Findings of the WMT24 General Machine Translation Shared Task: The LLM Era is Here but MT is Not Solved Yet

Tom Kocmi Microsoft
Eleftherios Avramidis DFKI
Rachel Bawden Inria, Paris, France
Ondřej Bojar Charles University
Anton Dvorkovich Dubformer
Christian Federmann Microsoft
Mark Fishel University of Tartu
Markus Freitag Google
Thamme Gowda Microsoft
Roman Grundkiewicz Microsoft
Barry Haddow University of Edinburgh
Marzena Karpinska UMass Amherst
Philipp Koehn Johns Hopkins University
Benjamin Marie The Kaitchup
Christof Monz University of Amsterdam
Kenton Murray JHU
Masaaki Nagata NTT
Martin Popel Charles University
Maja Popović DCU & IU
Mariya Shmatova Dubformer
Steinþór Steingrímsson The Árni Magnússon Institute
Vilém Zouhar ETH Zürich

Abstract

This overview paper presents the results of the General Machine Translation Task organised as part of the 2024 Conference on Machine Translation (WMT). In the general MT task, participants were asked to build machine translation systems for any of 11 language pairs, to be evaluated on test sets consisting of three to five different domains. In addition to participating systems, we collected translations from 8 different large language models (LLMs) and 4 online translation providers. We evaluate system outputs with professional human annotators using a new protocol called Error Span Annotations (ESA).

1 Introduction

The Ninth Conference on Machine Translation (WMT24)[1] was held at EMNLP 2024 and hosted a number of shared tasks on various aspects of machine translation (MT). This conference built on 18 previous editions as a workshop or a conference (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).

[1] www2.statmt.org/wmt24/

The goal of the General Machine Translation shared task is to explore the translation capabilities of current systems in diverse settings. We assess MT systems' ability to handle a broad range of translation and language use. How to test general MT performance is a research question in itself. Countless phenomena could be evaluated, the most important being:

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

Evaluating all phenomena is nearly impossible and creates numerous unforeseen problems. Therefore, we decided to simplify the problem and tackle only a selection of the phenomena.

We choose to evaluate different domains, this year focusing on the following ones: news, social/user-generated content, speech, literary, and educational. They were chosen to represent topics with different content styles and to be understandable for humans without specialist in-domain knowledge, thus not requiring specialized translators or human raters for evaluation. Due to limited access to monolingual data across all languages,

Proceedings of the Ninth Conference on Machine Translation, pages 1–46

the test set for each language direction contains at most four of the domains (Czech-Ukrainian uses different domains).

We evaluate a diverse set of languages pairs:

Czech→Ukrainian,
Japanese→Chinese – *new*,
English→Chinese,
English→Czech,
English→German,
English→Hindi,
English→Icelandic – *new*,
English→Japanese,
English→Russian,
English→Spanish (Latin America) – *new*,
English→Ukrainian,

We newly test an audio modality as an additional source in the speech domain. Participants in this domain were provided with audio files and automatic speech recognized (ASR) text. Submission could use the original audio as an additional cleaner source modality instead of the provided ASR text.

In contrast to previous years, we adopt the Error Span Annotation protocol (Kocmi et al., 2024b), ESA for evaluation. This protocol, described in Section 6, combines aspects of DA (Graham et al., 2013) and MQM (Lommel et al., 2014).

In a shift towards document-level evaluation, we no longer provide source texts segmented into individual sentences. Instead, we keep all paragraphs intact and evaluated together.

Finally, this year's shared task included an increased number of test suites (Section 8) under the motto "Help us break the LLMs", focusing on revealing issues in the LLM translations from different perspectives, including a range of linguistic phenomena, idiomatic expressions and proper names, complex sentence structures, multiple domains, translation isochrony, speaker-listener gender resolution, prompt injection attacks, and gender-diverse, queer-inclusive content.

All General MT task submissions, sources, references and human judgements are available in the dedicated Github repository.[2] The interactive visualization and comparison of differences between systems can be browsed online on an interactive leaderboard[3] using MT-ComparEval (Klejch et al., 2015; Sudarikov et al., 2016).

The structure of the paper is as follows. We describe the process of collecting, cleaning and translating the test sets in Section 2 followed by a summary of the permitted training data and pretrained models for the constrained track in Section 3. We list all submitted systems in Section 4. Automatic evaluation is described in Section 5. The human evaluation approach of ESA is described in Section 6. The main results can be found in Section 7 and their extended version in Appendix D. Finally, Section 8 describes the test suites and summarises their conclusions.

Findings of the WMT2024 General MT Task
Across the evaluation conditions, we observe the following:

- The best systems for English→Spanish produced close to flawless translations making it the easiest language pair (Section 6.4).
- The speech domain is the most challenging domain (likely due to the ASR) while the other three domains (news, literary, social) are similarly difficult (Section 6.4).
- Human references are in the winning cluster in 7 out of 11 language pairs. For one of the remaining 4 pairs (English→Hindi), we know the reference quality was suboptimal. This suggest that ESA protocol works well in our setting.
- ESA produced 37% more clusters than DA+SQM while using only half the number of human annotations (Section 6.5).
- The best performing system in the open and constrained system category is IOL-Research (wins 10 language pairs in this category). The best performing participating system is Unbabel-Tower70B, which wins in 8 language pairs. And the best performing system in general is Claude-3.5-Sonnet winning in 9 language pairs.
- Automatic scores are biased; although Unbabel-Tower70B placed first across all languages in automatic ranking it didn't perform as the winning system across the board of human evaluation. This is likely because we used the same metric (COMET) for automatic ranking as the system used during MBR highlighting the issue of hill-climbing on automatic metrics.
- We got a total of 28 participants, which nearly 50% more than last year. Most of the participants use an LLM as a part of their system, generally by fine-tuning it.

[2] github.com/wmt-conference/wmt24-news-systems
[3] wmt.ufal.cz

- Quality estimation metrics with fixed score for perfect translation and interpretable delta are promising for checking the quality of standalone human references.

2 Test Data

In this section, we describe the data collection process (Section 2.1), and the production of human reference translations (Section 2.3).

2.1 Collecting test data

As in previous years, the test sets consist of unseen translations created specifically for the shared task and released publicly to be used as translation benchmarks. Our aim was to collect public domain or open-licence source data covering a range of domains, and we also focused on using as recent data as possible to limit possible contamination (particularly relevant when using LLMs).

We chose four main domains from which to collect data (news, literary, speech and social), although we were not able to collect data in all domains for all three source languages (no social domain data is provided for Japanese→Chinese and Czech→Ukrainian data was collected separately, comprising news data and four other separate domains). For all language pairs, the test sets are "source-original", meaning that the text was originally written in the source language, which is then manually translated into the target languages. This is important to avoid "translationese" in the source texts, which can have a negative impact on evaluation accuracy (Toral et al., 2018; Freitag et al., 2019; Läubli et al., 2020; Graham et al., 2020). We aimed for a certain number of *tokens*[4] in each domain rather than a certain number of *sentences* (as in previous years) to better balance the domains and also because the document-level focus this year allowed avoid manual sentence splitting. We aimed for approximately 10,000 tokens per domain, adjusting this figure in cases where not all domains could be covered (this is notably the case for Japanese→Chinese, where the other domains are up-sampled to account for us not being able to provide data in the social domain). Basic statistics of each subdomain are given in Table 1.[5]

Note that by default, when languages are mentioned in this section, this refers to the source language of the texts.

News domain This domain contains data prepared in the same way as in previous years (Kocmi et al., 2023). We collected news articles from January 2024 extracted from online news sites, preserving document boundaries. We expect that news domain text will generally be of high quality.

For Japanese, the total amount of text data was determined by the number of characters since Japanese does not put spaces between words. Using the WMT23 Japanese test set and its translation into English, we found the ratio of the number of Japanese characters to English words was 2 to 1. Since the English news test set consisted of 8K words, we started making a Japanese news test set with a goal of 16K characters. After discovering that the Japanese social domain was unavailable, we set this goal to 24K characters.

Literary Domain The English source texts were manually selected from Archive of Our Own,[6] focusing on recent, high-quality stories.[7] The stories were divided into 1000-word segments, ensuring the preservation of entire paragraphs. In total, we obtained data from four stories (8K words).[8]

For the Japanese source texts, we selected five novels recently made public on Aozora Bunko,[9] a website that digitizes and publishes Japanese literary works whose copyright has expired. To maintain consistency with the English dataset, we tokenized the Japanese novels using MeCab (Kudo, 2005) and divided them into segments of up to 1000 tokens, while preserving paragraph boundaries. The final size of the Japanese literary test set was 15 chunks (22K characters).

Speech domain The speech data corpus was compiled from a diverse range of YouTube videos licensed under Creative Commons. These sources encompassed various domains, including documentaries, instructional (DIY) videos, tutorials, travel blogs, and film content. For this part of the test set, segments from 166 videos were selected and processed through automated speech recognition (ASR) systems. For the English-language source

[4]For Japanese source texts, we choose to use a certain number of characters, since words are not space-separated.

[5]Texts are sentence-segmented and tokenised using the language-specific Spacy models (Honnibal and Montani, 2017) optimised for accuracy where available. For Czech, we use the multilingual Spacy model, as a language-specific one is not available. Note that statistics, particularly for this language, are approximate.

[6]archiveofourown.org

[7]Texts were published between February and April 2024.

[8]For each, we select first two chunks of up to 1000 words.

[9]aozora.gr.jp

Language pair	News	Literary	Speech	Social	Education	Official	Personal	Voice
				#tokens				
English→*	9,268	9,601	9,611	9,829	-	-	-	-
Japanese→Chinese	14,896	14,541	11,025	-	-	-	-	-
Czech→Ukrainian	7,996	-	-	-	7,825	6,029	6,846	5,305
				#segs (% of total #segs for language pair)				
English→*	149 (14.9)	206 (20.7)	111 (11.1)	531 (53.3)	-	-	-	-
Japanese→Chinese	269 (37.3)	316 (43.8)	136 (18.9)	-	-	-	-	-
Czech→Ukrainian	175 (7.6)	-	-	-	1160 (50.1)	243 (10.5)	323 (13.9)	415 (17.9)
				#docs (#segments/doc)				
English→*	17 (8.8)	8 (25.8)	111 (1.0)	34 (15.6)	-	-	-	-
Japanese→Chinese	45 (6.0)	15 (21.1)	136 (1.0)	-	-	-	-	-
Czech→Ukrainian	23 (7.6)	-	-	-	166 (7.0)	23 (10.6)	29 (11.1)	61 (6.8)
				#sents (#sents/doc)				
English→*	333 (19.6)	607 (75.9)	685 (6.2)	759 (22.3)	-	-	-	-
Japanese→Chinese	634 (14.1)	875 (58.3)	332 (2.4)	-	-	-	-	-
Czech→Ukrainian	439 (19.1)	-	-	-	1166 (7.0)	412 (17.9)	571 (19.7)	462 (7.6)
				Type-token ratio of source texts				
English→*	0.30	0.23	0.24	0.27	-	-	-	-
Japanese→Chinese	0.22	0.20	0.19	-	-	-	-	-
Czech→Ukrainian	0.46	-	-	-	0.39	0.45	0.34	0.37

Table 1: Basic statistics concerning the subdomains of each test set. Statistics are calculated on the source side. Sentence segmentation and tokenisation are carried out automatically as described in Footnote 5.

material, we used the proprietary Dubformer engine developed in-house. Japanese-language content was processed using the Whisper ASR system (Radford et al., 2022).

For Japanese, We selected 136 segments from 56 YouTube videos. They include both monologues and dialogues, as well as a variety of speakers, both men and women, adults and children. Video content includes press conferences, interviews, cooking recipes, travel vlogs, DIY videos, tutorials, product reviews, etc. We decided the total amount of speech data based on the number of characters transcribed. We started creating the data with a target of 16K characters and eventually ended up with 18K characters.

Social domain The social domain data is sourced using the Mastodon Social API.[10] 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 the first four months of 2024, using the reported language ID label to target the source languages of interest. Unfortunately, there were too few good quality posts for Czech and Japanese, and we therefore only release social domain data for English.

Given the document-level nature of the task this year, 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 for a total of 1.5M toots scraped) 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). 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 between 5 and 76 segments of text, each segment corresponding either to 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 indi-

[10] mastodon.social/api/v1/timelines/public

cating 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.

Czech and Ukrainian source texts Source texts for Czech→Ukrainian translation included the news domain as described above, Educational domain collected from online exercises and three domains (Personal, Official and Voice) from texts collected through Charles Translator.[11] The Charles Translator mobile app supports voice input, which is converted to text using Google ASR. The texts collected this way were marked as the Voice domain. The remaining Czech inputs from the Charles Translator service were classified either as Official (formal communication) or Personal (personal communication, usually between a Czech and Ukrainian speaker).

The texts were filtered and pseudonymized in the same way as in the last two years (Kocmi et al., 2022). For example we asked the annotators not to delete or fix noisy inputs as long as they are comprehensible. The only exception was the voice domain, where the source texts were post-edited to fix ASR errors, including punctuation and casing.

The Educational domain includes selected exercises from an online portal *Škola s nadhledem*[12] for elementary-school students from various subjects (chemistry, geography, Czech language, etc.). Some segments are not full sentences but short phrases. The reference translations for this domain were created by professional translators within the EdUKate project.

2.2 Comparison between Domains

Due to the change to document-level translation this year, for each language direction, we measured the amount of text per domain by counting tokens, aiming for approximately the same number of tokens per domain (see Table 1 for statistics of the different domains). In one sense, this means that the amount of textual content is roughly balanced per domain, as opposed to taking the same number of sentences per domain, which would result in domains with longer sentence lengths (e.g. news or literary) being over-represented with respect to domains with shorter sentences (e.g. social). However, it is worth noting that the nature of documents, in terms of their length and structure, differs greatly depending on the domain. This can be exemplified at its most extreme by a comparison between the literary, social and speech domains for from-English language directions.

The literary domain has only 8 documents, each one containing a large number of segments (25.8 on average), with each segment containing an average of 75.9 sentences. A document represents an extract from a longer literary text and each segment represents a paragraph of text.

The speech domain is represented by a larger number of documents (111), each one containing a single segment, composed of an average of 6.2 sentences. Each document in this case corresponds to a short dialogue, provided without segmentation into dialogue turns.

The social domain is represented by a fair number of documents (34 in total), but the composition is very different from the other domains, as we made a choice to preserve the structure of the initial posts (new-line separated text is represented by multiple segments) and of the thread itself (separate posts are separate segments). This has the advantage of preserving post boundaries and formatting choices, but has the disadvantage of creating a large number of individual segments (531 in total, compared to 206 for the literary domain and 111 for the speech domain), each containing few sentences. This has two main consequences: (i) if segments are handled individually by systems, most sentences will be handled with little context, since the other sentences appear in separate segments, (ii) in terms of the overall number of segments evaluated in the human evaluation (see Section 6), the social domain represents over half of the total number of evaluated segments (53.3% compared to 20% for the literary domain and only 11.1% for the speech domain). This has consequences for the calculation of macro-average scores when computing human rankings, as discussed in Section 7.1. The formatting choice could be rethought for future years, although would have to take into account the particularities of non-standard text in order to not introduce extra noise (e.g. concatenating newline-separated sentences would have to take into account the potential lack of end-of-sentence punctuation, but it would also have to take into account instances where newlines are used with a single sentence for purely visualisation purposes. A possible solution would be to allow a linebreak symbol such as
 to appear in the segments.

[11] translator.cuni.cz
[12] skolasnadhledem.cz

2.3 Human References

The test sets were translated by professional translation agencies, according to the translation brief shown in Appendix C. Different partners sponsored each language pair and various translation agencies were therefore used, which could affect the differences and quality of translations.

The quality of human references is critical especially for reference-based metrics (Freitag et al., 2023), and getting high quality translations is challenging despite the use of professional translators. Therefore, we propose to use a quality estimation metric to assess the quality of translation. We need a metric whose score is interpretable in an absolute way, i.e. a metric that generates a fixed score for perfect translations (such as 0) and has an understandable delta (for example -1 means a single minor error as in MQM-based metrics). For that reason, we chose a GPT-4-based implementation of GEMBA-MQM (Kocmi and Federmann, 2023).

Table 2 shows the GEMBA scores for individual domains together with the ESA human cluster that was obtained a few months later in our official manual evaluation.

The two target languages with the lowest GEMBA scores were Russian and Hindi. The vendor providing Russian translations improved the initial quality of translations after being presented with the GEMBA scores. On the other hand, the vendor providing Hindi translators claimed that the translations were flawless.

When we compare the average GEMBA score to human rank in Table 2, we can see that human reference is ranked in the top cluster for all languages except of Hindi, Ukrainian, and Chinese. While the GEMBA score did not reflect lower quality of Ukrainian, its low score for Hindi was confirmed by ESA. This shows that using quality estimation metrics is a possible way of assessing the quality of human translations, although better approaches needs to be developed.

2.4 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*, i.e. 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 718,598 source test segments (detailed numbers can be found in Table 9).

	Literary	News	Social	Speech	Avg.	Hum.
En.→Czech	-2.4	-2.0	-1.9	-1.8	-2.03	1
En.→German$_A$	-2.1	-2.0	-2.3	-2.3	-2.18	1
En.→German$_B$	-2.7	-0.8	-1.7	-2.0	-1.80	1
En.→Spanish	-1.1	-1.6	-1.2	-1.6	-1.38	1
En.→Hindi	-3.4	-4.5	-2.5	-2.9	-3.33	3
En.→Icelandic	-2.6	-0.8	-1.9	-1.4	-1.68	1
En.→Japanese	-1.7	-1.6	-1.7	-1.7	-1.68	1
En.→Russian	-2.6	-2.8	-2.5	-2.3	-2.55	1
En.→Ukrainian	-1.8	-1.0	-2.0	-2.3	-1.78	3
En.→Chinese	-3.1	-1.7	-2.8	-2.2	-2.45	2

Table 2: GEMBA-MQM score for human references. The first four columns are scores for individual domains, the fifth column is the average. The last column is the human cluster assigned with ESA protocol. Czech→Ukrainian is not included because of different domains and source data.

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 10 and Table 11. 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), CCAligned (El-Kishky et al., 2020), UN Parallel Corpus (Ziemski et al., 2016), and language-specific corpora such as CzEng v2.0 (Kocmi et al., 2020), YandexCorpus,[13] 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. In addition, we have automated the data preparation pipeline using MTDATA (Gowda et al., 2021).[14] MTDATA downloads all the mentioned datasets, except CzEng v2.0, which required user authentication. 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.

