Findings of the WMT26 General Machine Translation Shared Task: Contrastive Dynamic Human Evaluation at Scale

Tom Kocmi Cohere
Ekaterina Artemova Toloka AI
Eleftherios Avramidis alangu, DFKI
Rachel Bawden Inria, Paris, France
Ondřej Bojar Charles University
Sergey Dukanov Dubformer
Anton Dvorkovich Dubformer
Mark Fishel University of Tartu
Markus Freitag Google
Samuel Frontull University of Innsbruck
Thamme Gowda Microsoft
Roman Grundkiewicz Microsoft
Barry Haddow University of Edinburgh
Stan Kharevich Alconost
Philipp Koehn Johns Hopkins University
Zheng Li Tencent
Jean Maillard Meta
Christof Monz University of Amsterdam
Alexander Murauski Alconost
Kenton Murray GMU
Masaaki Nagata NTT
Stefano Perrella Cohere
Martin Popel Charles University
Maja Popović DCU & IU
Lorenzo Proietti Sapienza University of Rome
Sara Rajaee University of Amsterdam
Parker Riley Google
Mariya Shmatova Toloka AI
Steinþór Steingrímsson The Árni Magnússon Institute
Lisa Yankovskaya University of Tartu
Vilém Zouhar ETH Zurich

Abstract

This paper presents the results of the General Machine Translation Task organized under the 2026 Conference on Machine Translation (WMT). Participants build systems for any of the 23 language pairs spanning four to five domains. This year we brought major changes to the human evaluation: (1) new annotation platform Pearmut for better reproducibility, (2) new human protocol "contrastive Error Span Annotation" for higher quality and side-by-side evaluation, and (3) Dynamic human evaluation that allows assessing all submitted models while evaluating higher-performing models more frequently. Beyond human evaluation, we (4) extended difficulty sampling with a human-driven stage, (5) added two new domains, (6) made the test sets fully document-level without requiring segment-level alignment and relying on HTML or JSON structure, (7) prepared some human references by post-editing open-weight model outputs, and (8) introduced contextual instructions governing formality and structural style. We evaluated 46 systems in total: 30 submitted by participants and 16 consisting of translations from large language models (LLMs) and industry translation providers.

Sections

1 Introduction
2 Test data
3 Training data
4 System Submissions
5 Human Evaluation Methodology
6 Human Evaluation Results
7 Test Suites Sub-task
8 Conclusions
A Annotator and Translator Details
B System Submission Summaries

Data & Code

github.com/wmt-conference/wmt26-general-mt

Website

www2.statmt.org/wmt26

Proceedings of the Eleventh Conference on Machine Translation, pages 873–923
October 28-29, 2026 ©2026 Association for Computational Linguistics

1 Introduction

The Eleventh Conference on Machine Translation (WMT26) was held in conjunction with the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP 2026). This 21st iteration of the conference builds on previous editions (2006; 2007; 2008; 2009; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019; 2020; 2021; 2022; 2023; 2024; 2025) and hosted thirteen shared tasks on various aspects of machine translation (MT). This paper describes two of these tasks, the General Machine Translation shared task and MT Test suites.

The goal of the General Machine Translation shared task is to explore the translation capabilities of current systems across different domains and a broad range of language pairs. We focus on the following domains: news, social/user-generated content, speech, software, educational, and fact-checking texts. These domains were chosen to cover diverse styles while remaining broadly accessible; thus allowing for evaluation by human annotators without in-domain expertise. However, due to limited access to monolingual data, the set of evaluated domains is not identical across all language pairs.

Our evaluation includes the following 23 language pairs:

Czech→Ukrainian,
Czech→German,
Czech→Vietnamese ***(new)***,
Simplified Chinese→Japanese,
English→Egyptian Arabic,
English→Armenian ***(new)***,
English→Belarusian ***(new)***,
English→Simplified Chinese,
English→Traditional Chinese,
English→Czech,
English→Estonian,
English→German,
English→Icelandic,
English→Indonesian ***(new)***,
English→Japanese,
English→Kazakh ***(new)***,
English→Korean,
English→Ladin ***(new)***,
English→Ligurian ***(new)***,
English→Northern Sami ***(new)***,
English→Russian,
English→Thai ***(new)***,
English→Ukrainian.

Furthermore, this year focused on improving the shared task definition, data preparation and evaluation process in the following aspects:

- Contrastive Error Span Annotation protocol replaces previously used ESA and MQM by showing multiple translations to the human annotators at the same time next to each other.
- Dynamic evaluation where better systems are human-evaluated on more items allows all submitted models to be human-evaluated.
- Pearmut replaces Appraise as the platform for human evaluation due to higher reproducibility, more features, and ease of use.
- Extend source-only difficulty sampling with an human-driven Artificial Crowd stage, selecting sources whose translations receive the lowest average human evaluation scores.
- Add two new domains: Software, with English UI strings in JSON format requiring key and placeholder preservation; and Fact-checking, with Czech journalistic texts debunking false claims.
- Test sets are fully document-level and we no longer require participants to keep segment-level alignment.
- We prepared human references in two setups: fully translated by humans from scratch as in the past, as well as via post-editing of open weight models.
- We introduce instruction following context which instruct systems to follow particular formality, structure style.
- We no longer publish automatic evaluation scores as we cannot prevent participating systems in hill-climbing any of the metrics (Kocmi et al., 2025; Lavie et al., 2025).

Finally, test suites remain the part of the General MT task and for the third year they focus on translations produced by LLMs. These specialised test sets focus on phenomena which cannot be detected by the generic evaluation set-up.

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

★ **The number of participants:** Participation slightly decreased this year, with a total of 30 submissions. Again, the majority of participants used LLMs in their systems, and one system fully relies on the rule-based approach. (Section 4).

- ★ **Human translations:** Human references reached the top cluster for 8 of the 12 provided language pairs (Section 6.4). Furthermore, while post-edited human translations consistently outperform their underlying base models, it remains inconclusive whether post-editing or translating from scratch is the superior strategy.
- ★ **Social domain was the most challenging:** The social domain texts were most challenging to translate, while software-related texts were the easiest (Section 6.4). The speech domain, being the most difficult last year and second most difficult this year, also remains as a challenge. The news domain seems to be more difficult than last year, probably due to the improved difficulty sampling.
- ★ **Lumen led the constrained track,** ranking in the top cluster for 20 of 23 language pairs in constrained track, followed by Qingqiu-MT-9B (Section 6.3).
- ★ **Language-specific models have potential:** Grammatist, LiLa, and CUNI-AR reached the top-performing participant cluster in every translation direction they entered, although other targeted submissions did not show the same pattern (Section 6.4).
- ★ **The best system** overall was Wayfinder, scoring in the top cluster for 17 out of 23 language pairs. It was followed by Gemini 3.1 Pro (15 language pairs), TRIVE (14 language pairs), and Lumen (13 language pairs). Notably, while the top three systems are all built utilizing Gemini, Lumen stands out as a fine-tuned model, making its competitive performance as a constrained system particularly encouraging.
- ★ **Significant cost variations:** The financial cost of translation varies widely and scales closely with model performance. Expenses span from a high of $252 for Gemini, through $83 for Google Translate, down to $4 for Tiny Aya.
- ★ **Low-resource languages** remain challenging. Although the top-ranked systems for some of these languages score above 80, most systems do not exceed 60. For some languages, challenges beyond data scarcity are also evident, including inconsistent use of language varieties and archaic forms (Section 6.4).
- ★ **Targeted test suites** reveal that current MT systems show consistent failures on figurative language, genre shifts, and terminological nuance, alongside pervasive stereotype-driven gender bias and a marked tendency toward hallucination and semantic detachment from source content – weaknesses that aggregated human evaluation scores largely fail to expose (Section 7).

