	72.2
3-5	GPT-4.1	76.2	5.4	85.9	73.3	75.3	70.2
3-5	Claude-4	75.9	8.7	94.8	73.5	71.2	65.6
6-9	DeepSeek-V3 ?	73.6	5.7	89.0	66.5	74.8	63.0
5-8	Algharb	73.3	5.2	62.8	83.6	70.2	77.0
6-9	CommandA-WMT	73.2	4.2	92.3	72.9	71.3	54.8
8-10	Yandex	72.0	4.5	90.0	77.9	63.8	58.0
9-11	Human	70.5		91.5	65.9	71.8	49.5
10-12	UvA-MT	69.1	4.5	79.0	75.9	63.3	58.6
11-14	Qwen3-235B	67.6	8.8	74.5	67.0	68.6	58.5
12-15	IRB-MT	65.4	10.1	77.2	65.5	63.7	54.3
12-15	Yolu	64.5	6.9	80.9	63.4	63.6	48.0
13-16	GemTrans	62.5	5.1	51.5	79.9	59.4	56.5
15-16	Gemma-3-27B	61.7	8.9	74.5	56.5	60.4	56.3
17-19	RuZh ?	57.9	9.6	54.4	58.3	65.0	48.1
17-19	SRPOL	56.9	10.6	75.4	59.8	53.0	38.2
17-19	Lanigo	56.2	8.8	58.3	56.6	57.8	50.1
20-42	22 systems not human-evaluated		...				

English→Ukrainian							
Rank	System	Human	AutoRank	literary	news	social	speech
1-3	Gemini-2.5-Pro	90.3	3.3	93.8	90.5	90.2	86.2
1-3	Algharb	90.0	4.2	91.5	91.2	89.6	87.2
1-3	Wenyiil	89.5	3.5	91.9	90.9	89.9	84.1
4-5	Shy-hunyuan-MT	88.4	1.0	90.8	90.2	89.0	82.0
4-5	GemTrans	88.2	4.6	89.9	90.8	88.2	82.4
6-7	GPT-4.1	87.9	3.5	90.3	88.9	88.5	82.7
5-8	Human	87.3		95.2	85.3	86.3	82.7
7-9	UvA-MT	86.4	4.4	86.0	88.0	87.9	81.5
8-13	CommandA-WMT	86.3	3.9	87.1	87.1	86.4	84.0
9-13	Llama-4-Maverick	86.2	8.8	91.2	86.0	87.2	78.8
9-13	DeepSeek-V3?	85.8	5.0	87.4	88.0	85.0	82.2
9-14	Claude-4?	85.6	7.0	87.3	85.3	86.5	81.9
9-13	Yolu	85.4	6.0	88.0	88.3	87.7	73.8
14-16	Mistral-Medium?	84.5	6.0	85.3	86.0	84.1	82.5
14-16	TowerPlus-9B[M]	84.2	8.8	86.3	86.4	84.7	77.6
14-16	CommandA	84.0	7.4	84.4	87.4	83.3	79.7
17-17	IRB-MT	82.9	8.2	83.8	87.1	83.4	74.8
18-19	SRPOL	79.9	8.4	76.5	83.1	84.3	71.1
18-19	Laniquo	79.8	7.7	81.2	82.2	82.0	70.6
20-44	24 systems not human-evaluated		...				

English→Italian							
Rank	System	Human	AutoRank	literary	news	social	speech
1-4	Gemini-2.5-Pro	79.4	4.4	74.4	86.1	80.6	71.9
1-4	GemTrans	79.4	5.2	85.8	79.0	81.7	68.0
1-4	GPT-4.1	79.0	4.5	87.0	73.9	83.3	69.3
1-4	Shy-hunyuan-MT	78.7	1.0	74.4	80.4	83.4	71.8
5-7	CommandA-WMT	75.5	2.6	77.9	79.4	77.0	63.3
5-8	Mistral-Medium?	73.8	7.1	79.1	67.8	79.9	65.4
5-10	CommandA	73.2	8.4	82.3	80.2	67.4	62.6
6-10	Claude-4	72.1	8.4	73.9	75.5	70.6	67.7
7-10	UvA-MT	71.8	5.3	68.4	74.1	77.5	60.7
7-10	DeepSeek-V3?	71.7	6.1	63.6	75.7	73.8	69.9
11-11	Qwen3-235B	67.0	7.2	60.8	71.3	71.8	57.4
12-13	TowerPlus-9B[M]	61.2	11.3	71.6	62.6	57.8	53.5
12-13	IRB-MT	60.3	10.2	53.7	67.1	62.1	53.2
14-16	SalamandraTA	57.5	10.3	45.5	69.9	62.2	41.6
14-16	AyaExpanse-8B	57.0	14.9	50.8	65.8	60.5	42.9
14-16	EuroLLM-9B[M]	56.6	15.2	58.4	57.1	57.1	52.7
17-18	Gemma-3-12B	53.6	15.5	25.5	59.2	64.3	55.9
17-18	Laniquo	53.4	7.6	37.0	61.2	57.4	51.5
19-34	15 systems not human-evaluated		...				

English→Icelandic							
Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Human	87.5		87.4	88.4	86.8	87.3
2-2	Gemini-2.5-Pro	77.6	1.8	79.8	68.8	83.1	77.9
3-4	Erlendur	68.3	2.2	69.4	61.2	72.0	71.2
3-4	GPT-4.1	68.0	1.9	74.7	67.4	63.6	69.1
5-5	Shy-hunyuan-MT	63.2	1.0	51.3	67.0	66.6	65.2
6-6	TowerPlus-9B[M]	57.4	3.9	46.0	57.1	65.5	56.3
7-7	ONLINE-B	51.8	4.4	43.4	45.6	59.3	57.3
8-10	Claude-4?	47.8	5.2	43.0	48.9	45.4	56.3
8-10	TowerPlus-72B[M]	46.3	5.7	39.5	39.2	52.1	54.5
8-10	TranssionTranslate?	46.2	5.8	29.1	45.5	52.6	55.6
11-11	AMI	39.9	7.4	47.8	35.5	39.8	37.5
12-12	GemTrans	34.8	7.0	25.0	32.4	39.5	41.4
13-14	SalamandraTA	31.3	8.6	28.0	23.9	33.9	41.6
13-15	UvA-MT	30.6	6.8	23.8	23.7	37.3	36.9
14-15	CommandA-WMT	29.0	6.8	9.6	36.0	31.6	36.5
16-16	NLLB	24.1	15.2	22.8	21.1	25.1	28.2
17-17	IRB-MT	20.7	11.9	6.2	21.2	24.7	29.5
18-18	Gemma-3-12B	16.5	13.8	8.4	12.8	19.1	26.6
19-19	Llama-3.1-8B	10.5	24.9	10.5	4.4	13.4	14.4
20-35	15 systems not human-evaluated		...				