[13] github.com/mashashma/WMT2022-data
[14] statmt.org/wmt24/mtdata

System	Language pairs	Architecture	Strategy
AIST-AIRC (Rikters and Miwa, 2024)	en→de, en→ja	dec, enc-dec, MEGA	sentence
AMI (Jasonarson et al., 2024)	en→is	enc-dec	hybrid
BJFU-LPT	cs→uk	–	–
CUNI-DocTransformer (Hrabal et al., 2024)	en→cs	enc-dec	paragraph
CUNI-Transformer (Hrabal et al., 2024)	cs→uk, en→cs	enc-dec	sentence
CUNI-DS (Semin and Bojar, 2024)	en→ru	dec	sentence
CUNI-GA (Hrabal et al., 2024)	en→cs	enc-dec	sentence
CUNI-MH (Hrabal et al., 2024)	en→cs	dec	sentence
CUNI-NL (Hrabal et al., 2024)	en→de	dec	sentence
CycleL (Dreano et al., 2024)	**All language pairs**	CycleGAN	–
CycleL2 (Dreano et al., 2024)	en→cs, en→de, en→ru, en→zh	CycleGAN	–
DLUT-GTCOM (Zong et al., 2024)	en→ja, ja→zh	enc-dec	–
Dubformer	en→de, en→es, en→is, en→ru, en→uk	–	–
HW-TSC (Wu et al., 2024)	en→zh	hybrid	sentence
IKUN (Liao et al., 2024)	**All language pairs**	dec	sentence
IKUN-C (Liao et al., 2024)	**All language pairs**	dec	sentence
IOL-Research (Zhang, 2024)	**All language pairs**	dec	paragraph
MSLC (Larkin et al., 2024)	en→de, en→es, ja→zh	enc-dec	sentence
NTTSU (Kondo et al., 2024)	en→ja, ja→zh	hybrid	paragraph
NVIDIA-NeMo	All except cs→uk, en→is and ja→zh	dec	paragraph
Occiglot (Avramidis et al., 2024)	en→de, en→es	dec	–
SCIR-MT (Li et al., 2024)	en→cs	dec	–
Team-J (Kudo et al., 2024)	en→ja, ja→zh	hybrid	hybrid
TranssionMT	All except en→ja, en→zh and ja→zh	enc-dec	–
TSU-HITs (Mynka and Mikhaylovskiy, 2024)	en→cs, en→de, en→es, en→is, en→ru	ddm	sentence
Unbabel-Tower70B (Rei et al., 2024)	**All language pairs**	dec	paragraph
UvA-MT (Tan et al., 2024)	en→ja, en→zh, ja→zh	hybrid	hybrid
YandexSubtitles (Elshin et al., 2024)	en→ru	dec	paragraph
Aya23 (Aryabumi et al., 2024)	**All language pairs**	dec	paragraph
Claude-3.5	**All language pairs**	dec	paragraph
CommandR+	**All language pairs**	dec	paragraph
GPT-4 (OpenAI, 2024)	**All language pairs**	dec	paragraph
Gemini-1.5-Pro (Team, 2024a)	All except en→is	dec	paragraph
Llama3-70B (Team, 2024b)	**All language pairs**	dec	paragraph
Mistral-Large (Jiang et al., 2023)	**All language pairs**	dec	paragraph
Phi-3-Medium (Team, 2024c)	**All language pairs**	dec	paragraph
ONLINE-A	**All language pairs**	–	–
ONLINE-B	**All language pairs**	–	–
ONLINE-G	**All language pairs**	–	–
ONLINE-W	All except en→is and en→hi	–	–

Table 3: Participating submissions in the General MT shared task. The top section covers the externally contributed submissions, the middle section lists the language models added by us and the lower section covers the online systems. Online system translations were not submitted by their respective companies but were obtained by us, and are therefore anonymized in a fashion consistent with previous editions of the task. Row coloring shows closed-track (dark gray), open-track (light gray) and constrained (white background) submissions. The Architecture column shows whether the submission used decoder-only language models (dec), sequence-to-sequence (enc-dec), hybrid between dec and enc-dec or other architectures. The Strategy column shows the approach used to handling paragraph-level test data: sentence-level training and translation (sentence), paragraph-level training and translation (paragraph), hybrid between both (hybrid). Some values are unknown (–) due to missing information or submission papers.

4 System Submissions

This year, we received a total of 105 primary submissions from 28 participants. The increase in number of participants from last year's 19 can be explained by the shift in the field and the ease with which LLMs can be fine-tuned. The increased number of primary submissions can be explained by the fact that most submissions are multilingual and therefore cover many translation directions.

In the same manner as previous years, we also collected translations from online MT systems for all language pairs. Online system outputs come from four public MT services and were anonymized as ONLINE-{A,B,G,W}, which resulted in further 42 system outputs. Finally, we added contrastive translations by 8 LLMs, which included closed commercial products (such as GPT-4) and open models (such as Llama3). This resulted in 95 more submissions, with the total number of submissions being 242.

All participating systems are listed in Table 3. Appendix B provides more detailed short descriptions of the submitted systems, as provided by the authors at submission time. Section 4.1 discusses the general trends in chosen architectures and approaches to paragraph-level translation. Section 4.2 presents details on LLM benchmark usage in the task. Section 4.3 describes the different tracks to which participants could submit outputs: constrained, open and closed track. Section 4.4 describes the submission system setup.

4.1 Architectures and Strategies

In addition to a reference to a description paper (if one was provided), the submission name and the list of language pairs covered, Table 3 includes columns for the architecture and strategy used to approach the task of paragraph-level translation. If we compare the frequency of usage of different architectures between the external participants (i.e,. excluding benchmarking LLMs and online systems), we can see that:

- 11 participants train decoder-only language models (*dec* in Table 3)
- 7 participants train encoder-decoder seq2seq transformer models (*enc-dec*)
- 4 participants use a hybrid of the decoder-only and encoder-decoder architectures (*hybrid*)
- 3 alternative architectures were used: MEGA (Ma et al., 2023) in AIST-AIRC, CycleGAN (Zhu et al., 2017) in CycleL and discrete diffusion models in TSU-HITs.

Not all description papers specified the strategy used to translate the test set paragraphs. Of those who did, 5 submissions approached it by explicitly training paragraph-level translation systems, while 11 submissions translated single sentences after sentence-splitting the paragraph. 3 submissions described a hybrid approach of, for example, translating single sentences but automatically post-editing at the paragraph level. Several papers do not mention the strategy at all. We plan to address this lack of information in future WMT editions by requesting that the information be provided at submission time.

Interestingly, the paragraph-level approach is not limited to a single architecture: for instance, the CUNI-DocTransformer team uses an encoder-decoder approach, but trains it on paragraph-level parallel data, which includes synthetic data. There are examples to the contrary: several submissions fine-tune a decoder-only language model, but apply it to translate single sentences (IKUN, AIST-AIRC, several CUNI submissions).

Finally, almost all submissions used an LLM as a part of their setup. The most common use is fine-tuning of a pretrained model, most often LLama. Other uses of LLMs are for generating or cleaning up training data with an LLM (Jasonarson et al., 2024) or using an LLM for automatic post-editing (Tan et al., 2024).

4.2 LLM Benchmark

Over the last year, many new LLMs claimed multilingual and translation capabilities. However, there is no systematic and reliable MT evaluation of the most popular LLMs using the same setup on blind test sets. We therefore decided to collect the translations of LLMs ourselves.

We design unified code for collecting the translations in an identical setup for all LLMs. We used a 3-shot approach, where three fixed examples are taken from the past WMT test sets. We set the temperature to zero to avoid introducing randomness into the process.[15]

We evaluated most of the popular LLMs, both closed-source models and those with open

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

Language model	Input tok.	Output tok.	Cost
Aya23	4.4 M	0.7 M	4.1 $
Claude-3.5	5.5 M	1.0 M	31.9 $
CommandR-plus	4.4 M	0.7 M	23.4 $
Gemini-1.5-Pro	3.9 M	0.6 M	40.3 $
GPT-4	5.9 M	1.0 M	240.4 $
Llama3-70B	5.0 M	0.7 M	5.1 $
Mistral-Large	6.0 M	1.1 M	37.0 $
Phi-3-Medium	5.9 M	1.1 M	4.5 $

Table 4: Number of input and output tokens and estimated pricing for translating the full WMT24 test set without test suites. The Gemini model refused to translate Icelandic, and the estimate is therefore lower. Pricing for the open models Aya23 and Llama3 was estimated via together.ai.

weights. Specifically, we collect translations from Aya23, Claude-3.5-Sonnet, Command R+, GPT-4, Gemini-1.5-Pro, Llama3-70B, Mistral-Large, Nvidia-NeMo and Phi-3-Medium. As most of the models do not claim multilingual capabilities for all languages covered, we looked into the original reports for these LLMs to see which languages are claimed to be supported. We check if both source and target language are mentioned or evaluated in any of their multilingual settings. We mark LLMs that do not officially claim a support for a given language with the § symbol in the tables. However, to avoid selection bias, we collect translations for all languages for all LLMs, even those not officially claimed to be supported.

We collect all translations via the API of the respective services, and all data was collected during the submission week. Table 4 shows the number of input and output tokens as segmented via the models' internal tokenizers. The estimated cost is for the whole test set without test suites. Note that the prices for more recent GPT models are significantly lower.

4.3 Constrained, Open, and Closed Tracks

We distinguish three types of MT systems participating in the shared task: constrained, open and closed systems. The main idea is to level the field for different setups. For the constrained setup, we only allow specific training data and pretrained models from a specified list. Open systems are those developed using publicly available data or models. The final group of closed systems corresponds to all other systems that are built at least partly with a non-replicable setup.

- **Constrained systems** are those using only the specifically allowed training data (see Section 3) and the following pretrained models: Llama-2-7B, Llama-2-13B, Mistral-7B, mBART, BERT, RoBERTa, XLM-RoBERTa, sBERT, LaBSE. Constrained systems may use any publicly available metric that was evaluated in past WMT Metrics shared tasks (e.g. COMET or Bleurt) and any basic linguistic tools (e.g. taggers, parsers, morphology analyzers).
- **Open systems** (marked in tables with a light gray background) are limited to using software, data and models that are freely available for research purposes, so that the subsequent work could be replicated by a research group.
- **Closed systems** (marked with dark gray) correspond to all the remaining (fully automatic) systems, with no limitations imposed on their training data (all ONLINE systems and LLMs released without binaries fit into this category).

4.4 OCELoT

We used the open-source OCELoT platform[16] to collect system submissions again this year. As in previous years, only registered and verified teams with correct contact information were allowed to submit their system outputs and each verified team was limited to 7 submissions per test set. Submissions on leaderboards with BLEU (Papineni et al., 2002) and CHRF (Popović, 2015) scores from SacreBLEU (Post, 2018) were displayed anonymously to avoid publishing rankings based on automatic scores during the submission period. Until one week after the submission period, teams could select a single primary submission per test set, specify if the primary submission followed a constrained, open or closed system setting, and submit a system description paper abstract. These were mandatory for a system submission to be included in the human evaluation campaign.

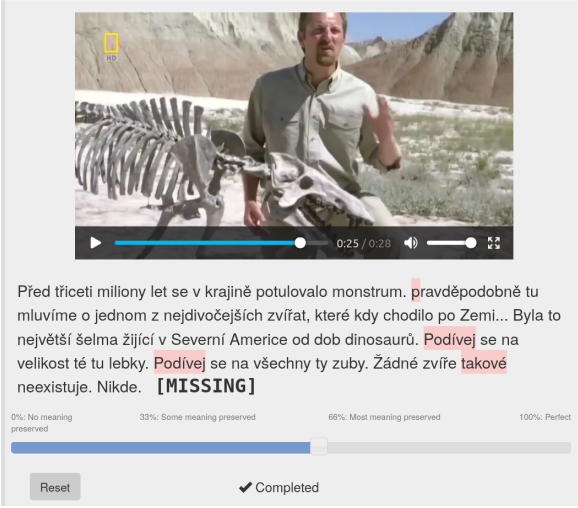
5 Automatic Evaluation

This year, we received an unusually high number of submitted systems and we were not able to provide manual evaluation for all of them. Therefore, we decided to use automatic metrics to preselect the best performing systems with a method we call AutoRank, which is based on two different metrics:

- MetricX-23-XL (Juraska et al., 2023), a reference-based metric built on top of the mT5 model (Xue, 2020).

[16] github.com/AppraiseDev/OCELoT

(a) Excerpt of two segments from a larger document. In the first segment, the name *"Kayel"* is omitted which is a major error. In the second segment, there are many minor errors.



(b) Example of a video to text translation with several minor errors. The annotator can control the video player.

Figure 1: Two screenshots of ESA (Kocmi et al., 2024b) and the annotator instructions. ESA shows multiple segments within a document at once as well as video sources. After marking the individual error spans, the annotator assigns the final segment score from 0 to 100. The tool is implemented in Appraise (Federmann, 2018).

- CometKiwi-DA-XL (Rei et al., 2023), a quality estimation metric built on the XLM-R XL model (Conneau, 2019).

Both metrics are top performing metrics (Freitag et al., 2023), and we intentionally select two distinct metrics (different underlying pretrained systems and architectures) to minimize their bias and potential problems. Although quality estimation is on average slightly worse than reference-based evaluation, it helps us to avoid a potential reference bias when human references are suboptimal (Freitag et al., 2023). Multilingual quality estimation can be fooled when the translation is in the incorrect language, which the reference-based metric will penalize.

To compute MetricX, we used the official implementation[17] and the "google/metricx-23-xl-v2p0" model. MetricX produces scores at the segment level. To obtain scores at the system level, we averaged the segment scores. To compute CometKiwi scores, we used the official implementation[18] with the "Unbabel/wmt23-cometkiwi-da-xl" model, a reference-free model, taking the translation hypothesis and the source segment as inputs. COMET can produce system-level scores so we use them directly.

To merge the two metrics, we first linearly scale the scores of each metric to a range between 1 and the number of systems for a given language pair. We then average both normalized scores to reach the final automatic ranking, which we refer to as AutoRank. We provide a Jupyter notebook in the WMT24 repository to reproduce the scores.[19]

5.1 Selecting Systems for Human Evaluation

When selecting the systems for human evaluation, we prioritize open and constrained systems while penalizing closed systems. We select a subset of 10 to 15 systems per language pair based on AutoRank and following two rules. First, we exclude closed systems that are not among the first third of all systems and we exclude open systems that are not among the top two thirds of all systems. Second, motivated by several very low-performing systems, we also define a hard cutoff point. After this point we do not evaluate any system from any category. The cutoff point is selected as the first gap between two neighboring system's ranks larger than 1.5 of AutoRank. This decision was discussed and published in more detail in Kocmi et al. (2024a).

6 Human Evaluation

This year's human evaluation features Error Span Annotation (ESA; Kocmi et al., 2024b) for most languages. For Japanese→Chinese and

[17] github.com/google-research/metricx
[18] github.com/Unbabel/COMET
[19] github.com/wmt-conference/wmt24-news-systems/blob/main/Automatic_Evaluation.ipynb

Language pairs	Annotators' profile	Tool
English→Chinese/Japanese/Hindi/Spanish	Microsoft annotators — bilingual target-language native speakers, professional translators or linguists, experienced in machine translation evaluation.	Appraise ESA
Czech→Ukrainian English→Czech	ÚFAL Charles University annotators — linguists, annotators, researchers, and students who were native speakers in the target language and had a very high proficiency in English (for English→Czech) and good knowledge of Czech (for Czech→Ukrainian).	Appraise ESA
English→Ukrainian/Russian/Spanish	Toloka AI paid expert crowd — Bilingual native target-language speakers who were high-performing on the platform.	Appraise ESA
English→Icelandic	The Árni Magnússon Institute for Icelandic Studies annotators — bilingual target-language native speakers, paid translators with 10–25 years of experience in Icelandic↔English translation.	Appraise ESA
English↔German Japanese→Chinese	Campaign managed by the 2024 metrics shared task.	Google MQM

Table 5: Annotators' profiles and annotation tools for each language pair in the human evaluation. English→Spanish was split between Microsoft and Toloka AI. All annotators were paid a fair wage in their respective countries.

English→German, we rely on the evaluation campaign from the metrics shared task 2024 (Freitag et al., 2024), which uses Multidimensional Quality Metrics (MQM; Lommel et al., 2014).

Annotation Protocol. ESA is based on highlighting/marking errors without classifying them into different error types (Kreutzer et al., 2020; Popović, 2020) and represents a compromise between overall scoring (such as direct assessment, DA; Graham et al. 2013) and error classification (such as MQM; Lommel et al. 2014).

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", as in Kocmi et al. (2024b); Popović (2020). In addition, the annotators were also asked to assign a score from 0 to 100, similar to DA, to the whole annotation segments (usually a sentence or 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 interface is shown in Figure 1 with annotator instructions and other changes from the original implementation by Kocmi et al. (2024b) given 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 length of the context given to the annotators varies depending on the domain, ranging from one to ten sentences, as discussed in Section 6.1. The source for the speech domain is a video which is shown in a native HTML video player.

	Systems	Annotators		
Language pair		Duplication		Assess./system
Cs→Uk	11	1.0	14	1299
En→Czech	15	1.3	20	751
En→Spanish	13	1.0	14	370
En→Hindi	10	1.3	15	775
En→Icelandic	10	1.0	4	376
En→Japanese	12	1.5	14	1212
En→Russian	13	1.0	7	370
En→Ukrainian	10	1.0	8	376
En→Chinese	12	1.5	12	1217

Table 6: Number systems, annotators, and number of assessments per system in a language pair. Duplication of d means that each segment is annotated by d annotators. All language pairs had 649 segments over 170 documents except for Czech→Ukrainian which had 1954 segments over 302 documents. In total we collected 57k segment-level annotations. English→German and Japanese→Chinese are managed by the metrics shared task 2024.

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.

Human Annotators Campaigns for different language pairs were managed by various vendors, as described in Table 6. In all cases, professional translators-cum-annotators are used. This is an increasingly strict requirement given the high quality of MT systems, which requires more expert annotators.

6.1 Data Preparation

Document Filtering. In our setup, all systems for a given language pair are evaluated on the same set of segments. On average, we start with 1092

lines per system, encompassing 184 documents. However, the distribution of document lengths is unbalanced. The majority of the documents (104) consist of just a single line, which is almost exclusively due to video translation segments (103), where each "document" contains strictly one segment. On average, 33 documents per language contain more than 10 segments. We limit these documents to the first 10 segments, motivated by the difficulty of annotating very long documents while maintaining relevant context in mind. After this adjustment, we arrive at an average of 744 lines per system. An overview is shown in Table 6.

Workload balancing We use the term "task" as a contained unit of 100 annotation segments. Each annotator is usually assigned to multiple tasks. This 100-segment constraint was kept for historical reasons and will be dropped in future iterations. In order to make it so that each task contains a comparable amount of work, we attempt to balance the number of words in each task to be as constant as possible.

For each task, we show a tutorial at the beginning consisting of 2 documents with 6 segments in total. The tutorial is for German→English translation but does not require any knowledge of German. Finally, we reserve 12 segments for quality control (Section 6.2) in each task. The resulting 82 segments are filled with full documents as much as possible. If that is not possible (i.e., because the next document is too long), a random document is drawn that is either duplicated or incomplete, in order to fill the 100 segments.

Annotation waves In order for a segment to be useful in the evaluation, we require that translations by all systems are evaluated. We therefore split (at the document level) the translated data for each language into two "waves", each of which covers a distinct set of source segments. The vendors are instructed to start the second campaign only after the first one is fully complete.

For some language pairs, the vendors finished early. In this case, we prepared an extra two waves, with a different coverage split of the same data, which they annotated afterwards. As a result, some language pairs have multiple annotations per source segment, was shown in Table 6. This is useful to compute inter-annotator agreement but also provides less noisy annotations.

6.2 Quality Control

Each task (100 segments) includes 12 quality control segments to ensure the high quality of the annotations. The tasks are created as follows:

1. The task (a maximum of 100 segments) is filled with machine-translated documents to be evaluated.
2. A random document is selected from the task.
3. Segments within the sampled document are perturbed.
4. The perturbed document is shuffled within the task at the document-level.
5. Steps 2-4 are repeated until 12 quality control segments are included in the task.

The segment perturbation is done by randomly sampling a span from the segment and replacing it with random text sampled from the entire corpus in the correct language. Since segment lengths vary and a single perturbation could be lost in a very large paragraph, we apply as many perturbations as there are sentences in the output. See Figure 2 for an example.

Source: *Sie haben gestern das Treffen wieder verschoben.*
Original: *He postponed the meeting again yesterday.*
Perturbed: *He postponed the meeting squirrels are never.*

Figure 2: An example of a perturbed translation based on the original system translation. In addition to the original error (the correct pronoun here is *They* and not *He*), we introduce the perturbed part.

After each task is completed, we check whether the perturbed segments received lower scores. Specifically, we compare the distribution of 12 original and 12 perturbed segments with a one-sided Mann-Whitney U test (Mann and Whitney, 1947). If the task fails to pass quality control (p>0.05), it is reset and reassigned to another annotator.[20] In the final data, 96% of perturbed segments have lower scores than their original counterparts.

6.3 Human Data Analysis

We briefly analyze the data from a broader perspective. The scores given by the annotators are largely concentrated near 100, with a small peak around 0 (see Figure 3). Most languages consistently had very few errors per segment, resulting in higher overall scores (see Table 7). For instance, for the Czech→Ukrainian, an average of 0.2 minor errors

[20]Task generation code: github.com/wmt-conference/ErrorSpanAnnotations/tree/main/preparation/wmt24

and 0.1 major errors per segment means there is approximately one minor error for every 5 segments and one major error for every 10 segments.

The annotation time, which is the primary focus of the analysis in Kocmi et al. (2024b), is similar across most languages with the exception of English→Icelandic. This could be caused either by more meticulous annotators or lower quality of submitted systems, which would require more annotation. The average time per segment is just 22 seconds (see Figure 4).

