Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here But Not Quite There Yet

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al.

Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here But Not Quite There Yet

Tom Kocmi
Microsoft

Eleftherios Avramidis
DFKI

Rachel Bawden
Inria, Paris

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

Philipp Koehn
Johns Hopkins University

Benjamin Marie
4i.ai

Christof Monz
University of Amsterdam

Makoto Morishita
NTT

Kenton Murray
Johns Hopkins University

Masaaki Nagata
NTT

Toshiaki Nakazawa
University of Tokyo

Martin Popel
Charles University

Maja Popović
Dublin City University

Mariya Shmatova
Dubformer

Jun Suzuki
Tohoku University

Abstract

This paper presents the results of the General Machine Translation Task organised as part of the 2023 Conference on Machine Translation (WMT). In the general MT task, participants were asked to build machine translation systems for any of 8 language pairs (covering 14 translation directions), to be evaluated on test sets consisting of up to four different domains. We evaluate system outputs with professional human annotators using a combination of source-based Direct Assessment and scalar quality metric (DA+SQM).

1 Introduction

The Eighth Conference on Machine Translation (WMT23)[1] was held at EMNLP 2023 and hosted a number of shared tasks on various aspects of machine translation (MT). This conference built on 17 previous editions of WMT 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).

Following last year's shift from focusing mainly on the news domain, we have continued to explore the capabilities of "General Machine Translation". While the news domain provided a clear and familiar benchmark, we responded to the need to test MT in more diverse settings. Our goal is to assess MT systems' ability to handle a broader range of language use. How to test general MT performance is a research question in itself. Countless phenomena could be evaluated, the most important being:

- various domains (news, medicine, IT, patents, legal, social, gaming, etc.)
- style of text (formal or spoken language, fiction, technical reports, etc.)
- robustness to non-standard (or noisy) user-generated content (grammatical errors, code-switching, abbreviations, etc.)

Evaluating all phenomena is nearly impossible and creates numerous unforeseen problems. Therefore, we decided to simplify the problem and start with an evaluation of different domains. We selected the following domains: news, e-commerce, social/user-generated content (UGC), speech, and manuals. They were chosen to represent topics with different content styles and to be understandable for humans without special in-domain knowledge, thus not requiring specialized translators or human raters for evaluation. Due to limited access

[1] http://www2.statmt.org/wmt23/

to monolingual data across all languages, each language direction contains only a subset of up to four domains.

In addition to language pairs evaluated last year:

Czech→Ukrainian,
English↔Chinese,
English→Czech,
English↔German,
English↔Japanese,
English↔Russian,
Ukrainian→English,

we introduce a new language pair to WMT, namely:

English↔Hebrew.

Other than language pairs, there are several differences with respect to last year's task. All language pairs are provided with the sentence boundaries marked except for English↔German, where we decided to experiment with paragraph-level translation. Another significant change for this year is the unification of our human evaluation protocol. We no longer rely on reference-based MTurk evaluation and move the evaluation towards source-based DA+SQM evaluation (introduced last year) with professional annotators. Finally, this year's shared task included an increased number of test suites (Section 6), allowing the evaluation of MT outputs from different perspectives, including a range of linguistic phenomena, purposely difficult sentences, specialist domains, gendered translations and non-standard UGC translation.

All General MT task submissions, sources, references and human judgements are available at Github.[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 for the constrained track Section 3. We list all submitted systems in Section 4. The human evaluation approach of DA+SQM is described in Section 5. Finally, Section 6 describes the test suites and summarises their conclusions.

[2] https://github.com/wmt-conference/wmt23-news-systems

[3] http://wmt.ufal.cz

Summary of the WMT2023 General MT task
The main findings are as follows:

- Large Language Models (LLMs) exhibit strong performance across the majority of language pairs, although this is based only on two LLM-based system submissions. Test suite analysis revealed that although GPT4 excelled in some areas (e.g. UGC translation) struggled with other aspects such as speaker gender translation and specific domains (e.g. legal), whereas it ranked lower than encoder-decoder systems when translating from English into less-represented languages (e.g. Czech and Russian)

- We have observed a decline in the number of submissions into the constrained track. Consequently, we plan to re-evaluate the definition and the incentives of the constrained track and consider incorporating open-source LLMs in future constrained evaluations.

- We demonstrate the feasibility of paragraph-level German↔English tasks, although more investigation would be required before generalising to all language pairs.

- Professional human translations do not always guarantee high quality. For Hebrew↔English, our references are likely to be post-edited MT, while for Chinese→English, the reference translation is worse than the majority of automatic translations.

- The manual evaluation results obtained from DA+SQM and MQM methods yield comparable cluster rankings.

2 Test Data

In this section, we describe the process of collecting data in Section 2.1, followed by the explanation of preprocessing steps in Section 2.2. Producing human references is summarized in Section 2.3 and lastly test set analysis is conducted in Section 2.4.

2.1 Collecting test data

As in the previous years, the test sets consist of unseen translations collected especially for the task. The guaranteed novelty of the test sets has become even more important with the rise of LLMs trained on unspecified training data. To prevent possible contamination, we focused on collecting as recent

Lang. pair	Domain name	Domain type	#docs	#segs	#segs/#docs
cs→uk	*	*	156	2017	12.93
	games	News	17	180	10.59
	news	News	35	567	16.20
	official	Social/UGC	26	347	13.35
	personal	Social/UGC	31	390	12.58
	voice	Speech	47	533	11.34
de→en	*	*	210	549	2.61
	manuals	Manuals	15	74	4.93
	mastodon	Social/UGC	95	103	1.08
	news	News	47	277	5.89
	user_review	E-commerce	53	95	1.79
en→{cs,he,ja,ru,uk,zh}	*	*	192	2074	10.80
	mastodon	Social/UGC	79	504	6.38
	news	News	30	516	17.20
	speech	Meeting notes	25	547	21.88
	user_review	E-commerce	58	507	8.74
en→de	*	*	192	557	2.90
	mastodon	Social/UGC	79	212	2.68
	news	News	30	139	4.63
	speech	Meeting notes	25	113	4.52
	user_review	E-commerce	58	93	1.60
he→en	*	*	94	1910	20.32
	news	News	68	1558	22.91
	reviews	Social/UGC	26	352	13.54
ja→en	*	*	282	1992	7.06
	ad	Social/UGC	53	245	4.62
	ec	Social/UGC	25	255	10.20
	news	News	37	495	13.38
	qa	Conversational	118	497	4.21
	user_review	E-commerce	49	500	10.20
ru→en	*	*	162	1723	10.64
	manuals	Manuals	15	505	33.67
	news	News	54	676	12.52
	reviews	Social/UGC	93	542	5.83
uk→en	*	*	132	1826	13.83
	clipboard	Social/UGC	30	504	16.80
	news	News	26	514	19.77
	other	Social/UGC	27	538	19.93
	voice	Speech	49	270	5.51
zh→en	*	*	179	1976	11.04
	manuals	Manuals	14	487	34.79
	news	News	38	763	20.08
	user_review	E-commerce	127	726	5.72

Table 1: Test set statistics per direction and domain (rows marked * are over all domains). Note that en→de shares source test data with the other from-English directions, but as translation and evaluation for both en→de and de→en were carried out on the paragraph level (a segment therefore being a paragraph rather than a sentence), this results in a lower number of segments per document. The domain name is as indicated in the released test sets and domain type indicates the broader domain category.

data as possible across various domains. This task is incredibly difficult and needs further investigation in future years. There are three main limitations:

- Finding sources with different domains.
- Finding data that are in the public domain or under open licenses.
- Finding recently created data to minimize the risk of them being part of the training pipelines.

The test sets are publicly released to be used as translation benchmarks. Here we describe the test sets' production and composition.

We decided to collect data from 5 domains (news, social/user-generated, e-commerce, manuals, and speech). For all language pairs, we aimed for a test set size of 2,000 sentences and to ensure that the test sets were "source-original", namely

that the source text was first written in the source language, and then the target text is the human translation. This is to avoid "translationese" effects in the source language, which can have a detrimental impact on the accuracy of evaluation (Toral et al., 2018; Freitag et al., 2019; Läubli et al., 2020; Graham et al., 2020). We collected roughly the same number of sentences for each domain. For some languages, we could not locate high quality data and therefore we selected more sentences from other domains. Note that descriptions in this section refer to source monolingual data when mentioning a language.

News domain For most languages this domain contains data prepared in the same way as in previous years (Akhbardeh et al., 2021). We collected news articles from February 2023 extracted from online news sites, preserving document boundaries. We expect that news domain text will generally be of high quality. The news in Hebrew was kindly provided by the Israeli Association of Human Language Technologies (IAHLT).[4] These are samples of originally Hebrew texts from news published in Israel Hayom[5] in 2022.

E-commerce domain (product reviews) This domain consists of user reviews of different Amazon products selected from the publicly available multilingual corpus (Keung et al., 2020). This corpus was designed for multilingual text classification and consists of reviews written in English, Japanese, German, French, Spanish, and Chinese, between 2015 and 2019. We used the test parts of the English, German, Japanese and Chinese corpora for extracting the source part of the WMT test set. The reviews were selected so that the resulting corpus covers each product, all rating scores for the product, and the lexical diversity is maximized. The lexical diversity was estimated as a simple ratio between the number of distinct words/characters (vocabulary) divided by the total number of words/characters.

Social/user-generated domain For English and German, we relied on the Mastodon Social API.[6] Mastodon is a federated social network that is compatible with the W3C standard ActivityPub (Webber et al., 2018). Users publish short-form content similar to tweets that are referred to as "toots" for historical reasons. As this is a decentralized social media network, different servers have very different data, policies, communities, and uses. We decided to use `mastodon.social`, the original server, as it has a large community as well as publicly available toots. We collected data in early May of 2023. We used the reported language ID label, but were only able to collect enough data in German and English. We only collected toots with more than 150 characters in length in order to allow for data that was more likely to be semantically interesting for evaluating translation systems.

For Hebrew, we used comments on news articles from the Israel Hayom site mentioned above. This data was also provided by IAHLT.

For Russian, we used data from the Geo Reviews Dataset containing reviews about organizations published on Yandex Maps and open for academic and research purposes.[7]

For Japanese, we used product descriptions of a b2b e-commerce site and search advertising text ads for the social and user-generated domain, because we could not obtain high-quality data for this domain type. MonotaRo Co., Ltd. provided product descriptions of their private label brands listed on their b2b e-commerce site.[8] We defined a document for a product description as a combination of a title, product description, and cautionary note. CyberAgent, Inc.[9] provided search advertising text ads with their client's consent. We defined a document for an ad as the longest possible combination of multiple titles and descriptions.

Manuals For this domain, we primarily sourced scanned versions of different mostly gaming manuals provided by Centific[10]. These were then converted to digital text format using Optical Character Recognition (OCR) technology. Given the inaccuracies of OCR, the digitized content underwent a subsequent post-editing phase, where humans reviewed and corrected any errors. The selection of manuals ranged across various sources, and none of them were older than five years.

Speech The exact data types used in the "conversational" or "speech" domain vary across language pairs.

[4] https://www.iahlt.org

[5] https://www.israelhayom.co.il

[6] https://mastodon.social/api/v1/timelines/public

[7] https://github.com/yandex/geo-reviews-dataset-2023

[8] https://www.monotaro.com/

[9] https://www.cyberagent.co.jp

[10] https://www.centific.com

For English→Czech, the data comes from the test set which was created for the 2023 instance of AutoMin 2023 (Ghosal et al., 2022).[11] The texts are manually curated transcripts of project meetings, same in style as released in ELITR Minuting Corpus (Nedoluzhko et al., 2022). The meetings were held mostly remotely or in a hybrid form, all meeting participants were non-native speakers of English and the meetings were always on rather technical and in-depth topics. Our manual curation corrected ASR errors (but not errors in English grammar or vocabulary) and de-identified the transcripts, replacing names with placeholders ("PERSONxy", "PROJECTxy" and similar). For person names, round brackets are used at the beginnings of lines to indicate the speaker and square brackets are used in the text when the person was mentioned. The data contain also some markup, e.g. "<`unintelligible/`>". These conventions are likely to be distorted by translation systems and we also noticed that they were distorted in the reference translation (the style of the brackets was ignored). This tiny detail can influence both manual and automatic scoring on this domain.

For Japanese, we used question-answer pairs from a community question-answering service. NTT Resonant Inc., which recently merged with NTT DOCOMO, INC., provided question-answer pairs from their website, *Oshiete! goo*.[12] For every question-answer pair, we defined a document as a combination of a question and its best answer marked by the user.

Czech and Ukrainian source texts Source texts for Czech→Ukrainian and Ukrainian→English translation included the News domain as described above and texts collected through the Charles Translator for Ukraine.[13] With users' consent, the service can log their inputs for the purpose of creating a dataset of real use cases. The datasets are extracted from the inputs collected from May 2022 to April 2023.

The Charles Translator mobile app supports voice input, which is converted to text using Google ASR (automatic speech recognition). The texts collected this way were marked as the `voice` domain. For Ukrainian→English, the remaining Ukrainian inputs were classified either as `clipboard` (texts inserted to the Charles Translator using the *Paste from clipboard* button) and `other`. The `clipboard` texts are more likely to include formal communication copied from web sites, but we noticed it includes personal communication (copied from chat applications) as well. Thus for Czech→Ukrainian, we decided to classify the remaining Czech inputs either as `official` (formal communication) or `personal` (personal communication), ignoring whether they were inserted from a clipboard or written using a keyboard.

The texts were filtered and pseudonymized in the same way as last year (Kocmi et al., 2022), so for example we asked the annotators not to delete or fix noisy inputs as long as they are comprehensible. There was one exception from this rule this year: the Czech `voice` domain data was post-edited to fix ASR errors, including missing punctuation and casing.

The source texts were translated by professional translators principally following the brief in Appendix C. Last year, parts of the Ukrainian→Czech test set was detected to be post-edited MT. Therefore this year, we decided to hire two professional translators directly without the mediation of a translation agency, we emphasised the rule that the translations must be done from scratch (without MT postediting and without translation memories). We could not detect any MT postediting in the resulting translations.

2.2 Human preprocessing of test data

Although testing of robustness of MT is an important task, the noisy data introduces problems for human translators and annotators. Therefore, we decided to discard data considered too noisy. Furthermore, publicly available data often contains inappropriate content, which can stress either human translators or human annotators, leading to a decrease in the quality (for example, translators refuse to translate political content considered censored in their countries).

Therefore, we asked humans to check collected data and carry out minor corrections (mainly checking sentence splits and discarding similar or repeated content). This was sufficient for the news domain because it was often clean and without serious problems. However, with the expansion towards general MT, we find ourselves running into an issue of source data being noisier and less well formatted and that therefore needs to be handled before translation. Furthermore, we asked them to

[11] https://ufal.github.io/automin-2023/
[12] https://oshiete.goo.ne.jp/
[13] http://translator.cuni.cz

remove shortest documents to keep longer context. The source data for test sets therefore goes through human validation checks involving linguists discarding inappropriate content altogether and carrying out minor textual corrections to the data. You can find the linguistic brief for prepossessing in Appendix B.

2.3 Test set translation

The translation of the test sets was performed by professional translation agencies, according to the brief in Appendix C. Different partners sponsored each language pair and various translation agencies were therefore used, which may affect the quality of the translation.

Regrettably, upon reviewing translations procured from one of the agencies (the one responsible for English to Hebrew and Hebrew to English translations), it appeared that the translations might have been post-edited from publicly available online translation systems. This observation contradicts the initial instruction provided for agency that precluded the use of any automated translation platforms. While the agency has asserted that their professional translations conducted translations from scratch, our evaluation suggested otherwise. Moving forward, we propose to build a step-by-step verification system to avoid such discrepancies.

Human translations would not be possible without the sponsorship of our partners: Microsoft, Toloka AI, Google, Charles University, NTT, and Dubformer.

2.4 Test set analysis

As described previously, the chosen domains, sources for the data and the number of sentences per domain was subject to the availability of high quality data in each language direction. For example, while the news domain was available for all language directions, social media data was only available for English, German (both from Mastodon) and Hebrew (from comments on news articles). The number of documents, segments, average document length and type-token ratio (of the source side of the test sets) are given in Table 1.

Document context Document context is available for all language directions, although the average document length varies both by domain and language direction. Manuals tend to represent the longest domains, followed by the news domain. The social media domain tends to represent the shortest documents. along with reviews. Note that this year, we piloted translation and evaluation of en→de and de→en at the paragraph level (with each segment therefore containing several sentences), with the aim of avoiding the constraint of having a one-to-one mapping at the level of the sentence between source texts and their translations. This is visible in the statistics in Table 1 as the number of segments is lower for these two directions, as is the average document length.

Lexical diversity We can compare the type-token ratio (TTR) to get an idea of the relative lexical diversity of (i) domains and (ii) original vs. translated sentences.[14,15] Raw TTRs for each language pair and domain are shown in Table 11 in Appendix D. Regarding domains, the TTR appears highest for texts mastodon, perhaps illustrating the diversity of conversational topics and also of the potentially non-standard nature of the texts. User reviews appear to have the lowest TTR, most likely due to the fact that similar vocabulary is used across reviews. The TTR of course differs according to the language in question, according to the differing morphological properties.