English→Serbian (Cyrilic)

Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Gemini-2.5-Pro	94.2	3.0	97.3	92.8	96.0	89.2
2-3	GPT-4.1	92.5	3.4	98.6	90.5	91.9	89.5
2-4	Shy-hunyuan-MT	92.2	1.0	94.4	90.0	94.3	88.8
3-4	ONLINE-B	90.6	6.1	97.7	90.9	90.6	81.1
5-5	Claude-4 ?	90.0	6.8	96.1	86.4	93.2	81.8
6-6	Human	88.7		83.8	93.5	88.4	86.9
7-7	TranssionTranslate ?	85.1	8.0	88.7	87.7	86.6	73.2
8-9	GemTrans	81.5	4.6	88.3	78.8	79.7	81.6
8-9	DeepSeek-V3 ?	78.7	8.6	89.8	87.0	61.7	84.5
10-11	IRB-MT	77.6	9.9	81.7	80.4	77.3	68.4
10-15	DLUT_GTCOM	77.2	9.3	72.3	80.0	78.4	75.9
11-14	CommandA-WMT ☹	76.5	7.0	59.1	78.8	84.8	77.9
10-15	UvA-MT	76.2	5.8	63.4	77.9	83.0	75.4
11-15	SalamandraTA	75.5	8.8	62.0	82.8	78.2	73.9
13-15	Gemma-3-12B	74.8	12.1	70.1	71.6	81.9	72.2
16-17	CUNI-SFT	60.9	13.5	65.6	58.5	61.8	57.4
16-17	Llama-3.1-8B ☹	58.4	19.4	53.7	63.6	60.8	50.3
18-18	NLLB	53.5	19.8	41.6	60.5	59.3	44.1
19-19	EuroLLM-9B[M] ☹	41.8	22.3	73.4	30.3	41.6	22.9
20-34	14 systems not human-evaluated		...				

Czech→German

Rank	System	Human	AutoRank	dialogue	edu	news	social	speech
1-1	Gemini-2.5-Pro	90.7	2.5	94.5	92.9	86.2	94.0	89.8
2-4	GPT-4.1	89.5	2.4	90.2	90.2	86.6	95.7	85.9
2-4	Claude-4	88.8	4.8	92.1	88.4	86.5	93.1	85.8
2-6	DeepSeek-V3 ?	88.1	3.5	92.3	91.5	85.0	93.2	81.0
4-7	Shy-hunyuan-MT	87.2	1.0	89.9	81.7	87.6	92.4	84.7
4-8	Mistral-Medium	87.0	4.2	91.8	86.6	87.0	90.1	80.7
5-7	CommandA	86.8	4.8	91.0	83.1	85.1	93.6	83.1
8-8	CommandA-WMT	85.6	2.1	87.6	86.0	83.8	91.6	80.0
9-12	Human	82.8		93.6	88.1	75.5	81.1	84.1
9-13	GemTrans	82.6	6.3	87.1	83.6	79.9	87.7	77.3
9-13	Gemma-3-27B	82.0	7.2	86.7	85.4	74.6	87.9	80.8
9-13	Wenyiil	82.0	10.9	88.2	72.3	86.9	86.6	74.9
10-15	Algharb	80.9	13.2	90.5	72.0	88.2	81.6	71.1
13-15	TowerPlus-9B[M]	79.8	10.3	81.2	81.5	74.9	89.6	73.9
13-15	UvA-MT	79.5	7.0	94.6	69.0	73.0	89.9	79.8
16-19	CUNI-MH-v2	77.2	14.2	77.1	73.0	73.8	87.9	75.6
16-18	Gemma-3-12B	76.8	11.5	76.2	69.0	75.5	89.0	74.2
16-18	SRPOL	76.7	11.0	79.7	69.1	73.8	90.8	71.9
19-19	Yolu	75.3	9.3	91.5	63.3	71.3	85.2	72.9
20-21	IRB-MT	71.7	12.4	63.0	70.9	65.2	86.5	72.3
20-21	Laniqo	70.0	10.3	76.3	70.0	66.7	74.4	66.0
22-42	20 systems not human-evaluated		...					

English→Czech

Rank	System	Human	AutoRank	dialogue	literary	news	social	speech
1-1	Gemini-2.5-Pro	88.7	3.4	91.4	96.1	86.5	84.4	87.6
2-2	Shy-hunyuan-MT	87.1	1.0	88.7	94.1	89.8	81.6	80.7
3-4	DeepSeek-V3 ?	85.1	5.1	91.0	90.4	85.6	84.0	75.0
3-4	Human	84.5		86.4	88.3	84.0	84.1	80.0
5-6	CommandA-WMT	82.6	3.6	90.1	83.5	84.1	82.7	72.8
5-6	Wenyiil	82.4	4.5	82.9	81.2	83.6	82.8	81.1
7-9	GPT-4.1	80.8	4.0	91.3	70.6	80.7	84.2	81.0
7-9	Mistral-Medium ?	80.4	7.1	86.6	88.1	78.7	77.4	74.0
7-10	Claude-4 ?	79.6	9.0	86.5	85.5	78.9	75.0	75.8
9-11	UvA-MT	78.6	6.5	85.6	86.4	70.6	84.2	68.7
10-14	Algharb	76.7	6.4	85.1	50.7	84.9	81.9	81.4
11-14	CommandA	76.4	8.8	88.1	75.6	77.9	73.2	71.4
11-15	Yolu	75.6	6.3	82.3	83.3	73.1	76.0	64.8
11-15	Gemma-3-27B	75.6	9.2	82.9	85.1	72.3	72.8	68.3
13-15	GemTrans	73.2	5.1	87.5	55.3	79.1	75.6	72.0
16-16	CUNI-MH-v2	71.0	12.1	75.7	77.4	76.1	65.7	58.8
17-18	SRPOL	67.5	8.7	74.9	67.7	75.3	58.9	61.5
17-19	Laniqo	66.1	8.8	51.1	79.6	67.7	64.3	59.1
18-19	TowerPlus-9B[M]	65.8	11.6	74.4	58.4	70.6	66.5	59.4
20-20	SalamandraTA	60.3	10.5	57.0	62.0	70.0	52.5	55.7
21-44	23 systems not human-evaluated		...					