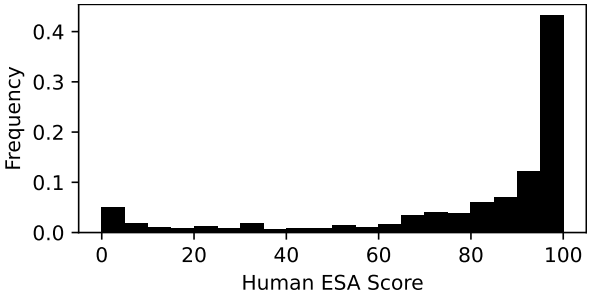


Figure 3: Distribution of final human segment-level scores. The ratings are dominated by the score close to 100.

Language pair	Minor	Major	Score	Time
Czech→Ukrainian	0.2	0.1	87.1	15.8s
English→Czech	0.6	0.2	86.2	25.3s
English→Spanish	0.7	0.4	87.1	22.0s
English→Hindi	0.5	0.2	87.3	25.7s
English→Icelandic	1.4	0.8	72.3	37.8s
English→Japanese	0.2	0.1	89.2	18.9s
English→Russian	0.5	0.3	83.4	23.0s
English→Ukrainian	0.4	0.3	84.4	21.8s
English→Chinese	0.2	0.1	87.6	16.8s

Table 7: Average number of minor and major errors per segment, average score and annotation time. Despite different annotation crowds, the statistics are balanced.

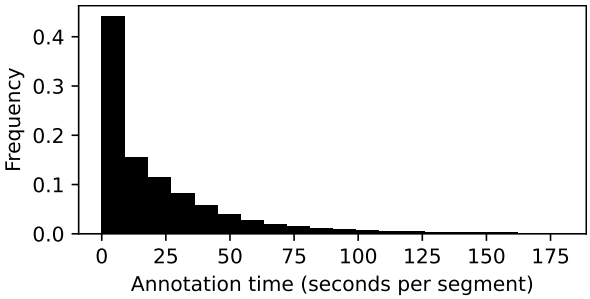


Figure 4: Distribution of annotation times per segment. The vast majority of segments is annotated under one minute.

6.4 Domain Difficulty across Languages

In Table 8 we present the maximal obtained score per domain per language. Although absolute scores are not comparable due to different sets of systems and different groups of annotators, we observe that, across the table, the speech domain obtains the lowest scores for nearly all language pairs suggesting it is the most difficult domain. This is reflected by the fact that the top-performing systems achieve lower scores in the speech domain compared to other domains. This difficulty likely arises from the reliance on ASR text rather than the original audio. This finding is consistent with MQM results from Freitag et al. (2024).

	Literary	News	Social	Speech	Average
En.→Czech	93.1	94.9	93.3	92.1	93.3
En.→Spanish	96.3	96.2	95.5	94.1	95.5
En.→Hindi	95.4	93.6	91.3	88.3	92.2
En.→Icelandic	92.2	92.6	95.0	92.4	93.1
En.→Japanese	92.4	93.7	91.3	92.4	92.5
En.→Russian	94.1	93.1	92.1	86.6	91.5
En.→Ukrainian	93.2	93.9	94.3	85.9	91.8
En.→Chinese	92.0	92.5	90.7	88.4	90.9
Average	93.6	93.8	93.0	90.0	92.6

Table 8: Maximal obtained score per language and per domain across languages evaluated with the same source data (English).

Secondly, we observe that the English→Spanish language pair receives the highest scores, suggesting that either the pair itself or the specific tested domains are relatively easy for top systems, which provide almost flawless translations. These results are consistent with the MQM results from Freitag et al. (2024) where the best system got only -0.12 MQM score, which is close to perfect, while the best German system got -1.58 MQM and the best Japanese-Chinese system an MQM score of -1.22.

Separate scores for each domain, system and language pair are presented in Appendix D.

6.5 Clustering of ESA compared to DA+SQM

This year, we revised the human evaluation protocol, ultimately moving from DA+SQM to ESA. In this section, we briefly compare several aspects of both methods. However, due to the absence of a direct head-to-head comparison on the same data and the many changes introduced since last year, this analysis cannot attribute all the improvements solely to the ESA protocol.

ESA produced 59 clusters across 114 systems. This compares to only 37 clusters produces by last year's DA+SQM approach for the same number of systems. In other words, ESA formed a cluster for every 1.9 systems, while DA+SQM created a cluster for every 3.1 systems. This increased clustering efficiency was achieved despite a decrease in

the number of collected samples. With DA+SQM, we collected an average of 1400 annotations per system, whereas ESA required only an average of 750 annotations per system to achieve greater discriminative power.

7 Official Ranking Results

We now describe how we compute the final ranking, then discuss the final results and some potential issues with our ranking method. The results are shown on the following two pages in tabular form.

7.1 Human Ranking Computation

We calculate three different scores: the human ESA score, rank, and the cluster.

The **human ESA score** is the macro-average of the segment-level ESA scores grouped over the domains. This represents a change compared to previous years, since we used to calculate a simple average over all data. However, with the change towards paragraph-level test sets, the average number of segments per domain is imbalanced and the social domain represents almost half of all segments (see Table 1). To circumvent this imbalance, we use the macro-average as the main human score.

For the statistical analysis and **clustering**, we use the Wilcoxon signed-rank test, a paired non-parametric test (Wilcoxon, 1945), as suggested by Kocmi et al. (2024b). However, given the domain-level imbalanced distribution, we adapted our approach by combining the results from independent domain-level experiments via Stouffer's Z-score method (Stouffer et al., 1949), which combines p-values from individual domain-level Wilcoxon tests. The method produces almost identical clustering as if we had used Wilcoxon over the whole dataset whilst ignoring the imbalance.

Rank ranges indicate the number of systems a particular system underperforms or outperforms: the top end of the rank range is $l + 1$, where l is the number of losses, while the bottom is $n - w$, where n is the total number of systems and w is the number of systems against which the system in question significantly wins.

Systems are grouped into ranks that are separated by thick lines, such that systems within the same group do not strictly outperform other systems within the group. In other words, it is not possible to clearly say which system in the cluster is better than the all others. The ranks and clusters are computed with $p < 0.05$.

We say that a system is winning if it ranks in the first cluster, while ignoring the human reference.

The official rankings shown in Section 7.4 are generated on the basis of the ESA scores. Tables with head-to-head comparisons between all systems are included in Appendix E.

7.2 Verbosity of LLMs

As pointed out by Briakou et al. (2024), some LLMs produce verbose outputs, including an attempt to explain the translation or a refusal to translate. This creates an issue for both automatic and human evaluation of how to treat such outputs.

During the collection of LLM outputs, we asked the LLM to wrap the translation in a particular type of quotes (```) and post-edited LLM outputs removing all extra details outside of these quotes (keeping the whole answer if no quotes have been found). Therefore LLMs that did not follow the expected output format and produced additional output were not considered in the evaluation.

For future work, we should instruct humans to penalize verbose outputs and strengthen the prompt used for collecting LLM translations.

7.3 Human Ranking Discussion

When investigating the official results in Section 7.4, we can make several observations.

The best performing system in the open and constrained systems category is IOL-Research, winning 10 LPs in this category.

The Unbabel-Tower70B system is the best performing participating system winning in 8 LPs. In contrast, this system was ranked the first in all LPs in the automatic evaluation.This highlights that systems can overfit on automatic scores, especially when using Minimum Bayes Risk (MBR; Freitag et al., 2022) with testing metric.

Over all, the best performing system in general seems to be Claude-3.5-Sonnet (wins in 9 LPs); it even outperforms GPT-4 (wins in 5 LPs), which is much more expensive model. Human references are ranked in the first place for 5 language pairs and in the winning cluster for 8 language pairs, suggesting that the reference quality is high and ESA is robust to our setting.

For English→Icelandic, it was almost the case that each system belonged to its own statistically significant cluster. This could be put down to a greater diversity in the quality of systems (also highlighted by more diverse AutoRank scores).

7.4 Official Ranking Results Tables

Results tables legend

The human score is the macro-average of human judgements, grouped by domain. The rank takes into consideration head-to-head wins and losses. AutoRank is calculated from automatic metrics.
Ranking and clustering on human scores is done using Wilcoxon signed rank test for each domain separately and final p-value is combined via Stouffer's Z-score method to align with macro average for human score.
Systems are either constrained (white), open-track (light gray), or closed-track (dark gray).
LLMs that do not officially claim a support a language pair are marked with §.

Czech→Ukrainian

Rank	System	Human	AutoRank
1-2	Claude-3.5 §	93.0	1.7
2-2	HUMAN-A	92.7	-
3-3	Gemini-1.5-Pro	92.6	2.0
3-4	Unbabel-Tower70B	92.2	1.0
5-5	IOL-Research	90.2	1.9
6-7	CommandR-plus §	89.7	1.9
6-8	ONLINE-W	88.7	2.3
7-9	GPT-4 §	88.6	2.0
8-9	IKUN	87.1	2.3
10-10	Aya23	86.6	2.5
11-11	CUNI-Transformer	85.3	3.0
12-12	IKUN-C	82.6	3.0

English→Czech

Rank	System	Human	AutoRank
1-2	HUMAN-A	92.9	-
2-2	Unbabel-Tower70B	91.6	1.0
2-3	Claude-3.5 §	91.2	2.1
4-5	ONLINE-W	89.0	2.8
4-6	CUNI-MH	88.4	2.1
6-6	Gemini-1.5-Pro	88.2	2.6
6-8	GPT-4 §	87.7	2.6
8-8	CommandR-plus §	86.9	2.9
8-9	IOL-Research	86.5	2.8
10-11	SCIR-MT	85.4	3.2
10-11	CUNI-DocTransformer	84.3	4.4
12-12	Aya23	84.2	4.3
13-13	CUNI-GA	82.1	2.3
14-14	IKUN	81.7	3.9
15-15	Llama3-70B §	77.4	4.1
16-16	IKUN-C	75.4	4.7

English→German

Rank	System	Human	AutoRank
1-11	GPT-4	-1.6	1.8
1-7	Dubformer	-1.8	1.8
2-10	ONLINE-B	-1.9	1.8
2-10	TranssionMT	-1.9	1.8
2-9	Unbabel-Tower70B	-1.9	1.0
1-9	HUMAN-B	-2.0	-
2-12	Mistral-Large	-2.1	2.0
4-11	CommandR-plus	-2.3	2.0
8-10	ONLINE-W	-2.3	2.2
2-12	Claude-3.5	-2.4	1.9
3-13	HUMAN-A	-2.5	-
10-12	IOL-Research	-2.5	2.3
5-13	Gemini-1.5-Pro	-2.8	2.2
14-15	Aya23	-3.2	2.7
14-17	ONLINE-A	-3.5	3.0
15-17	Llama3-70B §	-4.3	2.5
15-17	IKUN	-4.3	3.0
18-18	IKUN-C	-6.1	3.8
19-19	MSLC	-15.5	11.9

English→Spanish

Rank	System	Human	AutoRank
1-1	HUMAN-A	95.3	-
2-2	Dubformer	93.4	2.0
3-4	GPT-4	91.9	1.9
4-7	IOL-Research	91.4	2.3
5-8	Mistral-Large	89.3	2.2
5-9	Unbabel-Tower70B	88.9	1.0
3-8	Claude-3.5	88.8	2.1
5-8	Gemini-1.5-Pro	88.8	2.4
7-9	CommandR-plus	88.3	2.1
9-10	Llama3-70B §	87.2	2.6
11-11	ONLINE-B	85.6	2.7
12-13	IKUN	84.7	2.8
12-13	IKUN-C	80.4	3.4
14-14	MSLC	63.9	7.4

English→Hindi

Rank	System	Human	AutoRank
1-3	TranssionMT	91.3	1.3
1-4	Unbabel-Tower70B	90.5	1.0
3-3	Claude-3.5 §	90.2	1.2
3-4	ONLINE-B	90.1	1.4
3-5	Gemini-1.5-Pro §	90.0	1.6
6-6	GPT-4 §	88.5	2.1
7-8	HUMAN-A	88.5	-
8-8	IOL-Research	87.2	2.1
8-9	Llama3-70B §	86.7	2.1
10-10	Aya23	84.7	3.2
11-11	IKUN-C	70.7	5.5

English→Icelandic

Rank	System	Human	AutoRank
1-1	HUMAN-A	93.1	-
2-3	Dubformer	84.3	2.5
2-3	Claude-3.5 §	81.9	2.3
4-4	Unbabel-Tower70B	80.2	1.0
5-5	AMI	73.3	3.7
6-6	IKUN	71.0	3.2
7-7	ONLINE-B	68.0	4.2
8-9	GPT-4	66.3	3.4
8-9	IKUN-C	65.2	3.7
10-10	IOL-Research	58.0	4.3
11-11	Llama3-70B §	41.0	6.7

English→Ukrainian

Rank	System	Human	AutoRank
1-2	Claude-3.5	90.5	2.0
1-2	Unbabel-Tower70B	89.8	1.0
3-3	Dubformer	89.0	1.8
4-6	HUMAN-A	87.3	-
4-6	Gemini-1.5-Pro	87.1	2.2
5-8	ONLINE-W	86.0	2.1
5-9	GPT-4	84.6	2.3
6-9	CommandR-plus §	83.2	2.3
7-9	IOL-Research	83.2	2.4
10-10	IKUN	78.4	2.8
11-11	IKUN-C	67.9	3.9

English→Japanese

Rank	System	Human	AutoRank
1-1	HUMAN-A	91.8	-
2-4	ONLINE-B	91.1	1.4
3-4	CommandR-plus	91.0	1.9
4-4	GPT-4	90.8	1.7
4-5	Claude-3.5	90.8	1.5
4-7	Gemini-1.5-Pro	90.0	1.7
7-7	Unbabel-Tower70B	89.7	1.0
8-8	IOL-Research	89.7	2.3
8-9	Aya23	89.7	2.3
10-10	NTTSU	89.4	1.9
11-11	Team-J	88.5	1.9
12-12	Llama3-70B §	86.8	2.6
13-13	IKUN-C	81.7	3.9

English→Chinese

Rank	System	Human	AutoRank
1-1	GPT-4	89.6	2.0
2-4	Unbabel-Tower70B	89.6	1.0
2-4	HUMAN-A	89.4	-
4-4	Gemini-1.5-Pro	89.3	1.8
5-6	ONLINE-B	89.3	1.7
6-6	IOL-Research	89.0	1.8
6-7	Claude-3.5	88.9	1.7
6-8	CommandR-plus	88.3	2.2
9-9	Llama3-70B §	86.5	2.8
10-10	HW-TSC	86.2	2.3
11-11	IKUN	85.3	3.1
12-12	Aya23	85.2	3.0
13-13	IKUN-C	82.1	3.5

English→Russian

Rank	System	Human	AutoRank
1-1	HUMAN-A	89.2	-
2-3	Dubformer	89.1	1.9
3-4	Claude-3.5	88.2	2.0
3-5	Unbabel-Tower70B	88.1	1.0
3-7	Yandex	87.0	1.9
6-8	Gemini-1.5-Pro	85.5	2.3
6-9	GPT-4	85.0	2.3
6-9	ONLINE-G	84.6	2.2
5-9	CommandR-plus §	84.3	2.4
10-10	IOL-Research	82.1	2.6
11-11	IKUN	79.2	3.2
12-12	Aya23	78.6	3.3
13-13	Llama3-70B §	75.7	3.1
14-14	IKUN-C	69.8	3.9

Japanese→Chinese

Rank	System	Human	AutoRank
1-3	Claude-3.5	-1.4	1.7
1-3	HUMAN-A	-1.5	-
3-5	GPT-4	-1.7	2.1
2-5	DLUT-GTCOM	-1.7	2.0
4-8	Unbabel-Tower70B	-1.9	1.0
3-6	Gemini-1.5-Pro	-2.1	1.9
6-8	CommandR-plus	-2.2	2.8
6-8	IOL-Research	-2.4	2.2
9-10	Llama3-70B §	-3.4	3.1
9-10	Aya23	-3.5	3.7
11-12	Team-J	-4.5	2.8
11-12	NTTSU	-5.1	3.7
13-13	ONLINE-B	-5.8	5.2
14-14	IKUN-C	-7.7	5.5
15-15	MSLC	-10.7	8.9

8 Test Suites Sub-task: "Help us break LLMs"

The results in the previous tables indicate that the current evaluation methods, despite being more detailed and sophisticated, have difficulties in distinguishing MT output from human translations, or distinguishing the performance among different systems. Additionally, the appearance of LLMs has made it even more clear that generated translations, even those which seem to be fluent and surrounded by seemingly perfect content, can contain serious flaws. The increased interest in this new technology and the use of LLMs for translation, prompted us to set the theme of this year's test suite subtask as "Help us break LLMs". This was intended as a broader invitation to the NLP community to expose the weaknesses of LLM translations that are hidden behind the apparent overall high quality generation, but also to propose new innovative evaluation methods that may be of high interest for specific use cases. We are thrilled that this year's participation exceeded every precedent, with 11 test suites providing their valuable conclusions, which are presented below.

8.1 Setup of the sub-task

Each test suite is a customised extension of the standard test sets, focusing on specific aspects of the MT output. The evaluation of the MT output takes place in a decentralized manner, where test suite providers were invited to submit their customized test sets, following the setup introduced at the Third Conference on Machine Translation (Bojar et al., 2018). Each test suite provider submitted a source-side test set, which was appended by the organisers of the General MT Shared Task to its standard test sets. The corresponding outputs from the systems of the General MT Shared Task were returned to the test suite providers, who were responsible for carrying out the evaluation based on their own individual evaluation concept. The results of each test suite evaluation, together with the relevant analysis, appear in separate description papers, while a summary is given below.

This year's timeline gave the test suite contributors more time. We offered a pre-run in April, when test suite providers were given the opportunity to submit the current version of their corpus in order to receive translation output from online systems, which could help them to carry out the individual (often manual) evaluation in a more timely manner.

8.2 Submissions

The test suite sub-task received 11 submissions, out of which 9 completed the entire evaluation cycle. An overview of the test suites can be seen in Table 9. The descriptions of each submission and their main findings are given below.

Árni Magnússon Institute for Icelandic Studies (AMI; Ármannsson et al., 2024) The submission of the Árni Magnússon Institute's team to the WMT24 test suite subtask focuses on idiomatic expressions and proper names for the English→Icelandic translation direction. Intuitively and empirically, idioms and proper names are known to be a significant challenge for neural translation models. They create two different test suites. The first evaluates the competency of MT systems in translating common English idiomatic expressions, as well as testing whether systems can distinguish between those expressions and the same phrases when used in a literal context. The second test suite consists of place names that should be translated into their Icelandic exonyms (and correctly inflected) and pairs of Icelandic names that share a surface form between the male and female variants, so that incorrect translations impact meaning as well as readibility. The scores reported are relatively low, especially for idiomatic expressions and place names, and indicate considerable room for improvement.

Complex Sentence Structure Testset (CoST; vIIT_HYD; Mukherjee et al., 2024) This test suite presents an evaluation of 16 machine translation systems submitted to the Shared Task for the English-Hindi using our Complex Structures Test suite. Aligning with this year's test suite sub-task theme, "Help us break LLMs", the authors curated a comprehensive test suite encompassing diverse datasets across various categories, including autobiography, poetry, legal, conversation, play, narration, technical, and mixed genres. The evaluation reveals that all the systems struggle significantly with the archaic style of text like legal and technical writings or text with creative twist like conversation and poetry datasets, highlighting their weaknesses in handling complex linguistic structures and stylistic nuances inherent in these text types. This evaluation identifies the strengths and limitations of the models, pointing to specific areas where further research is needed to enhance their performance.[21]

[21] github.com/AnanyaCoder/CoST-WMT-24-Test-Suite-Task

Test suite	Institution	Focus	Language pair	Segments
AMI (Ármannsson et al., 2024)	AMI	idiomatic expressions, proper names	en→is	3,082
COST (Mukherjee et al., 2024)	IIIT_HYD	complex sentence structure	en→hi	3,908
DFKI (Manakhimova et al., 2024)	DFKI	110 linguistic phenomena	en→de, en→ru	54,736
GenderQueer (Friðriksdóttir, 2024)	UI	gender-diverse, queer-inclusive content	en→is	672
IITP (Bhattacharjee et al., 2024)	IITP	multi-domain dynamics	en→hi	4,198
Isochrony (Rozanov et al., 2024)	RaskAI, IC	isochrony of translations	en→de, en→es, en→ja, en→ru, en→zh	10,730
NRCC (Dawkins et al., 2024)	NRCC	speaker-listener gender resolution	en→cs, en→de, en→es, en→is	53,560
PIA_TQA (Miceli Barone and Sun, 2024)	UEDIN	prompt injection attacks	cs→uk, en→cs, en→de, en→es, en→hi, en→is, en→ja, en→ru, en→uk, en→zh, ja→zh	250,744
RoCS-MT (Bawden and Sagot, 2023)	Inria	robustness to non-standard user-generated texts	en→cs, en→de, en→es, en→hi, en→is, en→ja, en→ru, en→uk, en→zh	7883

Table 9: Overview of the participating test suites.