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

2 Test Data

In this section, we describe the test data collection process (Section 2.1) and the creation of human reference translations (Section 2.2). This year, we retain the difficulty-based sub-sampling procedure for source texts introduced last year and further extend it with an additional phase (Section 2.1.1). Motivated by the ever-increasing capabilities of modern MT systems, difficulty-based sub-sampling is designed to make our test sets more challenging by selecting source documents that are estimated to be more difficult to translate. This year, we implemented human-driven difficulty sampling in a form of artificial crowd. Furthermore, we include a toxicity-filtering step that prevents source documents containing toxic content from being included in the test set (Section 2.1.2).

2.1 Collecting test data

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

Domain and language coverage. We collected data from six domains (news, social, speech, software, education, and fact-checking) and three source languages (Chinese, Czech, and English). However, not all domains are represented in all three source languages. Detailed statistics on the

Language pair	News	Social	Speech	Software	Education	Fact-checking
			#words			
English→*	6,363	11,425	8,656	2,485	-	-
Chinese→Japanese	13,177	12,559	7,294	-	-	-
Czech→*	9,375	7,785	6,984	-	9,106	10,114
			#segments (#words/segment)			
English→*	192 (33.14)	186 (61.42)	133 (65.08)	180 (13.81)	-	-
Chinese→Japanese	208 (63.35)	219 (57.35)	114 (63.98)	-	-	-
Czech→*	191 (49.08)	200 (38.92)	114 (61.26)	-	183 (49.76)	142 (71.23)
			#documents (#segments/document)			
English→*	14 (13.71)	44 (4.23)	133 (1.00)	7 (25.71)	-	-
Chinese→Japanese	28 (7.43)	48 (4.56)	114 (1.00)	-	-	-
Czech→*	35 (5.46)	73 (2.74)	114 (1.00)	-	63 (2.90)	30 (4.73)
			#sentences (#sentences/segment)			
English→*	229 (1.19)	644 (3.46)	1,296 (9.74)	166 (0.92)	-	-
Chinese→Japanese	506 (2.43)	392 (1.79)	874 (7.67)	-	-	-
Czech→*	492 (2.58)	431 (2.15)	452 (3.96)	-	220 (1.20)	426 (3.00)
			Type-token ratio of source texts			
English→*	0.40	0.36	0.30	0.48	-	-
Chinese→Japanese	0.29	0.32	0.26	-	-	-
Czech→*	0.55	0.56	0.44	-	0.53	0.41

Table 1: Per-domain statistics of our test sets, calculated on the source side. To compute word counts, we used simple whitespace tokenization for Czech and English and the spacy-pkuseg word segmentation tool for Chinese (Luo et al., 2019). Sentence segmentation was carried out using the language-specific spaCy models for English and Chinese, and the multilingual spaCy model for Czech (Honnibal et al., 2020), given that a language-specific one is not available.

test sets, including the language coverage of each domain, are provided in Table 1.

2.1.1 Difficulty sampling

To identify documents that are particularly challenging for modern MT systems, we apply a two-step difficulty-based sub-sampling procedure based on the approach of Proietti et al. (2025). In Step 1, we estimate the difficulty of the collected source documents using their best-performing source-only estimator, sentinel-src-25. This model is an improved version of `sentinel-src`, a regression model based on XLM-RoBERTa Large (Conneau et al., 2020) and trained to predict translation quality using only the source text (Perrella et al., 2024). For each document, we estimate the difficulty of each segment and average the resulting scores to obtain a document-level difficulty score. We then retain the most difficult documents for inclusion in the test sets.

In Step 2, we estimate translation difficulty using the approach that Proietti et al. (2025) refer to as Artificial Crowd: (1) we translate the source texts using a crowd of translation systems, which serve as proxies for modern MT systems; (2) we ask human annotators to evaluate the generated translations; and (3) we subsample the most difficult documents, defined as those receiving the lowest human evaluation scores. Specifically, we translate the English source texts into Italian using a crowd of translation systems: Tower-Plus-9B (Rei et al., 2026), aya-expanse-8B (Dang et al., 2024), and NLLB-moe-54B (NLLB Team et al., 2022). We selected older or smaller models rather than more recent MT systems, aiming to minimize overlap with the participating systems and thereby confine this bias to models that are unlikely to be included in the final analyses, because artificial crowd-based difficulty sampling may bias the test set against the MT systems used as the crowd to generate translations, as it selects documents that are difficult for these systems to translate.

We then conduct a human evaluation campaign in which all three translations of the same input are evaluated side by side on a four-point Likert scale. We use the Pearmut annotation tool (Zouhar and Kocmi, 2026), and report the annotation guidelines in Table 2. Finally, we retain the most difficult documents, defined as those with the lowest aver-

Evaluate each Italian translation against the English source using the Acceptability slider.
4 clearly acceptable: faithful and natural, no meaningful edits needed.
3 mostly acceptable: meaning preserved, only minor issues.
2 problematic: noticeable errors or awkwardness, editing needed.
1 unacceptable: serious errors, omissions, or mistranslations.

Table 2: Guidelines provided to human annotators during Step 2 of the difficulty-based subsampling procedure (Section 2.1.1).

age translation scores. A document-level score is computed by averaging the scores of its segments,[1] while each segment-level score is computed as the average human score assigned to the translations produced by aya-expanse-8B and Tower-Plus-9B.[2]

Importantly, source-based difficulty subsampling (Step 1) is applied only to the news, social, and fact-checking domains, across all three source languages. The remaining domains either do not contain enough samples to support further subsampling or, in the case of the software domain, are included only in Step 2. Artificial Crowd-based difficulty subsampling (Step 2) is applied only to English source texts from the news, social, and software domains. Because Step 2 is more expensive than Step 1, as it involves human annotation, we restrict it to English source documents as English is the source language represented in the largest number of translation directions in our test set. The news and social domains are the only domains to undergo both stages of difficulty subsampling, with Step 1 followed by Step 2.