Rank	System	English→Chinese					
		Human	AutoRank	literary	news	social	speech
1-1	Algharb	88.4	4.2	95.0	87.7	88.4	81.9
2-4	Shy-hunyuan-MT	88.2	1.0	93.2	84.5	92.4	80.1
2-5	Claude-4	86.9	7.2	98.2	86.3	84.0	79.7
2-5	Wenyiil	86.3	4.0	89.5	80.3	91.4	82.0
3-6	DeepSeek-V3	85.0	7.3	94.5	83.8	82.5	80.1
5-10	GemTrans	84.4	5.0	94.2	80.7	85.3	76.7
6-11	Qwen3-235B	84.0	4.9	88.2	85.5	85.5	74.3
5-10	GPT-4.1	84.0	4.7	98.3	80.9	79.5	80.2
6-11	Gemini-2.5-Pro	83.8	4.0	82.1	83.2	85.5	83.7
5-10	UvA-MT	83.4	6.4	96.7	78.0	84.3	74.4
11-13	Human	82.1		92.8	74.2	83.7	78.3
11-15	CommandA-WMT	81.3	5.7	82.0	86.8	80.0	75.0
11-15	Llama-4-Maverick	80.7	8.1	83.9	81.1	82.3	73.5
12-16	Mistral-Medium?	79.9	5.0	78.0	83.2	78.7	79.7
12-16	Yolu	79.0	4.9	84.5	82.9	76.9	71.1
14-17	SRPOL	77.7	10.5	68.8	79.4	85.7	70.8
16-18	IRB-MT	76.5	9.5	90.3	70.6	77.5	67.4
17-18	RuZh?	75.7	10.6	84.1	73.2	77.1	66.9
19-19	Laniko	70.5	9.3	83.0	72.4	63.6	65.7
20-40	20 systems not human-evaluated		...				

Rank	System	English→Japanese					
		Human	AutoRank	literary	news	social	speech
1-1	Human	89.2		94.5	85.2	92.1	84.2
2-4	Gemini-2.5-Pro	85.8	2.5	87.3	82.5	87.7	86.0
2-6	Algharb	85.7	3.3	84.3	88.9	83.8	85.6
2-5	Mistral-Medium?	84.8	5.5	98.4	77.1	83.3	82.6
3-6	Wenyiil	84.4	3.0	88.6	80.9	85.6	82.5
5-6	GPT-4.1	83.7	2.9	95.4	77.0	80.7	84.9
7-7	CommandA-WMT	82.2	3.7	83.3	85.2	78.0	83.1
8-12	Shy-hunyuan-MT	79.9	1.0	75.6	78.2	81.8	84.3
8-13	DeepSeek-V3?	79.3	4.7	82.9	80.0	74.1	82.7
8-13	Claude-4	79.3	5.8	86.5	76.1	72.8	86.3
8-13	UvA-MT	79.3	6.5	74.9	79.7	81.7	80.1
8-14	ONLINE-B	78.0	6.3	82.5	78.1	76.3	75.4
9-16	In2x?	77.8	2.3	60.8	83.6	81.9	82.7
12-16	GemTrans	76.2	5.6	81.0	66.9	80.9	76.8
13-16	KIKIS	76.2	3.2	66.6	78.5	79.2	79.1
13-16	Systran	75.6	7.5	69.2	84.5	75.9	69.5
17-18	NTTSU	73.3	8.1	75.3	77.9	71.9	66.5
17-18	Yolu	72.6	6.1	72.0	76.4	70.7	71.0
19-19	Laniko	67.8	9.5	49.0	72.0	81.4	61.6
20-45	25 systems not human-evaluated		...				

Rank	System	Czech→Ukrainian					
		Human	AutoRank	edu	news	social	speech
1-2	Gemini-2.5-Pro	92.9	1.1	96.8	93.4	91.6	89.4
1-3	GPT-4.1	92.1	1.3	94.0	92.3	92.9	88.9
2-3	Shy-hunyuan-MT	91.8	1.0	91.7	94.7	90.1	89.0
4-8	GemTrans	90.2	4.4	92.9	91.0	89.5	86.8
4-6	Human	90.1		93.0	92.6	85.5	88.0
4-10	Mistral-Medium?	89.4	4.2	91.1	91.7	88.7	84.6
6-10	Claude-4?	89.1	3.7	91.4	92.4	88.7	81.3
4-10	DeepSeek-V3?	89.0	3.2	90.7	91.0	88.2	84.8
6-10	CommandA-WMT	88.7	1.3	87.3	89.6	91.2	85.7
6-10	Gemma-3-27B	88.6	5.0	89.1	91.3	88.5	83.7
11-12	CommandA	86.4	4.6	86.1	86.6	89.6	83.0
11-13	Wenyiil	85.7	5.4	72.9	93.6	89.6	81.3
12-15	TowerPlus-9B[M]	85.3	7.9	85.0	87.9	88.1	78.2
13-16	Algharb	84.1	7.2	74.7	93.8	87.2	73.9
13-17	UvA-MT	83.5	5.1	75.3	86.0	87.4	83.5
14-17	Laniko	83.4	7.7	79.6	89.3	84.7	75.7
15-17	IRB-MT	82.7	9.1	77.2	86.7	84.4	79.5
18-19	SRPOL	80.8	7.8	74.3	88.4	80.9	74.6
18-19	Yolu	80.1	6.0	66.4	88.6	82.6	77.2
20-44	24 systems not human-evaluated		...				