Anonymisation and markup A particularity of the 'speech' domain is the presence of placeholders for anonymised elements and markup (in the form of tags). For example, there are 35 placeholders surrounded either by square or rounded brackets to indicate different people, organisations and projects (e.g. (PERSON1), [PERSON9], [ORGANIZATION4], [PROJECT8], etc.). The 'person' tags are used both in-text to replace the names of people and at the beginning of lines to indicate who is talking. Markup is added to indicate speakers talking at the same time (<parallel_talk>), unintelligible passages (<unintelligible/>), laughter (<laugh/>) and other noise (<other_noise/>).

2.5 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

[14]The TTR is the ratio of unique tokens to total tokens, and it is higher the diverse the vocabulary of a text is. It is dependent on the morphological complexity of a language, but can also vary due to other factors.

[15]Texts are tokenised using the language-specific Spacy models (Honnibal and Montani, 2017) where available. For Hebrew, we took the multilingual Spacy model, since a language-specific one was not available.

MT translation. The test suites were contributed and evaluated by test suite providers as part of a decentralized sub-task, which will be detailed in Section 6.

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 Appendix in Table 9 and Table 10. Specifically, our parallel data selection include large multilingual corpora such as Europarl-v10 (Koehn, 2005), Paracrawl-v9 (Bañón et al., 2020), CommonCrawl, NewsCommentary-v18, WikiTitles-v3, WikiMatrix (Schwenk et al., 2021), TildeCorpus (Rozis and Skadiņš, 2017), OPUS (Tiedemann, 2012), UN Parallel Corpus (Ziemski et al., 2016), and language-specific corpora such as CzEng-v2.0 (Kocmi et al., 2020), YandexCorpus,[16] ELRC EU Acts, JParaCrawl (Morishita et al., 2020), Japanese-English Subtitle Corpus (Pryzant et al., 2018), KFTT(Neubig, 2011), TED (Cettolo et al., 2012), CCMT, and back-translated news. Links for downloading these datasets were provided on the task web page;[17] in addition, we automated the data preparation pipeline using MTDATA (Gowda et al., 2021).[18] MTDATA downloads all the mentioned datasets, except CCMT and 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.

4 System submissions

This year, we received a total of 72 primary submissions from 17 participants. In addition, we collected translations from online MT systems across all language pairs. Online system outputs come from 6 public MT services and were anonymized as ONLINE-{A,B,G,M,W,Y}, which added additional 77 system outputs. The participating systems are listed in Table 2 and detailed in the rest of this section.

[16] https://github.com/mashashma/WMT2022-data

[17] https://statmt.org/wmt23/translation-task.html

[18] http://www2.statmt.org/wmt23/mtdata

Finally, we added translations by three contrastive systems. Two of them are based on the NLLB translation model (NLLB Team et al., 2022) modified by (Freitag et al., 2023) to have a suboptimal performance, using (i) greedy search (NLLB_Greedy) and (ii) following minimum Bayes risk decoding (MBR) optimizing the BLEU metric (NLLB_MBR_BLEU). Neither of them is the official (and better performing) NLLB model. The third contrastive translation is produced by the large language model GPT4 using 5-shot prompting with fixed random translation examples, using the exact prompt by Hendy et al. (2023) together with their predefined few-shot examples. For languages not evaluated in their study, we took examples from the last WMT test sets.

Appendix E provides details of the submitted systems if the authors provided such details.

4.1 Constrained and unconstrained tracks

For presentation of the results, systems are treated as either constrained or unconstrained. A system is classified as constrained if the authors reported training only on the provided data and adhering to the rules describing the use of publicly available pre-trained models. The constrained track imposes restrictions on training data, metrics, and pretrained models, while the unconstrained track provides unrestrained flexibility.

The constrained track limitations are mainly around the training and testing data, together with the limitation on pretrained models:

- **Training data:** Only data specified for the current year are permissible, see Section 3. Multilingual systems can be used as long as they only use WMT23 data.
- **Metrics:** The training pipeline can use pretrained metrics evaluated in previous WMT Metrics shared tasks, e.g., COMET (Rei et al., 2022), Bleurt (Yan et al., 2023).
- **Pretrained models:** only the following list of models is allowed together with all their public sizes: mBART (Liu et al., 2020), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020), sBERT (Reimers and Gurevych, 2019), and LaBSE (Feng et al., 2022).
- **Linguistic tools:** Basic tools like taggers, parsers, and morphology analyzers are allowed.

Submission Name	Language Pairs	System Description
AIRC	de-en, en-ja, ja-en, en-de	(Rikters and Miwa, 2023)
ANVITA	ja-en, zh-en, en-ja, en-zh	(no associated paper)
CUNI-DocTransformer	en-cs	(Popel, 2020)
CUNI-GA	en-cs, cs-uk	(Jon et al., 2023)
CUNI-Transformer	en-cs, cs-uk	(Popel, 2020)
GPT4-5shot	**All language pairs**	(Hendy et al., 2023)
GTCOM	de-en, ja-en, he-en, en-cs, en-he, cs-uk, en-uk, uk-en	(Zong, 2023)
HW-TSC	de-en, en-zh, zh-en	(Wu et al., 2023b)
IOL-Research	zh-en, en-zh	(Zhang, 2023)
TeamKYB	ja-en, en-ja	(LI et al., 2023)
Lan-BridgeMT	**All language pairs**	(Wu and Hu, 2023)
MUNI-NLP	cs-uk	(Rychlý and Teslia, 2023)
NAIST-NICT	en-ja, ja-en	(Deguchi et al., 2023)
NLLB_Greedy	**All language pairs**	(Freitag et al., 2023)
NLLB_MBR_BLEU	**All language pairs**	(Freitag et al., 2023)
ONLINE-A	**All language pairs**	-
ONLINE-B	**All language pairs**	-
ONLINE-G	**All language pairs**	-
ONLINE-M	en-ru, zh-en, en-zh, de-en, en-cs, ja-en, en-de, en-ja, ru-en	-
ONLINE-W	en-uk, ja-en, de-en, en-ja, ru-en, en-de, uk-en, en-ru, zh-en, en-cs, en-zh, cs-uk	-
ONLINE-Y	**All language pairs**	-
PROMT	en-ru, ru-en	(Molchanov and Kovalenko, 2023)
SRPH	he-en, en-he	(Cruz, 2023)
SKIM	en-ja, ja-en	(Kudo et al., 2023)
UPCite-CLILLF	fr-en, en-fr	(no associated paper)
UvA-LTL	he-en, en-he	(Wu et al., 2023a)
YiShu	zh-en, en-zh	(Min et al., 2023)
LanguageX	en-zh, en-uk, ru-en, uk-en, en-de, he-en, ja-en, zh-en, en-he, de-en, en-cs, en-ja, en-ru	(Zeng, 2023)

Table 2: Participants in the General MT shared task. 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.

The online systems and contrastive systems are treated as unconstrained during the automatic and human evaluation.

4.2 OCELoT

We used the open-source OCELoT platform[19] to collect system submissions again this year. The platform provides anonymized public leaderboards[20] and was also used for two other WMT23 shared tasks: Biomedical (Neves et al., 2023) and Sign Language Translation (Müller et al., 2023). 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 or unconstrained setting, and submit a system description paper abstract. These were mandatory for a system submission to be included in the human evaluation campaign.

5 Human Evaluation

Human evaluation for all language translation directions is performed with source-based ("bilingual") Direct Assessment (DA, Graham et al., 2013) of individual segments in document context with Scalar Quality Metrics (SQM) guidelines, mostly following the setup established at WMT22 (DA+SQM, Kocmi et al., 2022). DA+SQM asks the annotators to provide a score between 0 and 100 on a sliding scale, but the slider is presented with seven labelled tick marks, as demonstrated in Figure 1.

Two different annotation platforms and four distinct pools of annotators (Table 3) are used for annotation of different language pairs. We use the open-source framework Appraise (Federmann, 2018) for the evaluation of English→Czech, English↔{Chinese, German, Japanese}, and Czech→Ukrainian. Toloka AI[21] hosts the evaluation of English↔{Hebrew, Russian, Ukrainian} using their own implementation of the source-based document-level DA+SQM task, which is as close as possible to the Appraise user interface.

We keep the selection process of documents for annotation mostly the same as in the previous year. The only change made in order to align closer with the MQM-based evaluation run at the Metrics shared task (Freitag et al., 2023) is to present the first 10 segments from a document instead of random 10 consecutive segments.

We again collect both segment-level scores and document-level scores, but compute rankings based on segment scores only.

5.1 Human annotators

Annotations for different language pairs are provided by four different parties with their pool of annotators of distinct profiles as presented in Table 3. We shift towards more professional or semi-professional annotators' pools and decide not to use MTurk annotations as in past years for reference-based DA evaluation for into-English language directions.

Assessments for English↔{Chinese, German, Japanese} are provided by Microsoft and their pool of bilingual target-language native speakers, professional translators or linguists, highly experienced in MT evaluation. Microsoft monitors the annotators' performance over time and permanently removes from the pool those who fail quality control, which increases the overall quality of the human assessment.

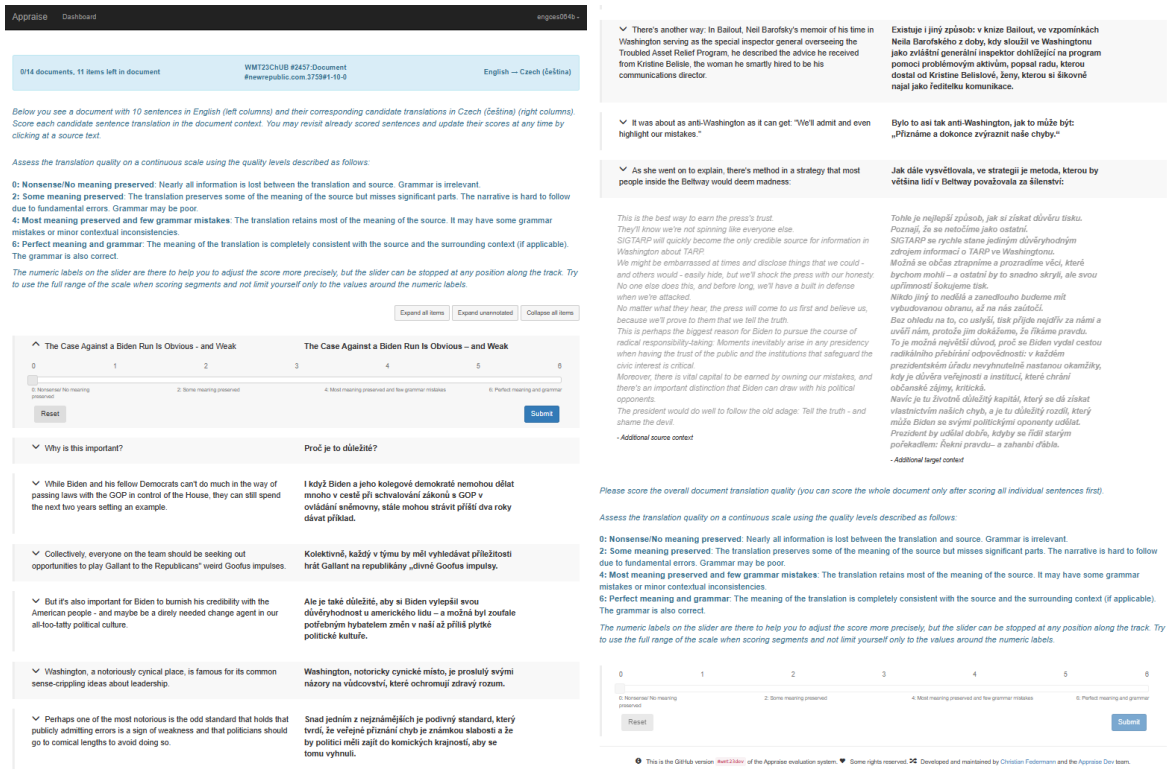
Charles University provides annotators for language pairs involving the Czech language, i.e., English→Czech and Czech→Ukrainian. Their annotators are linguists, translators, researchers and students who are native speakers of the target language with high proficiency in the source language.

DA scores for English↔{Hebrew, Russian, Ukrainian} are collected by Toloka AI using their paid crowd of bilingual target-language native speakers. Toloka AI tests proficiency of their annotator crowd across different NLP annotation tasks and allowed only annotators who deemed reliable according to their quality control measures.

5.2 Document selection and quality control

The document selection process remains the same as in the previous year with minor changes. We first randomly sample a subset of document snippets from each of the domains for annotations, sampling the domains with approximately the same number of segments per domain. This ensures that

[19] https://github.com/AppraiseDev/OCELoT
[20] https://ocelot-wmt23.mteval.org
[21] https://toloka.ai



(a) Top part of the screen with segment-level scoring.

(b) Bottom part of the screen with document-level scoring.

Figure 1: Screenshot of the document-level DA+SQM configuration in the Appraise interface for an example assessment from the human evaluation campaign for out of English language pairs. The annotator is presented with the entire translated document snippet randomly selected from competing systems (anonymized) with additional static contexts, and is asked to rate the translation of individual segments and then the entire document on sliding scales between 0 and 100.

all systems in the given language pairs are evaluated on the same subset of the test set, allowing fair comparison between them. As in previous years, we aim to collect approximately 1,500 assessments per system per language pair. Due to concerns about having sufficient annotations, we create two batches of HITs, each providing half of the required assessments, such that at least all segments in the first batch could be covered for all systems, with the second campaign completed if possible.

For HIT generation for English↔German, which feature paragraph-level test sets (documents consist of paragraphs instead of sentences), we simply consider a whole paragraph as a "segment", collecting paragraph-level assessments. In that regard, we collect fewer DA scores per system comparing to other language pairs, but the human evaluation covers a larger subset of the test sets.

Last year, we used snippets of at most 10 randomly selected consecutive segments from a document as "documents" for document-level annotation. This year, we use 10 first segments from a document instead, in order to align with the MQM-based evaluation used at the Metrics shared task (Freitag et al., 2023).

All HITs consist of exactly 100 segments and are generated as in the past:

1. Snippet-system pairs are randomly sampled (from the restricted set of pre-sampled snippets) to create up to 80 segments;

2. Random snippets for the remaining 20 (or more) segments are duplicated from the first 80 to serve as quality control items;

3. BAD references are introduced to the random segments in the duplicated snippets to have about 12-14% of quality control segments per HIT.

BAD translations are created by replacing an embedded sequences of tokens in the segment with a random phrase of the same length from a different reference segment.[22]

We perform quality control by measuring an annotator's ability to reliably score BAD translations

[22]For full details, see the HIT and batch generation code: https://github.com/wmt-conference/wmt23-news-systems

Language pairs	Annotators' profile	Tool
English↔Chinese/German/Japanese	Microsoft annotators: bilingual target-language native speakers, professional translators or linguists, experienced in MT evaluation	Appraise
Czech→Ukrainian	Paid translators and target-language native speakers	Appraise
English→Czech	Czech paid linguists, annotators, researchers, students with high proficiency in English	Appraise
English↔Hebrew/Russian/Ukrainian	Toloka AI paid crowd: bilingual target-language native speakers high-performing in other task types	Toloka.ai

Table 3: Annotators' profiles and annotation tools for each language pair in human evaluation.

Language Pair	Sys.	Assess.	Assess/Sys
Chinese→English	16	20,535	1283.4
Czech→Ukrainian	14	23,191	1656.5
German→English	14	13,573	969.5
English→Chinese	16	24,551	1534.4
English→Czech	16	25,527	1595.4
English→German	13	14,267	1097.5
English→Japanese	17	26,115	1536.2
Japanese→English	18	27,858	1547.7

Table 4: Amount of segments evaluated in the WMT23 manual evaluation campaign; including human references as systems; after excluding quality control items and document-level scores.

Language Pair	Ann.	HITs	HITs/Ann.
Chinese→English	13	128	9.8
Czech→Ukrainian	9	146	16.2
German→English	21	82	3.9
English→Czech	36	162	4.5
English→German	22	87	4.0
English→Japanese	21	164	7.8
English→Chinese	13	154	11.8
Japanese→English	20	174	8.7

Table 5: Numbers of individual annotators taking part in the WMT23 human evaluation campaign and the average number of HITs collected per annotator.

significantly lower than corresponding original system outputs using a paired significance test with $p < 0.05$. We pair two HITs into a single annotation task with about 24-28 quality control segments to ensure a sufficient sample size for the statistical test. In campaigns hosted on Appraise, if an annotator is not able to demonstrate reliability on BAD references, they are excluded from further annotations, the HITs are reset and annotated from scratch by another annotator if possible.

The total number of assessments collected for each language pair and the average number of assessments per system in WMT23 manual evaluation are presented in Table 4.

5.3 Calibration HITs

Last year we introduced calibration HITs, which this year we collect for all language pairs. A calibration HIT is a HIT with 100 randomly selected segments, which is identical for and completed by all annotators, in addition to their regular annotation HITs. We release these alongside the other annotations and the anonymized mapping between annotators and HITs in order to enable additional analysis. With a small set of sentences annotated by all annotators, we are better able to examine questions about inter-annotator consistency and provide data for future research in this area.

Table 5 shows the number of unique annotators per language pair along with the total number of HITs and average number of HITs per annotator. We leave more detailed analysis of collected calibration data to future work.