DFKI (Manakhimova et al., 2023b) This test suite offers a fine-grained linguistically motivated analysis of the shared task MT outputs for English–German and English–Russian, based on more than 11,500 manually devised test items, which cover up to 110 phenomena in 14 categories per language direction. Extending their previous test suite submissions (e.g. Avramidis et al., 2020; Macketanz et al., 2021, 2022; Manakhimova et al., 2023a), the submission of this year includes a considerable effort of manual linguistic annotation for the evaluation on 39 MT systems submitted at the Shared Task. Based on the results, LLMs are inferior to NMT in English–German when translating a few linguistic phenomena, though they show quite a competitive performance in English-Russian. Additionally, some LLMs generate very verbose or empty outputs, posing challenges to the evaluation process. Looking more closely at specific phenomena of English-German, LLMs seem to perform worse than the two best performing NMT systems in terms of punctuation, future verb tenses and stripping. For English-Russian, Yandex is weaker in named entities and terminology, Claude in function words, while Unbabel is weaker in verb valency. GPT-4 into Russian performs even worse than several commercial NMT-based systems.

Indian Institute of Technology Patna (IITP; domain dynamics; Bhattacharjee et al., 2024) LLMs have demonstrated impressive capabilities in machine translation, leveraging extensive pretraining on vast amounts of data. However, this generalist training often overlooks domain-specific nuances, leading to potential difficulties when translating specialized texts. This study presents a multi-domain dataset designed to challenge and evaluate the translation abilities of LLMs. The dataset encompasses diverse domains such as judicial, education, literature (specifically religious texts), and noisy user-generated content from online product reviews and forums like Reddit. Each domain consists of approximately 250–300 sentences, carefully curated and randomized in the final compilation. This English-to-Hindi dataset aims to evaluate and expose the limitations of LLM-based translation systems, offering valuable insights into areas requiring further research and development.

Inria (RoCS-MT; Bawden and Sagot, 2023), Robust Challenge Set for Machine Translation, 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). RoCS-MT was first submitted to the 2023 task, showing that many non-standard phenomena still pose problems for most systems, although more common phenomena are better handled by the larger, closed-source models, presumably due to the large quantity of web-based seen during training. This year's version is largely the same as last year but with some improvements, including modifications to normalisations and to the annotation typology used (all modifications are documented in the GitHub repository).[22] Systems varied greatly in terms of their handling of

[22] github.com/rbawden/RoCS-MT

non-standard sentences, with marked differences depending on the type of system. Constrained systems inevitably struggling most, particularly with phenomena affecting the spelling of words (resulting in frequent copying of non-standard source words), a problem also affecting online systems. LLMs exhibited some of the best quality translations, although behaviour varied between translating standard and non-standard input, and additional issues such as refusal to translate and usage notes pose new technical challenges.

Isochrony Translation (Rask AI, Imperial College; Rozanov et al., 2024) MT has come a long way and is readily employed in production systems to serve millions of users daily. With the recent advances in generative AI, a new form of translation is becoming possible – video dubbing. This work motivates the importance of isochronic translation, especially in the context of automatic dubbing, and introduces 'IsoChronoMeter' (ICM). ICM is a simple yet effective metric to measure isochrony of translations in a scalable and resource efficient way without the need for gold data, based on state-of-the-art text-to-speech (TTS) duration predictors. The authors motivate IsoChronoMeter and demonstrate its effectiveness. Using ICM, they demonstrate the short-comings of state-of-the-art translation systems and show the need for new methods. The code has been released.

National Research Council Canada (Speaker-Listener Gender Resolution; gender-res; Dawkins et al., 2024) This test suite assesses the gender resolution tendencies of MT systems in literary dialogue settings. That is, each instance contains dialogue interleaved with additional meta-context. The spoken dialogue refers to either the speaker or listener such that the gender of the referent, if known, must be inferred from the meta-context and informs the correct translation. They find that stereotype factors within the meta-context, such as character descriptions and manner of speaking, affect the gender agreement choices of words within the dialogue. Regression analysis is performed to evaluate the relative influence of these contextual factors compared to structural factors and known stereotype influences (e.g., the internal gender stereotype of an adjective).

University of Edinburgh Prompt Injection, TruthfulQA (PIA; Miceli Barone and Sun, 2024) LLM-based systems typically work by embedding their input data into prompt templates which contain instructions and/or in-context examples, creating queries which are submitted to a LLM, then parse the LLM response in order to generate the system outputs. Prompt Injection Attacks (PIAs) are a type of subversion of these systems where a malicious user crafts special inputs which interfere with the prompt templates, causing the LLM to respond in ways unintended by the system designer. Recently, Sun and Miceli Barone (2024) proposed a class of PIAs against LLM-based machine translation. Specifically, the task is to translate questions from the TruthfulQA test suite, where an adversarial prompt is prepended to the questions, instructing the system to ignore the translation instruction and answer the questions instead. In this test suite, the authors extend this approach to all the language pairs of the WMT 2024 General Machine Translation task. Moreover, they include additional attack formats in addition to the one originally studied.

University of Iceland (GenderQueer; Friðriksdóttir, 2024) This paper introduces the GenderQueer Test Suite, a novel evaluation set for assessing MT systems' capabilities in handling gender-diverse and queer-inclusive content, focusing on English to Icelandic translation. As MT quality improves, there is an increasing need for specialized evaluation methods that address nuanced aspects of language and identity. The suite evaluates MT systems on various aspects of gender-inclusive translation, including pronoun and adjective agreement, LGBTQIA+ terminology accuracy, and the impact of explicit gender specifications. Its authors evaluated 18 MT systems submitted to the WMT24 English-Icelandic track. Key findings reveal significant performance differences between large language model-based systems and smaller models in handling context for gender agreement. Challenges in translating singular "they" were widespread, while most systems performed well in translating LGBTQIA+ terminology. Accuracy in adjective gender agreement varies, with some models struggling particularly with feminine forms. This evaluation set contributes to the ongoing discussion about inclusive language in MT and natural language processing. By providing a tool for assessing MT systems' handling of gender-diverse content, it aims to enhance the inclusivity of language technology. The methodology and evaluation scripts are made available for adaptation to other languages, promoting further research in this critical area.

9 Conclusions

The WMT 2024 General Machine Translation Task covered 11 translation pairs, two of which are non-English: Czech→Ukrainian and Japanese→Chinese. We introduced ESA (Error Span Annotations) as the main human protocol for assessing the translation quality, which enabled more efficient collection of human judgements than MQM while keeping high quality of annotations. In total, 108 human (semi-)professional annotators contributed more than 57,000 judgments.

We received 105 primary submissions from 28 participants, 4 online systems and 8 production large language models, which is a large increase from last year's task. The majority of participants already use LLMs in their systems.

The best performing open system is IOL-Research (wins 10 LPs in it's category), the best performing participating system is Unbabel-Tower70B (wins 8 LPs), and the best performing system in general is Claude-3.5-Sonnet (wins 9 LPs).

While the best performing system based on automatic metrics is Unbabel-Tower70B, it was not the winner across the board in the human evaluation, with the mismatch between the results likely due to metric bias (Kovacs et al., 2024) in MBR. This shows that human evaluation should be used as the final judge of translation quality.

Lastly, we showed promising results in the multimodal evaluation of the speech domain, proving to be a challenging domain for MT systems. On the opposite side, systems were able to produce near-perfect translations in English→Spanish, for the domains that we tested.

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 modalities, we used audio and text.

Although we use human judgements 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.

Some models may have used Comet or MetricX during their training, for example, using Minimum Bayes Risk. Our automatic evaluation of such models will be biased, giving them artificially higher scores.

Automatic metrics are limited and biased (Karpinska et al., 2022; Moghe et al., 2024), especially in novel domains (Zouhar et al., 2024a), 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.

The ESA annotation interface implemented in Appraise is in English only with a tutorial in German→English. This caused difficulties to some of the Czech→Ukrainian annotators we hired, who could not understand English. One such annotator did not pass the initial tutorial and therefore did not participate in the annotation campaign. Next year, we plan to translate the annotation interface to either the source or target language for each translation direction.

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 randomly among pools of annotators and we do not store any personal information. We do store the mapping between which annotator (with pseudonym) annotated which account. Annotators received standard professional translator's or evaluator's wage with respect to their countries.

Sentences in the Czech→Ukrainian dataset (in Personal, Official and Voice domains) were collected with users' opt-in consent, and any personal data related to people other than well-known people was pseudonymized (using random first names and surnames). Sentences where such pseudonymization would not be enough to preserve reasonable anonymity of the users (e.g., describing events uniquely identifying the persons involved) were not included in the test set.

Acknowledgments

This task would not have been possible without the partnership with Microsoft, Charles University, Dubformer, Toloka, NTT, Google, Árni Magnússon Institute for Icelandic Studies, Custom.mt, Cohere, Together.ai, Unbabel and the German Research Center for AI (DFKI).

Additionally, we would like to thank Nikolay Bogoychev, Konstantin Dranch and many others who provided help, feedback, and recommendations, as well as all shared task participants, for their participation and for providing their system descriptions for the paper.

We would also like to thank Toshiaki Nakazawa, Yoshimasa Tsuruoka, Jun Suzuki, and Takehito Utsuro for their help and recommendations in creating the Japanese test set.

Barry Haddow's participation was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546 – HPLT].

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.

Maja Popović's participation was funded by the ADAPT SFI Centre for Digital Media Technology, funded by Science Foundation Ireland through the SFI Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

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

Ondřej Bojar acknowledges the support of the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO. The manual evaluations were also supported by the InCroMin FSTP under the HE grant UTTER (101070631 – HE, 0039436 – UKRI).

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 Espa
textasciitilde na-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.

Bjarki Ármannsson, Hinrik Hafsteinsson, Atli Jasonarson, and Steinthor Steingrimsson. 2024. Killing two flies with one stone: An attempt to break llms using english→icelandic idioms and proper names. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress.

Eleftherios Avramidis, Annika Grützner-Zahn, Manuel Brack, Patrick Schramowski, Pedro Ortiz Suarez, Malte Ostendorff, Fabio Barth, Shushen Manakhimova, Vivien Macketanz, Georg Rehm, and Kristian Kersting. 2024. Occiglot at WMT24: European open-source large language models evaluated on translation. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

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 Sarrías, 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.

Soham Bhattacharjee, Baban Gain, and Asif Ekbal. 2024. Domain dynamics: Evaluating large language models in english-hindi translation. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

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 Névéol, 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.

Eleftheria Briakou, Zhongtao Liu, Colin Cherry, and Markus Freitag. 2024. On the implications of verbose llm outputs: A case study in translation evaluation.

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.

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.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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.

Hillary Dawkins, Isar Nejadgholi, and Chi-kiu Lo. 2024. WMT24 test suite: Gender resolution in speaker-listener dialogue roles. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Sören Dreano, Derek Molloy, and Noel Murphy. 2024. Cyclegn: a cycle consistent approach for neural machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, USA. 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.

Denis Elshin, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, Ekaterina Enikeeva, Dmitry Popov, Anton Chekashev, Vladislav Negodin, Vera Frantsuzova, Alexander Chernyshev, and Kirill Denisov. 2024. From general LLM to translation: How we dramatically improve translation quality using human evaluation data for LLM finetuning. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88. 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, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are llms breaking mt metrics? results of the wmt24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

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.

Steinunn Rut Friðriksdóttir. 2024. The genderqueer test suite. In *Proceedings of the Ninth Conference on Machine Translation*, USA. 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).

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.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Miroslav Hrabal, Josef Jon, Martin Popel, Nam Luu, Danil Semin, and Ondřej Bojar. 2024. CUNI at WMT24 general translation task: Llms, (q)lora, CPO and model merging. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Atli Jasonarson, Hinrik Hafsteinsson, Bjarki Ármannsson, and Steinþór Steingrímsson. 2024. Cogs in a machine, doing what they're meant to do – the AMI submission to the WMT24 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: 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.

Ondrej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. MT-ComparEval: Graphical evaluation interface for machine translation development. *Prague Bull. Math. Linguistics*, 104:63–74.

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, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 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, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024a. Preliminary WMT24 Ranking of General MT Systems and LLMs. *arXiv preprint arXiv:2407.19884*.

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. 2023. 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, Martin Popel, and Ondrej Bojar. 2020. Announcing CzEng 2.0 Parallel Corpus with over 2 Gigawords. *CoRR*, abs/2007.03006.

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*, USA. 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. 2006. Proceedings on the workshop on statistical machine translation. New York, USA. Association for Computational Linguistics.

Minato Kondo, Ryo Fukuda, Xiaotian Wang, Katsuki Chousa, Masato Nishimura, Kosei Buma, Takatomo Kano, and Takehito Utsuro. 2024. NTTSU at WMT2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. 2024. Mitigating metric bias in minimum bayes risk

decoding. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. Correct me if you can: Learning from error corrections and markings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 135–144. European Association for Machine Translation.

Keito Kudo, Hiroyuki Deguchi, Makoto Morishita, Ryo Fujii, Takumi Ito, Shintaro Ozaki, Koki Natsumi, Kai Sato, Kazuki Yano, Ryosuke Takahashi, Subaru Kimura, Tomomasa Hara, Yusuke Sakai, and Jun Suzuki. 2024. Document-level translation with LLM reranking: Team-j at WMT 2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Taku Kudo. 2005. Mecab: Yet another part-of-speech and morphological analyzer. `https://taku910.github.io/mecab/`. Accessed: 2023-10-02.

Samuel Larkin, Chi-kiu Lo, and Rebecca Knowles. 2024. MSLC24 submissions to the general machine translation task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Baohang Li, Zekai Ye, yichong huang, Xiaocheng Feng, and Bing Qin. 2024. SCIR-MT's submission for WMT24 general machine translation task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Baohao Liao, Christian Herold, Shahram Khadivi, and Christof Monz. 2024. IKUN for WMT24 general MT task: Llms are here for multilingual machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. 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.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, 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.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2023. Mega: Moving average equipped gated attention. In *The Eleventh International Conference on Learning Representations*.

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. 2023a. 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, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023b. 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 (WMT)*, Singapore, Singapore (Hybrid). 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*, USA. Association for Computational Linguistics.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

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.

Antonio Valerio Miceli Barone and Zhifan Sun. 2024. A test suite of prompt injection attacks for LLM-based machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2024. Machine Translation

Meta Evaluation through Translation Accuracy Challenge Sets. *Computational Linguistics*, pages 1–60.

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.

Ananya Mukherjee, Saumitra Yadav, and Manish Shrivastava. 2024. Cost of breaking the llms. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Vladimir Aleksandrovich Mynka and Nikolay Mikhaylovskiy. 2024. TSU HITS's submissions to the WMT 2024 general machine translation shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 Biomedical Translation Shared Task: Test Sets on Abstract Level. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

OpenAI. 2024. GPT-4 Technical Report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.

Maja Popović. 2020. Informative manual evaluation of machine translation output. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5059–5069. International Committee on 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.

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. 2022. Robust speech recognition via large-scale weak supervision.

Ricardo Rei, Nuno M. Guerreiro, Jos
textasciitilde A© 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, Jose Maria Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. de Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2023 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Matiss Rikters and Makoto Miwa. 2024. AIST AIRC systems for the WMT 2024 shared tasks. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Nikolai Rozanov, Vikentiy Pankov, Dmitrii Mukhutdinov, and Dima Vypirailenko. 2024. Isochronometer: A simple and effective isochronic translation evaluation metric. In *Proceedings of the Ninth Conference on Machine Translation*, USA. 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.

Danil Semin and Ondřej Bojar. 2024. CUNI-DS submission: A naive transfer learning setup for english-to-russian translation utilizing english-to-czech data. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Samuel A Stouffer, Edward A Suchman, Leland C DeVinney, Shirley A Star, and Robin M Williams Jr. 1949. The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.

Roman Sudarikov, Martin Popel, Ondřej Bojar, Aljoscha Burchardt, and Ondřej Klejch. 2016. Using MT-ComparEval. In *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 76–82.

Zhifan Sun and Antonio Valerio Miceli Barone. 2024. Scaling behavior of machine translation with large language models under prompt injection attacks. In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 9–23, St. Julian's, Malta. Association for Computational Linguistics.

Shaomu Tan, David Stap, Seth Aycock, Christof Monz, and Di Wu. 2024. Uva-MT's participation in the WMT24 general translation shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Gemini Team. 2024a. Gemini: A family of highly capable multimodal models.

Llama-3 Team. 2024b. The Llama 3 Herd of Models.

Phi-3 Team. 2024c. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone.

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.

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.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Hengchao Shang, Jiaxin GUO, Shaojun Li, Zhiqiang Rao, Yuanchang Luo, Ning Xie, and Hao Yang. 2024. Choose the final translation from NMT and LLM hypotheses using MBR decoding: HW-TSC's submission to the WMT24 general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Wenbo Zhang. 2024. IOL research machine translation systems for WMT24 general machine translation shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232.

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, Huan Liu, Conghu Yuan, Wentao Chen, and Degen Huang. 2024. DLUT and GTCOM's neural machine translation systems for WMT24. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024a. 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, Bangkok, Thailand. Association for Computational Linguistics.

Vilém Zouhar, Věra Kloudová, Martin Popel, and Ondřej Bojar. 2024b. Evaluating optimal reference translations. *Natural Language Processing*, page 1–24.

A Error Span Annotation Miscellaneous

A.1 Annotation Guidelines

Higlighting errors: Highlight the text fragment where you have identified a translation error (drag or click start & end). Click repeatedly on the highlighted fragment to increase its severity level or to remove the selection.
- **Minor Severity:** Style/grammar/lexical choice could be better/more natural.
- **Major Severity:** Seriously changed meaning, difficult to read, decreases usability.

If something is missing from the text, mark it as an error on the **[MISSING]** word. The highlights do not have to have character-level precision. It's sufficient if you highlight the word or rough area where the error appears. Each error should have a separate highlight.

Score: After highlighting all errors, please set the overall segment translation scores. The quality levels associated with numerical scores on the slider:
- 0%: No meaning preserved: Nearly all information is lost in the translation.
- 33%: Some meaning preserved: Some of the meaning is preserved but significant parts are missing. The narrative is hard to follow due to errors. Grammar may be poor.
- 66%: Most meaning preserved and few grammar mistakes: The translation retains most of the meaning. It may have some grammar mistakes or minor inconsistencies.
- 100%: Perfect meaning and grammar: The meaning and grammar of the translation is completely consistent with the source.

A.2 Changes to Interface

Since the original study of Kocmi et al. (2024b), we used an updated version of the interface. Apart from minor quality of life changes, a noticeable change is the addition of a pop-up bubble that shows the exact score of the segment (see Figure 5). While it appears as a minor change, it might change the annotator behavior that prefer for example certain numbers, as annotators did in translation evaluation study of Zouhar et al. (2024b).

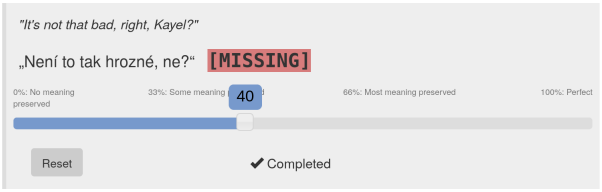


Figure 5: Interacting with the score slider shows the exact score to the annotator in the updated ESA interface.