Rank	System	Japanese→Chinese					
		Human	AutoRank	literary	news	social	speech
1-1	Human	-3.5		-3.6	-3.8	-3.0	-3.3
2-2	Gemini-2.5-Pro	-4.4	3.3	-3.9	-5.1	-2.2	-6.8
3-6	Algharb	-5.8	4.3	-6.5	-4.6	-5.1	-7.5
3-7	Claude-4	-5.9	6.4	-4.6	-4.6	-5.3	-11.5
3-7	Shy-hunyuan-MT	-6.1	1.0	-5.2	-5.4	-4.5	-11.1
3-7	GPT-4.1	-6.2	4.5	-4.5	-7.1	-4.7	-9.9
4-7	Wenyiil	-6.9	4.5	-6.4	-6.5	-5.4	-10.5
8-10	CommandA-WMT	-7.7	5.2	-7.1	-6.3	-4.5	-15.7
8-10	DeepSeek-V3	-8.1	6.5	-8.9	-5.9	-4.0	-16.3
8-13	Kaze-MT	-8.6	3.9	-8.1	-8.4	-6.0	-13.1
10-13	Mistral-Medium 	-10.0	6.6	-12.2	-7.3	-6.4	-15.8
10-13	In2x 	-10.0	3.0	-9.2	-10.4	-7.9	-13.8
10-13	Qwen3-235B	-10.9	7.6	-14.3	-7.5	-5.7	-17.9
14-15	GemTrans	-10.9	6.6	-11.0	-9.1	-8.4	-17.5
14-15	NTTSU	-11.3	5.9	-10.5	-9.4	-6.3	-22.8
16-17	Yolu	-12.6	7.1	-14.2	-7.6	-9.1	-23.8
16-17	TowerPlus-9B[M]	-13.3	11.5	-12.4	-9.4	-8.2	-29.3
18-18	IRB-MT	-13.9	12.4	-16.2	-10.8	-11.6	-18.9
19-19	Laniquo	-18.3	11.3	-20.4	-14.6	-14.6	-26.3
20-42	22 systems not human-evaluated		...				

Rank	System	English→Korean					
		Human	AutoRank	literary	news	social	speech
1-3	Human	-1.9		-2.4	-1.7	-1.7	-1.4
1-3	Shy-hunyuan-MT	-2.5	1.0	-3.0	-2.2	-1.0	-2.4
1-3	Gemini-2.5-Pro	-2.7	2.5	-3.5	-3.8	-0.7	-1.5
4-6	GPT-4.1	-3.3	2.9	-4.2	-3.7	-1.6	-2.1
4-7	Claude-4	-3.4	4.4	-3.1	-5.8	-2.2	-2.7
4-7	DeepSeek-V3 	-3.8	5.1	-4.1	-4.5	-3.2	-2.8
5-10	GemTrans	-4.1	5.0	-4.1	-8.0	-1.7	-2.2
7-12	CommandA-WMT	-4.3	2.9	-4.3	-5.5	-0.7	-4.7
5-12	Wenyiil	-4.3	3.0	-6.2	-4.5	-1.1	-2.4
5-12	Algharb	-4.4	3.1	-6.3	-5.1	-1.6	-1.6
8-15	Mistral-Medium 	-4.7	6.1	-5.6	-6.1	-1.8	-3.2
7-15	CommandA	-4.7	6.0	-3.9	-7.6	-2.2	-4.9
11-16	UvA-MT	-5.2	4.3	-5.5	-8.4	-1.2	-3.7
11-16	Qwen3-235B	-5.5	6.5	-6.3	-7.2	-1.9	-4.2
11-16	IRB-MT	-5.6	8.6	-6.3	-8.1	-3.2	-3.5
13-16	Gemma-3-12B	-5.9	9.2	-5.9	-8.4	-2.9	-4.9
17-18	TowerPlus-9B[M]	-7.2	10.1	-7.4	-8.2	-2.9	-7.8
17-18	Yolu	-7.3	7.0	-7.3	-11.3	-2.3	-6.3
19-19	Laniquo	-9.1	9.2	-10.6	-12.6	-3.6	-5.9
20-37	17 systems not human-evaluated		...				

E Analysis of Human Evaluation Scores

Figure 4 shows the correlation between ranks obtained from human evaluation ranks and automatic evaluation (AUTORANK) for each system. Figure 5 shows the distribution of human evaluation ranks across all systems. Finally, Figure 6 and Figure 7 break down the distribution of average human evaluation scores by language pair and by domain, respectively.