5.4 Human ranking computation

The official rankings shown in Table 6 are generated on the basis of the segment-level raw DA+SQM scores that are collected within document context for all language pairs.[23] Whole documents with at least one quality control segment (i.e., BAD references) and HITs that failed to pass quality control are removed prior to computing the rankings.[24]

In this year's evaluation, we have chosen not to normalize scores by discontinuing the use of z-scores, given their potential to exacerbate system comparisons (Knowles, 2021). While utilizing raw scores is not flawless—considering each annotator employs distinct annotation strategies — we have sought to counteract this by distributing

[23]The code used to generate the rankings in Table 6 can be found here: https://github.com/AppraiseDev/Appraise/blob/main/Campaign/management/commands/ComputeWMT23Results.py

[24]Two HITs for Czech→Ukrainian and one HIT for English→Czech.

German→English

Rank	Ave.	System
1-3	90.3	GPT4-5shot
1-3	89.9	Human-refA
1-5	89.6	ONLINE-A
3-6	89.1	ONLINE-B
3-6	88.8	ONLINE-W
4-7	88.0	ONLINE-Y
6-8	87.7	ONLINE-G
8-9	86.5	GTCOM_DLUT
7-9	85.3	ONLINE-M
10-11	81.8	LanguageX
10-13	80.0	Lan-BridgeMT
11-14	79.6	NLLB_MBR_BLEU
12-14	78.8	AIRC
11-14	77.9	NLLB_Greedy

English→German

Rank	Ave.	System
1-5	89.0	GPT4-5shot
1-5	88.8	ONLINE-B
1-4	88.3	ONLINE-W
2-6	88.1	ONLINE-A
4-6	88.0	ONLINE-Y
1-6	87.7	Human-refA
7-8	86.7	ONLINE-M
7-8	85.5	ONLINE-G
9	84.0	Lan-BridgeMT
10	82.7	LanguageX
11-12	76.8	NLLB_MBR_BLEU
11-12	75.7	NLLB_Greedy
13	73.6	AIRC

English→Czech

Rank	Ave.	System
1	85.4	Human-refA
2	84.1	ONLINE-W
3-5	81.8	GPT4-5shot
3-4	80.4	CUNI-GA
5-8	80.3	ONLINE-A
5-8	79.4	CUNI-DocTransformer
4-7	78.8	ONLINE-B
8-14	78.6	NLLB_MBR_BLEU
6-11	78.4	GTCOM_DLUT
8-12	77.4	CUNI-Transformer
10-14	76.8	NLLB_Greedy
9-14	75.7	ONLINE-M
10-15	75.2	ONLINE-G
13-15	75.0	ONLINE-Y
8-15	75.0	Lan-BridgeMT
16	74.1	LanguageX

Czech→Ukrainian

Rank	Ave.	System
1-3	83.7	ONLINE-B
1-3	83.6	GPT4-5shot
1-3	83.2	Human-refA
4-8	82.8	ONLINE-W
4-8	82.4	CUNI-GA
4-8	81.8	CUNI-Transformer
4-8	81.3	GTCOM_DLUT
4-8	80.6	ONLINE-A
9-11	79.5	ONLINE-G
9-13	78.7	ONLINE-Y
9-13	78.7	MUNI-NLP
10-13	77.4	Lan-BridgeMT
10-13	76.9	NLLB_MBR_BLEU
14	76.7	NLLB_Greedy

Chinese→English

Rank	Ave.	System
1-2	82.9	Lan-BridgeMT
1-2	80.9	GPT4-5shot
3-8	80.3	Yishu
3-7	80.2	ONLINE-W
5-10	80.0	ONLINE-G
3-7	79.8	ONLINE-B
4-9	79.7	ONLINE-Y
3-8	79.1	HW-TSC
6-10	77.8	ONLINE-A
10-11	77.7	IOL_Research
8-11	77.2	LanguageX
12-13	76.9	ONLINE-M
13-16	76.2	NLLB_MBR_BLEU
12-15	76.1	Human-refA
14-16	74.0	NLLB_Greedy
13-16	72.6	ANVITA

English→Chinese

Rank	Ave.	System
1-5	82.2	Yishu
1-5	82.1	Human-refA
1-7	82.1	GPT4-5shot
3-8	82.0	Lan-BridgeMT
1-6	81.8	ONLINE-B
1-8	81.5	HW-TSC
4-8	81.4	ONLINE-W
5-8	80.2	ONLINE-Y
9-10	79.8	IOL_Research
9-10	79.7	ONLINE-A
11-13	78.6	LanguageX
11-13	78.2	ONLINE-M
11-13	77.1	ONLINE-G
14	64.5	ANVITA
15	64.3	NLLB_Greedy
16	57.2	NLLB_MBR_BLEU

Japanese→English

Rank	Ave.	System
1	81.3	GPT4-5shot
2-4	80.6	SKIM
3-8	80.4	Human-refA
3-8	79.5	ONLINE-Y
2-8	79.4	ONLINE-B
3-9	79.2	ONLINE-A
2-8	78.8	ONLINE-W
3-8	78.4	NAIST-NICT
8-9	76.9	GTCOM_DLUT
10-13	76.4	Lan-BridgeMT
10-13	75.8	ANVITA
10-13	74.8	ONLINE-G
10-13	74.6	LanguageX
14-15	72.9	ONLINE-M
14-15	72.4	KYB
16	68.9	AIRC
17-18	66.7	NLLB_MBR_BLEU
17-18	66.1	NLLB_Greedy

English→Japanese

Rank	Ave.	System
1-2	80.7	Human-refA
2-6	79.5	GPT4-5shot
1-5	78.8	ONLINE-B
2-6	78.6	ONLINE-Y
2-5	78.5	SKIM
4-6	78.4	ONLINE-W
7-10	76.6	LanguageX
7-10	76.2	ONLINE-A
7-10	76.1	NAIST-NICT
7-10	75.2	Lan-BridgeMT
11-12	73.1	ANVITA
11-12	72.6	ONLINE-M
13-15	70.8	KYB
13-15	69.6	AIRC
13-15	69.6	ONLINE-G
16	64.5	NLLB_Greedy
17	61.3	NLLB_MBR_BLEU

Table 6: Official results of WMT23 General Translation Task. Systems ordered by DA score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$; rank ranges indicate the number of systems a system significantly underperforms or outperforms; grayed entry indicates resources that fall outside the constraints provided. All language pairs used document-level evaluation.

systems evenly across annotators. This approach aims to minimize the potential bias of a particularly stringent annotator disproportionately penalizing a single system. Ideally, every annotator would assess documents translated by all systems; however, this could introduce task repetitiveness concerns. For future considerations, employing calibration HITs (see Section 5.3) to normalize each annotator's behaviour could offer a promising solution.

All segment-level scores are averaged per system to compute the system-level scores. The clusters are computed using the Wilcoxon rank-sum test with $p < 0.05$. 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 that the system in questions significantly wins against.

Tables with head-to-head comparisons between all systems are included in Appendix G.

At the time of preparation of the camera-ready version of the paper, we have not been able to collect the required number of high-quality assessments for language pairs run through Toloka AI that would meet WMT standards for human evaluation. In that regard, we decided not to publish official rankings based on manual evaluation for English↔{Hebrew, Russian, Ukrainian} until the conference, we are planning to address it later.

5.5 Comparison of human evaluation methods

In collaboration with the metrics shared task (Freitag et al., 2023), human annotation data for the Chinese→English and English→German direction was collected using two different approaches: the source-based DA+SQM approach, and the Multi-dimensional Quality Metrics (MQM) framework (Freitag et al., 2021). We present the rankings produced by the two approaches in Table 7.

Upon examining the system rankings and individual clusters produced by both techniques, it is evident that DA+SQM produces fewer clusters. This suggests that it might not be sufficiently robust to differentiate smaller system differences, whereas MQM creates more detailed clusters. One potential explanation is that DA+SQM, constrained by budgetary restrictions, might be under-powered. As highlighted by Wei et al. (2022), the 1500 segments we gather per system might not suffice to segregate systems in a more detailed manner.

Conversely, the largest difference in the evaluation techniques is the cost. While MQM manages to establish more refined clusters, its deployment is significantly more costly and complex, especially when training professionals. An interesting question would be determining the number of MQM labels that could be procured within the budget allocated for DA+SQM.

It is also important to note that the set of data over which each of these rankings was produced may have differed slightly due to the sampling (e.g., the distribution over topic domains or the amount of coverage of the full test set), making it difficult to determine whether these differences in rankings represent differences due to data or due to different annotation methods.

6 Test Suites

As can be seen in the general MT task, the improvement of translation quality has made it difficult to discriminate MT output from human translation with the current evaluation methods. Nevertheless, there are still cases where MT has difficulties, delivering outputs which despite seeming fluent and being surrounded by other seemingly perfect translations, entail serious flaws. In general evaluation methods, such flaws can get "hidden in the average" or simply get missed altogether. In an effort to shed light to these cases, evaluation via test suites is embedded in the shared task.

6.1 Setup of the sub-task

Test suites are custom extensions to standard test sets, constructed so that they can focus on particular aspects of the MT output. Here, the evaluation of the MT outputs takes place in a decentralized manner as a part of a sub-task, where test suite providers were invited to submit their customized test sets, following the setting introduced at the Third Conference on Machine Translation (Bojar et al., 2018).

Every test suite provider submitted a source-side test set, which the shared task organizers appended to the standard test sets of the shared task. The corresponding outputs from the MT systems of the shared task were returned to the test suite providers, who were responsible for running the evaluation, based on their own custom evaluation methods. The results of each test suite evaluation, together with the relevant analysis, appear in separate description papers.

Rank	Ave. ↑	System (En-De)
1-5	89.0	GPT4-5shot
1-5	88.8	ONLINE-B
1-4	88.3	ONLINE-W
2-6	88.1	ONLINE-A
4-6	88.0	ONLINE-Y
1-6	87.7	Human-refA
7-8	86.7	ONLINE-M
7-8	85.5	ONLINE-G
9	84.0	Lan-BridgeMT
10	82.7	LanguageX
11-12	76.8	NLLB_MBR_BLEU
11-12	75.7	NLLB_Greedy
13	73.6	AIRC

System (En-De)	MQM ↓
refA	2.96
GPT4-5shot	3.72
ONLINE-W	3.95
ONLINE-B	4.71
ONLINE-Y	5.64
ONLINE-A	5.67
ONLINE-G	6.57
ONLINE-M	6.94
Lan-BridgeMT	8.67
LanguageX	9.25
NLLB_Greedy	9.54
NLLB_MBR_BLEU	10.79
AIRC	14.23

Rank	Ave. ↑	System (Zh-En)
1-2	82.9	Lan-BridgeMT
1-2	80.9	GPT4-5shot
3-8	80.3	Yishu
3-7	80.2	ONLINE-W
5-10	80.0	ONLINE-G
3-7	79.8	ONLINE-B
4-9	79.7	ONLINE-Y
3-8	79.1	HW-TSC
6-10	77.8	ONLINE-A
10-11	77.7	IOL_Research
8-11	77.2	LanguageX
12-13	76.9	ONLINE-M
13-16	76.2	NLLB_MBR_BLEU
12-15	76.1	Human-refA
14-16	74.0	NLLB_Greedy
13-16	72.6	ANVITA

System (Zh-En)	MQM ↓
Lan-BridgeMT	2.10
GPT4-5shot	2.31
Yishu	3.23
ONLINE-B	3.39
HW-TSC	3.40
ONLINE-A	3.79
ONLINE-Y	3.79
ONLINE-G	3.86
ONLINE-W	4.06
LanguageX	4.23
IOL_Research	4.59
refA	4.83
ONLINE-M	5.43
ANVITA	6.08
NLLB_MBR_BLEU	6.36
NLLB_Greedy	6.57

Table 7: Comparison of system clustering as done by DA+SQM and MQM technique. Top two tables are for English to German, while bottom two are for Chinese to German.

6.2 Submissions

The test suite sub-task received 5 submissions with 6 test suites, whose overview can be seen in Table 8. The descriptions of each submission and their main findings are given below.

DFKI (Manakhimova et al., 2023) test suite offers a fine-grained linguistically motivated analysis of the shared task MT outputs, 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 efforts (e.g. Avramidis et al., 2018; Macketanz et al., 2022), the submission of this year includes an updated test set featuring new linguistic phenomena and focuses additionally on the participating LLMs. The evaluation spans German→English, English→German, and English→Russian language directions.

Some of the phenomena with the lowest accuracies for German→English are *idioms* and *resultative predicates*. For English→German, these include *mediopassive voice*, and *noun formation(er)*. As for English→Russian, these include *idioms* and *semantic roles*. GPT4 performs equally or comparably to the best systems in German→English and English→German but falls in the second significance cluster for English→Russian.

HW-TSC (Chen et al., 2023) propose a systematic approach to select test sentences with high-level of difficulty from the Wiki Corpus. The strategy considers the difficulty level of a sentence from four dimensions: *word difficulty, length difficulty, grammar difficulty* and *model learning difficulty*. They open-source two Multifaceted Challenge Sets for Chinese→English and English→Chinese, each of them containing 2,000 sentences. Then, they use these challenge sets to test the shared task systems, presenting results by three automatic metrics.

The resulting system ranks are quite different from the official results. The authors point out that systems that perform well on average test sets may not perform as well on sets with high difficulty. If the ranking difference is caused by domain issues, the top-ranked systems on the official test sets may not be so general. GPT4 is ranked in the first two positions in Chinese→English but its rank in

Test suite	Directions	Phenomena	#Sentences	Citation	Link
DFKI	de–en, en–de, en–ru	110 linguistic phenomena	11,517	Manakhimova et al. (2023)	DFKI-NLP
HW-TSC	zh–en, en–zh	4 difficulty dimensions	4,000	Chen et al. (2023)	HwTsc
IIIT HYD	en–de	5 domains, 5 writing styles	2,268	Mukherjee and Shrivastava (2023)	wmt23
INES	en–de	Inclusive language forms	162	Savoldi et al. (2023)	fbk.eu
MuST-SHE	en–de	Binary gender bias	200	Savoldi et al. (2023)	fbk.eu
RoCS-MT	en–de, en–cs, en–uk, en–ru	Non-standard user-generated content	1,922	Bawden and Sagot (2023)	RoCS-MT

Table 8: Overview of the participating test suites.

English→Chinese is much lower (ranks 4-9).

IIIT HYD (Mukherjee and Shrivastava, 2023) This test suite covers five specific domains (entertainment, environment, health, science, legal) and spans five distinct writing styles (descriptive, judgments, narrative, reporting, technical-writing) for English–German. The authors conduct their analysis through a combination of au- tomated assessments and manual evaluations.

Based on their evaluation, it is evident that both ONLINE-B and ONLINE-Y consistently surpassed other MT systems in performance across a diverse array of writing styles and domains. When focusing on GPT4, whereas it performs comparably to the best systems for most domains and writing styles, it gives considerably worse results when applied to the legal domain, and the writing style of judgments.

MuST-SHE^{WMT23} and INES (Savoldi et al., 2023) By focusing on the en-de and de-en language pairs, the authors rely on these newly created test suites to investigate systems' ability to translate feminine and masculine gender and produce gender-inclusive translations. Furthermore, they discuss metrics associated with the test suites and validate them by means of human evaluations.

The results indicate that systems achieve reasonable and comparable performance in correctly translating both feminine and masculine gender forms for naturalistic gender phenomena. Instead, the generation of inclusive language forms in translation emerges as a challenging task for all the evaluated MT models, indicating room for future improvements and research on the topic.

Concerning GPT 4, it is noticeable that its overall accuracy is 2% worse than the best MT system, whereas it achieves a relatively low accuracy with regard to the feminine gender, when evaluating whether the first-person singular references to the speaker are translated according to the speaker's linguistic expression of gender.

RoCS-MT (Bawden and Sagot, 2023) The RoCS-MT Challenge Set is designed to test MT systems' robustness to user-generated content (UGC) displaying non-standard characteristics, such as spelling errors, devowelling, acronymisation, etc. It is composed of non-standard English comments from Reddit, manually normalised and professionally translated into four of the WMT 2023 target languages, German, Czech, Ukrainian and Russian, and also French.

Through automatic and manual analysis of system outputs, we find that many of the phenomena remain challenging for most systems, but to varying degrees depending on the phenomenon, the particular instance (notably how frequent the non-standard word is) and the system, especially with respect to the quantity of training data. For example, non-standard instances of words (e.g. through devowelling or through phonetically inspired spelling) are often either omitted in the translation or copied unchanged. When non-standard words are translated, it is often in their standard form, but with some exceptions, for example capitalisation is sometimes preserved. However, there is often inconsistency within a same system's outputs.

GPT4-5shot has a clear lead over all other systems, correctly translating even some of the most challenging examples. It sometimes (although inconsistently) reproduces non-standardness in its outputs, but also does not always remain entirely faithful to the source sentence. However, aside the huge disparity in the amount of training data compared to other systems, notably the constrained ones, the lack of access to its training data is a serious obstacle to any meaningful scientific comparison; we cannot know which phenomena were seen during training and how frequently, and more

crucially, we cannot verify whether RoCS-MT sentences were seen during training.

7 Conclusions

The General Machine Translation Task at WMT 2023 covered 14 translation pairs, where the only non-English language pair was Czech→Ukrainian. Source based DA+SQM was the main human golden truth. The evaluation included 72 primary submissions from 17 participants, 6 online systems and 3 additional contrastive systems including GPT4. It was performed by 155 human (semi-)professional annotators, who contributed more than 175,000 judgments altogether. For most language pairs (apart from English→Czech), MT systems produce outputs that cannot be identified as being worse than the manually produced references translations in a statistically significant way, using our current evaluation methods.