Dataset	Segments	Tokens		Characters	
		Source	Target	Source	Target
Czech→Ukrainian	Segs	Czech	Ukrainian	Czech	Ukrainian
OPUS	9.8M	103.0M	102.9M	752.0M	1.3B
Facebook-wikimatrix-1	849.0k	10.4M	10.1M	76.0M	127.3M
ELRC	130.0k	2.5M	2.6M	19.6M	35.3M
(Total)	10.8M	115.9M	115.6M	847.6M	1.4B
English→Czech	Segs	English	Czech	English	Czech
ParaCrawl-paracrawl-9	50.6M	692.1M	626.3M	4.3B	4.7B
Facebook-wikimatrix-1	2.1M	33.6M	29.7M	206.8M	216.6M
Tilde	2.1M	42.3M	38.3M	276.5M	303.7M
Statmt-europarl-10	644.4k	15.6M	13.0M	94.3M	98.1M
Statmt-wikititles-3	410.9k	1.0M	965.6k	7.5M	7.6M
Statmt-news_commentary-18.1	265.4k	5.7M	5.2M	36.2M	39.8M
Statmt-commoncrawl_wmt13-1	161.8k	3.3M	2.9M	20.7M	20.7M
(Total)	56.3M	793.7M	716.3M	5.0B	5.4B
English→German	Segs	English	German	English	German
ParaCrawl-paracrawl-9	278.3M	4.3B	4.0B	26.4B	29.5B
Facebook-wikimatrix-1	6.2M	100.5M	97.0M	623.7M	701.2M
Tilde	5.2M	107.4M	102.7M	698.6M	822.1M
Statmt-commoncrawl_wmt13-1	2.4M	51.4M	47.0M	314.2M	340.5M
Statmt-europarl-10	1.8M	45.5M	42.4M	272.9M	312.1M
Statmt-wikititles-3	1.5M	3.6M	3.1M	26.5M	25.5M
Statmt-news_commentary-18.1	437.5k	9.6M	9.8M	61.2M	74.3M
(Total)	295.9M	4.6B	4.3B	28.4B	31.7B
English→Hindi	Segs	English	Hindi	English	Hindi
AllenAi-nllb-1	33.2M	327.0M	311.6M	1.8B	3.8B
OPUS	12.1M	147.6M	165.7M	919.3M	2.2B
AI4Bharath-samananthar-0.2	8.5M	135.8M	152.3M	819.0M	2.0B
Statmt-ccaligned-1	8.2M	114.5M	129.8M	724.3M	1.7B
Anuvaad	3.0M	58.5M	61.6M	359.5M	836.2M
IITB-hien_train-1.5	1.6M	19.8M	21.4M	114.7M	283.6M
Facebook-wikimatrix-1	696.1k	12.0M	13.5M	74.0M	182.4M
Statmt-pmindia-1	56.8k	1.1M	1.2M	6.7M	16.6M
JoshuaDec-indian_training-1	37.7k	562.6k	659.1k	3.4M	8.9M
Neulab-tedtalks_train-1	18.8k	372.6k	491.2k	1.9M	4.4M
Statmt-news_commentary-18.1	4.9k	149.7k	167.7k	963.6k	2.3M
ELRC	245	4.9k	6.3k	31.6k	85.7k
(Total)	67.3M	817.3M	858.4M	4.9B	11.1B
English→Icelandic	Segs	English	Icelandic	English	Icelandic
OPUS	16.4M	174.9M	166.5M	1.0B	1.1B
ParaCrawl-paracrawl-9	3.0M	45.1M	42.7M	266.1M	292.2M
ParIce-eea_train-20.05	1.7M	26.7M	24.2M	170.4M	179.5M
Statmt-ccaligned-1	1.2M	18.6M	17.8M	115.6M	124.4M
Tilde	420.7k	6.3M	6.1M	41.7M	45.3M
ParIce-ema_train-20.05	399.1k	6.1M	5.9M	40.4M	43.9M
Facebook-wikimatrix-1	313.9k	5.7M	4.8M	34.5M	34.0M
Statmt-wikititles-3	50.2k	99.0k	88.4k	722.2k	763.3k
EU	4.7k	54.4k	52.3k	369.0k	398.5k
(Total)	23.4M	283.7M	268.2M	1.7B	1.8B
English→Russian	Segs	English	Russian	English	Russian
Statmt-backtrans_ruen-wmt20	39.4M	746.5M	596.3M	4.5B	7.8B
OPUS	25.2M	563.8M	520.7M	3.7B	7.3B
ParaCrawl-paracrawl-1_bonus	5.4M	101.3M	80.4M	632.5M	1.1B
Facebook-wikimatrix-1	5.2M	86.8M	76.5M	537.7M	1.0B
Statmt-wikititles-3	1.2M	3.1M	2.9M	22.8M	39.3M
Statmt-yandex-wmt22	1.0M	21.3M	18.7M	131.0M	250.8M
Statmt-commoncrawl_wmt13-1	878.4k	18.8M	17.4M	116.2M	214.6M
Statmt-news_commentary-18.1	377.7k	8.7M	8.1M	55.7M	112.1M
Tilde	34.3k	752.7k	702.8k	4.8M	10.0M
(Total)	78.6M	1.6B	1.3B	9.7B	17.7B

Table 10: 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→Spanish	Segs	English	Spanish	English	Spanish
ParaCrawl-paracrawl-9	269.4M	4.4B	4.8B	26.7B	30.0B
OPUS	223.4M	4.1B	4.6B	26.3B	30.0B
Statmt-ccaligned-1	98.4M	1.2B	1.3B	7.7B	8.6B
LinguaTools-wikititles-2014	16.6M	41.3M	46.0M	304.8M	335.2M
Facebook-wikimatrix-1	6.5M	120.1M	137.4M	742.9M	854.5M
Tilde	3.8M	80.0M	92.9M	521.0M	603.4M
EU	3.7M	70.7M	80.6M	457.1M	519.6M
Statmt-europarl-7	2.0M	49.1M	51.6M	294.5M	324.6M
Statmt-commoncrawl_wmt13-1	1.8M	40.8M	43.5M	248.8M	272.8M
Statmt-news_commentary-18.1	500.2k	11.1M	13.1M	71.1M	83.5M
Neulab-tedtalks_train-1	196.0k	4.1M	3.9M	20.4M	20.6M
(Total)	626.2M	10.2B	11.2B	63.4B	71.6B
English→Ukrainian	Segs	English	Ukrainian	English	Ukrainian
ParaCrawl-paracrawl-1_bonus	13.4M	505.8M	487.5M	3.3B	6.0B
Statmt-ccaligned-1	8.5M	119.4M	104.1M	755.4M	1.3B
Facebook-wikimatrix-1	2.6M	41.5M	35.6M	257.6M	447.3M
ELRC	129.9k	3.0M	2.6M	19.6M	35.7M
Tilde	1.6k	36.1k	34.2k	238.0k	477.9k
(Total)	24.6M	669.8M	629.8M	4.3B	7.8B
English→Japanese	Segs	English		English	Japanese
KECL-paracrawl-3	25.7M	599.0M		3.7B	4.6B
Facebook-wikimatrix-1	3.9M	61.6M		379.1M	455.0M
StanfordNLP-jesc_train-1	2.8M	19.3M		104.0M	119.6M
Statmt-wikititles-3	757.0k	1.9M		14.0M	18.7M
Phontron-kftt_train-1	440.3k	9.7M		59.9M	49.1M
Statmt-ted-wmt20	241.7k	4.0M		23.0M	27.3M
Statmt-news_commentary-18.1	1.9k	40.3k		253.2k	318.5k
(Total)	33.9M	695.7M		4.3B	5.2B
English→Chinese	Segs	English		English	Chinese
Statmt-backtrans_enzh-wmt20	19.8M	364.2M		2.2B	2.0B
OPUS	17.5M	417.3M		2.7B	2.1B
ParaCrawl-paracrawl-1_bonus	14.2M	217.6M		1.3B	1.2B
Facebook-wikimatrix-1	2.6M	49.9M		311.1M	277.8M
Statmt-wikititles-3	922.0k	2.4M		17.8M	16.3M
Statmt-news_commentary-18.1	442.9k	9.8M		62.7M	55.2M
(Total)	55.3M	1.1B		6.6B	5.6B
Japanese→Chinese	Segs			Japanese	Chinese
OPUS	19.6M			1.4B	1.1B
KECL-paracrawl-2wmt24	4.6M			1.0B	705.0M
LinguaTools-wikititles-2014	1.7M			35.2M	27.5M
Facebook-wikimatrix-1	1.3M			145.1M	113.6M
KECL-paracrawl-2	83.9k			18.9M	14.1M
Neulab-tedtalks_train-1	5.2k			490.9k	376.0k
Statmt-news_commentary-18.1	1.6k			272.8k	197.3k
(Total)	27.2M			2.6B	1.9B

Table 11: Training dataset statistics (continued from Table 10 on previous page).

B System Submission Summaries

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

B.1 AIST-AIRC (Rikters and Miwa, 2024)

At WMT 2024 AIST AIRC participated in the General Machine Translation shared task and the Biomedical Translation task (Neves et al., 2024). We trained constrained track models for translation between English, German, and Japanese. Before training the final models, we first filtered the parallel data, then performed iterative back-translation as well as parallel data distillation. We experimented with training baseline Transformer models, Mega models, and fine-tuning open-source T5 and Gemma model checkpoints using the filtered parallel data. Our primary submissions contain translations from ensembles of two Mega model checkpoints and our contrastive submissions are generated by our fine-tuned T5 model checkpoints.

B.2 AMI (Jasonarson et al., 2024)

This paper presents the submission of the Arni Magnusson Institute's team to the WMT24 General translation task. We work on the English→Icelandic translation direction. Our system comprises four translation models and a grammar correction model. For training our systems we carefully curate our datasets, aggressively filtering out sentence pairs that may detrimentally affect the quality of our systems output. Some of our data are collected from human translations and some are synthetically generated. A part of the synthetic data is generated using an LLM, and we find that it increases the translation capability of our system significantly.

B.3 CUNI-DS (Semin and Bojar, 2024)

We present a naive transfer learning approach for English-to-Russian translation, leveraging English-to-Czech data within the constrained track of WMT24. Utilizing the Mistral-7B-0.1 model in its 4-bit quantized variant, we employ QLoRA adapter training. The approach involves two phases: first, training the adapters on the English-to-Czech CzEng 2.0 dataset, followed by fine-tuning the adapters further for English-to-Russian translation with additional corpora. The training spans a total of 48 hours. Evaluation is performed using WMT22 and WMT23 datasets, including the paragraph-level version of the latter.

Phase 1: Training on English-to-Czech Data

Dataset: CzEng 2.0, with examples packed into chunks of sequence length 2048.

Parameters: Warmup Steps: 20, Learning Rate: 2e-5, Weight Decay: 1e-2, Cumulative Batch Size: 32

Instructions: Alpaca-like instructions

Duration: 24 hours on a single A100 GPU, using the Unsloth library.

Phase 2: Fine-Tuning for English-to-Russian

Data: Yandex Corpus and News Commentary v18.1, with the latter divided into chunks of 10 sentences.

Regimen: Training with parameters similar to Phase 1.

Duration: An additional 24 hours, totaling 48 hours of training.

B.4 CUNI-{Transformer, DocTransformer, GA, MH, NL} (Hrabal et al., 2024)

This paper presents the contributions of Charles University teams to the WMT24 General Translation task (English to Czech, German and Russian, and Czech to Ukrainian), and the WMT24 Translation into Low-Resource Languages of Spain task.

Our most elaborate submission, CUNI-MH for English→Czech, is the result of fine-tuning Mistral 7B v0.1 for translation using a three-stage process: Supervised fine-tuning using QLoRA, Contrastive Preference Optimization, and merging of model checkpoints. We also describe the CUNI-GA, CUNI-Transformer and CUNI-DocTransformer submissions, which are based on our systems from the previous year.

Our en2ru system CUNI-DS uses a similar first stage as CUNI-MH (QLoRA for English→Czech) and follows with transferring to en2ru.

For en2de (CUNI-NL), we experimented with a LLM-based speech translation system, to translate without the speech input.

For the Translation into Low-Resource Languages of Spain task, we performed QLoRA fine-tuning of a large LLM on a small amount of synthetic (backtranslated) data.

B.5 CycleL and CycleL2 (Dreano et al., 2024)

CycleGN is a fully self-supervised Neural Machine Translation framework relying on the Transformer architecture that does not require parallel data. Its approach is similar to a Discriminator-less CycleGAN, hence the "non-adversarial" name, specifically tailored for non-parallel text datasets. The foundational concept of our research posits that in an ideal scenario, retro-translations of generated translations should revert to the original source sentences. Consequently, a pair of models can be trained using a Cycle Consistency Loss (CCL) only, with one model translating in one direction and the second model in the opposite direction.

In the context of this research, two sub-categories of non-parallel datasets are introduced. A "permuted" dataset is defined as a parallel dataset wherein the sentences of one language have been systematically rearranged. Consequently, this results in a non-parallel corpus where it is guaranteed that each sentence has a corresponding translation located at an unspecified index within the dataset. A "non-intersecting" dataset is a non-parallel dataset for which it is guaranteed that no sentence has an exact translation.

Masked Language Modeling (MLM) is a pre-training strategy implemented in BERT, where a specified proportion of the input tokens are substituted with a unique $mask$ token. The objective of the neural network under this paradigm is to accurately reconstruct the original sentence from this degraded input.

In inference mode, Transformers are able to generate sentences without labels. Thus, the first step is to generate pseudo-labels in inference, that are then used as labels during training. However, the models consistently converge towards a trivial solution in which the input, the generated pseudo-labels and the output are identical, achieving an optimal outcome on the CCL function, registering a value of zero. CycleGN demonstrates how MLM pre-training can be leveraged to move away from this trivial path and perform actual text translation.

As a contribution to the WMT24 challenge, this study explores the efficacy of the CycleGN architectural framework in learning translation tasks across eleven language pairs under the permuted condition and four under the non-intersecting condition.

Moreover, two additional language pairs from the previous WMT edition were trained and the evaluations demonstrate the robust adaptability of CycleGN in learning translation tasks.

B.6 DLUT-GTCOM (Zong et al., 2024)

This paper presents the submission from Global Tone Communication Co., Ltd. and Dalian University of Technology for the WMT24 shared general Machine Translation (MT) task at the Conference on Empirical Methods in Natural Language Processing (EMNLP). Our participation encompasses two language pairs: English to Japanese and Japanese to Chinese. The systems are developed without particular constraints or requirements, facilitating extensive research in machine translation. We emphasize back-translation, utilize multilingual translation models, and apply fine-tuning strategies to improve performance. Additionally, we integrate both human-generated and machine-generated data to fine-tune our models, leading to enhanced translation accuracy. The automatic evaluation results indicate that our system ranks first in terms of BLEU score for the Japanese to Chinese translation.

B.7 HW-TSC (Wu et al., 2024)

This paper presents the submission of Huawei Translate Services Center (HW-TSC) to the WMT24 general machine translation (MT) shared task, where we participate in the English to Chinese (en→zh) language pair. Similar to previous years' work, we use training strategies such as regularized dropout, bidirectional training, data diversification, forward translation, back translation, alternated training, curriculum learning, and transductive ensemble learning to train the neural machine translation (NMT) model based on the deep Transformer-big architecture. The difference is that we also use continue pre-training, supervised fine-tuning, and contrastive preference optimization to train the large language model (LLM) based MT

model. By using Minimum Bayesian risk (MBR) decoding to select the final translation from multiple hypotheses for NMT and LLM-based MT models, our submission receives competitive results in the final evaluation.

B.8 IKUN and IKUN-C (Liao et al., 2024)

This paper introduces two multilingual systems, IKUN and IKUN-C, developed for the general machine translation task in WMT24. IKUN and IKUN-C represent an open system and a constrained system, respectively, built on Llama-3-8b and Mistral-7B-v0.3. Both systems are designed to handle all 11 language directions using a single model. According to automatic evaluation metrics, IKUN-C achieved 6 first-place and 3 second-place finishes among all constrained systems, while IKUN secured 1 first-place and 2 second-place finishes across both open and constrained systems. These encouraging results suggest that large language models (LLMs) are nearing the level of proficiency required for effective multilingual machine translation. The systems are based on a two-stage approach: first, continuous pre-training on monolingual data in 10 languages, followed by fine-tuning on high-quality parallel data for 11 language directions. The primary difference between IKUN and IKUN-C lies in their monolingual pre-training strategy. IKUN-C is pre-trained using constrained monolingual data, whereas IKUN leverages monolingual data from the OSCAR dataset. In the second phase, both systems are fine-tuned on parallel data sourced from NTREX, Flores, and WMT16-23 for all 11 language pairs.

B.9 IOL-Research (Zhang, 2024)

This paper illustrates the submission system of the IOL Research team for the WMT24 General Machine Translation shared task. We submitted translations for all translation directions in the general machine translation task. According to the official track categorization, our system qualifies as an open system due to the utilization of open-source resources in developing our machine translation model. With the growing prevalence of large language models (LLMs) as a conventional approach for managing diverse NLP tasks, we have developed our machine translation system by leveraging the capabilities of LLMs. Overall, We first performed continued pretraining using the open-source LLMs with tens of billions of parameters to enhance the model's multilingual capabilities. Subsequently, we employed open-source Large Language Models, equipped with hundreds of billions of parameters, to generate synthetic data. This data was then blended with a modest quantity of additional open-source data for precise supervised fine-tuning. In the final stage, we also used ensemble learning to improve translation quality.

B.10 MSLC (Larkin et al., 2024)

The MSLC (Metric Score Landscape Challenge) submissions for English–German, English–Spanish, and Japanese–Chinese are constrained systems built using Transformer models for the purpose of better evaluating metric performance in the WMT24 Metrics Task. They are intended to be representative of the performance of systems that can be built relatively simple using constrained data and with minimal modifications to the translation training pipeline.

B.11 NTTSU (Kondo et al., 2024)

The NTTSU team's submission leverages several large language models developed through a training procedure that includes continual pre-training and supervised fine-tuning. For paragraph-level translation, we generated synthetic paragraph-aligned data and utilized this data for training.

In the task of translating Japanese to Chinese, we particularly focused on the speech domain translation. Specifically, we built Whisper models for Japanese automatic speech recognition (ASR). We used YODAS dataset for Whisper training. Since this data contained many noisy data pairs, we combined the Whisper outputs using ROVER for polishing the transcriptions. Furthermore, to enhance the robustness of the translation model against errors in the transcriptions, we performed data augmentation by forward translation from audio, using both ASR and base translation models.

To select the best translation from multiple hypotheses of the models, we applied Minimum Bayes Risk decoding + reranking, incorporating scores such as COMET-QE, COMET, and cosine similarity by LaBSE.

B.12 Occiglot (Avramidis et al., 2024)

This document describes the submission of the very first version of the Occiglot open-source large language model to the General MT Shared Task of the 9th Conference of Machine Translation (WMT24). Occiglot is an open-source, community-based LLM based on Mistral-7B, which went through language-specific continual pre-training and subsequent instruction tuning, including instructions relevant to machine translation. We examine the automatic metric scores for translating the WMT24 test set and provide a detailed linguistically-motivated analysis.

Despite Occiglot performing worse than many of the other system submissions, we observe that it performs better than Mistral7B, which has been based upon, which indicates the positive effect of the language specific continual-pretraining and instruction tuning.

We see the submission of this very early version of the model as a motivation to unite community forces and pursue future LLM research on the translation task.

B.13 SCIR-MT (Li et al., 2024)

This paper introduces the submission of SCIR research center of Harbin Institute of Technology participating in the WMT24 machine translation evaluation task of constrained track for English to Czech. Our approach involved a rigorous process of cleaning and deduplicating both monolingual and bilingual data, followed by a three-stage model training recipe. During the testing phase, we used the beam serach decoding method to generate a large number of candidate translations. Furthermore, we employed COMET-MBR decoding to identify optimal translations.

B.14 Team-J (Kudo et al., 2024)

We participated in the constrained track for English-Japanese and Japanese-Chinese translations at the WMT 2024 General Machine Translation Task. Our approach was to generate a large number of sentence-level translation candidates and select the most probable translation using minimum Bayes risk (MBR) decoding and document-level large language model (LLM) re-ranking. We first generated hundreds of translation candidates from multiple translation models and retained the top 30 candidates using MBR decoding. In addition, we continually pre-trained LLMs on the target language corpora to leverage document-level information. We utilized LLMs to select the most probable sentence sequentially in context from the beginning of the document.