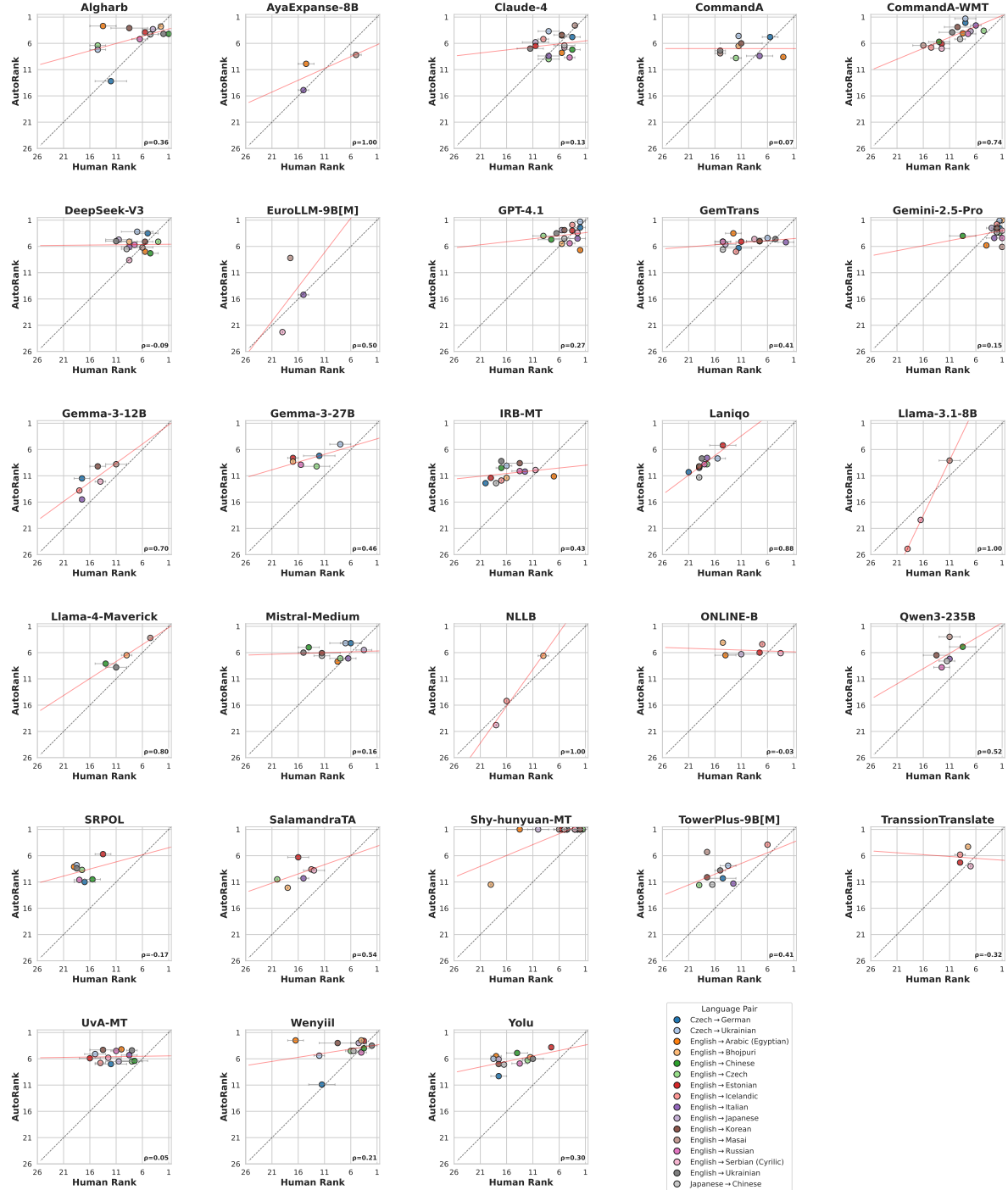


Figure 4: Correlation between automatic (AUTORANK) and human evaluation ranks by model (lower=better). Whiskers indicate the range of human ranks. The gray diagonal represents perfect correlation; points above this line mean AUTORANK ranked a model higher than humans, and vice versa. Colors denote language pairs, and the red line shows the Spearman correlation (ρ).

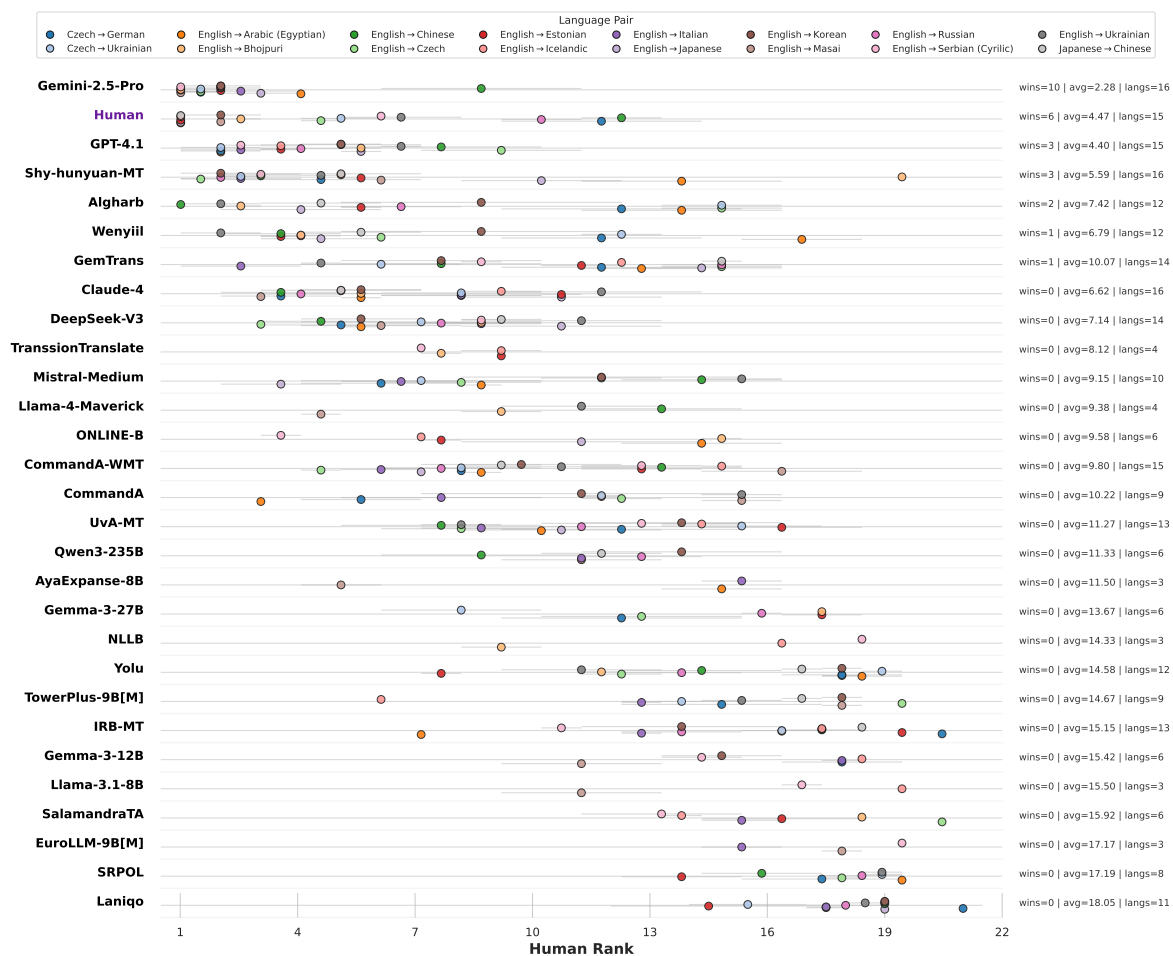


Figure 5: Distribution of ranks from human evaluation for each system, with whiskers indicating the assigned ranges. Systems are sorted by the number of “wins” (which refers to the situation when a system is being ranked first or has a rank range that includes the first position) and then by average rank.