It is apparent that this year, the amount of unconstrained submissions are lower thank in past years (27 submissions by 11 participants). Additionally, for some language pairs there are only few submissions by participants, and therefore they are dominated by many online systems, of whom we have no technical descriptions. We are therefore considering ways to encourage participation in the future, whereas redefining the constrained setting may be needed.

It is the first time that Large Language Models (LLMs) are included in the Shared Task as translation systems. Although the technology is very apparent in NLP research, we received only one submission using LLM methods (Lan-BridgeMT), whereas one dominant commercial LLM (GPT4) was included via our own efforts. GPT4 was in the first significance cluster for all systems translating towards English, but fell in the second significance cluster (rank 3-5) for English→Czech, whereas a similar sign was given by one of the test suites for English→Russian (rank 3; Manakhimova et al., 2023). Additionally, test suites providers noted that GPT4 outputs are not always faithful to the source sentence (Bawden and Sagot, 2023) and that they have some issues with speaker gender translation (Savoldi et al., 2023) and specific domains (Mukherjee and Shrivastava, 2023, e.g. legal;). Due to the closed-source nature of commercial tools, it is hard to know the exact reasons for these findings, although they confirm previous observations that GPT models have difficulties with under-represented languages (Hendy et al., 2023). We believe that a more transparent comparison including open source LLMs should be sought for the future.

8 Limitations

We investigated a research question of testing general capabilities of MT systems. However, we have simplified this approach. Firstly, we only used four domains that are not specialized. Secondly, we used only cleaner sentences, avoiding noisy in the source sentences.

Although we accept human judgement as a 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 quality of other evaluated systems (Mathur et al., 2020).

Different annotators are using different ranking strategy which may have an effect on the system ranking as we are using raw scores.

9 Ethical Consideration

Several of the domains contained texts that included personal data, for example the speech data (See Section 2.4 for more details). Entities were replaced by anonymisation tags (e.g. #NAME#, #EMAIL#) to preserve the anonymity of the users behind the content.

The sentences in Ukrainian datasets 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.

As described in Section 2.2 and in the linguistic brief (Appendix Section B), inappropriate, controversial and/or explicit content was filtered out prior to translation, particularly keeping in mind the translators and not exposing them to such content or obliging them to translate it. A few sentences containing explicit content managed to escape the filter, and we removed these sentences from the test sets without translation.

Human evaluation using Appraise for collecting human judgements was fully anonymous. Automatically generated accounts associated with annotation tasks with single-sign-on URLs were dis-

tributed randomly among pools of annotators and did not allow for storing personal information. For language pairs for which we used calibration HITs, we received lists of tasks completed by an individual anonymous annotator. Annotators have been well paid in respect to their countries.

Acknowledgments

This task would not have been possible without the sponsorship of monolingual data, test sets translation and evaluation from our partners. Namely Microsoft, Charles University, Toloka AI, Google, NTT Resonant, Dubformer, and Centific.

Additionally, we would like to thank Rebecca Knowles, Sergio Bruccoleri, Mariia Anisimova and many others who provided help and recommendations.

Barry Haddow's participation was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10039436 – UTTER].

Rachel Bawden's participation was funded by her chair position in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and by her Emergence project, DadaNMT, funded by Sorbonne Université.

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 GAČR EXPRO grant LUSyD (GX20-16819X).

Eleftherios Avramidis's participation was funded by the German Research Foundation (DFG) through the project TextQ (grant num. MO 1038/31-1, 436813723), and by the German Federal Ministry of Education and Research (BMBF) through the project SocialWear (grant num. 01IW2000).

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

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2018. Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 243–248, Boston, MA. Association for Machine Translation in the Americas.

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, Online. 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, Online. 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, Florence, Italy. 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 (WMT)*, Singapore, Singapore (Hybrid). 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, Sofia, Bulgaria. 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, Baltimore, Maryland, USA. 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, Copenhagen, Denmark. 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, Berlin, Germany. 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, Lisbon, Portugal. 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, Belgium, Brussels. Association for Computational Linguistics.

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, Prague, Czech Republic. 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, Columbus, Ohio. 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, Uppsala, Sweden. 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, Montréal, Canada. 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, Athens, Greece. 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, Edinburgh, Scotland. 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, Trento, Italy. European Association for Machine Translation.

Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang, Zongyao Li, Jiaxin GUO, Ning Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. Multifaceted Challenge Set for Evaluating Machine Translation Performance. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Jan Christian Blaise Cruz. 2023. Samsung R&D Institute Philippines at WMT 2023. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hiroyuki Deguchi, Kenji Imamura, Yuto Nishida, Yusuke Sakai, Justin Vasselli, and Taro Watanabe. 2023. NAIST-NICT WMT'23 general mt task submission. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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, Santa Fe, New Mexico. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. 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, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Chi kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric 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 Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tirthankar Ghosal, Marie Hledíková, Muskaan Singh, Anna Nedoluzhko, and Ondřej Bojar. 2022. The second automatic minuting (AutoMin) challenge: Generating and evaluating minutes from multi-party meetings. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 1–11, Waterville, Maine, USA and virtual meeting. 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, Istanbul, Turkey. 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, Online. 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, Sofia, Bulgaria. 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, Online. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

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

Josef Jon, Martin Popel, and Ondřej Bojar. 2023. CUNI-GA submission at WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. 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.

Rebecca Knowles. 2021. On the stability of system rankings at WMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 464–477, Online. Association for Computational Linguistics.

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

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. 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.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn and Christof Monz, editors. 2006. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City.

Keito Kudo, Takumi Ito, Makoto Morishita, and Jun Suzuki. 2023. SKIM at WMT 2023 general translation task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Ben LI, Yoko Matsuzaki, and Shivam Kalkar. 2023. KYB general machine translation systems for WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

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 Zettlemoyer Luke. 2022. Mega: Moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*.

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, Marseille, France. European Language Resources Association.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can ChatGPT Outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Luo Min, yixin tan, and Qiulin Chen. 2023. Yishu: Yishu at WMT2023 translation task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Alexander Molchanov and Vladislav Kovalenko. 2023. PROMT systems for WMT23 shared general translation task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

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, Marseille, France. European Language Resources Association.

Ananya Mukherjee and Manish Shrivastava. 2023. IIIT HYD's Submission for WMT23 Test-suite task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, Katja Tissi, and Davy Van Landuyt. 2023. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eight Conference on Machine Translation (WMT)*, Singapore. Association for Computational Linguistics.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR minuting corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3174–3182, Marseille, France. European Language Resources Association.

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

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 Biomedical Translation Shared Task: Evaluation of ChatGPT 3.5 as a Comparison System. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

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

Stephan Peitz, Sarthak Garg, Udhay Nallasamy, and Matthias Paulik. 2019. Cross+Self-Attention for Transformer Models. `https://github.com/pytorch/fairseq/files/3561282/paper.pdf`.

Martin Popel. 2020. CUNI English-Czech and English-Polish systems in WMT20: Robust document-level training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 269–273, Online. Association for Computational Linguistics.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Matīss Rikters. 2018. Impact of Corpora Quality on Neural Machine Translation. In *In Proceedings of the 8th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2018)*, Tartu, Estonia. IOS Press.

Matīss Rikters and Makoto Miwa. 2023. AIST AIRC submissions to the WMT23 shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). 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, Gothenburg, Sweden. Association for Computational Linguistics.

Pavel Rychlý and Yuliia Teslia. 2023. MUNI-NLP submission for Czech-Ukrainian translation task at WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test Suites Task: Evaluation of Gender Fairness in MT with MuST-SHE and INES. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). 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, Online. Association for Computational Linguistics.

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.

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, Istanbul, Turkey. 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, Brussels, Belgium. 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, Online. Association for Computational Linguistics.

Johnny Wei, Tom Kocmi, and Christian Federmann. 2022. Searching for a higher power in the human evaluation of MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 129–139, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Di Wu and Christof Monz. 2023. Beyond shared vocabulary: Increasing representational word similarities across languages for multilingual machine translation. *arXiv preprint arXiv:2305.14189*.

Di Wu, Shaomu Tan, David Stap, Ali Araabi, and Christof Monz. 2023a. UvA-MT's participation in the WMT 2023 general translation shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe YU, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin GUO, Yuhao Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023b. Treating general mt shared task as a multi-domain adaptation problem: Hw-tsc's submission to the WMT23 general mt shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

Hui Zeng. 2023. Achieving state-of-the-art multilingual translation model with minimal data and parameters. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Wenbo Zhang. 2023. IOL Research machine translation systems for WMT23 general machine translation shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

Hao Zong. 2023. GTCOM neural machine translation systems for WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

A Statistics of training data

This section describes statistics of the training corpora.

Dataset ID	Segs	Tokens		Chars	
eng-ces		**eng**	**ces**	**eng**	**ces**
Facebook-wikimatrix-1-ces-eng	2.09M	33.56M	29.66M	206.82M	216.62M
ParaCrawl-paracrawl-9-eng-ces	50.63M	692.12M	626.34M	4.33B	4.68B
Statmt-commoncrawl_wmt13-1-ces-eng	161.84k	3.35M	2.93M	20.66M	20.75M
Statmt-europarl-10-ces-eng	644.43k	15.63M	13.00M	94.31M	98.14M
Statmt-news_commentary-16-ces-eng	253.27k	5.46M	4.96M	34.58M	37.97M
Statmt-wikititles-3-ces-eng	410.94k	1.03M	965.62k	7.47M	7.57M
Tilde-ecb-2017-ces-eng	3.10k	52.12k	45.21k	327.57k	339.24k
Tilde-eesc-2017-ces-eng	1.33M	28.78M	25.63M	188.53M	205.14M
Tilde-ema-2016-ces-eng	495.23k	7.64M	7.28M	50.31M	57.01M
Tilde-rapid-2019-ces-eng	263.29k	5.79M	5.30M	37.36M	41.26M
(Total)	56.29M	793.41M	716.10M	4.97B	5.36B
eng-deu		**eng**	**deu**	**eng**	**deu**
Facebook-wikimatrix-1-deu-eng	6.23M	100.50M	96.95M	623.66M	701.23M
ParaCrawl-paracrawl-9-eng-deu	278.31M	4.27B	3.99B	26.37B	29.46B
Statmt-commoncrawl_wmt13-1-deu-eng	2.40M	51.40M	47.05M	314.18M	340.51M
Statmt-europarl-10-deu-eng	1.82M	45.51M	42.41M	272.94M	312.14M
Statmt-news_commentary-16-deu-eng	388.48k	8.55M	8.77M	54.40M	65.94M
Statmt-wikititles-3-deu-eng	1.47M	3.61M	3.08M	26.48M	25.50M
Tilde-airbaltic-1-deu-eng	0.84k	17.60k	15.08k	104.34k	105.52k
Tilde-czechtourism-1-deu-eng	6.76k	128.29k	114.44k	769.04k	829.41k
Tilde-ecb-2017-deu-eng	4.15k	85.52k	74.81k	545.51k	582.63k
Tilde-eesc-2017-deu-eng	2.86M	61.47M	58.28M	400.37M	469.94M
Tilde-ema-2016-deu-eng	347.63k	5.09M	5.01M	33.48M	39.43M
Tilde-rapid-2016-deu-eng	1.03M	20.65M	19.85M	134.26M	158.13M
Tilde-rapid-2019-deu-eng	939.81k	19.90M	19.30M	129.03M	153.08M
(Total)	295.81M	4.59B	4.29B	28.36B	31.73B
eng-heb		**eng**	**heb**	**eng**	**heb**
ELRC-wikipedia_health-1-eng-heb	3.16k	69.71k	54.76k	442.38k	583.87k
Facebook-wikimatrix-1-eng-heb	2.04M	35.83M	28.96M	218.77M	300.61M
Neulab-tedtalks_train-1-eng-heb	211.82k	4.45M	3.44M	22.36M	29.00M
OPUS-bible_uedin-v1-eng-heb	62.20k	1.55M	830.23k	8.16M	7.46M
OPUS-ccmatrix-v1-eng-heb	25.23M	313.87M	249.49M	1.81B	2.45B
OPUS-elrc_2922-v1-eng-heb	3.16k	69.73k	54.77k	442.40k	583.54k
OPUS-elrc_3065_wikipedia_health-v1-eng-heb	3.16k	69.71k	54.76k	442.31k	583.51k
OPUS-elrc_wikipedia_health-v1-eng-heb	3.16k	69.71k	54.76k	442.31k	583.51k
OPUS-globalvoices-v2018q4-eng-heb	1.03k	20.31k	15.03k	122.39k	158.63k
OPUS-gnome-v1-eng-heb	0.15k	0.42k	0.40k	2.89k	3.96k
OPUS-kde4-v2-eng-heb	79.32k	338.22k	347.35k	2.09M	3.13M
OPUS-multiccaligned-v1-eng-heb	5.33M	60.55M	52.81M	380.74M	518.33M
OPUS-opensubtitles-v2018-eng-heb	29.89M	195.98M	154.25M	1.03B	1.40B
OPUS-php-v1-eng-heb	27.82k	83.46k	93.03k	498.72k	789.34k
OPUS-qed-v2.0a-eng-heb	464.35k	6.37M	4.48M	34.70M	42.34M
OPUS-tatoeba-v20220303-eng-heb	164.20k	1.02M	806.38k	5.41M	7.37M
OPUS-tatoeba-v2-eng-heb	54.36k	357.09k	277.32k	1.87M	2.56M
OPUS-ubuntu-v14.10-eng-heb	1.44k	6.13k	5.78k	38.78k	54.69k
OPUS-wikimedia-v20210402-eng-heb	226.83k	8.51M	7.56M	57.58M	78.26M
OPUS-wikipedia-v1.0-eng-heb	139.85k	2.69M	2.27M	16.45M	22.43M
OPUS-xlent-v1.1-eng-heb	3.19M	9.61M	7.93M	60.53M	73.11M
Statmt-ccaligned-1-eng-heb_IL	5.33M	60.55M	52.81M	380.76M	518.34M
(Total)	72.46M	702.05M	566.59M	4.04B	5.45B

Table 9: Statistics for parallel training set provided for General/News Translation Task. Suffixes, k, M, and B, are short for thousands, millions, and billions, respectively. Dataset ID is the unique identifier created by MTData, example `mtdata echo <dataset_id>`.