2.1.2 Toxicity filtering

To identify documents containing toxic content, we used the MuTox toxicity detector (Costa-jussà et al., 2024). We assigned a toxicity score to each segment using MuTox; we then discarded a document if any of its segments received a toxicity score above a predefined threshold, which we heuristically set to 0.95.

2.1.3 Instruction following context

With the increasing popularity of instruction-following LLMs, translation systems can be asked not only to preserve the meaning of the source text, but also to follow natural-language requirements concerning style, register, or output structure. We therefore extended the test sets with explicit translation instructions to use as additional context.

For each test item, we added a field to the additional context containing a translation instruction. Besides specifying the target language and requesting only the translation, the instruction described how the text should be translated for the given domain. For example, it requested a formal journalistic style for news and fact-checking, colloquial language for speech, informal and expressive language for social media, preservation of JSON keys and placeholders for software, and terminology suitable for educational material. Reformulated variants of these instructions were provided to the human annotators and are listed in Section A.1.

2.1.4 Test sets statistics

Source text segmentation. To balance the evaluation across domains and source languages, we aimed for a consistent size for each source-language–domain combination, namely approximately 150 to 200 segments per domain (see Table 1). A segment contains one or more paragraphs, where each paragraph is defined by a line break (a `<br>` HTML tag) in the source document. In multi-segment documents, we enclose segments in `<p>`, `</p>` HTML tags, separated by a single newline `\n`. Whenever possible, we also try to keep the average segment length at approximately 60 words. Maintaining approximately the same number of segments per domain enables micro-averaging of segment-level scores without any single category disproportionately influencing the final scores.

The composition of the test sets remains domain-specific because of the different characteristics of the source texts. For example, the software domain, which consists of JSON data, contains only 2,485 words, while the speech domain contains fewer segments (approximately 120 per source language) than the other domains. Furthermore, the speech domain consists of single-segment documents, whereas the other domains contain multi-segment documents.

In the following paragraphs, we provide additional details about the data collection process for each domain.

News domain. (all translation directions) For this domain, the data was prepared similarly to previous years (Kocmi et al., 2025). We collected

[1] The only exception is the software domain, for which we assembled the final documents ourselves by combining the JSON key-value pairs with the lowest average translation scores. Thus, the software domain was subsampled at the key-value level rather than at the document level.

[2] NLLB-moe-54B achieved a particularly low average score of 1.98 on the news domain, compared with 2.76 for aya-expanse-8B and 3.17 for Tower-Plus-9B. We considered NLLB-moe-54B insufficiently representative and excluded it.

news articles published in early 2026 on online news sites and extracted the text while preserving paragraph boundaries. We specifically aimed to extract articles that were more challenging to translate. For English, we limited the news crawl to opinion pieces on the basis that they tend to use a more complex, literary style than the straightforward event reporting. For Czech and Chinese, we extracted a larger pool of news to follow up with down-sampling.

Social domain. (all translation directions) Mastodon is a federated social network that is compatible with the W3C standard ActivityPub (Webber et al., 2018). Users publish short-form content known as "toots", with the possibility of replying to other toots to form threads. This year, we focused on different servers with the aim of increasing a broader coverage of domains and languages. Though a particular server may claim to focus on specific languages or regions, in practice many languages and dialects are frequently present, reflecting the true nature of how humans use language. Regardless, we focused on mastodon.sg in an attempt to get Chinese data and witter.cz to get native Czech data. These servers do contain native Chinese and Czech data, but also contain other languages including lots of English. We also collected data from funami.tech, masto.nu, mastodon.com.pl, sakurajima.moe, and mastodon.ee which were nominally in Japanese, Dutch, Spanish, Polish, and Estonian, but in practice contained lots of English data. All data was scraped intermittently over a 6 week period in March and April of 2026.

As in the previous edition of the task, we aimed to collect threads made up of multiple toots in order to create documents. Therefore, in addition to scraping random toots from the previous hour, we also did targeted scraping of missing toots within partially sourced threads (identified using the 'in_reply_to_id' attribute of already sourced toots). To avoid spam and uninformative toots, we removed empty toots, texts that appeared several times (probable spam) and toots from accounts we heuristically identified as being news outlets or bots (containing the keywords 'bot', 'news', 'weather', 'sports', 'feeds' or 'press' in their handle). We created threads from the individual toots with a minimum and maximum length of 2 and 15 respectively, ignoring toots that had already been included thread (in cases where toots appear in several forks). Individual toots are separated with double newlines, and newlines within toots are preserved. As with the previous edition, we aimed to provide screenshots capturing the thread and including any accompanying images. The screenshots were used by human evaluators to aid translation. Unfortunately, automatically capturing screenshots was more difficult than in the previous edition, and several issues prevented us from obtaining screenshots for all scraped toots. These included toots published on other servers but visible on Mastodon, which had different and unexpected formatting, and toots that were no longer available on the platform. These issues caused most of the difficulties in obtaining screenshots for Chinese and Czech toots. In following years, we will need to overcome these challenges, one possibility being to artificially construct the screenshots using a uniform format from the downloaded metadata of toots.

Speech domain. (all translation directions) The speech corpus was compiled from Creative Commons (CC BY) licensed videos retrieved through the YouTube Data API in English, Czech and Mandarin Chinese. In WMT25 the retrieval queries named specific topics, and the resulting corpus was dominated by a few narrow domains; this year we used around one hundred deliberately broad queries (e.g. *short film drama*, *stage play*) so that the difficulty would come from the naturalness of spontaneous speech rather than from terminology. Candidate channels were filtered by view and subscriber counts, which removes low-quality uploads effectively, and by the language of the title and description; the spoken language was then verified from the audio.

Roughly 15,000 videos per language entered the pipeline. Segment boundaries were placed on silence: utterances were grouped into blocks separated by pauses of at least two seconds, and adjacent blocks were merged greedily into segments of 20 to 35 seconds in which speech occupies more than 30% of the duration, with half a second of padding on each side. At most five candidate segments were taken from any one video. Automatic filtering and LLM scores over the audio and the transcript were used only to rank the resulting candidates — by how self-contained a segment is without its neighbours, how safe and how everyday its subject matter is, how much hearing or seeing the scene would help resolve translation ambiguities, and how expressive the delivery is — leaving about 1,500 candidates per language. **The final selec-**

tion was manual: every remaining candidate was watched and its transcript read, both to check copyright and admissibility and to discard clips where the ASR had broken down. Transcription and diarisation were performed with the Speechmatics batch transcription, using the per-language model at the *enhanced* operating point with speaker diarisation.