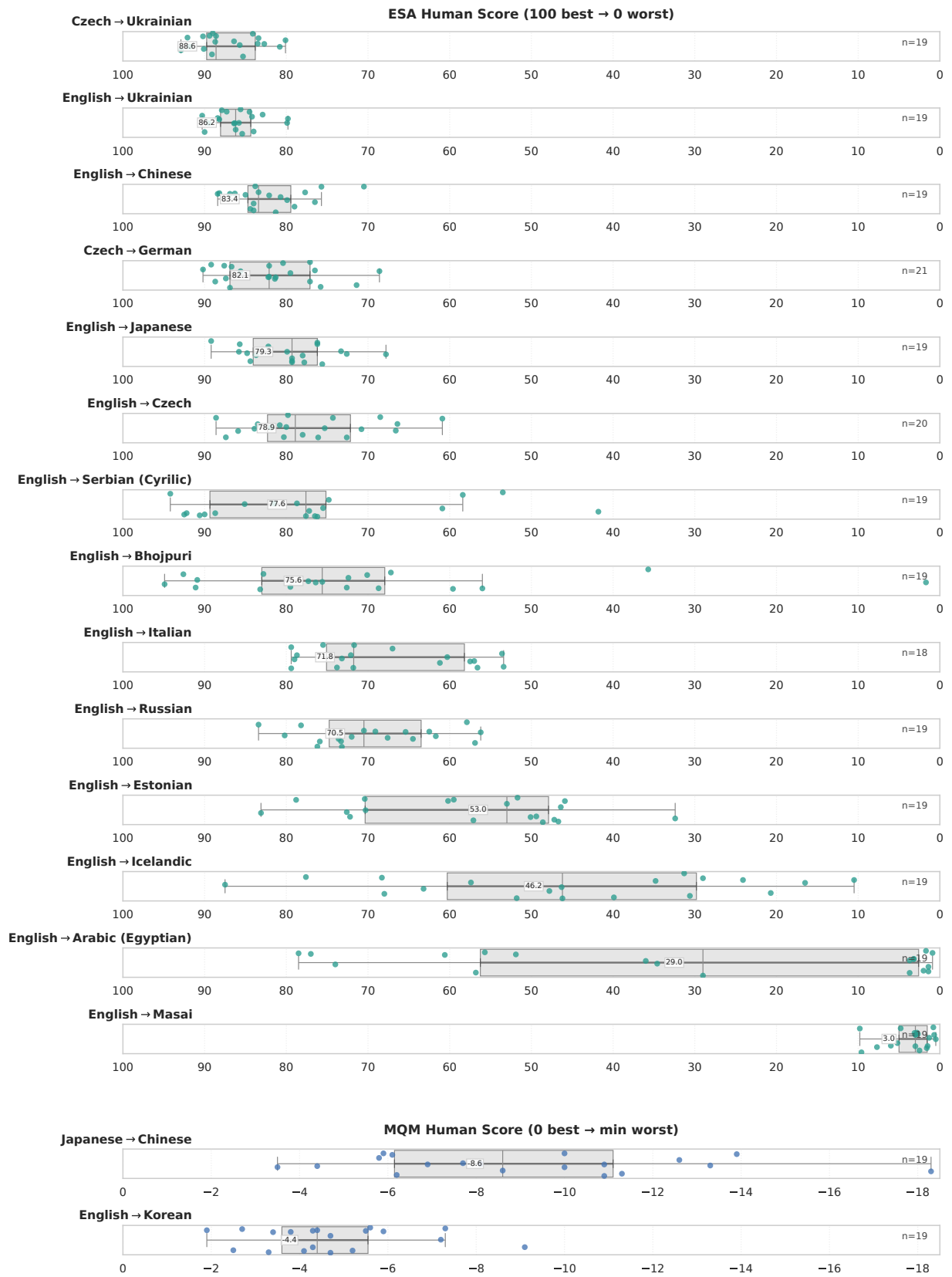


Figure 6: The distribution of human evaluation scores for each language pair. Pairs are grouped by their evaluation protocol, with ESA at the top and MQM at the bottom.

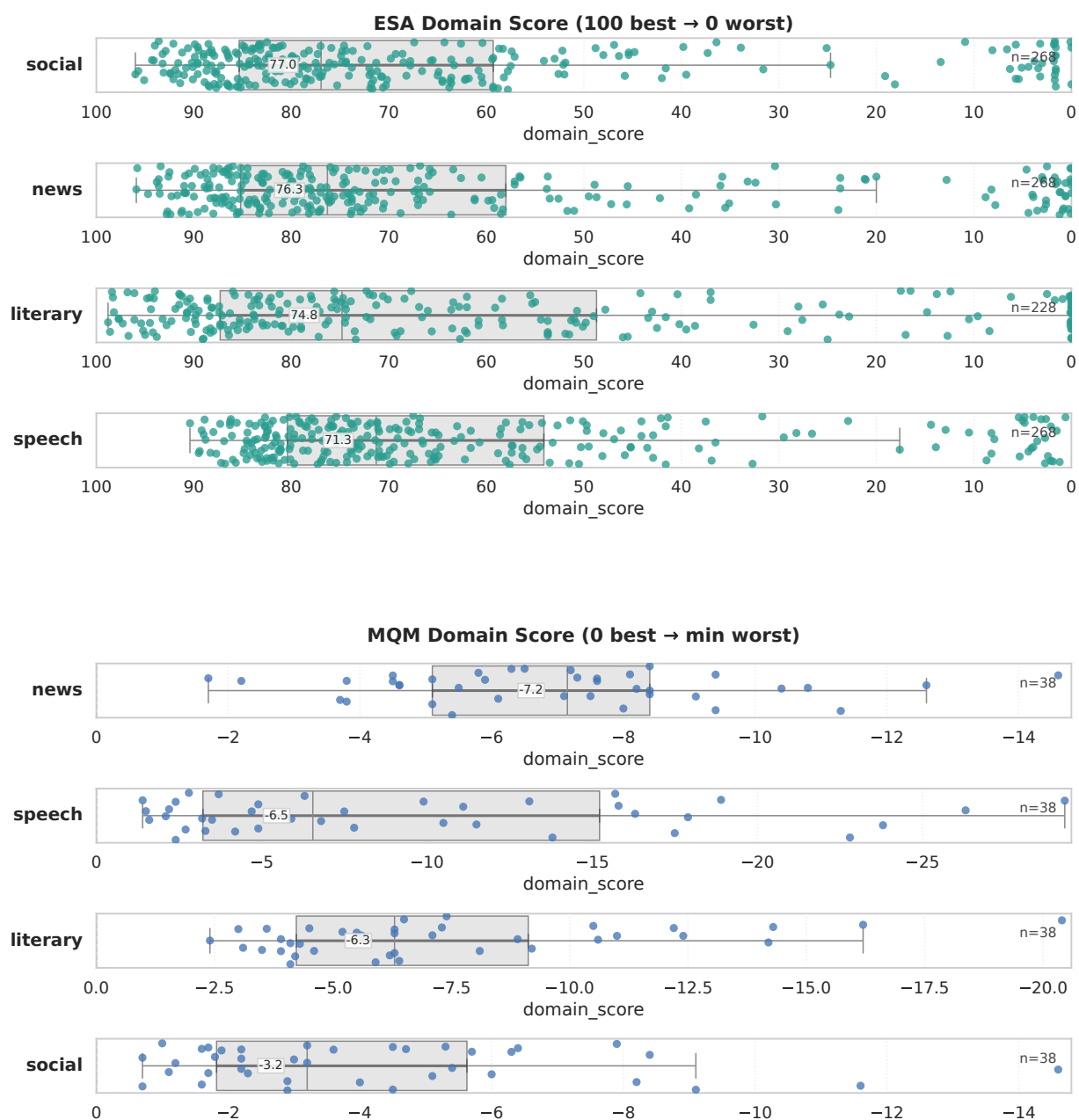


Figure 7: Distribution of scores by domain for four main domains. ESA scores are presented at the top, while MQM scores are presented at the bottom.