B.15 TranssionMT

Hyper-SNMT represents a cutting-edge approach in the field of machine translation. Hyper-SNMT is based on embedding sentences in a hyperbolic space, where distances naturally reflect language hierarchy and dependencies. This novel embedding space enables the model to achieve more accurate translations, especially for languages with complex grammatical structures and rich morphology. Both speed and accuracy are significantly improved compared to existing models. This submission is highlighting the portential of Hyper-SNMT to revolutionize the field of neural machine translation.

B.16 TSU-HITs (Mynka and Mikhaylovskiy, 2024)

This paper describes the TSU HITS team's submission system for the WMT'24 general translation task. We focused on exploring the capabilities of discrete diffusion models for the English-to-Russian, German, Czech, Spanish translation tasks in the constrained track. Our submission system consists of a set of discrete diffusion models for each language pair. The main advance is using a separate length regression model to determine the length of the output sequence more precisely.

B.17 Unbabel-Tower70B (Rei et al., 2024)

In this work, we present Tower v2, an improved iteration of the state-of-the-art open-weight Tower models, and the backbone of our submission to the WMT24 General Translation shared task. Tower v2 introduces key improvements including expanded language coverage, enhanced data quality, and increased model capacity up to 70B parameters. Our final submission combines these advancements with quality-aware decoding strategies, selecting translations based on multiple translation quality signals. The resulting

system demonstrates significant improvement over previous versions, outperforming closed commercial systems like GPT-4o, Claude 3.5, and DeepL even at a smaller 7B scale.

B.18 UvA-MT (Tan et al., 2024)

Fine-tuning Large Language Models (FT-LLMs) with parallel data has emerged as a promising paradigm in recent machine translation research. In this paper, we explore the effectiveness of FT-LLMs and compare them to traditional encoder-decoder Neural Machine Translation (NMT) systems under the WMT24 general MT shared task across three high-resource directions: English to Chinese, English to Japanese, and Japanese to Chinese. We implement several techniques, including Quality Estimation (QE) data filtering, supervised fine-tuning, and post-editing that integrate NMT systems with LLMs. We demonstrate that fine-tuning LLaMA2 on a high-quality but relatively small bitext dataset (100K) yields COMET results comparable to much smaller encoder-decoder NMT systems trained on over 22 million bitexts. However, this approach largely underperforms on surface-level metrics like BLEU and ChrF. We further control the data quality using the COMET-based quality estimation method. Our experiments show that 1) filtering low COMET scores largely improves encoder-decoder systems, but 2) no clear gains are observed for LLMs when further refining the fine-tuning set. Finally, we show that combining NMT systems with LLMs via post-editing generally yields the best performance in our experiments.

B.19 Yandex (Elshin et al., 2024)

In this paper, we present the methodology employed by the NLP team at Yandex LLC for participating in the WMT 2024 General MT Translation track, focusing on English-to-Russian translation. Our approach involves training a YandexGPT LLM-based model for translation tasks using a multi-stage process to ensure high-quality and contextually accurate translations.

Initially, we utilize a pre-trained model, trained on a large corpus of high-quality monolingual texts in various languages, crawled from various open sources, not limited to English and Russian. This extensive pre-training allows the model to capture a broad spectrum of linguistic nuances and structures. Following this, the model is fine-tuned on a substantial parallel corpus of high-quality texts collected from diverse open sources, including websites, books, and subtitles. These texts are meticulously aligned at both the sentence and paragraph levels to enhance the model's contextual understanding and translation accuracy.

In the subsequent stage, we employ p-tuning on an internal high-quality corpus of paragraph-aligned data. This step ensures that the model is finely adjusted to handle complex paragraph-level translations with greater fluency and coherence.

Next, we apply the Contrastive Pretraining Objective (CPO) method, as described in the paper CPO, using a human-annotated translation corpus. This stage focuses on refining the model's performance based on metrics evaluated at the paragraph level, emphasizing both the accuracy of the translation and the fluency of the resulting texts. The CPO method helps the model to better distinguish between subtle contextual differences, thereby improving translation quality.

In the final stage, we address the importance of preserving the content structure in translations, which is crucial for the General MT test set. To achieve this, we introduce a synthetic corpus based on web pages and video subtitles, and use it during HE markup finetune training. This encourages the model to maintain the original text's tag structure. This step ensures that the translated output retains the structural integrity of the source web pages, providing a seamless user experience.

Our multi-stage approach, combining extensive pre-training, targeted fine-tuning, advanced p-tuning, and structure-preserving techniques, ensures that our model delivers high-quality, fluent, and structurally consistent translations suitable for practical applications and competitive benchmarks.

C Translator Brief

In this project we wish to translate data from several domains for use in 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. Using post-editing would bias the evaluation, so we need to avoid it. We can detect post-editing and will reject translations that are post-edited.
- Translation should preserve the paragraph boundaries but may change number of sentences per paragraph. The source texts contain one paragraph per line and the translations should be the same.
- 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 contain errors, typos, or other problems, do not change the source sentences, instead try to prepare correct translation as if the error wouldn't be in the source.
- The data contain several domains, each folder containing one domain source.

The source files will be delivered as text files (sometimes known as "notepad" files), with one paragraph per 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 The texts are the transcriptions of audio, edited by native speakers. Each file represents one segment of audio (you are also provided with correspondent audio in WAW format). Phrases said by different speakers are located on different lines. Audios 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 text.

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 presented on different lines in the file, although 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 leave the whole line blank and let us know that you have not translated it. 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
- Marks of expressivity (e.g. asterisks *wow*, capitals letters WOW) should be conserved as best as possible. However, we recommend not to attempt to reproduce repeated characters (e.g. woooow) 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 worse). 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 have been pseudo-anonymised (e.g. @user1, @user2). These 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 as appropriate 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.

A file entitled README-social-domain-translation-notes.pdf has been distributed with the texts to translate. This file should not be translated. It contains some notes to provide additional context on the topic and terms used in some of the texts.

D Official Ranking Results (extends Section 7.4)

Results tables legend

The human score is the macro-average of human judgements, grouped by domain. The rank takes into consideration head-to-head wins and losses. AutoRank is calculated from automatic metrics.
Ranking and clustering on human scores is done using Wilcoxon signed rank test for each domain separately and final p-value is combined via Stouffer's Z-score method to align with macro average for human score.
Systems are either constrained (white), open-track (light gray), or closed-track (dark gray).
LLMs that do not officially claim a support a language pair are marked with §.

Human scores for individual domains are marked by an up arrow ↑ if their difference from the system domain score is larger than the standard deviation over all domains for given system (row) and down arrow ↓ indicates that the domain score is worse than the overall score.
Underlined domain scores indicate that the domain score is better than the domain score of system above it (of a better ranked system).

Czech→Ukrainian

Rank	System	Human	AutoRank	CometKiwi	MetricX	education	news	official	personal	voice
1-2	Claude-3.5 §	93.0	1.7	-0.7	1.0	↓ 90.4	91.7	↑ 95.3	↑ 95.4	92.2
2-2	HUMAN-A	92.7	-	-	-	92.6	93.0	92.0	↑ 94.9	↓ 91.1
3-3	Gemini-1.5-Pro	92.6	2.0	-0.7	1.2	↓ 88.6	94.7	94.5	93.6	91.9
3-4	Unbabel-Tower70B	92.2	1.0	-0.7	0.9	↓ 86.8	93.5	94.8	94.1	91.8
5-5	IOL-Research	90.2	1.9	-0.7	1.3	↓ 80.8	89.9	92.7	94.6	93.0
6-7	CommandR-plus §	89.7	1.9	-0.7	1.3	↓ 83.4	89.6	↑ 93.8	92.1	89.4
6-8	ONLINE-W	88.7	2.3	-0.7	1.4	↓ 84.4	89.4	87.9	↑ 91.3	90.4
7-9	GPT-4 §	88.6	2.0	-0.7	1.4	↓ 83.2	87.9	89.0	↑ 92.4	90.3
8-9	IKUN	87.1	2.3	-0.7	1.6	↓ 77.6	86.8	89.7	91.2	90.3
10-10	Aya23	86.6	2.5	-0.7	1.9	↓ 77.4	91.1	88.5	87.6	88.3
11-11	CUNI-Transformer	85.3	3.0	-0.6	2.0	↓ 83.2	85.2	84.8	↑ 88.0	85.3
12-12	IKUN-C	82.6	3.0	-0.6	2.4	79.6	↓ 70.0	87.2	88.4	87.8
	Mistral-Large §	-	2.3	-	-	-	-	-	-	-
	TranssionMT	-	2.6	-	-	-	-	-	-	-
	ONLINE-B	-	2.6	-	-	-	-	-	-	-
	ONLINE-A	-	2.6	-	-	-	-	-	-	-
	Llama3-70B §	-	2.6	-	-	-	-	-	-	-
	ONLINE-G	-	2.8	-	-	-	-	-	-	-
	Phi-3-Medium §	-	9.1	-	-	-	-	-	-	-
	BJFU-LPT	-	11.5	-	-	-	-	-	-	-
	CycleL	-	21.0	-	-	-	-	-	-	-

English→Czech

Rank	System	Human	AutoRank	CometKiwi	MetricX	literary	news	social	speech
1-2	HUMAN-A	92.9	-	-	-	93.1	↑ 94.5	92.0	92.1
2-2	Unbabel-Tower70B	91.6	1.0	-0.7	1.8	91.7	94.1	93.3	↓ 87.5
2-3	Claude-3.5 §	91.2	2.1	-0.7	2.4	91.2	↑ 94.9	91.6	↓ 87.2
4-5	ONLINE-W	89.0	2.8	-0.7	2.8	91.0	↑ 92.1	88.2	↓ 84.9
4-6	CUNI-MH	88.4	2.1	-0.7	2.3	89.7	↑ 91.9	88.0	↓ 84.1
6-6	Gemini-1.5-Pro	88.2	2.6	-0.7	2.8	88.6	89.3	↓ 85.2	89.6
6-8	GPT-4 §	87.7	2.6	-0.7	2.9	↓ 85.2	89.5	↑ 90.1	86.1
8-8	CommandR-plus §	86.9	2.9	-0.7	2.9	↓ 85.2	87.5	↑ 88.6	86.2
8-9	IOL-Research	86.5	2.8	-0.7	3.0	84.7	↑ 90.4	86.3	84.5
10-11	SCIR-MT	85.4	3.2	-0.7	3.3	85.0	↑ 92.4	82.2	82.1
10-11	CUNI-DocTransformer	84.3	4.4	-0.6	4.0	83.1	↑ 90.7	80.9	82.4
12-12	Aya23	84.2	4.3	-0.6	4.0	81.6	↑ 89.9	84.9	↓ 80.3
13-13	CUNI-GA	82.1	2.3	-0.7	3.7	82.8	↑ 88.5	81.7	↓ 75.3
14-14	IKUN	81.7	3.9	-0.6	3.7	80.2	↑ 87.0	82.2	↓ 77.5
15-15	Llama3-70B §	77.4	4.1	-0.6	4.0	↓ 65.4	83.0	82.4	78.8
16-16	IKUN-C	75.4	4.7	-0.6	4.3	↓ 70.5	77.7	77.5	75.7
	TranssionMT	-	3.5	-	-	-	-	-	-
	ONLINE-A	-	3.6	-	-	-	-	-	-
	Mistral-Large §	-	3.7	-	-	-	-	-	-
	ONLINE-B	-	4.0	-	-	-	-	-	-
	CUNI-Transformer	-	4.7	-	-	-	-	-	-
	ONLINE-G	-	5.7	-	-	-	-	-	-
	NVIDIA-NeMo	-	7.6	-	-	-	-	-	-
	Phi-3-Medium §	-	15.0	-	-	-	-	-	-
	TSU-HITs	-	19.5	-	-	-	-	-	-
	CycleL2	-	24.2	-	-	-	-	-	-
	CycleL	-	27.0	-	-	-	-	-	-

English→German

Rank	System	Human	AutoRank	CometKiwi	MetricX	literary	news	social	speech
1-11	GPT-4	-1.6	1.8	-0.7	1.4	-0.7	-1.4	-0.9	↓ -3.6
1-7	Dubformer	-1.8	1.8	-0.7	1.2	-1.2	-1.3	-0.6	↓ -4.2
2-10	ONLINE-B	-1.9	1.8	-0.7	1.4	-1.3	-1.5	-1.2	↓ -3.6
2-10	TranssionMT	-1.9	1.8	-0.7	1.4	-1.3	-1.2	-1.2	↓ -3.9
2-9	Unbabel-Tower70B	-1.9	1.0	-0.7	1.1	-1.4	-2.0	↑ -0.8	↓ -3.5
1-9	HUMAN-B	-2.0	-	-	-	-0.8	-1.4	-0.8	↓ -4.9
2-12	Mistral-Large	-2.1	2.0	-0.7	1.5	-1.5	-1.9	-1.1	↓ -3.9
4-11	CommandR-plus	-2.3	2.0	-0.7	1.4	-1.7	-2.4	↑ -1.1	↓ -3.9
8-10	ONLINE-W	-2.3	2.2	-0.7	1.5	-2.1	-1.3	-1.7	↓ -4.1
2-12	Claude-3.5	-2.4	1.9	-0.7	1.4	-1.1	-1.0	-1.2	↓ -6.0
3-13	HUMAN-A	-2.5	-	-	-	-2.0	-1.8	-1.0	↓ -5.0
10-12	IOL-Research	-2.5	2.3	-0.7	1.6	-2.0	-1.7	-1.6	↓ -4.9
5-13	Gemini-1.5-Pro	-2.8	2.2	-0.7	1.5	↓ -5.0	↑ -1.3	-1.9	-2.9
14-15	Aya23	-3.2	2.7	-0.7	1.8	-2.3	-2.7	-2.2	↓ -5.7
14-17	ONLINE-A	-3.5	3.0	-0.7	1.8	-2.8	-1.9	-2.3	↓ -6.9
15-17	Llama3-70B §	-4.3	2.5	-0.7	1.7	-4.8	-2.9	↑ -2.3	↓ -7.1
15-17	IKUN	-4.3	3.0	-0.7	1.8	-3.5	-4.3	↑ -2.4	↓ -7.1
18-18	IKUN-C	-6.1	3.8	-0.6	2.0	-7.6	-3.4	-3.3	↓ -9.9
19-19	MSLC	-15.5	11.9	-0.4	4.4	-15.3	-11.5	↑ -8.2	↓ -26.8
	Phi-3-Medium §	-	3.4	-	-	-	-	-	-
	ONLINE-G	-	3.5	-	-	-	-	-	-
	CUNI-NL	-	4.2	-	-	-	-	-	-
	AIST-AIRC	-	7.2	-	-	-	-	-	-
	NVIDIA-NeMo	-	7.4	-	-	-	-	-	-
	Occiglot	-	8.2	-	-	-	-	-	-
	TSU-HITs	-	13.3	-	-	-	-	-	-
	CycleL2	-	27.0	-	-	-	-	-	-
	CycleL	-	27.0	-	-	-	-	-	-

English→Spanish

Rank	System	Human	AutoRank	CometKiwi	MetricX	literary	news	social	speech
1-1	HUMAN-A	95.3	-	-	-	95.2	↑ 96.2	95.5	↓ 94.1
2-2	Dubformer	93.4	2.0	-0.7	2.2	95.3	94.5	94.4	↓ 89.4
3-4	GPT-4	91.9	1.9	-0.7	2.5	93.5	94.0	93.2	↓ 87.0
4-7	IOL-Research	91.4	2.3	-0.7	2.8	↑ 96.3	92.5	90.9	↓ 86.0
5-8	Mistral-Large	89.3	2.2	-0.7	2.7	90.5	90.4	91.0	↓ 85.2
5-9	Unbabel-Tower70B	88.9	1.0	-0.7	1.9	86.2	↑ 93.7	91.1	↓ 84.6
3-8	Claude-3.5	88.8	2.1	-0.7	2.6	91.5	92.8	90.4	↓ 80.5
5-8	Gemini-1.5-Pro	88.8	2.4	-0.7	2.8	89.6	↑ 92.3	87.0	↓ 86.2
7-9	CommandR-plus	88.3	2.1	-0.7	2.6	88.2	89.3	↑ 90.8	↓ 84.8
9-10	Llama3-70B §	87.2	2.6	-0.7	3.0	↑ 89.4	87.1	87.9	↓ 84.2
11-11	ONLINE-B	85.6	2.7	-0.7	3.1	87.4	88.6	86.8	↓ 79.4
12-13	IKUN	84.7	2.8	-0.7	3.3	85.4	↑ 92.4	82.8	↓ 78.3
12-13	IKUN-C	80.4	3.4	-0.7	3.5	83.3	↑ 85.6	79.0	↓ 73.6
14-14	MSLC	63.9	7.4	-0.5	6.4	59.3	↑ 78.8	55.9	61.7
	ONLINE-W	-	2.7	-	-	-	-	-	-
	TranssionMT	-	2.8	-	-	-	-	-	-
	Phi-3-Medium §	-	3.0	-	-	-	-	-	-
	ONLINE-A	-	3.0	-	-	-	-	-	-
	Aya23	-	3.1	-	-	-	-	-	-
	ONLINE-G	-	3.2	-	-	-	-	-	-
	NVIDIA-NeMo	-	4.5	-	-	-	-	-	-
	Occiglot	-	5.9	-	-	-	-	-	-
	TSU-HITs	-	16.3	-	-	-	-	-	-
	CycleL	-	24.0	-	-	-	-	-	-

English→Hindi

Rank	System	Human	AutoRank	CometKiwi	MetricX	literary	news	social	speech
1-3	TranssionMT	91.3	1.3	-0.6	3.3	↑ 94.0	93.0	89.8	↓ 88.2
1-4	Unbabel-Tower70B	90.5	1.0	-0.7	3.1	90.9	↑ 92.7	90.7	↓ 87.7
3-3	Claude-3.5 §	90.2	1.2	-0.6	3.3	95.4	93.6	91.0	↓ 81.1
3-4	ONLINE-B	90.1	1.4	-0.6	3.3	91.8	90.4	91.3	↓ 86.9
3-5	Gemini-1.5-Pro §	90.0	1.6	-0.6	3.6	90.3	↑ 91.9	89.4	↓ 88.3
6-6	GPT-4 §	88.5	2.1	-0.6	4.5	89.9	90.4	89.2	↓ 84.4
7-8	HUMAN-A	88.5	-	-	-	88.8	↓ 88.1	↑ 88.9	88.2
8-8	IOL-Research	87.2	2.1	-0.6	4.3	87.2	↑ 88.9	87.7	↓ 84.9
8-9	Llama3-70B §	86.7	2.1	-0.6	4.6	86.4	87.1	↓ 86.1	87.1
10-10	Aya23	84.7	3.2	-0.6	5.4	83.3	↑ 86.9	↓ 83.1	85.7
11-11	IKUN-C	70.7	5.5	-0.5	7.1	71.2	↓ 59.2	↑ 80.2	72.4
	CommandR-plus §	-	2.3	-	-	-	-	-	-
	ONLINE-A	-	3.5	-	-	-	-	-	-
	ONLINE-G	-	4.2	-	-	-	-	-	-
	Mistral-Large §	-	5.0	-	-	-	-	-	-
	NVIDIA-NeMo	-	5.8	-	-	-	-	-	-
	Phi-3-Medium §	-	7.4	-	-	-	-	-	-
	IKUN	-	7.7	-	-	-	-	-	-
	ONLINE-empty	-	15.3	-	-	-	-	-	-
	CycleL	-	20.0	-	-	-	-	-	-