Inspecting some 500 high-scoring candidates showed that segments whose translation depends on the *video* — what is on screen, gestures, objects, actions — are considerably easier to find than segments where the *audio* alone, through prosody, emotion and delivery, carries the decisive signal. Since few participants exploited the video modality in the previous edition, the manual selection in all three source languages deliberately favoured material for which an ASR transcript alone is not sufficient; for English, where the pool was largest, we estimate that about one half of the segments require the video and about one quarter the audio. At four participating systems this year report using visual or audio-visual context (AdaptiveMT, Cohere CAT+, Dubformer, PseudoRef). The same criteria were applied throughout, but what they could select from differed: in the Czech subset the video and the audio contribute less, simply because Czech Creative Commons material on YouTube covers a much narrower range. Its value lies instead in genuinely spontaneous speech, whose naturalness MT systems tend not to preserve.

For the same reason the speech domain is exempt from the difficulty sampling of Section 2.1.1. Text-only difficulty estimation cannot see what makes these segments hard — an utterance that reads as plain text may be translated far better once the video is available — so it would select against exactly the material the domain is built around. A further filter would also have shrunk a corpus that was, in every language, only just large enough.

The resulting corpus contains 133 English documents (9,108 tokens, 59 minutes of audio), 114 Czech (6,937 tokens, 53 minutes) and 114 Chinese (6,900 tokens, 54 minutes);[3]. The median clip lasts 26 to 28 seconds. Each document is a single segment, in contrast to the multi-segment documents of the other domains.

The automatic transcripts were subsequently corrected against the video by annotators following the guidelines in Section A.7. Participating systems received the *automatic* transcript together with the original video, while the corrected transcripts were used to produce the human reference translations and the translations offered for post-editing. Both versions are released, which makes the transcription error of this domain directly measurable: against the corrected transcript, the released automatic transcript has a median word error rate of 2.9% for English and 7.6% for Czech, and a median character error rate of 12.7% for Chinese.

Two conventions of this domain are worth stating explicitly, as both are enforced through the translation instructions rather than through the data. First, the transcripts preserve spontaneous speech as it was spoken — repetitions, false starts and unfinished phrases — but translations are asked to omit disfluencies. ASR output is inconsistent in whether it renders them at all, so rewarding their translation would systematically penalise systems that work from the audio directly for material absent from the transcript; and the choice of whether to render a hesitation is a configuration decision rather than a translation difficulty. Second, systems are asked to place each sentence on its own line, and speaker labels, although used when preparing the references, are not exposed to participants or to annotators — in both cases so that systems which do not rely on our ASR segmentation are not disadvantaged during evaluation.

Software Domain. (translating from English) The software domain subset was collected and provided by Alconost. It consists of short UI texts aggregated from several different applications, and introduces a typical software localization challenge. To preserve contextual clues for the translation systems, the data is structured in a JSON format utilizing key-value pairs. The keys act as descriptive identifiers detailing the string's purpose in the UI, such as a "marketing button submission", while the corresponding values hold the actual English source strings. A special feature is the presence of embedded placeholders within the text. Systems must successfully carry these placeholders and JSON keys over into the target language without translating or altering them.

Education domain. (translation from Czech into German, Ukrainian, Vietnamese)
The Educational domain includes selected exercises from an online portal Škola s nadhledem for elementary-school students from various sub-

[3]Tokens are counted as maximal word-character sequences for English and Czech and with jieba for Chinese, punctuation excluded.

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

Fact Checking domain. (translation from Czech into German, Ukrainian, Vietnamese)
The fact-checking domain contains debunking analyses from the Demagog.cz website, which regularly monitors Czech media, including spoken debates, and selects specific statements that can be verified. For each such statement, an analysis assesses its factual accuracy. This analysis is typically a few paragraphs long, cites supporting materials, and concludes by classifying the statement as correct, incorrect, misleading, or unverifiable. See a sample analysis.

2.2 Human Translations

The human translations were prepared by professional translators coordinated by the partners responsible for the individual language pairs. The partners received the Translator Brief provided in Appendix A.5. Unlike in WMT25, where all the human translations were produced from scratch (Kocmi et al., 2025), this year we also encouraged controlled human post-editing of automatic translations. While the Translator Brief describes the post-editing setup, partners were explicitly given the option to depart from post-editing and produce translations from scratch when, based on their preliminary analysis and knowledge, they expected this to result in higher-quality human translations. The overall goal was therefore to give partners sufficient flexibility to choose the procedure they considered most appropriate for producing the highest-quality possible human translations.

For Czech→Vietnamese and Czech→German, *two* human translations were created independently: translation from scratch and translation by post-editing. This allows us to directly compare humans, machines and the combination of humans and machines in the translation task.

Post-editing setup. For each language pair, we provided candidate translations generated by two open-weight models: Qwen3.5-397B-A17B and Gemma-4-31B-IT. Our preferred procedure was for partners to first inspect the source documents together with the outputs of both models, select the model that appeared to provide the better starting point for that language pair, and then use the selected model across all documents in the pair.

This was a preferred rather than mandatory procedure. Partners were explicitly allowed to deviate from it when they expected another strategy to produce higher-quality human translations, including translating entirely from scratch, using the output of a different MT system, or choosing different starting models for different documents. For instance, partners could use a system specifically adapted to the language pair in particularly low-resource settings. Whenever human post-editing was used, we required partners to return the automatic translations that had actually served as the starting point. This allowed us to retain the provenance of the final human translations while prioritizing their quality over strict procedural uniformity.

Selection of open-weight models for human post-editing. Before producing the human translations, we compared several open-weight models on WMT25 English-source test data covering the 12 target languages shown in Table 3. We evaluated their translations using reference-less GEMBA-ESA with GPT-5.4 as the judge. Qwen3.5-397B-A17B and Gemma-4-31B-IT achieved the two highest macro-average scores, 82.9 and 80.3, respectively, outperforming the remaining open-weight candidates. We therefore selected them for generating the default automatic translations provided as possible starting points for human post-editing. Their per-language results also differ substantially in some cases, motivating the decision to provide both candidates and let partners determine which output was most suitable for their language pair.