Dataset ID	Segs	Tokens		Chars	
eng-jpn		**eng**		**eng**	**jpn**
Facebook-wikimatrix-1-eng-jpn	3.90M	61.63M		379.09M	454.97M
KECL-paracrawl-3-eng-jpn	25.74M	599.02M		3.69B	4.58B
Phontron-kftt_train-1-eng-jpn	440.29k	9.74M		59.91M	49.08M
StanfordNLP-jesc_train-1-eng-jpn	2.80M	19.34M		104.00M	119.62M
Statmt-news_commentary-16-eng-jpn	1.84k	39.50k		247.70k	310.56k
Statmt-ted-wmt20-eng-jpn	241.74k	4.03M		23.02M	27.32M
Statmt-wikititles-3-jpn-eng	757.04k	1.94M		13.96M	18.67M
(Total)	33.88M	695.74M		4.27B	5.25B
eng-rus		**eng**	**rus**	**eng**	**rus**
Facebook-wikimatrix-1-eng-rus	5.20M	86.79M	76.48M	537.73M	965.44M
OPUS-unpc-v1.0-eng-rus	25.17M	563.82M	520.71M	3.70B	7.31B
ParaCrawl-paracrawl-1_bonus-eng-rus	5.38M	101.31M	80.41M	632.54M	1.06B
Statmt-backtrans_enru-wmt20-eng-rus	36.77M	736.20M	670.93M	4.31B	7.73B
Statmt-commoncrawl_wmt13-1-rus-eng	878.39k	18.77M	17.40M	116.16M	214.59M
Statmt-news_commentary-16-eng-rus	331.51k	7.67M	7.13M	48.79M	97.41M
Statmt-wikititles-3-rus-eng	1.19M	3.13M	2.88M	22.80M	39.34M
Statmt-yandex-wmt22-eng-rus	1.00M	21.25M	18.68M	130.99M	250.76M
Tilde-airbaltic-1-eng-rus	1.09k	23.98k	18.79k	142.52k	252.73k
Tilde-czechtourism-1-eng-rus	7.33k	140.09k	110.10k	838.09k	1.50M
Tilde-worldbank-1-eng-rus	25.85k	588.58k	573.93k	3.85M	8.21M
(Total)	75.96M	1.54B	1.40B	9.50B	17.67B
eng-ukr		**eng**	**ukr**	**eng**	**ukr**
ELRC-acts_ukrainian-1-eng-ukr	129.94k	3.04M	2.60M	19.55M	35.69M
Facebook-wikimatrix-1-eng-ukr	2.58M	41.55M	35.59M	257.56M	447.33M
ParaCrawl-paracrawl-1_bonus-eng-ukr	13.35M	505.83M	487.47M	3.28B	6.04B
Tilde-worldbank-1-eng-ukr	1.63k	36.07k	34.18k	237.96k	477.91k
(Total)	16.06M	550.46M	525.68M	3.55B	6.52B
eng-zho		**eng**		**eng**	**zho**
Facebook-wikimatrix-1-eng-zho	2.60M	49.87M		311.07M	277.84M
OPUS-unpc-v1.0-eng-zho	17.45M	417.25M		2.75B	2.14B
ParaCrawl-paracrawl-1_bonus-eng-zho	14.17M	217.60M		1.34B	1.18B
Statmt-backtrans_enzh-wmt20-eng-zho	19.76M	364.22M		2.16B	1.96B
Statmt-news_commentary-16-eng-zho	313.67k	6.92M		44.14M	38.83M
Statmt-wikititles-3-zho-eng	921.96k	2.37M		17.82M	16.28M
(Total)	55.22M	1.06B		6.62B	5.61B
ces-ukr		**ces**	**ukr**	**ces**	**ukr**
ELRC-acts_ukrainian-1-ces-ukr	130.00k	2.48M	2.56M	19.61M	35.26M
Facebook-wikimatrix-1-ces-ukr	848.96k	10.43M	10.07M	75.97M	127.31M
OPUS-bible_uedin-v1-ces-ukr	7.95k	140.03k	132.06k	904.31k	1.33M
OPUS-ccmatrix-v1-ces-ukr	3.99M	45.13M	45.10M	330.68M	566.27M
OPUS-elrc_5179_acts_ukrainian-v1-ces-ukr	130.00k	2.48M	2.56M	19.61M	35.26M
OPUS-elrc_wikipedia_health-v1-ces-ukr	0.19k	3.23k	3.18k	24.27k	41.63k
OPUS-eubookshop-v2-ces-ukr	1.51k	23.71k	19.15k	187.30k	275.14k
OPUS-gnome-v1-ces-ukr	0.15k	0.42k	0.41k	3.53k	5.82k
OPUS-kde4-v2-ces-ukr	133.67k	593.82k	677.35k	4.45M	7.97M
OPUS-multiccaligned-v1.1-ces-ukr	1.61M	19.75M	19.77M	146.44M	244.36M
OPUS-multiparacrawl-v9b-ces-ukr	2.20M	25.62M	25.55M	188.08M	325.50M
OPUS-opensubtitles-v2018-ces-ukr	730.80k	3.88M	3.90M	24.20M	40.62M
OPUS-qed-v2.0a-ces-ukr	161.02k	2.02M	2.04M	13.44M	22.80M
OPUS-tatoeba-v20220303-ces-ukr	2.93k	10.85k	11.40k	68.70k	118.67k
OPUS-ted2020-v1-ces-ukr	114.23k	1.57M	1.56M	10.70M	17.93M
OPUS-ubuntu-v14.10-ces-ukr	0.23k	1.67k	1.76k	13.02k	20.86k
OPUS-wikimedia-v20210402-ces-ukr	1.96k	39.18k	34.91k	285.74k	414.20k
OPUS-xlent-v1.1-ces-ukr	695.41k	1.78M	1.58M	12.92M	18.30M
(Total)	10.76M	115.95M	115.57M	847.58M	1.44B

Table 10: Statistics for parallel training set provided for General/News Translation Task. Suffixes, k, M, and B, are short for thousands, millions, and billions, respectively.

B Preprocessing cleanup brief for linguists

Human check briefing

In this task, we wish to check the data to remove all inappropriate content, remove repetitive content, or correct minor problems with the text.

The data is automatically broken down into individual sentences, which may contain wrong sentence splitting that needs to be fixed. Each paragraph is separated by empty lines. Keep the document-separators intact.

We ask you to read each document and either:

- Delete document completely if it contains any of following issues. Be on the save side, rather remove documents where you are uncertain
    - Remove documents written in different language (natural code-switching is fine)
    - Remove inappropriate content (such as sexually explicit, vulgar, or otherwise inappropriate)
    - Remove controversial content (propagandist, controversial political topics, etc.)
    - Remove content that is too noisy or doesn't resemble natural text (such as documents badly formatted, hard to understand, containing unusual language, lists of numbers/data, or other structured data generated automatically)
- Keep document while checking
    - Fix sentence-breaking, each line must be one sentence (do not reformulate, simply remove or add end of lines on a proper place).
    - Remove or move fragments of sentences to previous or following sentence (for example emoticons, one or few words sentences)
    - Fix minor issues and keep it (do not spent too much time on fixing it).
        - It is fine to keep some errors or problems
        - Remove boilerplates (segments that break the document, for example ads, page numbers, signatures, artefacts, …)
    - If a given document has more than around 30 sentences, consider splitting it by adding an empty line on a meaningful place splitting it into paragraphs

This task shouldn’t take much longer than reading through documents.

C Translator Brief for General MT

Translator Brief

In this project we wish to translate online news articles 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 news written directly in the target language. However, there are some constraints imposed by the intended usage:

- All translations should be "from scratch", without post-editing from MT. Using post-editing would bias the evaluation, so we need to avoid it. We can detect post-editing so will reject translations that are post-edited.
- Translation should preserve the sentence boundaries. The source texts are provided with exactly one sentence per line, and the translations should be the same, one sentence per line. Blank lines should be preserved in the translation.
- 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 a sample of the translations for quality, and we will check the entire set for evidence of post-editing.
- Please do not translate the anonymization tags (e.g. #NAME#), but use the same form as in the source text. These tags are used to de-identify names and various other sensitive data. In other words, translation must contain given tag #NAME# on a position where it would naturally be placed before anonymization.
- If the original data contain errors, typos, or other problems, do not try to fix them (or introduce them in the translation), instead try to prepare correct translation as if the error wouldn't be in the source.

The source files will be delivered as text files (sometimes known as "notepad" files), with one sentence per line. We need the translations to be returned in the same format. If you prefer to receive the text in a different format, then please let us know as we may be able to accommodate it.

D Additional statistics of the test sets

Table 11 shows the type-token ratios for the source and target side of each of the test sets, shown for the four main domains. As mentioned previously, texts are tokenised using the language-specific Spacy models (Honnibal and Montani, 2017) where available. For Hebrew, we use the multilingual Spacy model as no language-specific model is available. The type-token ratio is calculated as the number of unique tokens divided by the total number of tokens. The absolute value depends not only on the lexical diversity of the text but also on the morphological complexity of the language in question.

	manuals		mastodon		news		user_review	
	src	trg	src	trg	src	trg	src	trg
From English								
en–cs	–	–	0.30	0.42	0.27	0.39	0.22	0.35
en–de	–	–	0.30	0.32	0.27	0.29	–	–
en–he	–	–	0.30	0.30	0.27	0.29	0.22	0.24
en–ja	–	–	0.30	0.23	0.27	0.19	0.22	0.17
en–ru	–	–	0.30	0.41	0.27	0.38	0.22	0.33
en–uk	–	–	0.30	0.41	0.27	0.38	0.22	0.34
en–zh	–	–	0.30	0.29	0.27	0.26	0.22	0.21
Other language directions								
cs–uk	–	–	–	–	0.43	0.41	–	–
de–en	0.32	0.23	0.49	0.42	0.34	0.26	–	–
he–en	–	–	–	–	0.34	0.09	–	–
ja–en	–	–	–	–	0.22	0.23	0.22	0.21
ru–en	0.47	0.28	–	–	0.40	0.24	–	–
uk–en	–	–	–	–	0.36	0.21	–	–
zh–en	0.25	0.25	–	–	0.23	0.19	0.22	0.17

Table 11: Type-token ratio for individual source languages used in the general translation test sets.

E News Task System Submission Summaries

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

E.1 AIRC (Rikters and Miwa, 2023)

AIRC trained constrained track models for translation between English, German, and Japanese. Before training the final models we first filtered the parallel and monolingual data (Rikters, 2018), then performed iterative back-translation as well as parallel data distillation to be used for non-autoregressive model training. We experimented with training Transformer models, Mega (Ma et al., 2022) models, and custom non-autoregressive sequence-to-sequence models with encoder and decoder weights initialised by multilingual BERT base. Our primary submissions contain translations from ensembles of two Mega model checkpoints and our contrastive submissions are generated by our non-autoregressive models.

E.2 ANVITA (no associated paper)

ANVITA-ZhJa Machine Translation system for WMT2023 Shared Task:General MT(News). This paper describes ANVITA-ZhJa MT system, architected for submission to WMT 2023 General Machine Translation(News) shared task by the ANVITA team, where the team participated in 4 translation directions: Chinese, Japanese→English and English→Chinese, Japanese. ANVITA-ZhJa MT system comprised of four NMT models.Chinese, Japanese→English and English→Chinese, Japanese multilingual models for primary and Chinese→English and English→Chinese bilingual models for contrastive submissions. Base MT models are built using transformer(base) architecture, trained over the organizer provided parallel corpus and subsequently used deep transformer with added layers and other parameters. We also distilled corpus using heuristics based filtering and used model ensemble for enhanced performance.

E.3 CUNI-DocTransformer (Popel, 2020)

Exactly the same system as submitted in WMT20, document-level Transformer trained with Block Backtranslation.

E.4 CUNI-GA (Jon et al., 2023)

Our submission is a result of applying a novel n-best list reranking and modification method on translation candidates produced by two other competing systems, CUNI-Transformer and CUNI-DocTransformer. Our method uses a genetic algorithm and MBR decoding to search for optimal translation under a given metric (in our case, a weighted combination of ChrF, BLEU, COMET22-DA, and COMET22-QE-DA).

E.5 CUNI-Transformer (Popel, 2020)

The English↔Czech sentence-level models are exactly the same as submitted in WMT20 (Popel, 2020). The Ukrainian↔Czech models are very similar, also trained with Block Backtranslation.

E.6 GTCOM (Zong, 2023)

GTCOM uses transformer as the basic architecture and leverages multilingual models to improve translation quality. Besides, GTCOM does a lot of data cleaning and data augmentation work.

E.7 HW-TSC (Wu et al., 2023b)

HW-TSC's submission is a standard Transformer model equipped with our recent technique.

E.8 IOL-Research (Zhang, 2023)

This paper describes the IOL Research team's submission system for the WMT23 General Machine Translation shared task. We participate in two language translation directions, including English-to-Chinese and Chinese-to-English. Our final primary submissions belong to constrained systems, which means for both translation directions we only use officially provided monolingual and bilingual data to train the translation systems. Our systems are based on Transformer architecture with pre-norm or deep-norm, which has been proven to be helpful for training deeper models. We employ methods such as back-translation, data diversification, domain fine-tuning and model ensemble to build our translation systems. Another important aspect is that we carefully conduct data cleaning and use as much monolingual data as possible for data augmentation.

E.9 TeamKYB (LI et al., 2023)

We here describe our neural machine translation system for the general machine translation shared task in WMT 2023. Our systems are based on the Transformer with base settings. We trained our model with preprocessed train data. We collect multiple checkpoint from our model and performed inference with several hyperparameter settings. Collected translations were processed via some rule-based corrections. We chose best translation from the results by using N-best ranking method.

E.10 Lan-BridgeMT (Wu and Hu, 2023)

With the emergence of large-scale models, various industries have undergone significant transformations, particularly in the realm of document-level machine translation. This has introduced a novel research paradigm that we have embraced in our participation in the WMT23 competition. Focusing on advancements in models such as chatGPT and GPT4, we have undertaken numerous prompt-based experiments. Our objective is to achieve optimal human evaluation results for document-level machine translation, resulting in our submission of the final outcomes in the general track.

E.11 MUNI-NLP (Rychlý and Teslia, 2023)

MUNI-NLP system is a standard transformer.

E.12 NAIST-NICT (Deguchi et al., 2023)

In this paper, we describe our NAIST-NICT submission to the WMT'23 English-Japanese general machine translation task. Our system generates diverse translation candidates and reranks them with a two-stage reranking system to find the best translation. We first generate 50 candidates each from 18 different translation methods using a variety of techniques to increase the diversity of the translation candidates. We trained 7 different models per language direction using different combinations of hyperparameters. From these models we used various decoding algorithms, ensembling the models, and using kNN-MT. The 900 translation candidates go through a two-stage reranking system in order to find the most promising candidate.The first step compares the 50 candidates from each translation method using DrNMT and returns the one with the highest score. The final 18 candidates are ranked using COMET-MBR, and the highest scoring is returned as the system output. We found that generating diverse translation candidates improves the translation quality by using the well-designed relanker model.

E.13 PROMT (Molchanov and Kovalenko, 2023)

This paper describes the PROMT submissions for the WMT23 Shared General Translation Task. This year we participated in two directions of the Shared Translation Task: English to Russian and Russian to English. Our models are trained with the MarianNMT toolkit using the transformer-big configuration. We use BPE for text encoding, both models are unconstrained. We achieve competitive results according to automatic metrics in both directions.

E.14 SRPH (Cruz, 2023)

We submit single-model encode-decoder Transformer systems for the constrained English to Hebrew and Hebrew to English translation directions. Our dataset is cleaned and filtered via a combination of heuristic-based, ratio-based, and embedding-based (LaBSE) methods, resulting in a dataset with high alignment. We train models with heavy use of back-translation and decode using Noisy Channel Reranking using a reverse model and a language model trained with contest data.

E.15 SKIM (Kudo et al., 2023)

The SKIM team submission took a standard procedure of building ensemble Transformer models, including base-model training, data augmentation using back-translation of base models, and retraining several final models using back-translated training data. Each final model has its own architecture and configuration, including a 10.5B parameter at most, substituting self and cross sublayers in decoder with cross+self-attention sub-layer (Peitz et al., 2019). We select the best candidate from large candidate pools, namely 70 translations generated from 16 distinct models for each sentence, with an MBR reranking method using COMET and COMET-QE (Fernandes et al., 2022). We also applied data augmentation and selection techniques to training data of the Transformer models.

E.16 UPCite-CLILLF (no associated paper)

In this biomedical shared task, we have created data filters to better "choose" relevant training data for fine-tuning, among provided training data sources. In particular, we have used the textometric analysis tool ITRAMEUR to filter the segments and terms that characterize the test set and then extracted them from training data to fine-tune MBart-50 baseline (`decoder_attention_heads: 16`, `decoder_ffn_dim: 4096`, `decoder_layers: 12`, `encoder_attention_heads: 16`, `encoder_ffn_dim: 4096`, `encoder_layers: 12`, `num_hidden_layers: 12`, `max_length: 200`, `epoch: 3`). In doing so, we hope to meet several objectives : to build feasible fine-tuning strategy to train biomedical in-domain fr<->en models ; to specify filtering criteria of in-domain training data and to compare models' predictions, fine-tuning data and test set in order to better understand how neural machine translation systems work. We will also compare the pipeline of the shared task of this year to those of the past 2 years to evaluate the benefits of our training strategies of in-domain machine translation models.

E.17 UvA-LTL (Wu et al., 2023a)

We present our WMT system, UvA-MT, in the WMT 2023 shared general translation task. This year, we developed a single Multilingual Machine Translation (MMT) system to participate in the two-directional translation track between English and Hebrew. The main architecture is based on the prior work of Beyond Shared Vocabulary (Wu and Monz, 2023). We scaled it up to a transformer-large level (422M parameters). Additionally, we employed back translation to generate synthetic data and labeled them with a new language tag. After convergence, we further fine-tuned the system without using synthetic data. Several domain shift techniques were also introduced, such as the domain-aware language model, to filter monolingual data.

E.18 YiShu (Min et al., 2023)

Yishu's team participated in WMT23 Machine Translation Competition and adopted the most advanced neural machine translation method. They use Transformer model structure and use large-scale parallel corpus for training. In order to improve the translation quality, the team adopted cutting-edge data preprocessing technology, various attention mechanisms and improved decoding strategies. In addition, they also carried out in-depth parameter adjustment and model optimization. Yishu team incorporated evaluation indicators such as BLEU and TER into the training constraints of the model to achieve better translation performance. They strive for high accuracy and fluency in the competition, and strive to achieve excellent results in the field of translation.

E.19 LanguageX (Zeng, 2023)

LanguageX's submission is a many-to-many encoder decoder transformer model.

F Automatic scores

This section contains automatic metric scores. While human judgement is the official ranking of systems and their performance, we share automatic scores to show expected system performance for various testsets.

We use COMET (Rei et al., 2020) as the primary metric and chrF (Popović, 2015) as the secondary metric, following recommendation by (Kocmi et al., 2021). We also present BLEU (Papineni et al., 2002) scores as it is still a widely used metric. The COMET scores are calculated with the default model `Unbabel/wmt22-comet-da`. The chrF and BLEU scores are calculated using SacreBLEU (Post, 2018). Scores are multiplied by 100. We ranked the systems according to their scores. Unconstrained systems are indicated with a grey background in the tables.