English→Icelandic

Rank	System	Human	AutoRank	CometKiwi	MetricX	literary	news	social	speech
1-1	HUMAN-A	93.1	-	-	-	92.2	92.6	↑ 95.0	92.4
2-3	Dubformer	84.3	2.5	-0.7	3.4	84.1	83.1	↑ 87.5	82.5
2-3	Claude-3.5 §	81.9	2.3	-0.7	3.6	80.2	83.9	↑ 87.2	↓ 76.4
4-4	Unbabel-Tower70B	80.2	1.0	-0.7	2.5	↓ 76.6	80.6	↑ 84.3	79.2
5-5	AMI	73.3	3.7	-0.7	4.9	↑ 75.2	72.8	74.1	↓ 71.1
6-6	IKUN	71.0	3.2	-0.7	4.3	↓ 66.8	↑ 74.7	73.6	69.1
7-7	ONLINE-B	68.0	4.2	-0.7	5.5	70.5	↓ 59.4	↑ 74.0	67.9
8-9	GPT-4	66.3	3.4	-0.7	4.7	66.5	65.5	↑ 69.5	↓ 63.9
8-9	IKUN-C	65.2	3.7	-0.7	4.9	↓ 59.6	68.2	↑ 69.3	63.8
10-10	IOL-Research	58.0	4.3	-0.7	5.7	↓ 49.4	59.6	61.4	61.4
11-11	Llama3-70B §	41.0	6.7	-0.6	8.0	39.8	40.0	↑ 44.0	40.3
	TranssionMT	-	4.2	-	-	-	-	-	-
	ONLINE-A	-	5.5	-	-	-	-	-	-
	ONLINE-G	-	6.9	-	-	-	-	-	-
	CommandR-plus §	-	9.8	-	-	-	-	-	-
	Mistral-Large §	-	10.4	-	-	-	-	-	-
	Aya23 §	-	15.2	-	-	-	-	-	-
	Phi-3-Medium §	-	16.2	-	-	-	-	-	-
	ONLINE-empty	-	18.1	-	-	-	-	-	-
	TSU-HITs	-	19.2	-	-	-	-	-	-
	CycleL	-	21.0	-	-	-	-	-	-

English→Japanese

Rank	System	Human	AutoRank	CometKiwi	MetricX	literary	news	social	speech
1-1	HUMAN-A	91.8	-	-	-	92.4	93.0	↓ 89.5	92.4
2-4	ONLINE-B	91.1	1.4	-0.8	2.4	91.7	↑ 92.6	91.1	↓ 88.9
3-4	CommandR-plus	91.0	1.9	-0.7	2.7	92.2	↑ 93.7	89.5	↓ 88.5
4-4	GPT-4	90.8	1.7	-0.7	2.7	↑ 91.9	91.3	↓ 89.9	90.1
4-5	Claude-3.5	90.8	1.5	-0.7	2.3	91.4	↑ 92.8	91.3	↓ 87.6
4-7	Gemini-1.5-Pro	90.0	1.7	-0.7	2.5	91.1	↑ 92.2	↓ 88.1	88.7
7-7	Unbabel-Tower70B	89.7	1.0	-0.8	2.0	↓ 88.2	↑ 91.6	89.8	89.2
8-8	IOL-Research	89.7	2.3	-0.7	3.1	91.0	90.6	90.3	↓ 86.9
8-9	Aya23	89.7	2.3	-0.7	3.1	90.1	↑ 92.1	88.4	↓ 87.9
10-10	NTTSU	89.4	1.9	-0.7	2.6	90.0	↑ 93.2	88.4	↓ 86.2
11-11	Team-J	88.5	1.9	-0.7	2.9	↓ 85.0	90.1	↑ 91.3	87.5
12-12	Llama3-70B §	86.8	2.6	-0.7	3.5	89.3	↑ 89.8	85.2	↓ 82.7
13-13	IKUN-C	81.7	3.9	-0.7	4.3	↓ 77.5	↑ 88.5	81.2	79.8
	DLUT-GTCOM	-	2.6	-	-	-	-	-	-
	Phi-3-Medium §	-	2.8	-	-	-	-	-	-
	ONLINE-W	-	2.9	-	-	-	-	-	-
	Mistral-Large §	-	2.9	-	-	-	-	-	-
	ONLINE-A	-	3.0	-	-	-	-	-	-
	IKUN	-	3.1	-	-	-	-	-	-
	ONLINE-G	-	6.4	-	-	-	-	-	-
	AIST-AIRC	-	6.6	-	-	-	-	-	-
	UvA-MT	-	6.7	-	-	-	-	-	-
	NVIDIA-NeMo	-	6.9	-	-	-	-	-	-
	CycleL	-	24.0	-	-	-	-	-	-

English→Russian

Rank	System	Human	AutoRank	CometKiwi	MetricX	literary	news	social	speech
1-1	HUMAN-A	89.2	-	-	-	↑ 94.0	88.3	87.7	86.6
2-3	Dubformer	89.1	1.9	-0.7	2.8	90.7	88.5	↑ 92.1	↓ 84.9
3-4	Claude-3.5	88.2	2.0	-0.7	3.0	↑ 94.1	93.1	85.7	↓ 80.0
3-5	Unbabel-Tower70B	88.1	1.0	-0.7	2.4	87.5	91.2	90.6	↓ 83.2
3-7	Yandex	87.0	1.9	-0.7	2.9	89.6	↑ 91.8	84.5	↓ 82.0
6-8	Gemini-1.5-Pro	85.5	2.3	-0.7	3.2	↑ 90.7	84.9	83.4	82.9
6-9	GPT-4	85.0	2.3	-0.7	3.4	↑ 89.3	85.4	84.6	↓ 80.7
6-9	ONLINE-G	84.6	2.2	-0.7	3.3	88.3	88.8	84.6	↓ 76.6
5-9	CommandR-plus §	84.3	2.4	-0.7	3.4	86.7	84.5	85.7	↓ 80.5
10-10	IOL-Research	82.1	2.6	-0.7	3.7	84.8	86.4	84.2	↓ 73.1
11-11	IKUN	79.2	3.2	-0.7	4.1	80.2	↑ 87.2	78.5	↓ 70.9
12-12	Aya23	78.6	3.3	-0.7	4.2	77.8	↑ 82.9	78.5	↓ 75.3
13-13	Llama3-70B §	75.7	3.1	-0.7	4.1	77.0	↑ 80.1	76.3	↓ 69.5
14-14	IKUN-C	69.8	3.9	-0.6	4.7	65.1	↑ 78.3	72.9	↓ 62.6
	ONLINE-W	-	2.6	-	-	-	-	-	-
	Mistral-Large §	-	2.7	-	-	-	-	-	-
	ONLINE-B	-	3.1	-	-	-	-	-	-
	TranssionMT	-	3.1	-	-	-	-	-	-
	ONLINE-A	-	3.4	-	-	-	-	-	-
	Phi-3-Medium §	-	3.9	-	-	-	-	-	-
	CUNI-DS	-	5.9	-	-	-	-	-	-
	NVIDIA-NeMo	-	7.2	-	-	-	-	-	-
	TSU-HITs	-	10.8	-	-	-	-	-	-
	CycleL	-	24.3	-	-	-	-	-	-
	CycleL2	-	25.0	-	-	-	-	-	-

English→Ukrainian

Rank	System	Human	AutoRank	CometKiwi	MetricX	literary	news	social	speech
1-2	Claude-3.5	90.5	2.0	-0.7	3.0	93.2	93.9	92.2	↓ 82.7
1-2	Unbabel-Tower70B	89.8	1.0	-0.7	2.2	92.5	92.8	91.1	↓ 82.9
3-3	Dubformer	89.0	1.8	-0.7	2.7	↓ 84.4	91.3	↑ 94.3	85.9
4-6	HUMAN-A	87.3	-	-	-	89.6	↑ 91.5	↓ 83.8	84.1
4-6	Gemini-1.5-Pro	87.1	2.2	-0.7	3.0	↑ 90.1	88.8	85.3	↓ 84.4
5-8	ONLINE-W	86.0	2.1	-0.7	2.8	86.7	↑ 88.9	86.8	↓ 81.8
5-9	GPT-4	84.6	2.3	-0.7	3.3	81.2	↑ 90.3	84.5	82.4
6-9	CommandR-plus §	83.2	2.3	-0.7	3.2	79.6	↑ 89.1	83.6	80.4
7-9	IOL-Research	83.2	2.4	-0.7	3.4	80.6	↑ 90.2	83.1	↓ 78.8
10-10	IKUN	78.4	2.8	-0.7	3.7	83.2	↑ 88.2	72.7	↓ 69.7
11-11	IKUN-C	67.9	3.9	-0.6	4.7	↓ 65.2	69.0	68.3	69.2
	ONLINE-G	-	2.3	-	-	-	-	-	-
	Mistral-Large §	-	2.4	-	-	-	-	-	-
	ONLINE-B	-	3.1	-	-	-	-	-	-
	TranssionMT	-	3.1	-	-	-	-	-	-
	Llama3-70B §	-	3.2	-	-	-	-	-	-
	Aya23	-	3.3	-	-	-	-	-	-
	ONLINE-A	-	3.3	-	-	-	-	-	-
	NVIDIA-NeMo	-	6.2	-	-	-	-	-	-
	Phi-3-Medium §	-	11.1	-	-	-	-	-	-
	CycleL	-	21.0	-	-	-	-	-	-

English→Chinese

Rank	System	Human	AutoRank	CometKiwi	MetricX	literary	news	social	speech
1-1	GPT-4	89.6	2.0	-0.7	3.3	88.7	↑ 91.2	90.3	↓ 88.4
2-4	Unbabel-Tower70B	89.6	1.0	-0.7	2.3	90.0	↑ 92.3	90.2	↓ 85.8
2-4	HUMAN-A	89.4	-	-	-	89.9	90.1	90.7	↓ 86.8
4-4	Gemini-1.5-Pro	89.3	1.8	-0.7	3.1	92.0	↑ 92.5	↓ 85.2	87.5
5-6	ONLINE-B	89.3	1.7	-0.7	2.9	↑ 91.9	89.7	90.3	↓ 85.0
6-6	IOL-Research	89.0	1.8	-0.7	3.1	91.0	90.8	88.3	↓ 86.1
6-7	Claude-3.5	88.9	1.7	-0.7	3.0	92.0	90.8	89.5	↓ 83.4
6-8	CommandR-plus	88.3	2.2	-0.7	3.3	85.9	↑ 90.8	90.4	85.9
9-9	Llama3-70B §	86.5	2.8	-0.7	3.9	87.5	86.8	87.0	↓ 84.6
10-10	HW-TSC	86.2	2.3	-0.7	3.4	87.1	↑ 91.5	84.9	↓ 81.4
11-11	IKUN	85.3	3.1	-0.6	4.0	88.6	↑ 89.1	82.1	↓ 81.5
12-12	Aya23	85.2	3.0	-0.7	4.1	85.4	↑ 88.3	85.5	↓ 81.7
13-13	IKUN-C	82.1	3.5	-0.6	4.2	81.0	↑ 85.9	83.1	↓ 78.6
	ONLINE-W	-	2.2	-	-	-	-	-	-
	Mistral-Large §	-	2.8	-	-	-	-	-	-
	Phi-3-Medium §	-	3.1	-	-	-	-	-	-
	ONLINE-A	-	3.3	-	-	-	-	-	-
	UvA-MT	-	4.3	-	-	-	-	-	-
	ONLINE-G	-	4.8	-	-	-	-	-	-
	NVIDIA-NeMo	-	7.3	-	-	-	-	-	-
	CycleL	-	20.1	-	-	-	-	-	-
	CycleL2	-	22.0	-	-	-	-	-	-

Japanese→Chinese

Rank	System	Human	AutoRank	CometKiwi	MetricX	literary	news	speech
1-3	Claude-3.5	-1.4	1.7	-0.6	3.5	-0.5	-0.8	↓ -3.0
1-3	HUMAN-A	-1.5	-	-	-	-0.7	-0.8	↓ -3.2
3-5	GPT-4	-1.7	2.1	-0.6	3.8	-1.0	-0.8	↓ -3.2
2-5	DLUT-GTCOM	-1.7	2.0	-0.6	3.3	-0.5	-1.1	↓ -3.7
4-8	Unbabel-Tower70B	-1.9	1.0	-0.6	3.2	-1.0	-1.2	↓ -3.5
3-6	Gemini-1.5-Pro	-2.1	1.9	-0.6	3.5	-1.6	-0.8	↓ -3.8
6-8	CommandR-plus	-2.2	2.8	-0.6	4.1	-0.7	-1.3	↓ -4.6
6-8	IOL-Research	-2.4	2.2	-0.6	3.9	-1.4	-1.1	↓ -4.8
9-10	Llama3-70B §	-3.4	3.1	-0.6	4.7	-2.0	-2.2	↓ -6.2
9-10	Aya23	-3.5	3.7	-0.6	5.0	-2.1	-1.9	↓ -6.4
11-12	Team-J	-4.5	2.8	-0.6	4.0	-3.1	-2.0	↓ -8.5
11-12	NTTSU	-5.1	3.7	-0.6	5.3	-2.8	-2.1	↓ -10.5
13-13	ONLINE-B	-5.8	5.2	-0.5	5.5	-4.2	-3.7	↓ -9.5
14-14	IKUN-C	-7.7	5.5	-0.5	6.2	-5.1	-3.4	↓ -14.4
15-15	MSLC	-10.7	8.9	-0.5	8.8	-9.1	↑ -4.0	↓ -19.0
	Mistral-Large §	-	3.5	-	-	-	-	-
	Phi-3-Medium §	-	4.0	-	-	-	-	-
	IKUN	-	4.4	-	-	-	-	-
	UvA-MT	-	5.2	-	-	-	-	-
	ONLINE-W	-	5.3	-	-	-	-	-
	ONLINE-A	-	6.8	-	-	-	-	-
	ONLINE-G	-	10.3	-	-	-	-	-
	CycleL	-	23.0	-	-	-	-	-

E Head to head comparisons

Following tables show differences in average human scores for each language pair. The number in each of cell shows the difference in average human scores for the systems in the column and row.

Because there are many systems and data conditions, the significance of each pairwise comparison needs to be quantified. We apply Wilcoxon signed-rank test to measure the likelihood that such differences could occur simply by chance. In the following tables $\star$ indicates statistical significance at $p < 0.05$, † indicates statistical significance at $p < 0.01$, and ‡ indicates statistical significance at $p < 0.001$.

Each table contains final rows showing the macro-average score achieved by that system and the rank range. Gray lines separate clusters based on non-overlapping rank ranges.

Head to head comparison for Czech→Ukrainian systems

	Claude-3.5	refA	Gemini-1.5-Pro	Unbabel-Tower70B	IOL-Research	CommandR-plus	ONLINE-W	GPT-4	IKUN	Aya23	CUNI-Transformer	IKUN-C
Claude-3.5	–	0.3†	0.4‡	0.8	2.8‡	3.3‡	4.3‡	4.4‡	5.9‡	6.4‡	7.7‡	10.4‡
refA	–	–	0.1‡	0.5†	2.5‡	3.0‡	4.0‡	4.2‡	5.6‡	6.1‡	7.4‡	10.1‡
Gemini-1.5-Pro	–	–	–	0.4⋆	2.4‡	3.0‡	4.0‡	4.1‡	5.5‡	6.1‡	7.3‡	10.1‡
Unbabel-Tower70B	–	–	–	–	2.0‡	2.5‡	3.5‡	3.6‡	5.1‡	5.6‡	6.9‡	9.6‡
IOL-Research	–	–	–	–	–	0.5†	1.5‡	1.6‡	3.1‡	3.6‡	4.9‡	7.6‡
CommandR-plus	–	–	–	–	–	–	1.0	1.1⋆	2.5‡	3.1‡	4.4‡	7.1‡
ONLINE-W	–	–	–	–	–	–	–	0.1	1.6‡	2.1‡	3.4‡	6.1‡
GPT-4	–	–	–	–	–	–	–	–	1.4	2.0‡	3.3‡	6.0‡
IKUN	–	–	–	–	–	–	–	–	–	0.6‡	1.8‡	4.5‡
Aya23	–	–	–	–	–	–	–	–	–	–	1.3‡	4.0†
CUNI-Transformer	–	–	–	–	–	–	–	–	–	–	–	2.7‡
IKUN-C	–	–	–	–	–	–	–	–	–	–	–	–
Scores	93.0	92.7	92.6	92.2	90.2	89.7	88.7	88.6	87.1	86.6	85.3	82.6
Ranks	1-2	2-2	3-3	3-4	5-5	6-7	6-8	7-9	8-9	10-10	11-11	12-12

Head to head comparison for English→Czech systems

	refA	Unbabel-Tower70B	Claude-3.5	ONLINE-W	CUNI-MH	Gemini-1.5-Pro	GPT-4	CommandR-plus	IOL-Research	SCIR-MT	CUNI-DocTransformer	Aya23	CUNI-GA	IKUN	Llama3-70B	IKUN-C
refA	–	1.3‡	1.7	3.9‡	4.5‡	4.7‡	5.2‡	6.0‡	6.4‡	7.5‡	8.7‡	8.8‡	10.8‡	11.2‡	15.5‡	17.5‡
Unbabel-Tower70B	–	–	0.4†	2.6†	3.2‡	3.4‡	3.9‡	4.7‡	5.2‡	6.2‡	7.4‡	7.5‡	9.5‡	9.9‡	14.2‡	16.3‡
Claude-3.5	–	–	–	2.2‡	2.8‡	3.0‡	3.5‡	4.3‡	4.7‡	5.8‡	6.9‡	7.0‡	9.1‡	9.5‡	13.8‡	15.8‡
ONLINE-W	–	–	–	–	0.6	0.8‡	1.3†	2.1‡	2.6‡	3.6‡	4.8‡	4.9‡	6.9‡	7.3‡	11.6‡	13.7‡
CUNI-MH	–	–	–	–	–	0.2‡	0.7	1.5‡	2.0†	3.0‡	4.2‡	4.3‡	6.3‡	6.7‡	11.0‡	13.1‡
Gemini-1.5-Pro	–	–	–	–	–	–	0.4‡	1.3‡	1.7‡	2.8‡	3.9‡	4.0‡	6.1‡	6.5‡	10.8‡	12.8‡
GPT-4	–	–	–	–	–	–	–	0.9⋆	1.3	2.3‡	3.5‡	3.6‡	5.7‡	6.0‡	10.3‡	12.4‡
CommandR-plus	–	–	–	–	–	–	–	–	0.4‡	1.5‡	2.6‡	2.7‡	4.8‡	5.2‡	9.5‡	11.5‡
IOL-Research	–	–	–	–	–	–	–	–	–	1.0†	2.2†	2.3⋆	4.4‡	4.7‡	9.1‡	11.1‡
SCIR-MT	–	–	–	–	–	–	–	–	–	–	1.2	1.3‡	3.3‡	3.7‡	8.0‡	10.1‡
CUNI-DocTransformer	–	–	–	–	–	–	–	–	–	–	–	0.1‡	2.2‡	2.5‡	6.9‡	8.9‡
Aya23	–	–	–	–	–	–	–	–	–	–	–	–	2.1‡	2.4†	6.8‡	8.8‡
CUNI-GA	–	–	–	–	–	–	–	–	–	–	–	–	–	0.4†	4.7‡	6.7‡
IKUN	–	–	–	–	–	–	–	–	–	–	–	–	–	–	4.3‡	6.4‡
Llama3-70B	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2.0‡
IKUN-C	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Scores	92.9	91.6	91.2	89.0	88.4	88.2	87.7	86.9	86.5	85.4	84.3	84.2	82.1	81.7	77.4	75.4
Ranks	1-2	2-2	2-3	4-5	4-6	6-6	6-8	8-8	8-9	10-11	10-11	12-12	13-13	14-14	15-15	16-16