Domain-specific context. For the Speech domain, the automatic translations provided for possible post-editing were generated from the cleaned transcripts after human correction of the automatic transcriptions, following the transcript cleanup guidelines reported in Appendix A.7. For the Social domain, we additionally provided screenshots of the original conversations, where available, so

Model	Avg.	ar_{egy}	cs	de	et	id	is	ja	ko	th	uk	vi	zh_{Hans}
Qwen3.5-397B-A17B	**82.9**	78.7	**83.1**	87.0	**80.4**	84.5	**70.0**	85.5	85.3	83.1	84.4	86.1	86.2
Gemma-4-31B-IT	80.3	**85.7**	77.1	**88.6**	52.1	**86.9**	43.1	**89.3**	**89.8**	**88.4**	**86.7**	**87.1**	**88.4**
Command-A-Translate	69.9	72.7	76.4	81.4	51.0	76.0	39.7	77.1	78.0	57.5	74.5	76.6	77.4
DeepSeek-V4-Pro	68.9	54.7	66.4	72.8	61.8	76.7	49.8	74.6	69.7	68.8	70.5	73.7	87.0
TranslateGemma-27B	67.1	61.0	67.2	70.6	58.2	71.1	51.4	67.5	64.0	70.9	73.4	75.0	75.2

Table 3: Reference-less GEMBA-ESA scores (Kocmi and Federmann, 2023b,a), using GPT-5.4 as the judge (OpenAI Team, 2026), for the open-weight models considered in our preliminary comparison on WMT25 English-source test data. Columns report scores for English→target translation, and Avg. is the macro-average across the 12 target languages. Bold scores indicate the best result in each column. Bold model names indicate the two open-weight models selected to generate the default automatic translations provided as possible starting points for human post-editing.

that translators could use their visual and conversational context when producing the final translations. The domain-specific prompts used to generate the Qwen3.5-397B-A17B and Gemma-4-31B-IT translations provided as possible starting points for human post-editing are reported in Appendix A.6. These automatic translations were generated using document-level translation.

Final production strategies. Because partners were free to select the procedure they considered most appropriate for their language pair, the final production strategies differed across languages. Table 4 summarizes the choices for the language pairs for which completed human translations were obtained. Five language pairs used only human post-editing, while five were translated only from scratch. For Czech→German and Czech→Vietnamese, the partners instead provided two separate sets of human translations: one produced entirely from scratch and one produced through human post-editing. Both sets were included separately in the human evaluation.

The results (Section 6.2) indicate that across the evaluated languages, post-edits rank higher than the primary initial system, Gemma-4-31B. We unfortunately did not include Qwen3.5-397B-A17B in the human evaluation as it was never exclusively selected by our partners, limiting a direct comparison. For Czech→Vietnamese, the post-edited reference tops the ranking, significantly outperforming the initial system.

For Czech→Vietnamese and Czech→German, we prepared human translations both from scratch and via post-editing. The results show that both approaches yield strong translations, but different ranking makes it inconclusive whether one strategy strictly outperforms the other.

Overall, the 12 language pairs therefore resulted in 14 independently produced human translations: seven based on human post-editing and seven translated from scratch. Since each language pair was coordinated by a different partner and the translation production strategy was deliberately flexible, some variability in the resulting human translations across languages is expected.

Language pair	Human translation(s)	PE starting system
Czech→German	Scratch + PE (two refs.)	Gemma/Qwen, per document
Czech→Vietnamese	Scratch + PE (two refs.)	Gemma/Qwen, per document
Chinese→Japanese	PE	Gemma
English→Arabic (EG)	PE	Gemma
English→Czech	PE	Gemma/Qwen, per document
English→Estonian	Scratch	—
English→German	PE	Gemma/Qwen, per document
English→Icelandic	Scratch	—
English→Japanese	PE	Gemma/Qwen, per document
English→Ladin	Scratch	—
English→Ligurian	Scratch	—
English→Northern Sámi	Scratch	—

Table 4: Production strategies used for the completed human translations. PE denotes post-editing and Scratch denotes translation from scratch. For Czech→German and Czech→Vietnamese, two sets of human translations were produced and included in the human evaluation. "Gemma/Qwen, per document" indicates that the automatic translation used as the starting point for post-editing was selected separately for each document.

Synthetic references. We release one synthetic reference for each of the documents with retained human-evaluation evidence. As a translation can be assessed in several contrastive screens, we first average its retained cESA scores and major error span counts for each source segment and system. For each candidate translation in a document, we then

average these per-segment cESA values over the document and select the highest-scoring candidate, breaking ties by fewer average major error spans. All selected candidates have cESA evidence for every segment in their document. To preserve document coherence, we retain the selected complete translation by default. We consider a segment for repair when its per-segment average cESA is below 80, or its average major error count is above zero; it is replaced only if another translation displayed and human-assessed for that same segment has a higher average cESA score, or the same average cESA score and fewer average major error spans. This repairs 862 segments (6.8%) and raises the in-sample mean cESA score from 91.88 to 93.34, while lowering the mean major error count from 0.137 to 0.092. The resulting system-level cESA scores for all evaluated directions are reported in Table 5.

3 Training Data

Similar to the previous years, we provide a selection of parallel and monolingual corpora for model training. The parallel data selection include large multilingual corpora such as Europarl-v10 (Koehn, 2005), Paracrawl-v9 (Bañón et al., 2020), CommonCrawl, NewsCommentary-v18.1, WikiTitles-v3, WikiMatrix (Schwenk et al., 2021a), TildeCorpus (Rozis and Skadiņš, 2017), OPUS (Tiedemann, 2012), CCAligned (El-Kishky et al., 2020), CCMatrix (Schwenk et al., 2021b), UN Parallel Corpus (Ziemski et al., 2016), NLLB Data (NLLB Team et al., 2022), OPUS-100 (Zhang et al., 2020), and language-specific corpora such as Yandex-Corpus, ELRC EU Acts, JParaCrawl (Morishita et al., 2020), Japanese-English Subtitle Corpus (Pryzant et al., 2018), KFTT (Neubig, 2011), TED (Cettolo et al., 2012), and back-translated news. This year's monolingual data include the following: News Crawl, News Commentary, CommonCrawl, Europarl-v10 (Koehn, 2005), Leipzig Corpora (Goldhahn et al., 2012), and language specific resources such as LangUk. We have automated the data preparation pipeline using MTDATA (Gowda et al., 2021) and instructions to download datasets were provided on the task webpage at www2.statmt.org/wmt26/mtdata.

4 System Submissions

This year, 30 systems participated in the General MT Task, slightly fewer than in the previous year (36 systems). Most systems covered two or more language pairs, while seven focused on a single primary language pair.

All participating systems are listed in Table 6. Participating teams that did not submit a system description are marked as Anonymous in all tables as they did not fulfill the shared task criteria to be presented. Section 4.1 discusses general trends in the architectures and training strategies used by the participating systems.

In line with previous years, we included translations from 16 systems, comprising a publicly available MT service (Google Translate), proprietary LLMs such as GPT-5.5, and open-weight LLMs such as Mistral Medium 3.5. The systems added by organizers are listed in Table 7.

4.1 Architectures and Strategies

Each participating team was required to submit a form detailing their approach along with a description paper. This section discusses the submissions with a summary of strategic details, such as the approach or base model, provided in Table 6.

Architectures. As in the previous year, LLMs were the predominant approach: 27 of the 30 submissions involve LLMs, including one mixed-architecture system. Three systems use encoder–decoder architectures, either fully or as one component of the system. One hybrid system combines statistical machine translation (SMT) with encoder–decoder NMT, while another system relies entirely on rule-based machine translation (RBMT).