F Dataset Statistics

Statistics for the parallel training data provided for the shared task are shown in Tables 18 and 19.

Dataset	Segments	Tokens		Characters	
		Source	Target	Source	Target
Bhojpuri→English	Segs	Bhojpuri	English	Bhojpuri	English
OPUS	2.43M	23.07M	19.10M	270.18M	105.16M
Czech→German	Segs	Czech	German	Czech	German
OPUS	136.54M	1.47B	1.61B	10.65B	11.42B
LinguaTools-wikititles-2014	2.39M	4.65M	4.28M	40.00M	40.23M
Tilde	2.04M	36.30M	38.02M	288.88M	307.45M
Facebook-wikimatrix-1	1.60M	20.74M	22.47M	151.14M	162.61M
Statmt-news_commentary-18.1	244.83k	4.82M	5.45M	37.02M	41.07M
(Total)	142.82M	1.54B	1.68B	11.16B	11.97B
Czech→Ukrainian	Segs	Czech	Ukrainian	Czech	Ukrainian
OPUS	17.15M	138.66M	137.78M	0.97B	1.65B
Facebook-wikimatrix-1	848.96k	10.43M	10.07M	75.97M	127.31M
ELRC	130.00k	2.48M	2.56M	19.61M	35.26M
(Total)	18.13M	151.57M	150.41M	1.07B	1.81B
English→Arabic	Segs	English	Arabic	English	Arabic
OPUS	304.22M	4.65B	4.21B	28.48B	44.10B
Statmt-ccaligned-1	25.31M	355.78M	343.52M	2.27B	3.58B
LinguaTools-wikititles-2014	4.82M	11.15M	10.91M	84.51M	129.17M
Facebook-wikimatrix-1	1.97M	38.55M	35.77M	242.74M	376.25M
Statmt-tedtalks-2_clean	341.89k	6.17M	5.41M	34.54M	54.49M
Statmt-news_commentary-18.1	193.67k	8.94M	11.70M	57.33M	127.15M
(Total)	336.86M	5.07B	4.61B	31.17B	48.37B
English→Czech	Segs	English	Czech	English	Czech
OPUS	237.54M	2.85B	2.48B	17.02B	17.82B
ParaCrawl-paracrawl-9	50.63M	692.12M	626.34M	4.33B	4.68B
Statmt-ccaligned-1	12.73M	148.71M	135.81M	936.99M	1.01B
LinguaTools-wikititles-2014	4.81M	11.36M	9.67M	83.77M	81.29M
Facebook-wikimatrix-1	2.09M	33.56M	29.66M	206.82M	216.62M
Tilde	2.09M	42.26M	38.26M	276.52M	303.75M
ELRC	1.96M	37.18M	33.00M	243.79M	262.52M
EU	1.92M	34.27M	30.09M	222.84M	232.92M
Statmt-europarl-10	644.43k	15.63M	13.00M	94.31M	98.14M
Statmt-wikititles-3	410.94k	1.03M	965.62k	7.47M	7.57M
Statmt-news_commentary-18.1	265.37k	5.71M	5.19M	36.22M	39.81M
Statmt-commoncrawl_wmt13-1	161.84k	3.35M	2.93M	20.66M	20.75M
Neulab-tedtalks_train-1	103.09k	2.10M	1.77M	10.58M	10.39M
(Total)	315.37M	3.88B	3.40B	23.49B	24.78B
English→Estonian	Segs	English	Estonian	English	Estonian
OPUS	121.36M	1.83B	1.38B	11.18B	11.02B
ELRC	9.09M	201.49M	144.73M	1.29B	1.25B
ParaCrawl-paracrawl-9	8.54M	136.60M	103.32M	846.64M	840.74M
Statmt-ccaligned-1	4.11M	54.21M	43.28M	339.16M	338.17M
Tilde	2.06M	41.65M	30.28M	272.67M	271.35M
EU	2.03M	36.68M	26.85M	237.87M	231.57M
Facebook-wikimatrix-1	955.55k	15.41M	11.78M	96.18M	95.33M
Statmt-europarl-7	649.59k	15.68M	11.21M	94.64M	91.44M
Neulab-tedtalks_train-1	10.74k	215.97k	171.65k	1.09M	1.04M
(Total)	148.81M	2.33B	1.76B	14.36B	14.14B
English→Icelandic	Segs	English	Icelandic	English	Icelandic
OPUS	24.26M	292.15M	274.41M	1.70B	1.84B
ParaCrawl-paracrawl-9	2.97M	45.10M	42.66M	266.09M	292.17M
ParIce-eea_train-20.05	1.70M	26.75M	24.24M	170.36M	179.49M
Statmt-ccaligned-1	1.19M	18.63M	17.80M	115.58M	124.36M
Tilde	420.71k	6.31M	6.10M	41.71M	45.26M
ParIce-ema_train-20.05	399.09k	6.13M	5.94M	40.41M	43.90M
Facebook-wikimatrix-1	313.88k	5.66M	4.77M	34.53M	34.04M
Statmt-wikititles-3	50.18k	98.99k	88.35k	722.24k	763.33k
EU	4.72k	54.43k	52.31k	369.04k	398.50k
(Total)	31.31M	400.87M	376.06M	2.37B	2.56B

Table 18: Statistics for parallel training data provided for General/News Translation Task. Suffixes, k, M, and B, are short for thousands, millions, and billions, respectively.