Head to head comparison for English→German systems

	GPT-4	Dubformer	ONLINE-B	TranssionMT	Unbabel-Tower70B	refB	Mistral-Large	CommandR-plus	ONLINE-W	Claude-3.5	refA	IOL-Research	Gemini-1.5-Pro	Aya23	ONLINE-A	Llama3-70B	IKUN	IKUN-C	MSLC
GPT-4	–	0.2	0.3	0.3	0.3	0.3	0.5	0.6	0.7†	0.7	0.8	0.9‡	1.1	1.6‡	1.8‡	2.6‡	2.7‡	4.4‡	13.8‡
Dubformer	–	–	0.1⋆	0.1⋆	0.1	0.2	0.3	0.5†	0.5⋆	0.5	0.6	0.7‡	1.0⋆	1.4‡	1.7‡	2.4‡	2.5‡	4.2‡	13.7‡
ONLINE-B	–	–	–	0.0	0.0	0.1	0.2	0.4⋆	0.4⋆	0.5	0.6	0.6‡	0.9	1.3‡	1.6‡	2.4‡	2.4‡	4.2‡	13.6‡
TranssionMT	–	–	–	–	0.0†	0.1	0.2	0.4	0.4‡	0.4	0.6	0.6‡	0.9	1.3‡	1.6‡	2.4‡	2.4‡	4.2‡	13.6‡
Unbabel-Tower70B	–	–	–	–	–	0.1	0.2	0.4	0.4†	0.4	0.6⋆	0.6‡	0.9⋆	1.3‡	1.6‡	2.3‡	2.4‡	4.2‡	13.6‡
refB	–	–	–	–	–	–	0.1⋆	0.3‡	0.3‡	0.4	0.5	0.6‡	0.8	1.2‡	1.5‡	2.3‡	2.3‡	4.1‡	13.5‡
Mistral-Large	–	–	–	–	–	–	–	0.2	0.2	0.2	0.4	0.4⋆	0.7	1.1‡	1.4‡	2.1‡	2.2‡	3.9‡	13.4‡
CommandR-plus	–	–	–	–	–	–	–	–	0.0⋆	0.1	0.2	0.3⋆	0.5	0.9‡	1.2‡	2.0‡	2.0‡	3.8‡	13.2‡
ONLINE-W	–	–	–	–	–	–	–	–	–	0.0†	0.1⋆	0.2	0.5‡	0.9‡	1.2‡	1.9‡	2.0‡	3.7‡	13.1‡
Claude-3.5	–	–	–	–	–	–	–	–	–	–	0.1	0.2‡	0.4	0.9‡	1.1‡	1.9‡	2.0‡	3.7‡	13.1‡
refA	–	–	–	–	–	–	–	–	–	–	–	0.1	0.3	0.8‡	1.0‡	1.8‡	1.9‡	3.6‡	13.0‡
IOL-Research	–	–	–	–	–	–	–	–	–	–	–	–	0.2‡	0.7⋆	0.9⋆	1.7‡	1.8‡	3.5‡	12.9‡
Gemini-1.5-Pro	–	–	–	–	–	–	–	–	–	–	–	–	–	0.4‡	0.7‡	1.5‡	1.5‡	3.3‡	12.7‡
Aya23	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.3	1.0⋆	1.1†	2.8‡	12.3‡
ONLINE-A	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.8	0.8	2.6‡	12.0‡
Llama3-70B	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.1	1.8‡	11.2‡
IKUN	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.7‡	11.1‡
IKUN-C	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	9.4‡
MSLC	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Scores	-1.6	-1.8	-1.9	-1.9	-1.9	-2.0	-2.1	-2.3	-2.3	-2.4	-2.5	-2.5	-2.8	-3.2	-3.5	-4.3	-4.3	-6.1	-15.5
Ranks	1-11	1-7	2-10	2-10	2-9	1-9	2-12	4-11	8-10	2-12	3-13	10-12	5-13	14-15	14-17	15-17	15-17	18-18	19-19

Head to head comparison for English→Spanish systems

	refA	Dubformer	GPT-4	IOL-Research	Mistral-Large	Unbabel-Tower70B	Claude-3.5	Gemini-1.5-Pro	CommandR-plus	Llama3-70B	ONLINE-B	IKUN	IKUN-C	MSLC
refA	–	1.9†	3.3‡	3.8‡	6.0‡	6.3‡	6.5‡	6.5‡	7.0‡	8.1‡	9.7‡	10.5‡	14.9‡	31.3‡
Dubformer	–	–	1.5⋆	2.0†	4.1‡	4.5‡	4.6‡	4.6‡	5.1‡	6.2‡	7.8‡	8.7‡	13.0‡	29.5‡
GPT-4	–	–	–	0.5‡	2.7†	3.0†	3.1	3.2⋆	3.6‡	4.8‡	6.4‡	7.2‡	11.6‡	28.0‡
IOL-Research	–	–	–	–	2.2†	2.5	2.7	2.7	3.2†	4.3‡	5.9‡	6.7‡	11.1‡	27.5‡
Mistral-Large	–	–	–	–	–	0.4⋆	0.5	0.5⋆	1.0	2.1	3.7†	4.5‡	8.9‡	25.4‡
Unbabel-Tower70B	–	–	–	–	–	–	0.1	0.1	0.6	1.7†	3.3‡	4.2‡	8.5‡	25.0‡
Claude-3.5	–	–	–	–	–	–	–	0.0	0.5⋆	1.6†	3.2‡	4.0‡	8.4‡	24.9‡
Gemini-1.5-Pro	–	–	–	–	–	–	–	–	0.5⋆	1.6†	3.2‡	4.0‡	8.4‡	24.9‡
CommandR-plus	–	–	–	–	–	–	–	–	–	1.1‡	2.7†	3.5‡	7.9‡	24.4‡
Llama3-70B	–	–	–	–	–	–	–	–	–	–	1.6⋆	2.4‡	6.8‡	23.2‡
ONLINE-B	–	–	–	–	–	–	–	–	–	–	–	0.8⋆	5.2‡	21.7‡
IKUN	–	–	–	–	–	–	–	–	–	–	–	–	4.4	20.8‡
IKUN-C	–	–	–	–	–	–	–	–	–	–	–	–	–	16.4‡
MSLC	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Scores	95.3	93.4	91.9	91.4	89.3	88.9	88.8	88.8	88.3	87.2	85.6	84.7	80.4	63.9
Ranks	1-1	2-2	3-4	4-7	5-8	5-9	3-8	5-8	7-9	9-10	11-11	12-13	12-13	14-14

Head to head comparison for English→Hindi systems

	TranssionMT	Unbabel-Tower70B	Claude-3.5	ONLINE-B	Gemini-1.5-Pro	GPT-4	refA	IOL-Research	Llama3-70B	Aya23	IKUN-C
TranssionMT	–	0.7	1.0‡	1.1‡	1.3	2.7‡	2.8‡	4.1‡	4.6‡	6.5‡	20.5‡
Unbabel-Tower70B	–	–	0.3‡	0.4	0.5	2.0†	2.0‡	3.4‡	3.8‡	5.8‡	19.8‡
Claude-3.5	–	–	–	0.1‡	0.3‡	1.7‡	1.8‡	3.1‡	3.6‡	5.5‡	19.5‡
ONLINE-B	–	–	–	–	0.1†	1.6‡	1.6‡	3.0‡	3.4‡	5.4‡	19.4‡
Gemini-1.5-Pro	–	–	–	–	–	1.5⋆	1.5‡	2.8‡	3.3‡	5.2‡	19.3‡
GPT-4	–	–	–	–	–	–	0.0‡	1.3†	1.8‡	3.8‡	17.8‡
refA	–	–	–	–	–	–	–	1.3⋆	1.8	3.7‡	17.7‡
IOL-Research	–	–	–	–	–	–	–	–	0.5†	2.4‡	16.4‡
Llama3-70B	–	–	–	–	–	–	–	–	–	1.9†	16.0‡
Aya23	–	–	–	–	–	–	–	–	–	–	14.0‡
IKUN-C	–	–	–	–	–	–	–	–	–	–	–
Scores	91.3	90.5	90.2	90.1	90.0	88.5	88.5	87.2	86.7	84.7	70.7
Ranks	1-3	1-4	3-3	3-4	3-5	6-6	7-8	8-8	8-9	10-10	11-11

Head to head comparison for English→Icelandic systems

	refA	Dubformer	Claude-3.5	Unbabel-Tower70B	AMI	IKUN	ONLINE-B	GPT-4	IKUN-C	IOL-Research	Llama3-70B
refA	–	8.8‡	11.1‡	12.9‡	19.8‡	22.0‡	25.1‡	26.7‡	27.9‡	35.1‡	52.0‡
Dubformer	–	–	2.3	4.1‡	11.0‡	13.2‡	16.3‡	17.9‡	19.1‡	26.3‡	43.3‡
Claude-3.5	–	–	–	1.8‡	8.7‡	10.9‡	14.0‡	15.6‡	16.7‡	24.0‡	40.9‡
Unbabel-Tower70B	–	–	–	–	6.9‡	9.1‡	12.2‡	13.8‡	15.0‡	22.2‡	39.1‡
AMI	–	–	–	–	–	2.2⋆	5.3‡	6.9‡	8.1‡	15.3‡	32.3‡
IKUN	–	–	–	–	–	–	3.1†	4.7‡	5.9‡	13.1‡	30.0‡
ONLINE-B	–	–	–	–	–	–	–	1.6‡	2.8‡	10.0‡	26.9‡
GPT-4	–	–	–	–	–	–	–	–	1.2	8.4‡	25.3‡
IKUN-C	–	–	–	–	–	–	–	–	–	7.2‡	24.2‡
IOL-Research	–	–	–	–	–	–	–	–	–	–	16.9‡
Llama3-70B	–	–	–	–	–	–	–	–	–	–	–
Scores	93.1	84.3	81.9	80.2	73.3	71.0	68.0	66.3	65.2	58.0	41.0
Ranks	1-1	2-3	2-3	4-4	5-5	6-6	7-7	8-9	8-9	10-10	11-11

Head to head comparison for English→Japanese systems

	refA	ONLINE-B	CommandR-plus	GPT-4	Claude-3.5	Gemini-1.5-Pro	Unbabel-Tower70B	IOL-Research	Aya23	NTTSU	Team-J	Llama3-70B	IKUN-C
refA	–	0.7‡	0.9⋆	1.0‡	1.0‡	1.8†	2.1‡	2.1‡	2.2‡	2.4‡	3.3‡	5.1‡	10.1‡
ONLINE-B	–	–	0.1†	0.3‡	0.3	1.1	1.4‡	1.4‡	1.4⋆	1.7‡	2.6‡	4.3‡	9.3‡
CommandR-plus	–	–	–	0.1‡	0.2†	0.9	1.2‡	1.3‡	1.3‡	1.5‡	2.5‡	4.2‡	9.2‡
GPT-4	–	–	–	–	0.0‡	0.8†	1.1‡	1.1‡	1.2†	1.4‡	2.3‡	4.0‡	9.1‡
Claude-3.5	–	–	–	–	–	0.8⋆	1.1‡	1.1‡	1.1‡	1.4‡	2.3‡	4.0‡	9.1‡
Gemini-1.5-Pro	–	–	–	–	–	–	0.3‡	0.3‡	0.4	0.6‡	1.5‡	3.2‡	8.3‡
Unbabel-Tower70B	–	–	–	–	–	–	–	0.0‡	0.1‡	0.3‡	1.2‡	2.9‡	8.0‡
IOL-Research	–	–	–	–	–	–	–	–	0.1†	0.3†	1.2‡	2.9‡	8.0‡
Aya23	–	–	–	–	–	–	–	–	–	0.2⋆	1.2‡	2.9‡	7.9‡
NTTSU	–	–	–	–	–	–	–	–	–	–	0.9‡	2.7‡	7.7‡
Team-J	–	–	–	–	–	–	–	–	–	–	–	1.7‡	6.8‡
Llama3-70B	–	–	–	–	–	–	–	–	–	–	–	–	5.0‡
IKUN-C	–	–	–	–	–	–	–	–	–	–	–	–	–
Scores	91.8	91.1	91.0	90.8	90.8	90.0	89.7	89.7	89.7	89.4	88.5	86.8	81.7
Ranks	1-1	2-4	3-4	4-4	4-5	4-7	7-7	8-8	8-9	10-10	11-11	12-12	13-13

Head to head comparison for English→Russian systems

	refA	Dubformer	Claude-3.5	Unbabel-Tower70B	Yandex	Gemini-1.5-Pro	GPT-4	ONLINE-G	CommandR-plus	IOL-Research	IKUN	Aya23	Llama3-70B	IKUN-C
refA	–	0.1‡	0.9⋆	1.0‡	2.2†	3.7⋆	4.1‡	4.6†	4.8‡	7.1‡	10.0‡	10.5‡	13.4‡	19.4‡
Dubformer	–	–	0.8‡	0.9	2.1†	3.6‡	4.0‡	4.5‡	4.7‡	6.9‡	9.8‡	10.4‡	13.3‡	19.3‡
Claude-3.5	–	–	–	0.1‡	1.3	2.8†	3.2‡	3.7†	3.9†	6.1‡	9.0‡	9.6‡	12.5‡	18.5‡
Unbabel-Tower70B	–	–	–	–	1.1	2.7‡	3.1‡	3.6‡	3.8‡	6.0‡	8.9‡	9.5‡	12.4‡	18.4‡
Yandex	–	–	–	–	–	1.5⋆	1.9⋆	2.4	2.6	4.9‡	7.8‡	8.3‡	11.2‡	17.2‡
Gemini-1.5-Pro	–	–	–	–	–	–	0.4	0.9†	1.1	3.3‡	6.2‡	6.8‡	9.7‡	15.7‡
GPT-4	–	–	–	–	–	–	–	0.5	0.7	2.9†	5.8‡	6.4‡	9.3‡	15.3‡
ONLINE-G	–	–	–	–	–	–	–	–	0.2	2.5⋆	5.3‡	5.9‡	8.8‡	14.8‡
CommandR-plus	–	–	–	–	–	–	–	–	–	2.2†	5.1‡	5.7‡	8.6‡	14.6‡
IOL-Research	–	–	–	–	–	–	–	–	–	–	2.9†	3.5‡	6.4‡	12.4‡
IKUN	–	–	–	–	–	–	–	–	–	–	–	0.6⋆	3.5⋆	9.5‡
Aya23	–	–	–	–	–	–	–	–	–	–	–	–	2.9⋆	8.9‡
Llama3-70B	–	–	–	–	–	–	–	–	–	–	–	–	–	6.0‡
IKUN-C	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Scores	89.2	89.1	88.2	88.1	87.0	85.5	85.0	84.6	84.3	82.1	79.2	78.6	75.7	69.8
Ranks	1-1	2-3	3-4	3-5	3-7	6-8	6-9	6-9	5-9	10-10	11-11	12-12	13-13	14-14

Head to head comparison for English→Ukrainian systems

	Claude-3.5	Unbabel-Tower70B	Dubformer	refA	Gemini-1.5-Pro	ONLINE-W	GPT-4	CommandR-plus	IOL-Research	IKUN	IKUN-C
Claude-3.5	–	0.6	1.5†	3.2‡	3.3†	4.4‡	5.9‡	7.3‡	7.3‡	12.1‡	22.6‡
Unbabel-Tower70B	–	–	0.8⋆	2.6†	2.7⋆	3.8†	5.2‡	6.7‡	6.7‡	11.4‡	21.9‡
Dubformer	–	–	–	1.7‡	1.8‡	2.9†	4.4‡	5.8‡	5.8‡	10.6‡	21.1‡
refA	–	–	–	–	0.1	1.2⋆	2.7	4.1⋆	4.1⋆	8.8‡	19.4‡
Gemini-1.5-Pro	–	–	–	–	–	1.1	2.5⋆	4.0⋆	4.0‡	8.7‡	19.2‡
ONLINE-W	–	–	–	–	–	–	1.4	2.9	2.9⋆	7.6‡	18.1‡
GPT-4	–	–	–	–	–	–	–	1.4	1.4	6.2‡	16.7‡
CommandR-plus	–	–	–	–	–	–	–	–	0.0	4.8‡	15.3‡
IOL-Research	–	–	–	–	–	–	–	–	–	4.7‡	15.3‡
IKUN	–	–	–	–	–	–	–	–	–	–	10.5‡
IKUN-C	–	–	–	–	–	–	–	–	–	–	–
Scores	90.5	89.8	89.0	87.3	87.1	86.0	84.6	83.2	83.2	78.4	67.9
Ranks	1-2	1-2	3-3	4-6	4-6	5-8	5-9	6-9	7-9	10-10	11-11

Head to head comparison for English→Chinese systems

	GPT-4	Unbabel-Tower70B	refA	Gemini-1.5-Pro	ONLINE-B	IOL-Research	Claude-3.5	CommandR-plus	Llama3-70B	HW-TSC	IKUN	Aya23	IKUN-C
GPT-4	–	0.0‡	0.3⋆	0.3‡	0.4‡	0.6‡	0.7†	1.4⋆	3.2‡	3.4‡	4.3‡	4.4‡	7.5‡
Unbabel-Tower70B	–	–	0.2	0.3‡	0.3‡	0.6‡	0.7‡	1.3	3.1‡	3.4‡	4.3‡	4.4‡	7.5‡
refA	–	–	–	0.1‡	0.1‡	0.3‡	0.5⋆	1.1	2.9‡	3.1‡	4.1‡	4.2‡	7.2‡
Gemini-1.5-Pro	–	–	–	–	0.1‡	0.3†	0.4‡	1.1‡	2.9‡	3.1‡	4.0‡	4.1‡	7.2‡
ONLINE-B	–	–	–	–	–	0.2⋆	0.3	1.0†	2.8‡	3.0‡	3.9‡	4.0‡	7.1‡
IOL-Research	–	–	–	–	–	–	0.1⋆	0.8‡	2.6‡	2.8‡	3.7‡	3.8‡	6.9‡
Claude-3.5	–	–	–	–	–	–	–	0.6†	2.5‡	2.7‡	3.6‡	3.7‡	6.8‡
CommandR-plus	–	–	–	–	–	–	–	–	1.8‡	2.0‡	2.9‡	3.0‡	6.1‡
Llama3-70B	–	–	–	–	–	–	–	–	–	0.2‡	1.1‡	1.2†	4.3‡
HW-TSC	–	–	–	–	–	–	–	–	–	–	0.9‡	1.0‡	4.1‡
IKUN	–	–	–	–	–	–	–	–	–	–	–	0.1‡	3.2‡
Aya23	–	–	–	–	–	–	–	–	–	–	–	–	3.1‡
IKUN-C	–	–	–	–	–	–	–	–	–	–	–	–	–
Scores	89.6	89.6	89.4	89.3	89.3	89.0	88.9	88.3	86.5	86.2	85.3	85.2	82.1
Ranks	1-1	2-4	2-4	4-4	5-6	6-6	6-7	6-8	9-9	10-10	11-11	12-12	13-13

Head to head comparison for Japanese→Chinese systems

	Claude-3.5	refA	GPT-4	DLUT-GTCOM	Unbabel-Tower70B	Gemini-1.5-Pro	CommandR-plus	IOL-Research	Llama3-70B	Aya23	Team-J	NTTSU	ONLINE-B	IKUN-C	MSLC
Claude-3.5	–	0.1	0.3⋆	0.3	0.5‡	0.7⋆	0.8‡	1.0‡	2.0‡	2.0‡	3.1‡	3.7‡	4.4‡	6.2‡	9.3‡
refA	–	–	0.1⋆	0.2	0.4‡	0.5⋆	0.7‡	0.9‡	1.9‡	1.9‡	3.0‡	3.6‡	4.3‡	6.1‡	9.2‡
GPT-4	–	–	–	0.1⋆	0.2	0.4	0.5‡	0.8†	1.7‡	1.8‡	2.9‡	3.4‡	4.1‡	6.0‡	9.0‡
DLUT-GTCOM	–	–	–	–	0.2⋆	0.3	0.5†	0.7‡	1.7‡	1.7‡	2.8‡	3.4‡	4.1‡	5.9‡	9.0‡
Unbabel-Tower70B	–	–	–	–	–	0.2	0.3	0.5	1.5‡	1.6‡	2.6‡	3.2‡	3.9‡	5.7‡	8.8‡
Gemini-1.5-Pro	–	–	–	–	–	–	0.1‡	0.4†	1.4‡	1.4‡	2.5‡	3.1‡	3.8‡	5.6‡	8.7‡
CommandR-plus	–	–	–	–	–	–	–	0.3	1.2‡	1.3‡	2.3‡	2.9‡	3.6‡	5.5‡	8.5‡
IOL-Research	–	–	–	–	–	–	–	–	1.0‡	1.0‡	2.1‡	2.7‡	3.4‡	5.2‡	8.3‡
Llama3-70B	–	–	–	–	–	–	–	–	–	0.0	1.1†	1.7‡	2.4‡	4.2‡	7.3‡
Aya23	–	–	–	–	–	–	–	–	–	–	1.1‡	1.7⋆	2.4‡	4.2‡	7.3‡
Team-J	–	–	–	–	–	–	–	–	–	–	–	0.6	1.3‡	3.1‡	6.2‡
NTTSU	–	–	–	–	–	–	–	–	–	–	–	–	0.7‡	2.5‡	5.6‡
ONLINE-B	–	–	–	–	–	–	–	–	–	–	–	–	–	1.8†	4.9‡
IKUN-C	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3.1‡
MSLC	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Scores	-1.4	-1.5	-1.7	-1.7	-1.9	-2.1	-2.2	-2.4	-3.4	-3.5	-4.5	-5.1	-5.8	-7.7	-10.7
Ranks	1-3	1-3	3-5	2-5	4-8	3-6	6-8	6-8	9-10	9-10	11-12	11-12	13-13	14-14	15-15