System	COMET
CUNI-GA	90.9
GPT4-5shot	90.8
ONLINE-W	89.4
GTCOM_Peter	88.9
ONLINE-B	88.8
ONLINE-A	88.2
CUNI-Transformer	88.0
ONLINE-G	87.7
MUNI-NLP	87.0
ONLINE-Y	86.5
NLLB_Greedy	86.3
NLLB_MBR_BLEU	86.3
Lan-BridgeMT	86.0

System	chrF
GPT4-5shot	61.0
CUNI-GA	57.9
GTCOM_Peter	57.6
CUNI-Transformer	57.4
MUNI-NLP	57.0
Lan-BridgeMT	55.7
ONLINE-W	55.0
ONLINE-B	54.7
ONLINE-A	54.4
ONLINE-G	53.7
ONLINE-Y	53.4
NLLB_Greedy	52.5
NLLB_MBR_BLEU	52.3

System	BLEU
GPT4-5shot	32.8
CUNI-Transformer	30.2
GTCOM_Peter	29.8
CUNI-GA	29.5
MUNI-NLP	28.3
Lan-BridgeMT	27.5
ONLINE-W	26.8
ONLINE-B	25.7
ONLINE-A	25.4
NLLB_MBR_BLEU	25.1
NLLB_Greedy	24.9
ONLINE-G	24.8
ONLINE-Y	24.2

Table 12: Scores for the cs→uk translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-W	91.8
CUNI-GA	90.8
ONLINE-B	89.9
GPT4-5shot	89.4
ONLINE-A	88.4
CUNI-DocTransformer	88.3
GTCOM_Peter	87.7
ONLINE-M	87.4
Lan-BridgeMT	87.3
CUNI-Transformer	87.2
NLLB_Greedy	87.1
ONLINE-Y	87.0
NLLB_MBR_BLEU	86.9
ONLINE-G	85.9
ZengHuiMT	85.4

System	chrF
ONLINE-W	76.3
ONLINE-B	70.4
ZengHuiMT	67.5
ONLINE-A	66.3
CUNI-GA	65.9
GTCOM_Peter	65.4
CUNI-DocTransformer	65.1
ONLINE-Y	64.6
CUNI-Transformer	63.9
Lan-BridgeMT	63.8
ONLINE-G	63.7
ONLINE-M	63.2
GPT4-5shot	62.3
NLLB_Greedy	60.0
NLLB_MBR_BLEU	59.1

System	BLEU
ONLINE-W	59.4
ONLINE-B	50.1
ONLINE-A	43.4
CUNI-GA	43.3
ZengHuiMT	43.1
CUNI-DocTransformer	42.5
GTCOM_Peter	42.3
CUNI-Transformer	41.4
ONLINE-Y	40.8
Lan-BridgeMT	40.7
ONLINE-G	39.6
ONLINE-M	39.6
GPT4-5shot	37.8
NLLB_Greedy	35.9
NLLB_MBR_BLEU	35.1

Table 13: Scores for the en→cs translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
GPT4-5shot	86.3
ONLINE-W	86.0
ONLINE-B	85.6
ONLINE-A	85.5
ONLINE-Y	84.9
ONLINE-M	84.8
ONLINE-G	84.6
GTCOM_Peter	82.7
NLLB_MBR_BLEU	81.4
ZengHuiMT	81.1
Lan-BridgeMT	80.9
NLLB_Greedy	79.9
AIRC	78.7

System	chrF
ONLINE-W	72.1
ONLINE-A	70.0
GPT4-5shot	69.8
ONLINE-B	69.1
ONLINE-G	69.1
ONLINE-Y	68.4
ZengHuiMT	67.6
Lan-BridgeMT	66.7
GTCOM_Peter	66.6
ONLINE-M	66.5
NLLB_MBR_BLEU	57.6
NLLB_Greedy	57.3
AIRC	57.2

System	BLEU
ONLINE-W	51.8
GPT4-5shot	47.9
ONLINE-A	47.9
ONLINE-B	46.3
ONLINE-G	46.0
ONLINE-Y	43.9
GTCOM_Peter	42.2
Lan-BridgeMT	42.1
ONLINE-M	41.3
ZengHuiMT	40.8
NLLB_Greedy	33.1
AIRC	32.4
NLLB_MBR_BLEU	32.4

Table 14: Scores for the de→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-W	85.5
GPT4-5shot	85.0
ONLINE-B	84.8
ONLINE-Y	84.1
ONLINE-A	83.7
ONLINE-G	82.5
ONLINE-M	81.7
Lan-BridgeMT	80.4
ZengHuiMT	79.4
NLLB_MBR_BLEU	78.0
NLLB_Greedy	77.9
AIRC	72.9

System	chrF
ONLINE-W	71.8
ONLINE-A	69.7
ZengHuiMT	69.4
GPT4-5shot	69.1
ONLINE-B	69.1
ONLINE-Y	69.1
ONLINE-G	69.0
ONLINE-M	66.9
Lan-BridgeMT	66.1
NLLB_Greedy	56.2
NLLB_MBR_BLEU	55.4
AIRC	52.2

System	BLEU
ONLINE-W	47.8
ONLINE-A	43.7
GPT4-5shot	43.6
ONLINE-Y	43.6
ONLINE-G	43.2
ONLINE-B	42.7
ONLINE-M	40.5
ZengHuiMT	40.5
Lan-BridgeMT	39.4
NLLB_Greedy	31.1
NLLB_MBR_BLEU	29.6
AIRC	26.5

Table 15: Scores for the en→de translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-B	89.9
ONLINE-A	87.0
GPT4-5shot	86.9
GTCOM_Peter	86.7
ONLINE-G	85.6
ZengHuiMT	85.6
ONLINE-Y	84.9
UvA-LTL	84.7
NLLB_MBR_BLEU	82.9
NLLB_Greedy	82.8
Samsung_Research_Philippines	82.6
Lan-BridgeMT	82.4

System	chrF
ONLINE-B	87.5
ZengHuiMT	76.3
GTCOM_Peter	76.2
ONLINE-A	73.3
GPT4-5shot	71.4
UvA-LTL	70.9
ONLINE-Y	70.5
ONLINE-G	69.8
NLLB_Greedy	64.4
Lan-BridgeMT	63.5
NLLB_MBR_BLEU	63.0
Samsung_Research_Philippines	55.5

System	BLEU
ONLINE-B	76.5
GTCOM_Peter	59.2
ZengHuiMT	56.6
ONLINE-A	53.9
GPT4-5shot	51.2
UvA-LTL	51.0
ONLINE-Y	49.8
ONLINE-G	49.3
NLLB_Greedy	42.5
Lan-BridgeMT	41.4
NLLB_MBR_BLEU	40.7
Samsung_Research_Philippines	34.0

Table 16: Scores for the he→en (refA) translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6 |nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
GPT4-5shot	86.4
ONLINE-B	85.6
ONLINE-A	85.3
GTCOM_Peter	84.5
ONLINE-G	84.0
UvA-LTL	83.3
ZengHuiMT	83.3
ONLINE-Y	82.9
NLLB_MBR_BLEU	81.8
NLLB_Greedy	81.7
Lan-BridgeMT	81.3
Samsung_Research_Philippines	81.3

System	chrF
GPT4-5shot	69.5
ONLINE-B	66.5
ONLINE-A	65.6
GTCOM_Peter	65.3
ZengHuiMT	65.1
UvA-LTL	63.3
ONLINE-G	62.8
ONLINE-Y	62.0
NLLB_Greedy	59.6
Lan-BridgeMT	59.0
NLLB_MBR_BLEU	58.6
Samsung_Research_Philippines	51.3

System	BLEU
GPT4-5shot	50.4
ONLINE-B	45.0
GTCOM_Peter	44.4
ONLINE-A	44.4
UvA-LTL	41.7
ZengHuiMT	41.7
ONLINE-G	40.9
ONLINE-Y	38.5
NLLB_Greedy	37.1
Lan-BridgeMT	36.2
NLLB_MBR_BLEU	36.2
Samsung_Research_Philippines	29.8

Table 17: Scores for the he→en (refB) translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6 |nw:0|space:no|version:2.3.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-B	86.4
ONLINE-A	85.7
GPT4-5shot	84.9
GTCOM_Peter	84.7
ONLINE-Y	84.7
UvA-LTL	84.2
Samsung_Research_Philippines	83.7
Lan-BridgeMT	83.0
NLLB_Greedy	82.9
ZengHuiMT	82.7
NLLB_MBR_BLEU	82.5
ONLINE-G	82.2

System	chrF
ONLINE-B	66.4
ZengHuiMT	62.1
ONLINE-A	61.7
GTCOM_Peter	61.1
ONLINE-Y	60.4
UvA-LTL	59.0
ONLINE-G	58.1
Samsung_Research_Philippines	57.3
Lan-BridgeMT	54.9
NLLB_Greedy	54.8
NLLB_MBR_BLEU	54.3
GPT4-5shot	54.0

System	BLEU
ONLINE-B	47.8
ONLINE-A	38.9
GTCOM_Peter	37.2
ONLINE-Y	37.2
ZengHuiMT	36.5
UvA-LTL	35.0
Samsung_Research_Philippines	33.3
ONLINE-G	33.2
NLLB_MBR_BLEU	30.8
Lan-BridgeMT	30.5
NLLB_Greedy	30.3
GPT4-5shot	27.0

Table 18: Scores for the en→he translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
SKIM	84.0
GPT4-5shot	83.4
ONLINE-W	82.3
NAIST-NICT	81.9
ONLINE-Y	81.6
ONLINE-B	81.5
ONLINE-A	81.0
GTCOM_Peter	80.2
ANVITA	79.5
Lan-BridgeMT	79.3
ZengHuiMT	79.2
ONLINE-G	77.8
ONLINE-M	77.5
KYB	76.6
NLLB_MBR_BLEU	75.2
AIRC	74.5
NLLB_Greedy	74.3

System	chrF
ONLINE-W	51.4
GPT4-5shot	51.2
SKIM	51.1
ONLINE-A	49.6
NAIST-NICT	49.5
ONLINE-Y	49.5
ZengHuiMT	49.5
ONLINE-B	49.3
GTCOM_Peter	48.7
Lan-BridgeMT	47.3
ANVITA	46.7
ONLINE-G	45.5
KYB	43.9
ONLINE-M	43.9
AIRC	40.5
NLLB_MBR_BLEU	39.2
NLLB_Greedy	39.0

System	BLEU
ONLINE-W	25.9
SKIM	24.8
GPT4-5shot	24.1
ONLINE-B	23.9
NAIST-NICT	23.0
ONLINE-A	23.0
ZengHuiMT	22.6
GTCOM_Peter	22.3
ONLINE-Y	22.3
ANVITA	20.9
Lan-BridgeMT	20.2
ONLINE-G	18.3
KYB	17.6
ONLINE-M	17.2
AIRC	14.9
NLLB_MBR_BLEU	14.7
NLLB_Greedy	14.2

Table 19: Scores for the ja→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-B	88.2
ONLINE-W	87.5
ONLINE-Y	87.3
GPT4-5shot	87.0
SKIM	86.6
NAIST-NICT	86.2
ZengHuiMT	85.3
ONLINE-A	85.2
Lan-BridgeMT	84.5
ONLINE-M	13.3
ANVITA	82.7
KYB	80.8
AIRC	80.7
ONLINE-G	80.4
NLLB_Greedy	79.3
NLLB_MBR_BLEU	77.7

System	chrF
ONLINE-B	35.2
ONLINE-Y	34.1
ONLINE-W	33.5
SKIM	33.5
ZengHuiMT	32.9
NAIST-NICT	32.0
ONLINE-A	31.4
GPT4-5shot	31.0
Lan-BridgeMT	30.4
ONLINE-M	29.6
ANVITA	29.3
KYB	27.7
AIRC	27.6
ONLINE-G	27.3
NLLB_Greedy	20.9
NLLB_MBR_BLEU	18.7

System	BLEU
ONLINE-B	25.3
ONLINE-W	24.5
ONLINE-Y	24.5
SKIM	24.3
NAIST-NICT	22.6
ZengHuiMT	22.6
ONLINE-A	21.4
GPT4-5shot	21.3
Lan-BridgeMT	20.5
ONLINE-M	19.8
ANVITA	19.4
KYB	17.8
AIRC	17.6
ONLINE-G	17.2
NLLB_Greedy	11.3
NLLB_MBR_BLEU	9.0

Table 20: Scores for the en→ja translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:ja-mecab-0.996-IPA|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
GPT4-5shot	83.5
ONLINE-Y	82.5
ONLINE-B	82.3
ONLINE-W	82.2
ONLINE-G	82.0
ONLINE-A	81.9
PROMT	80.9
ONLINE-M	80.7
NLLB_MBR_BLEU	80.5
NLLB_Greedy	80.1
Lan-BridgeMT	79.9
ZengHuiMT	79.5

System	chrF
GPT4-5shot	60.4
ONLINE-G	59.6
ONLINE-A	59.4
ONLINE-B	59.4
ZengHuiMT	58.9
ONLINE-Y	58.6
PROMT	58.4
ONLINE-W	58.3
Lan-BridgeMT	57.4
ONLINE-M	56.7
NLLB_MBR_BLEU	55.8
NLLB_Greedy	55.5

System	BLEU
ONLINE-B	34.5
GPT4-5shot	34.4
ONLINE-G	34.0
ONLINE-A	33.8
ONLINE-Y	33.2
ONLINE-W	33.1
PROMT	32.8
Lan-BridgeMT	31.8
ZengHuiMT	31.3
NLLB_MBR_BLEU	31.0
ONLINE-M	30.7
NLLB_Greedy	30.3

Table 21: Scores for the ru→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-G	86.6
ONLINE-W	86.6
ONLINE-B	86.2
GPT4-5shot	86.1
ONLINE-Y	85.5
ONLINE-A	85.3
ONLINE-M	83.2
Lan-BridgeMT	83.1
NLLB_Greedy	82.9
NLLB_MBR_BLEU	82.7
PROMT	82.3
ZengHuiMT	81.3

System	chrF
ONLINE-B	61.9
ONLINE-A	59.0
ONLINE-G	58.9
ZengHuiMT	58.8
ONLINE-W	56.6
ONLINE-Y	56.4
GPT4-5shot	56.2
Lan-BridgeMT	55.7
PROMT	55.4
ONLINE-M	55.1
NLLB_Greedy	53.3
NLLB_MBR_BLEU	53.1

System	BLEU
ONLINE-B	40.4
ONLINE-A	34.8
ONLINE-G	32.9
ONLINE-Y	32.0
ZengHuiMT	31.6
ONLINE-W	31.4
ONLINE-M	30.9
Lan-BridgeMT	30.7
GPT4-5shot	30.6
PROMT	30.5
NLLB_MBR_BLEU	28.4
NLLB_Greedy	28.2

Table 22: Scores for the en→ru translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-W	87.5
GPT4-5shot	87.1
ONLINE-B	86.8
GTCOM_Peter	86.3
ONLINE-A	86.3
ONLINE-G	86.2
ONLINE-Y	85.8
Lan-BridgeMT	84.8
ZengHuiMT	84.4
NLLB_MBR_BLEU	84.3
NLLB_Greedy	84.2

System	chrF
GTCOM_Peter	69.3
ONLINE-W	69.2
ONLINE-B	69.0
ZengHuiMT	68.5
ONLINE-A	68.3
ONLINE-Y	68.2
GPT4-5shot	68.1
ONLINE-G	68.0
Lan-BridgeMT	66.2
NLLB_Greedy	62.4
NLLB_MBR_BLEU	62.4

System	BLEU
ONLINE-W	47.4
GTCOM_Peter	46.4
ONLINE-B	46.0
ONLINE-A	45.9
ONLINE-Y	45.7
ONLINE-G	44.9
GPT4-5shot	43.9
ZengHuiMT	43.5
Lan-BridgeMT	42.3
NLLB_MBR_BLEU	38.1
NLLB_Greedy	37.8

Table 23: Scores for the uk→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-W	86.7
ONLINE-B	85.6
GPT4-5shot	85.3
ONLINE-G	85.3
ONLINE-A	83.2
ONLINE-Y	82.9
GTCOM_Peter	82.1
NLLB_Greedy	82.1
NLLB_MBR_BLEU	81.7
Lan-BridgeMT	80.4
ZengHuiMT	79.0

System	chrF
ONLINE-B	61.7
ONLINE-W	59.2
ZengHuiMT	56.4
ONLINE-G	56.1
ONLINE-A	55.8
ONLINE-Y	55.4
GTCOM_Peter	54.4
GPT4-5shot	53.0
Lan-BridgeMT	51.9
NLLB_Greedy	50.8
NLLB_MBR_BLEU	50.5

System	BLEU
ONLINE-B	39.8
ONLINE-W	34.9
ONLINE-A	30.3
ONLINE-Y	29.5
ONLINE-G	28.6
ZengHuiMT	27.8
GTCOM_Peter	27.5
GPT4-5shot	25.2
NLLB_MBR_BLEU	24.9
Lan-BridgeMT	24.6
NLLB_Greedy	24.5

Table 24: Scores for the en→uk translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
HW-TSC	82.8
ONLINE-B	82.7
Yishu	82.7
GPT4-5shot	81.6
Lan-BridgeMT	81.2
ONLINE-G	80.9
ONLINE-Y	80.6
ONLINE-A	80.3
ZengHuiMT	79.6
ONLINE-W	79.3
IOL_Research	79.2
ONLINE-M	77.7
NLLB_MBR_BLEU	76.8
ANVITA	76.6
NLLB_Greedy	76.4

System	chrF
HW-TSC	57.5
ONLINE-B	57.5
Yishu	57.4
ZengHuiMT	54.6
ONLINE-G	53.9
ONLINE-A	53.4
GPT4-5shot	53.1
Lan-BridgeMT	53.1
ONLINE-W	52.5
IOL_Research	52.4
ONLINE-Y	52.3
ONLINE-M	49.7
ANVITA	47.1
NLLB_Greedy	46.1
NLLB_MBR_BLEU	45.8

System	BLEU
HW-TSC	33.6
ONLINE-B	33.5
Yishu	33.4
ONLINE-A	28.3
Lan-BridgeMT	27.3
IOL_Research	27.2
ZengHuiMT	27.0
GPT4-5shot	26.8
ONLINE-G	26.6
ONLINE-W	26.4
ONLINE-Y	25.0
ONLINE-M	23.5
ANVITA	21.8
NLLB_Greedy	20.5
NLLB_MBR_BLEU	19.8

Table 25: Scores for the zh→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-B	88.1
Yishu	88.1
HW-TSC	87.3
GPT4-5shot	87.1
ONLINE-W	86.8
Lan-BridgeMT	86.6
ONLINE-Y	86.5
ONLINE-A	86.2
IOL_Research	85.3
ZengHuiMT	84.3
ONLINE-M	84.2
ONLINE-G	83.8
NLLB_Greedy	75.7
ANVITA	75.6
NLLB_MBR_BLEU	71.5

System	chrF
HW-TSC	53.8
Yishu	53.0
ONLINE-B	52.9
ONLINE-A	52.8
IOL_Research	51.9
ONLINE-M	50.6
ONLINE-Y	49.8
ONLINE-G	49.4
ONLINE-W	47.3
ZengHuiMT	47.0
Lan-BridgeMT	46.8
GPT4-5shot	46.5
ANVITA	36.9
NLLB_Greedy	26.3
NLLB_MBR_BLEU	21.1

System	BLEU
HW-TSC	58.6
ONLINE-A	58.5
Yishu	57.6
ONLINE-B	57.5
IOL_Research	56.9
ONLINE-M	54.9
ONLINE-Y	54.2
ONLINE-G	54.1
ZengHuiMT	52.9
ONLINE-W	52.1
Lan-BridgeMT	50.2
GPT4-5shot	49.6
ANVITA	38.9
NLLB_Greedy	27.4
NLLB_MBR_BLEU	19.1

Table 26: Scores for the en→zh translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

G Head to head comparisons

Following tables show differences in average human scores for each language pair. The numbers in each of the tables' cells indicate the difference in average human scores for the system in that column and the system in that row.