Dataset	Segments	Tokens		Characters	
		Source	Target	Source	Target
English→Korean	Segs	English	Korean	English	Korean
OPUS	138.12M	1.64B	1.31B	9.84B	12.28B
Statmt-ccaligned-1	9.03M	98.69M	84.80M	635.05M	744.99M
LinguaTools-wikititles-2014	4.83M	11.62M	9.32M	84.86M	90.51M
ParaCrawl-paracrawl-1_bonus	4.00M	61.96M	48.70M	371.75M	433.95M
Facebook-wikimatrix-1	1.35M	21.63M	15.66M	135.00M	161.17M
Neulab-tedtalks_train-1	205.64k	4.29M	2.97M	21.55M	26.31M
ELRC	3.27k	67.72k	45.95k	424.80k	471.77k
(Total)	157.54M	1.84B	1.48B	11.09B	13.74B
English→Russian	Segs	English	Russian	English	Russian
OPUS	479.12M	7.32B	6.39B	44.88B	83.67B
Statmt-ccaligned-1	69.26M	0.97B	864.09M	6.18B	11.32B
Statmt-backtrans_ruen-wmt20	39.36M	746.47M	596.28M	4.47B	7.75B
LinguaTools-wikititles-2014	13.57M	33.05M	28.99M	245.88M	421.65M
ParaCrawl-paracrawl-1_bonus	5.38M	101.31M	80.41M	632.54M	1.06B
Facebook-wikimatrix-1	5.20M	86.79M	76.48M	537.73M	0.97B
Statmt-wikititles-3	1.19M	3.13M	2.88M	22.80M	39.34M
Statmt-yandex-wmt22	1.00M	21.25M	18.68M	130.99M	250.76M
Statmt-commoncrawl_wmt13-1	878.39k	18.77M	17.40M	116.16M	214.59M
Statmt-news_commentary-18.1	377.66k	8.72M	8.11M	55.68M	112.13M
Neulab-tedtalks_train-1	208.46k	4.37M	3.69M	21.96M	36.77M
ELRC	39.50k	891.98k	792.00k	5.73M	10.87M
Tilde	34.27k	752.66k	702.81k	4.83M	9.97M
(Total)	615.62M	9.31B	8.09B	57.31B	105.86B
English→Serbian	Segs	English	Serbian	English	Serbian
OPUS	127.45M	1.33B	1.17B	7.57B	9.99B
Statmt-ccaligned-1	1.99M	38.73M	34.34M	235.07M	399.09M
Facebook-wikimatrix-1	1.21M	20.95M	18.81M	129.91M	209.19M
Neulab-tedtalks_train-1	136.90k	2.79M	2.38M	14.05M	14.40M
Tilde	2.02k	46.81k	45.16k	303.95k	491.17k
ELRC	856	14.50k	13.28k	93.28k	149.56k
(Total)	130.79M	1.39B	1.22B	7.95B	10.62B
English→Ukrainian	Segs	English	Ukrainian	English	Ukrainian
OPUS	151.87M	2.68B	2.33B	16.50B	29.37B
ParaCrawl-paracrawl-1_bonus	13.35M	505.83M	487.47M	3.28B	6.04B
Statmt-ccaligned-1	8.55M	119.38M	104.10M	755.38M	1.33B
Facebook-wikimatrix-1	2.58M	41.55M	35.59M	257.56M	447.33M
ELRC	1.16M	16.65M	13.15M	110.37M	194.76M
Neulab-tedtalks_train-1	108.50k	2.25M	1.94M	11.33M	18.45M
Tilde	1.63k	36.07k	34.18k	237.96k	477.91k
(Total)	177.62M	3.36B	2.97B	20.92B	37.40B
English→Chinese	Segs	English	Chinese	English	Chinese
OPUS	221.88M	3.25B	392.85M	19.99B	17.76B
Statmt-backtrans_enzh-wmt20	19.76M	364.22M	32.72M	2.16B	1.96B
Statmt-ccaligned-1	15.18M	155.93M	42.42M	1.04B	1.13B
ParaCrawl-paracrawl-1_bonus	14.17M	217.60M	46.40M	1.34B	1.18B
LinguaTools-wikititles-2014	6.66M	16.16M	7.79M	118.50M	112.12M
Facebook-wikimatrix-1	2.60M	49.87M	5.00M	311.07M	277.84M
Statmt-wikititles-3	921.96k	2.37M	973.44k	17.82M	16.28M
Statmt-news_commentary-18.1	442.93k	9.80M	799.74k	62.67M	55.16M
Neulab-tedtalks_train-1	5.54k	95.63k	23.52k	476.98k	399.81k
ELRC	2.98k	91.23k	7.36k	591.36k	644.17k
(Total)	281.63M	4.07B	528.99M	25.05B	22.49B
Japanese→Chinese	Segs	Japanese	Chinese	Japanese	Chinese
OPUS	19.74M	46.43M	46.87M	1.44B	1.08B
KECL-paracrawl-2wmt24	4.60M	27.88M	29.51M	0.97B	704.98M
LinguaTools-wikititles-2014	1.66M	1.97M	1.97M	35.18M	27.48M
Facebook-wikimatrix-1	1.33M	2.36M	2.12M	145.10M	113.60M
KECL-paracrawl-2	83.89k	552.50k	633.77k	18.86M	14.11M
Neulab-tedtalks_train-1	5.16k	19.57k	22.30k	490.89k	375.98k
Statmt-news_commentary-18.1	1.62k	2.59k	2.17k	272.83k	197.25k
(Total)	27.42M	79.23M	81.13M	2.61B	1.94B

Table 19: Statistics for parallel training data provided for General/News Translation Task. Suffixes, k, M, and B, are short for thousands, millions, and billions, respectively.

G Analysis of English→Serbian Outputs

For the English→Serbian language direction, we tested generation of translations in both Cyrillic and Latin scripts, and we can therefore compare the use of the two scripts for each system. Table 20 shows the amount of overlap between generation in the Latin and Cyrillic scripts, measured using the word bigram F1-score (w2F).

SYSTEM	w2F
ONLINE-B (c)	99.8
ONLINE-G (c)	98.0
GemTrans	92.2
UvA-MT (l)	82.6
Claude-4	79.2
GPT-4.1	75.2
Gemini-2.5-Pro	73.7
CUNI-SFT (l)	73.5
Llama-3.1-8B (l)	73.0
EuroLLM-22B	73.0
Gemma-3-12B	71.5
Gemma-3-27B	68.0
Llama-4-Maverick	66.8
DeepSeek-V3	66.5
IRB-MT	65.7
AyaExpanses-8B	62.6
TowerPlus-9B	60.8
Qwen2.5-7B	60.2
CommandA	59.9
Shy	58.5
SalamandraTA	57.9
TranssionTranslate	57.7
Qwen3-235B	54.9
IR-MultiagentMT	54.4
CommandR7B	50.3
Mistral-7B	47.2
TowerPlus-72B	47.1
EuroLLM-9B	46.5
AyaExpanses-32B	41.2

Table 20: Content overlap between Cyrillic and Latin script translations for English→Serbian, measured with the word bigram F1-score (w2F).