Base models. The most frequently reported base-model family was Qwen (9 submissions), followed by Gemma/TranslateGemma, OpenAI GPT, and Gemini (6 submissions each). Eighteen submissions described training or adapting their models, while 11 used existing models without additional training; the remaining submission used a rule-based system.

Post-editing and inference-time techniques. Teams often combined multiple approaches; the most commonly reported techniques for post-editing and inference-time processing were prompt engineering and rule-based validation or correction (14 submissions each), followed by format reconstruction or checking (12).

Clear differences emerged between the two tracks. Agentic orchestration and ensembling were

Direction	cESA	Direction	cESA	Direction	cESA
Czech→German	95.30	English→Armenian	96.43	English→Russian	96.41
Czech→Ukrainian	94.32	English→Indonesian	96.41	English→Northern Sámi	75.98
Czech→Vietnamese	98.16	English→Icelandic	96.50	English→Thai	94.81
English→Arabic (Egyptian)	87.61	English→Japanese	98.47	English→Ukrainian	97.34
English→Belarusian	93.03	English→Kazakh	94.39	English→Simplified Chinese	90.48
English→Czech	90.23	English→Korean	91.49	English→Traditional Chinese	92.55
English→German	92.52	English→Ligurian	91.14	Simplified Chinese→Japanese	89.98
English→Estonian	92.65	English→Ladin	86.26		

Table 5: System-level cESA score of the hybrid synthetic reference in every evaluated direction. For each source segment, we use the arithmetic mean of retained human scores for the selected translation; the reported direction score is the arithmetic mean over those segment scores, exactly as for the official system-level cESA results.

reported exclusively by unconstrained-track systems (9 and 7 submissions, respectively). Automatic post-editing was also more common in the unconstrained track (6/17; 35.3%) than in the constrained track (3/13; 23.1%).

Data preparation. Overall, 20 of the 30 submissions reported using some form of data filtering, including basic filtering, elaborate filtering, or both. Synthetic data generation was reported by 18 submissions. The two tracks differed considerably in their reported data-preparation practices. Most constrained-track systems used data filtering and synthetic data (12/13), compared with 8/17 and 6/17 unconstrained-track systems, respectively.

4.2 LLM Benchmark

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

Evaluated models. The LLMs included in the evaluation were selected as follows: we chose popular LLM families that performed reasonably well in last year's shared task or exhibit strong multilingual capabilities and, for each family, selected the most recent available version for the constrained track and one for the unconstrained track. In addition, we included Google Translate as an industry-standard point of comparison. The final list of systems is presented in Table 7.

Prompting LLMs for translation. We developed a unified translation pipeline to collect outputs from all LLMs under an identical evaluation setup. We adopted a zero-shot, instruction-following approach and translated complete documents in a single pass using the officially provided prompt instructions.[4] Specifically, for open-weight LLMs, we used vLLM and Huggingface Transformers with Accelerate (fallback for models not supported in vLLM) to run the inference. We used the sampling temperatures recommended in the corresponding documentation and usage examples. We also provided a system prompt for models that supported this functionality. Although we initially evaluated models in thinking mode when available, these models were unable to generate translations within the practical constraints of our hardware resources (80 GB of GPU memory). Consequently, thinking mode was disabled for all models except for DeepSeek and Command A+.

API inference and cost. We collect the following systems via their respective service APIs: Gemini 3.1 Pro, GPT 5.5, Google Translate, Command A+, DeepSeek V4 Pro, Tiny Aya Global. Table 8 shows the number of input and output tokens as determined by each model's tokenizers. The estimated costs shown are for the main test set containing 23 language pairs and do not include the test suites. For the systems collected via API, we left their setup to the default parameters including high reasoning being used.

The cost per translation varies greatly, from $252 for Gemini to $4 for Tiny Aya, which also reflects the translation quality.

4.3 Constrained and Unconstrained Tracks

To promote fair comparison and encourage replicability, the WMT26 General MT Task distinguished between two tracks: *Constrained* and *Unconstrained*. These tracks differ in terms of model size, licensing, and reproducibility requirements.

[4]The code used to collect translations is available at: github.com/wmt-conference/wmt-collect-translations.