Because there were so many systems and data conditions the significance of each pairwise comparison needs to be quantified. We applied Wilcoxon rank-sum 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 $p < 0.001$, according to Wilcoxon rank-sum test.

Each table contains final rows showing the average score achieved by that system and the rank range according to Wilcoxon rank-sum test ($p < 0.05$). Gray lines separate clusters based on non-overlapping rank ranges.

Czech→Ukrainian

	ONLINE-B	GPT4-5shot	Human-refA	ONLINE-W	CUNI-GA	CUNI-Transformer	GTCOM_DLUT	ONLINE-A	ONLINE-G	ONLINE-Y	MUNI-NLP	Lan-BridgeMT	NLLB_MBR_BLEU	NLLB_Greedy
ONLINE-B	—	0.1	0.4	0.9⋆	1.3‡	1.8†	2.4⋆	3.1⋆	4.1‡	5.0‡	5.0‡	6.2‡	6.7‡	7.0‡
GPT4-5shot	-0.1	—	0.4	0.8†	1.2‡	1.8‡	2.3†	3.1‡	4.1‡	4.9‡	4.9‡	6.2‡	6.7‡	6.9‡
Human-refA	-0.4	-0.4	—	0.5‡	0.9‡	1.4‡	1.9‡	2.7‡	3.7‡	4.6‡	4.6‡	5.8‡	6.3‡	6.6‡
ONLINE-W	-0.9	-0.8	-0.5	—	0.4	0.9	1.5	2.2	3.2‡	4.1‡	4.1‡	5.3‡	5.8‡	6.1‡
CUNI-GA	-1.3	-1.2	-0.9	-0.4	—	0.6	1.1	1.8	2.9†	3.7‡	3.7†	5.0‡	5.5‡	5.7‡
CUNI-Transformer	-1.8	-1.8	-1.4	-0.9	-0.6	—	0.5	1.3	2.3†	3.1‡	3.2‡	4.4‡	4.9‡	5.1‡
GTCOM_DLUT	-2.4	-2.3	-1.9	-1.5	-1.1	-0.5	—	0.8	1.8‡	2.6‡	2.6‡	3.9‡	4.4‡	4.6‡
ONLINE-A	-3.1	-3.1	-2.7	-2.2	-1.8	-1.3	-0.8	—	1.0‡	1.9‡	1.9‡	3.1‡	3.6‡	3.9‡
ONLINE-G	-4.1	-4.1	-3.7	-3.2	-2.9	-2.3	-1.8	-1.0	—	0.8	0.9	2.1⋆	2.6†	2.8‡
ONLINE-Y	-5.0	-4.9	-4.6	-4.1	-3.7	-3.1	-2.6	-1.9	-0.8	—	0.0	1.3	1.8	2.0†
MUNI-NLP	-5.0	-4.9	-4.6	-4.1	-3.7	-3.2	-2.6	-1.9	-0.9	-0.0	—	1.2	1.7	2.0‡
Lan-BridgeMT	-6.2	-6.2	-5.8	-5.3	-5.0	-4.4	-3.9	-3.1	-2.1	-1.3	-1.2	—	0.5	0.7⋆
NLLB_MBR_BLEU	-6.7	-6.7	-6.3	-5.8	-5.5	-4.9	-4.4	-3.6	-2.6	-1.8	-1.7	-0.5	—	0.2⋆
NLLB_Greedy	-7.0	-6.9	-6.6	-6.1	-5.7	-5.1	-4.6	-3.9	-2.8	-2.0	-2.0	-0.7	-0.2	—
score	83.7	83.6	83.2	82.8	82.4	81.8	81.3	80.6	79.5	78.7	78.7	77.4	76.9	76.7
rank	1-3	1-3	1-3	4-8	4-8	4-8	4-8	4-8	9-11	9-13	9-13	10-13	10-13	14

Table 27: Head to head comparison for Czech→Ukrainian systems

German→English

	GPT4-5shot	Human-refA	ONLINE-A	ONLINE-B	ONLINE-W	ONLINE-Y	ONLINE-G	GTCOM_DLUT	ONLINE-M	LanguageX	Lan-BridgeMT	NLLB_MBR_BLEU	AIRC	NLLB_Greedy
GPT4-5shot	—	0.4	0.8	1.2†	1.5†	2.3†	2.6‡	3.8‡	5.0‡	8.5‡	10.3‡	10.7‡	11.5‡	12.4‡
Human-refA	-0.4	—	0.4	0.8⋆	1.1⋆	1.9†	2.2‡	3.4‡	4.6‡	8.1‡	9.9†	10.3‡	11.1‡	12.0‡
ONLINE-A	-0.8	-0.4	—	0.4	0.7	1.6⋆	1.9†	3.0‡	4.2‡	7.7‡	9.6‡	9.9‡	10.8‡	11.7‡
ONLINE-B	-1.2	-0.8	-0.4	—	0.3	1.1	1.4⋆	2.6‡	3.8‡	7.3‡	9.2‡	9.5‡	10.3‡	11.2‡
ONLINE-W	-1.5	-1.1	-0.7	-0.3	—	0.8	1.1⋆	2.3‡	3.5‡	7.0‡	8.9‡	9.2‡	10.0‡	10.9‡
ONLINE-Y	-2.3	-1.9	-1.6	-1.1	-0.8	—	0.3	1.5‡	2.7†	6.2‡	8.0‡	8.4‡	9.2‡	10.1‡
ONLINE-G	-2.6	-2.2	-1.9	-1.4	-1.1	-0.3	—	1.2‡	2.4	5.9‡	7.7‡	8.1‡	8.9‡	9.8‡
GTCOM_DLUT	-3.8	-3.4	-3.0	-2.6	-2.3	-1.5	-1.2	—	1.2	4.7‡	6.6‡	6.9‡	7.8‡	8.6‡
ONLINE-M	-5.0	-4.6	-4.2	-3.8	-3.5	-2.7	-2.4	-1.2	—	3.5‡	5.3‡	5.7‡	6.5‡	7.4‡
LanguageX	-8.5	-8.1	-7.7	-7.3	-7.0	-6.2	-5.9	-4.7	-3.5	—	1.9	2.2†	3.0‡	3.9†
Lan-BridgeMT	-10.3	-9.9	-9.6	-9.2	-8.9	-8.0	-7.7	-6.6	-5.3	-1.9	—	0.3	1.2⋆	2.1
NLLB_MBR_BLEU	-10.7	-10.3	-9.9	-9.5	-9.2	-8.4	-8.1	-6.9	-5.7	-2.2	-0.3	—	0.8	1.7
AIRC	-11.5	-11.1	-10.8	-10.3	-10.0	-9.2	-8.9	-7.8	-6.5	-3.0	-1.2	-0.8	—	0.9
NLLB_Greedy	-12.4	-12.0	-11.7	-11.2	-10.9	-10.1	-9.8	-8.6	-7.4	-3.9	-2.1	-1.7	-0.9	—
score	90.3	89.9	89.6	89.1	88.8	88.0	87.7	86.5	85.3	81.8	80.0	79.6	78.8	77.9
rank	1-3	1-3	1-5	3-6	3-6	4-7	6-8	8-9	7-9	10-11	10-13	11-14	12-14	11-14

Table 28: Head to head comparison for German→English systems

English→Czech

	Human-refA	ONLINE-W	GPT4-5shot	CUNI-GA	ONLINE-A	CUNI-DocTransformer	ONLINE-B	NLLB_MBR_BLEU	GTCOM_DLUT	CUNI-Transformer	NLLB_Greedy	ONLINE-M	ONLINE-G	ONLINE-Y	Lan-BridgeMT	LanguageX
Human-refA	—	1.3⋆	3.6‡	5.0‡	5.1‡	6.0‡	6.6‡	6.8‡	7.0‡	8.0‡	8.6‡	9.7‡	10.2‡	10.4‡	10.4‡	11.3‡
ONLINE-W	-1.3	—	2.3‡	3.7‡	3.8‡	4.7‡	5.3‡	5.5‡	5.7‡	6.7‡	7.3‡	8.4‡	8.9‡	9.1‡	9.1‡	10.0‡
GPT4-5shot	-3.6	-2.3	—	1.4	1.5†	2.5⋆	3.0	3.2‡	3.4†	4.4‡	5.1‡	6.1‡	6.6‡	6.8‡	6.8‡	7.7‡
CUNI-GA	-5.0	-3.7	-1.4	—	0.0‡	1.0‡	1.5⋆	1.8‡	2.0‡	3.0‡	3.6‡	4.7‡	5.1‡	5.3‡	5.4‡	6.3‡
ONLINE-A	-5.1	-3.8	-1.5	-0.0	—	1.0	1.5	1.7‡	1.9	2.9⋆	3.6‡	4.7†	5.1‡	5.3‡	5.4⋆	6.3‡
CUNI-DocTransformer	-6.0	-4.7	-2.5	-1.0	-1.0	—	0.5	0.7‡	0.9	1.9†	2.6‡	3.7‡	4.1‡	4.3‡	4.4†	5.3‡
ONLINE-B	-6.6	-5.3	-3.0	-1.5	-1.5	-0.5	—	0.2‡	0.4⋆	1.4‡	2.1‡	3.2‡	3.6‡	3.8‡	3.9‡	4.8‡
NLLB_MBR_BLEU	-6.8	-5.5	-3.2	-1.8	-1.7	-0.7	-0.2	—	0.2	1.2	1.9	3.0	3.4	3.6⋆	3.7	4.5‡
GTCOM_DLUT	-7.0	-5.7	-3.4	-2.0	-1.9	-0.9	-0.4	-0.2	—	1.0	1.7‡	2.8†	3.2‡	3.4‡	3.5	4.3‡
CUNI-Transformer	-8.0	-6.7	-4.4	-3.0	-2.9	-1.9	-1.4	-1.2	-1.0	—	0.7⋆	1.7	2.2†	2.4‡	2.4	3.3‡
NLLB_Greedy	-8.6	-7.3	-5.1	-3.6	-3.6	-2.6	-2.1	-1.9	-1.7	-0.7	—	1.1	1.5	1.7⋆	1.8	2.7‡
ONLINE-M	-9.7	-8.4	-6.1	-4.7	-4.7	-3.7	-3.2	-3.0	-2.8	-1.7	-1.1	—	0.4	0.6†	0.7	1.6‡
ONLINE-G	-10.2	-8.9	-6.6	-5.1	-5.1	-4.1	-3.6	-3.4	-3.2	-2.2	-1.5	-0.4	—	0.2	0.3	1.1†
ONLINE-Y	-10.4	-9.1	-6.8	-5.3	-5.3	-4.3	-3.8	-3.6	-3.4	-2.4	-1.7	-0.6	-0.2	—	0.1	1.0⋆
Lan-BridgeMT	-10.4	-9.1	-6.8	-5.4	-5.4	-4.4	-3.9	-3.7	-3.5	-2.4	-1.8	-0.7	-0.3	-0.1	—	0.9‡
LanguageX	-11.3	-10.0	-7.7	-6.3	-6.3	-5.3	-4.8	-4.5	-4.3	-3.3	-2.7	-1.6	-1.1	-1.0	-0.9	—
score	85.4	84.1	81.8	80.4	80.3	79.4	78.8	78.6	78.4	77.4	76.8	75.7	75.2	75.0	75.0	74.1
rank	1	2	3-5	3-4	5-8	5-8	4-7	8-14	6-11	8-12	10-14	9-14	10-15	13-15	8-15	16

Table 29: Head to head comparison for English→Czech systems

English→German

	GPT4-5shot	ONLINE-B	ONLINE-W	ONLINE-A	ONLINE-Y	Human-refA	ONLINE-M	ONLINE-G	Lan-BridgeMT	LanguageX	NLLB_MBR_BLEU	NLLB_Greedy	AIRC
GPT4-5shot	—	0.1	0.7	0.8	1.0†	1.3	2.3‡	3.4‡	5.0‡	6.3‡	12.1‡	13.2‡	15.4‡
ONLINE-B	-0.1	—	0.6	0.7	0.8⋆	1.2	2.2‡	3.3‡	4.8‡	6.2‡	12.0‡	13.1‡	15.2‡
ONLINE-W	-0.7	-0.6	—	0.2⋆	0.3‡	0.6	1.6‡	2.7‡	4.3‡	5.6‡	11.5‡	12.5‡	14.7‡
ONLINE-A	-0.8	-0.7	-0.2	—	0.1	0.5	1.4‡	2.6‡	4.1‡	5.5‡	11.3‡	12.4‡	14.5‡
ONLINE-Y	-1.0	-0.8	-0.3	-0.1	—	0.3	1.3†	2.5⋆	4.0‡	5.3‡	11.2‡	12.3‡	14.4‡
Human-refA	-1.3	-1.2	-0.6	-0.5	-0.3	—	1.0‡	2.1‡	3.7‡	5.0‡	10.8‡	11.9‡	14.1‡
ONLINE-M	-2.3	-2.2	-1.6	-1.4	-1.3	-1.0	—	1.1	2.7†	4.0‡	9.9‡	10.9‡	13.1‡
ONLINE-G	-3.4	-3.3	-2.7	-2.6	-2.5	-2.1	-1.1	—	1.5‡	2.9‡	8.7‡	9.8‡	11.9‡
Lan-BridgeMT	-5.0	-4.8	-4.3	-4.1	-4.0	-3.7	-2.7	-1.5	—	1.4⋆	7.2‡	8.3‡	10.4‡
LanguageX	-6.3	-6.2	-5.6	-5.5	-5.3	-5.0	-4.0	-2.9	-1.4	—	5.8‡	6.9‡	9.1‡
NLLB_MBR_BLEU	-12.1	-12.0	-11.5	-11.3	-11.2	-10.8	-9.9	-8.7	-7.2	-5.8	—	1.1	3.2‡
NLLB_Greedy	-13.2	-13.1	-12.5	-12.4	-12.3	-11.9	-10.9	-9.8	-8.3	-6.9	-1.1	—	2.2‡
AIRC	-15.4	-15.2	-14.7	-14.5	-14.4	-14.1	-13.1	-11.9	-10.4	-9.1	-3.2	-2.2	—
score	89.0	88.8	88.3	88.1	88.0	87.7	86.7	85.5	84.0	82.7	76.8	75.7	73.6
rank	1-5	1-5	1-4	2-6	4-6	1-6	7-8	7-8	9	10	11-12	11-12	13