Team	Approach	Model	Size	Data	Training	Post-proc	Best rank
RULE-BASED MACHINE TRANSLATION (RBMT)							
apertium-sme-eng	RBMT	Apertium	0	dict	—	—	16-16 (en-se)
TRAINED ENCODER–DECODER (ENC-DEC) AND HYBRID SYSTEMS							
LiLa	enc-dec	NLLB	1.3B	S, spell	SFT	format, rules	2-3 (en-lij)
ReMova	enc-dec, LLM	NLLB, TranslateGemma it, Gemma 4 it	3.3B, 12B, 12B	B, E, S, spell	SFT, Ada	—	18-22 (en-be)
[Anonymous]	SMT, enc-dec	NLLB	7B	B, E, S	CPT, SFT, CO	MBR, QE, ensemble, rules	1-5 (en-ru)
TRAINED LLMS							
AMI	LLM	Qwen 3	4B	E, S, dict	CPT, Ada, merge	rules, retry, prompt	20-22 (en-is)
Cohere CAT+	LLM (MoE)	Command A+	218B (24B A)	B, E, S	CPT, SFT, CO, RL	—	1-4 (en-cs)
CUNI-AR	LLM ensemble	Gemma 4 it	~12B	E, S	SFT, CO, Ada	QE, prompt, selection	1-5 (en-arz)
CUNI-EdUKate	LLM	Granite 4.1	9.96B total	E, S	SFT, CO, Ada	format	19-24 (cs-de)
CUNI-MH-v3	LLM	Qwen3	16.84B total	E, S	SFT, CO, Ada	format	18-21 (cs-uk)
Grial-SalamandraTA	LLM	SalamandraTA instruct	7B	B, S	SFT	—	15-19 (en-lld)
HW-TSC	LLM agent	openPangu-Embedded, HY-MT2, DeepSeek-V4	7B + 7B + 284B	B, E, S	SFT, CO	APE, prompt, agent, format	3-4 (cs-uk)
Lumen	LLM pipeline	Qwen3, Qwen3.5	14B + 4B	B, E, S	CPT, SFT, CO, Ada, RL, merge	APE, rules, format, retry	1-2 (cs-uk)
Qingqiu-MT-9B	LLM	Qwen3.5	9B	B, E, S	SFT, RL	format	1-11 (cs-vi)
RFMT	LLM	Marco-MT-Algharb	18.795B total	B, E, S, dict	SFT, CO	MBR, QE, APE, rules, retry, format	15-20 (en-ja)
SalamandraTA	LLM	Salamandra	7B	E, S	CPT, SFT, CO, Ada	MBR, QE, prompt, rules	7-8 (en-lij)
TartuNLP/Tähetorn	LLM	Tower+	9B	B, E, S	SFT	—	13-14 (en-se)
TransCore	LLM	Qwen 3	8B	B, E, S	SFT	prompt	9-15 (en-id)
Yandex	LLM (MoE)	Qwen 3	235B (22B A)	B, E, S	CPT, SFT	prompt	8-11 (en-ru)
Yandex-8B	LLM	YandexGPT-5-Lite pretrain	8B	B, E, S	SFT	prompt	10-13 (en-ru)
NO TRAINING, INFERENCE ONLY (NO WEIGHT UPDATES)							
AdaptiveMT	LLM pipeline	GPT-4.1, GPT-5-mini, Solar Pro 3	—	—	—	reasoning, ensemble, APE, prompt, agent	9-15 (en-ko)
Dubformer	LLM pipeline	Claude Sonnet 4.6, Claude Opus 4.8, Gemini 3.1 Pro	—	—	—	reasoning, agent, judge, rules, format, retry	1-5 (en-ru)
Grammatist	LLM agent	GPT-5.4	—	dict	—	agent, prompt, pivot, rules, retry	2-3 (en-lij)
Krosny	LLM pipeline	TranslateGemma	12B	B, dict	—	APE, prompt, rules, selection, judge, retry, format	15-18 (en-be)
MakotoAI	LLM pipeline	Claude Opus 4.6, GPT-5.4/5.5, Gemini 2.5 Flash, Qwen3.7-Max, Gemma-4-31B-it	—	—	—	agent, prompt, rules, format	1-4 (en-cs)
PseudoRef	LLM ensemble	GPT, Gemini, EuroLLM, TranslateGemma, Aya-expanse, Tower+, HY-MT1.5	—	E	—	MBR, QE, reasoning, ensemble, APE, prompt, rules	5-7 (en-hy)
SCIR-TG-MT	LLM agent	DeepSeek-V4-Pro	—	—	—	prompt, agent, format, rules, retry	1-4 (en-ja)
SRPOL	LLM agent, LLM ensemble	EuroLLM, HY-MT1.5/MT2, Hunyuan, MiLMMT, Qwen3.5/3.6, Tower+, TranslateGemma	7–122B	—	—	QE, ensemble	9-10 (en-kk)
TRIVE	LLM agent	Gemini-3.5-Flash	—	E	—	agent, APE, rules, retry, judge	1-1 (en-zh-Hant-TW)
UvA-MT	LLM ensemble	GPT-5.4, Gemini-3.5-Flash	—	—	—	ensemble, APE, prompt, judge	1-1 (en-zh-Hans)
Wayfinder	LLM agent, LLM ensemble	Gemini 3.1 Pro, Gemini 3.5 Flash, GPT-5.5, Qwen 3.7 Max	—	—	—	agent, ensemble, judge, APE, rules, format, retry	1-1 (en-et)

Table 6: Participant submissions grouped per type. Dark-gray/white rows mark unconstrained/constrained tracks. Sizes are in billions of parameters (**B**); **A** denotes active parameters in mixture-of-experts (**MoE**) models. In Model/Size, — means unavailable; in method columns, none or not applicable. Abbreviations: **SMT** (statistical machine translation), **it** (instruction-tuned model), **B/E** (basic/elaborate data filtering in Data column), **S** (synthetic data), **dict** (dictionary augmentation), **spell** (spell checking), **CPT** (continual pre-training), **SFT** (supervised fine-tuning), **CO** (contrastive optimization/preference tuning), **Ada** (adaptation such as low-rank adaptation, LoRA), **RL** (reinforcement learning), **merge** (model merging), **MBR** (minimum Bayes risk decoding), **QE** (quality estimation), **APE** (automatic post-editing), **rules** (rule-based validation/correction), **prompt** (prompt engineering), **ensemble** (system/candidate ensembling), **agent** (agentic orchestration), **format** (format reconstruction/checking), **retry** (regeneration after failed validation), **selection** (candidate selection), **judge** (LLM judging), **pivot** (relay translation via a pivot language).

Model	Size	Best Rank	Model	Size	Best Rank
Command A+	25B / 218B	9-13 (en-de)	DeepSeek V4 Pro	49B / 1600B	3-12 (en-zh-Hans)
Gemini 3.1 Pro	—	1-1 (en-uk)	Gemma 4	4B	14-23 (cs-vi)
Gemma 4	31B	2-5 (en-zh-Hant-TW)	GLM 4	9B	17-21 (en-lld)
Google Translate	—	4-4 (en-lij)	GPT 5.5	—	1-3 (cs-de)
GPT OSS	20B	16-20 (en-zh-Hant-TW)	GPT OSS	120B	14-17 (en-lld)
Ministral 3	14B	19-20 (en-hy)	Mistral Medium 3.5	128B	12-14 (en-cs)
Qwen 3.5	9B	19-28 (cs-vi)	Qwen 3.6	27B	10-19 (en-zh-Hant-TW)
Tiny Aya Global	3B	7-14 (en-cs)	Tower+	9B	16-17 (en-is)

Table 7: Systems added by the organizers to the General MT shared task, all of them pretrained models used at inference time without weight updates. Model sizes are reported in billions of parameters (B); mixture-of-expert is denoted as active / total parameters; — denotes unavailable information.

System	Tokens (in/out)	Cost
Gemini 3.1 Pro	2.1M / 20.6M	$251.7
GPT 5.5	2.0M / 4.1M	$133.5
GoogleTranslate (approx.)	4.1M chars	$82.9
Command A+	1.9M / 5.8M	$62.6
DeepSeek V4 Pro	2.1M / 10.9M	$41.5
Tiny Aya Global (approx.)	2.0M / 1.9M	$3.9
GoogleTranslate (21 lps)	3.8M chars	$76.1
Tiny Aya Global (18 lps)	1.6M / 1.5M	$3.0

Table 8: Input and output token usage, language-pair coverage, and estimated standard API cost using July 2026 prices. Google Translate and Tiny Aya Global do not support all languages, so we linearly approximated cost for all 23 languages while sharing the real cost below the table.

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

- Any training data, models, and frameworks may be used, provided that their use permits the final model weights to be publicly released
- The final model must contain fewer than 20 billion parameters. Larger models may be used during intermediate stages (e.g., for distillation), but the submitted outputs must be produced by a final model that complies with the size limit. For mixture-of-experts models, the total parameter count is considered; for ensembles, the parameter counts of all component models are summed.
- Model weights must be released under an open-source license allowing at least unrestricted non-commercial use and further research use.