Table 30: Head to head comparison for English→German systems

English→Japanese

	Human-refA	GPT4-5shot	ONLINE-B	ONLINE-Y	SKIM	ONLINE-W	LanguageX	ONLINE-A	NAIST-NICT	Lan-BridgeMT	ANVITA	ONLINE-M	KYB	AIRC	ONLINE-G	NLLB_Greedy	NLLB_MBR_BLEU
Human-refA	—	1.2‡	1.9	2.1†	2.2⋆	2.3‡	4.1‡	4.5‡	4.6‡	5.5‡	7.6‡	8.1‡	9.9‡	11.1‡	11.1‡	16.2‡	19.4‡
GPT4-5shot	-1.2	—	0.7	0.9	1.0	1.1	2.9⋆	3.3†	3.4†	4.3‡	6.4‡	6.9‡	8.8‡	9.9‡	10.0‡	15.0‡	18.3‡
ONLINE-B	-1.9	-0.7	—	0.2	0.3	0.4⋆	2.3‡	2.7‡	2.7‡	3.6‡	5.7‡	6.2‡	8.1‡	9.3‡	9.3‡	14.3‡	17.6‡
ONLINE-Y	-2.1	-0.9	-0.2	—	0.1	0.2	2.0‡	2.4‡	2.5‡	3.4‡	5.5‡	6.0‡	7.8‡	9.0‡	9.0‡	14.1‡	17.3‡
SKIM	-2.2	-1.0	-0.3	-0.1	—	0.1⋆	1.9‡	2.3‡	2.4‡	3.3‡	5.4‡	5.9‡	7.7‡	8.9‡	8.9‡	13.9‡	17.2‡
ONLINE-W	-2.3	-1.1	-0.4	-0.2	-0.1	—	1.8†	2.2†	2.3‡	3.2‡	5.3‡	5.8‡	7.6‡	8.8‡	8.8‡	13.8‡	17.1‡
LanguageX	-4.1	-2.9	-2.3	-2.0	-1.9	-1.8	—	0.4	0.5	1.4	3.5‡	4.0‡	5.8‡	7.0‡	7.0‡	12.0‡	15.3‡
ONLINE-A	-4.5	-3.3	-2.7	-2.4	-2.3	-2.2	-0.4	—	0.0	1.0	3.1‡	3.5‡	5.4‡	6.6‡	6.6‡	11.6‡	14.9‡
NAIST-NICT	-4.6	-3.4	-2.7	-2.5	-2.4	-2.3	-0.5	-0.0	—	0.9	3.0‡	3.5‡	5.4‡	6.5‡	6.6‡	11.6‡	14.9‡
Lan-BridgeMT	-5.5	-4.3	-3.6	-3.4	-3.3	-3.2	-1.4	-1.0	-0.9	—	2.1†	2.6‡	4.5‡	5.6‡	5.6‡	10.7‡	14.0‡
ANVITA	-7.6	-6.4	-5.7	-5.5	-5.4	-5.3	-3.5	-3.1	-3.0	-2.1	—	0.5	2.3‡	3.5‡	3.5‡	8.5‡	11.8‡
ONLINE-M	-8.1	-6.9	-6.2	-6.0	-5.9	-5.8	-4.0	-3.5	-3.5	-2.6	-0.5	—	1.9†	3.0‡	3.1†	8.1‡	11.4‡
KYB	-9.9	-8.8	-8.1	-7.8	-7.7	-7.6	-5.8	-5.4	-5.4	-4.5	-2.3	-1.9	—	1.2	1.2	6.2‡	9.5‡
AIRC	-11.1	-9.9	-9.3	-9.0	-8.9	-8.8	-7.0	-6.6	-6.5	-5.6	-3.5	-3.0	-1.2	—	0.0	5.0‡	8.3‡
ONLINE-G	-11.1	-10.0	-9.3	-9.0	-8.9	-8.8	-7.0	-6.6	-6.6	-5.6	-3.5	-3.1	-1.2	-0.0	—	5.0‡	8.3‡
NLLB_Greedy	-16.2	-15.0	-14.3	-14.1	-13.9	-13.8	-12.0	-11.6	-11.6	-10.7	-8.5	-8.1	-6.2	-5.0	-5.0	—	3.3‡
NLLB_MBR_BLEU	-19.4	-18.3	-17.6	-17.3	-17.2	-17.1	-15.3	-14.9	-14.9	-14.0	-11.8	-11.4	-9.5	-8.3	-8.3	-3.3	—
score	80.7	79.5	78.8	78.6	78.5	78.4	76.6	76.2	76.1	75.2	73.1	72.6	70.8	69.6	69.6	64.5	61.3
rank	1-2	2-6	1-5	2-6	2-5	4-6	7-10	7-10	7-10	7-10	11-12	11-12	13-15	13-15	13-15	16	17

Table 31: Head to head comparison for English→Japanese systems

English→Chinese

	Yishu	Human-refA	GPT4-5shot	Lan-BridgeMT	ONLINE-B	HW-TSC	ONLINE-W	ONLINE-Y	IOL_Research	ONLINE-A	LanguageX	ONLINE-M	ONLINE-G	ANVITA	NLLB_Greedy	NLLB_MBR_BLEU
Yishu	—	0.0	0.1	0.2⋆	0.3	0.7	0.8⋆	2.0⋆	2.3‡	2.5‡	3.6‡	4.0‡	5.0‡	17.7‡	17.9‡	25.0‡
Human-refA	-0.0	—	0.0	0.1†	0.3	0.7	0.8⋆	1.9†	2.3‡	2.5‡	3.6‡	3.9‡	5.0‡	17.7‡	17.8‡	25.0‡
GPT4-5shot	-0.1	-0.0	—	0.1	0.3	0.6	0.7	1.9⋆	2.3‡	2.4‡	3.5‡	3.9‡	5.0‡	17.6‡	17.8‡	24.9‡
Lan-BridgeMT	-0.2	-0.1	-0.1	—	0.2	0.5	0.6	1.8	2.2†	2.3‡	3.4‡	3.8‡	4.9‡	17.5‡	17.7‡	24.8‡
ONLINE-B	-0.3	-0.3	-0.3	-0.2	—	0.3	0.4†	1.6†	2.0‡	2.2‡	3.2‡	3.6‡	4.7‡	17.3‡	17.5‡	24.7‡
HW-TSC	-0.7	-0.7	-0.6	-0.5	-0.3	—	0.1	1.3	1.7‡	1.8‡	2.9‡	3.3‡	4.4‡	17.0‡	17.2‡	24.3‡
ONLINE-W	-0.8	-0.8	-0.7	-0.6	-0.4	-0.1	—	1.2	1.6†	1.7‡	2.8‡	3.2‡	4.3‡	16.9‡	17.1‡	24.2‡
ONLINE-Y	-2.0	-1.9	-1.9	-1.8	-1.6	-1.3	-1.2	—	0.4⋆	0.5†	1.6‡	2.0‡	3.1‡	15.7‡	15.9‡	23.0‡
IOL_Research	-2.3	-2.3	-2.3	-2.2	-2.0	-1.7	-1.6	-0.4	—	0.2	1.2†	1.6†	2.7‡	15.3‡	15.5‡	22.7‡
ONLINE-A	-2.5	-2.5	-2.4	-2.3	-2.2	-1.8	-1.7	-0.5	-0.2	—	1.1⋆	1.5⋆	2.5‡	15.2‡	15.4‡	22.5‡
LanguageX	-3.6	-3.6	-3.5	-3.4	-3.2	-2.9	-2.8	-1.6	-1.2	-1.1	—	0.4	1.5	14.1‡	14.3‡	21.4‡
ONLINE-M	-4.0	-3.9	-3.9	-3.8	-3.6	-3.3	-3.2	-2.0	-1.6	-1.5	-0.4	—	1.1	13.7‡	13.9‡	21.0‡
ONLINE-G	-5.0	-5.0	-5.0	-4.9	-4.7	-4.4	-4.3	-3.1	-2.7	-2.5	-1.5	-1.1	—	12.6‡	12.8‡	20.0‡
ANVITA	-17.7	-17.7	-17.6	-17.5	-17.3	-17.0	-16.9	-15.7	-15.3	-15.2	-14.1	-13.7	-12.6	—	0.2‡	7.3‡
NLLB_Greedy	-17.9	-17.8	-17.8	-17.7	-17.5	-17.2	-17.1	-15.9	-15.5	-15.4	-14.3	-13.9	-12.8	-0.2	—	7.1‡
NLLB_MBR_BLEU	-25.0	-25.0	-24.9	-24.8	-24.7	-24.3	-24.2	-23.0	-22.7	-22.5	-21.4	-21.0	-20.0	-7.3	-7.1	—
score	82.2	82.1	82.1	82.0	81.8	81.5	81.4	80.2	79.8	79.7	78.6	78.2	77.1	64.5	64.3	57.2
rank	1-5	1-5	1-7	3-8	1-6	1-8	4-8	5-8	9-10	9-10	11-13	11-13	11-13	14	15	16

Table 32: Head to head comparison for English→Chinese systems

Japanese→English

	GPT4-5shot	SKIM	Human-refA	ONLINE-Y	ONLINE-B	ONLINE-A	ONLINE-W	NAIST-NICT	GTCOM_DLUT	Lan-BridgeMT	ANVITA	ONLINE-G	LanguageX	ONLINE-M	KYB	AIRC	NLLB_MBR_BLEU	NLLB_Greedy
GPT4-5shot	—	0.7⋆	0.9‡	1.8‡	1.9‡	2.1‡	2.5†	2.9‡	4.4‡	4.8‡	5.5‡	6.5‡	6.7‡	8.4‡	8.9‡	12.4‡	14.6‡	15.2‡
SKIM	-0.7	—	0.2†	1.0⋆	1.2	1.3†	1.7	2.2⋆	3.6‡	4.1‡	4.7‡	5.8‡	5.9‡	7.7‡	8.1‡	11.6‡	13.8‡	14.5‡
Human-refA	-0.9	-0.2	—	0.9	1.0	1.1	1.5	2.0	3.5⋆	3.9‡	4.5‡	5.6‡	5.7‡	7.5‡	7.9‡	11.4‡	13.7‡	14.3‡
ONLINE-Y	-1.8	-1.0	-0.9	—	0.1	0.3	0.7	1.1	2.6†	3.1‡	3.7‡	4.7‡	4.9‡	6.6‡	7.1‡	10.6‡	12.8‡	13.4‡
ONLINE-B	-1.9	-1.2	-1.0	-0.1	—	0.2	0.6	1.0	2.5†	2.9‡	3.6‡	4.6‡	4.8‡	6.5‡	7.0‡	10.5‡	12.7‡	13.3‡
ONLINE-A	-2.1	-1.3	-1.1	-0.3	-0.2	—	0.4	0.8	2.3	2.8‡	3.4‡	4.4‡	4.6‡	6.3‡	6.8‡	10.3‡	12.5‡	13.2‡
ONLINE-W	-2.5	-1.7	-1.5	-0.7	-0.6	-0.4	—	0.4	1.9†	2.4‡	3.0‡	4.0‡	4.2‡	6.0‡	6.4‡	9.9‡	12.1‡	12.8‡
NAIST-NICT	-2.9	-2.2	-2.0	-1.1	-1.0	-0.8	-0.4	—	1.5†	2.0‡	2.6‡	3.6‡	3.8‡	5.5‡	6.0‡	9.5‡	11.7‡	12.3‡
GTCOM_DLUT	-4.4	-3.6	-3.5	-2.6	-2.5	-2.3	-1.9	-1.5	—	0.5‡	1.1†	2.1‡	2.3†	4.0‡	4.5‡	8.0‡	10.2‡	10.9‡
Lan-BridgeMT	-4.8	-4.1	-3.9	-3.1	-2.9	-2.8	-2.4	-2.0	-0.5	—	0.6	1.7	1.8	3.6‡	4.0‡	7.5‡	9.7‡	10.4‡
ANVITA	-5.5	-4.7	-4.5	-3.7	-3.6	-3.4	-3.0	-2.6	-1.1	-0.6	—	1.1	1.2	3.0‡	3.4‡	6.9‡	9.1‡	9.8‡
ONLINE-G	-6.5	-5.8	-5.6	-4.7	-4.6	-4.4	-4.0	-3.6	-2.1	-1.7	-1.1	—	0.2	1.9‡	2.4‡	5.9‡	8.1‡	8.7‡
LanguageX	-6.7	-5.9	-5.7	-4.9	-4.8	-4.6	-4.2	-3.8	-2.3	-1.8	-1.2	-0.2	—	1.8‡	2.2‡	5.7‡	7.9‡	8.6‡
ONLINE-M	-8.4	-7.7	-7.5	-6.6	-6.5	-6.3	-6.0	-5.5	-4.0	-3.6	-3.0	-1.9	-1.8	—	0.5	4.0‡	6.2‡	6.8‡
KYB	-8.9	-8.1	-7.9	-7.1	-7.0	-6.8	-6.4	-6.0	-4.5	-4.0	-3.4	-2.4	-2.2	-0.5	—	3.5‡	5.7‡	6.4‡
AIRC	-12.4	-11.6	-11.4	-10.6	-10.5	-10.3	-9.9	-9.5	-8.0	-7.5	-6.9	-5.9	-5.7	-4.0	-3.5	—	2.2†	2.9†
NLLB_MBR_BLEU	-14.6	-13.8	-13.7	-12.8	-12.7	-12.5	-12.1	-11.7	-10.2	-9.7	-9.1	-8.1	-7.9	-6.2	-5.7	-2.2	—	0.6
NLLB_Greedy	-15.2	-14.5	-14.3	-13.4	-13.3	-13.2	-12.8	-12.3	-10.9	-10.4	-9.8	-8.7	-8.6	-6.8	-6.4	-2.9	-0.6	—
score	81.3	80.6	80.4	79.5	79.4	79.2	78.8	78.4	76.9	76.4	75.8	74.8	74.6	72.9	72.4	68.9	66.7	66.1
rank	1	2-4	3-8	3-8	2-8	3-9	2-8	3-8	8-9	10-13	10-13	10-13	10-13	14-15	14-15	16	17-18	17-18

Table 33: Head to head comparison for Japanese→English systems

Chinese→English

	Lan-BridgeMT	GPT4-5shot	Yishu	ONLINE-W	ONLINE-G	ONLINE-B	ONLINE-Y	HW-TSC	ONLINE-A	IOL_Research	LanguageX	ONLINE-M	NLLB_MBR_BLEU	Human-refA	NLLB_Greedy	ANVITA
Lan-BridgeMT	—	1.9	2.6‡	2.7‡	2.9‡	3.1‡	3.2‡	3.8‡	5.1‡	5.2‡	5.6‡	6.0‡	6.7‡	6.8‡	8.9‡	10.3‡
GPT4-5shot	-1.9	—	0.6‡	0.8‡	1.0‡	1.1†	1.2‡	1.9†	3.1‡	3.3‡	3.7‡	4.1‡	4.7‡	4.9‡	6.9‡	8.3‡
Yishu	-2.6	-0.6	—	0.2	0.3⋆	0.5	0.6	1.3	2.5	2.6‡	3.1‡	3.5‡	4.1‡	4.3‡	6.3‡	7.7‡
ONLINE-W	-2.7	-0.8	-0.2	—	0.2⋆	0.4	0.5	1.1	2.3⋆	2.5‡	2.9‡	3.3‡	4.0‡	4.1‡	6.2‡	7.6‡
ONLINE-G	-2.9	-1.0	-0.3	-0.2	—	0.2	0.3	0.9	2.2	2.3⋆	2.8	3.1‡	3.8‡	3.9‡	6.0‡	7.4‡
ONLINE-B	-3.1	-1.1	-0.5	-0.4	-0.2	—	0.1†	0.8	2.0‡	2.1‡	2.6‡	3.0‡	3.6‡	3.8‡	5.8‡	7.2‡
ONLINE-Y	-3.2	-1.2	-0.6	-0.5	-0.3	-0.1	—	0.7	1.9	2.0†	2.5⋆	2.9‡	3.5‡	3.7‡	5.7‡	7.1‡
HW-TSC	-3.8	-1.9	-1.3	-1.1	-0.9	-0.8	-0.7	—	1.2‡	1.4‡	1.8‡	2.2‡	2.8‡	3.0‡	5.0‡	6.5‡
ONLINE-A	-5.1	-3.1	-2.5	-2.3	-2.2	-2.0	-1.9	-1.2	—	0.1⋆	0.6	1.0‡	1.6‡	1.8‡	3.8‡	5.2‡
IOL_Research	-5.2	-3.3	-2.6	-2.5	-2.3	-2.1	-2.0	-1.4	-0.1	—	0.4	0.8†	1.5‡	1.6‡	3.7‡	5.1‡
LanguageX	-5.6	-3.7	-3.1	-2.9	-2.8	-2.6	-2.5	-1.8	-0.6	-0.4	—	0.4†	1.0‡	1.2‡	3.2‡	4.6‡
ONLINE-M	-6.0	-4.1	-3.5	-3.3	-3.1	-3.0	-2.9	-2.2	-1.0	-0.8	-0.4	—	0.6†	0.8	2.8‡	4.3†
NLLB_MBR_BLEU	-6.7	-4.7	-4.1	-4.0	-3.8	-3.6	-3.5	-2.8	-1.6	-1.5	-1.0	-0.6	—	0.2	2.2	3.6
Human-refA	-6.8	-4.9	-4.3	-4.1	-3.9	-3.8	-3.7	-3.0	-1.8	-1.6	-1.2	-0.8	-0.2	—	2.0⋆	3.4
NLLB_Greedy	-8.9	-6.9	-6.3	-6.2	-6.0	-5.8	-5.7	-5.0	-3.8	-3.7	-3.2	-2.8	-2.2	-2.0	—	1.4
ANVITA	-10.3	-8.3	-7.7	-7.6	-7.4	-7.2	-7.1	-6.5	-5.2	-5.1	-4.6	-4.3	-3.6	-3.4	-1.4	—
score	82.9	80.9	80.3	80.2	80.0	79.8	79.7	79.1	77.8	77.7	77.2	76.9	76.2	76.1	74.0	72.6
rank	1-2	1-2	3-8	3-7	5-10	3-7	4-9	3-8	6-10	10-11	8-11	12-13	13-16	12-15	14-16	13-16

Table 34: Head to head comparison for Chinese→English systems