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

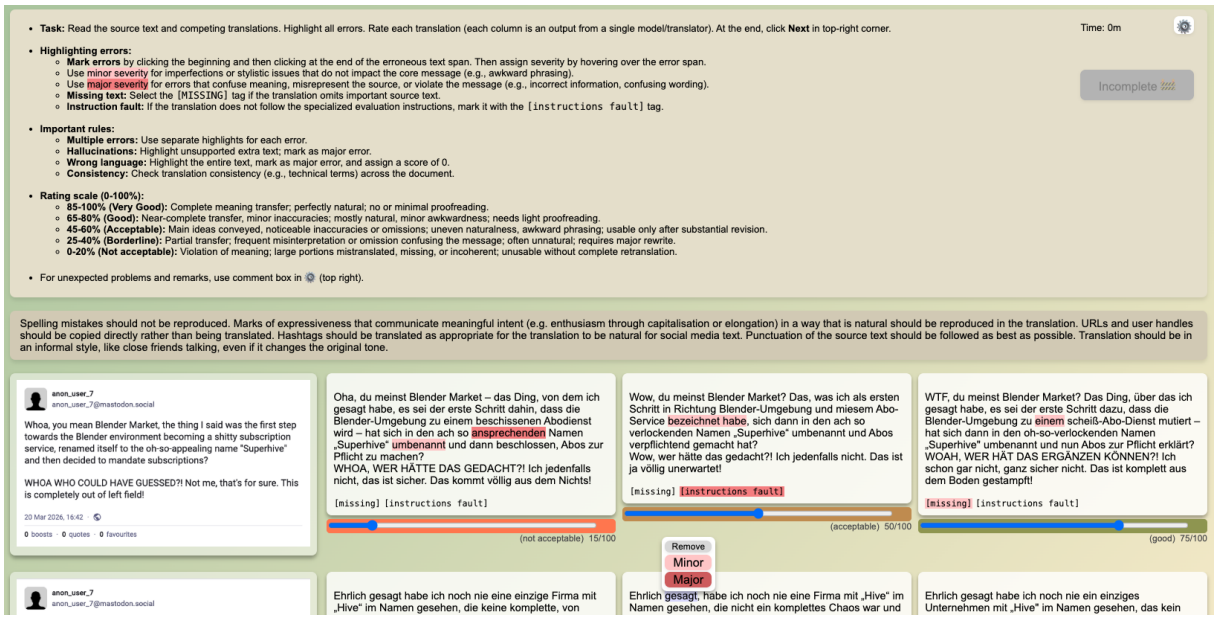
Unconstrained Track: This track imposes no restrictions on model size, training data, or licensing. It includes systems built with proprietary tools, closed-source models, or undisclosed training pipelines, such as commercial LLMs (e.g., GPT-based systems). While participation in this track is open,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

Although this year we did not restrict training data for either track, we curated a set of corpora that covers the majority of publicly available resources to support participants.[5]

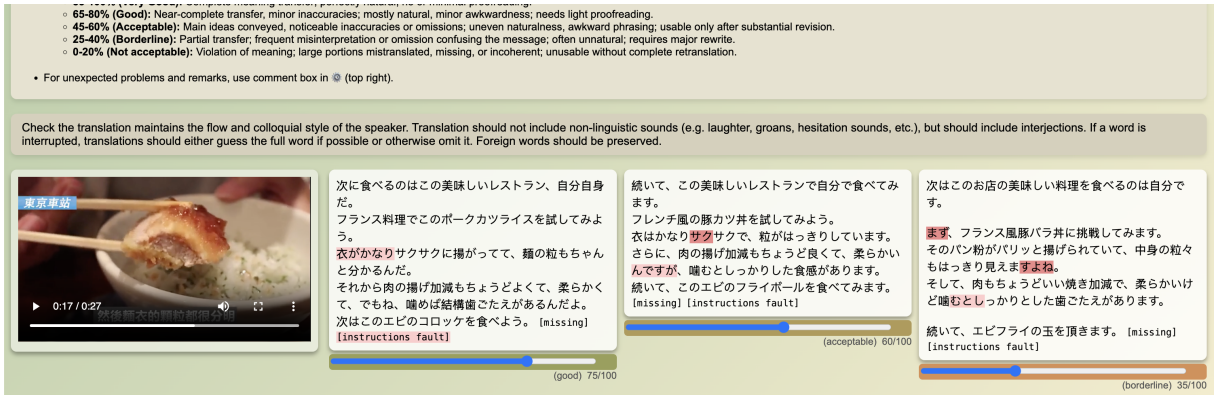
To assist participants in the constrained track, we provided a non-exhaustive list of suggested models under the 20B parameter limit, including: textual models: Aya Expanse 8B, Aya 101 (13B), Cohere R 7B, Gemma 3 12B, GPT-OSS 20B, LLaMA 3.1 8B, Qwen 2.5 7B, Ministral 3 14B, Mistral 7B, EuroLLM, NLLB; and multimodal models: Whisper, Seamless M4T.

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

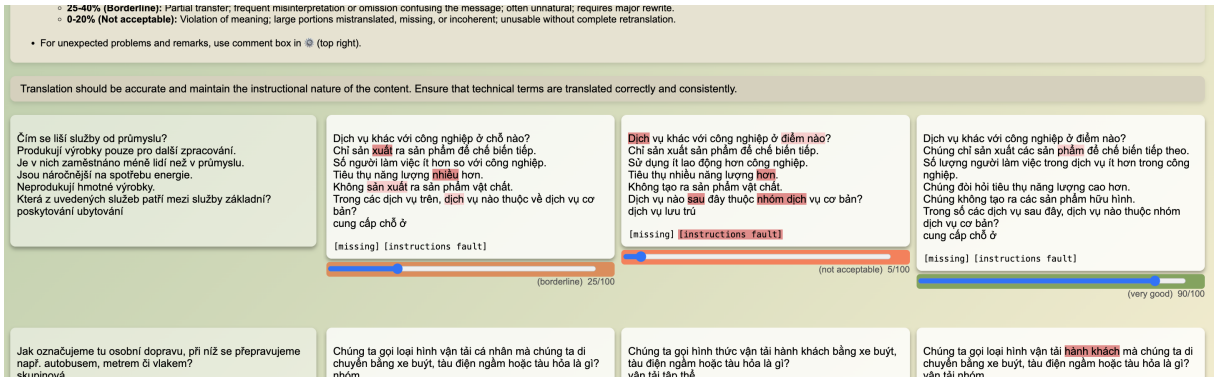
[5] www2.statmt.org/wmt26/mtdata



(a) Screenshot of English→German annotations in the social domain with instructions at top.



(b) Screenshot of Chinese(Simplified)→Japanese annotations in the speech domain. The video can be paused and replayed.



(c) Screenshot of Czech→Vietnamese annotations in the education domain. The image partially shows the next segment which provides context.

Figure 1: Three screenshots of cESA (Zouhar et al., 2026c) annotations. cESA shows multiple translations for the same source, multiple segments within a document at once as well as video or image sources. After marking the individual error spans, the annotator assigns the final segment score from 0 to 100.

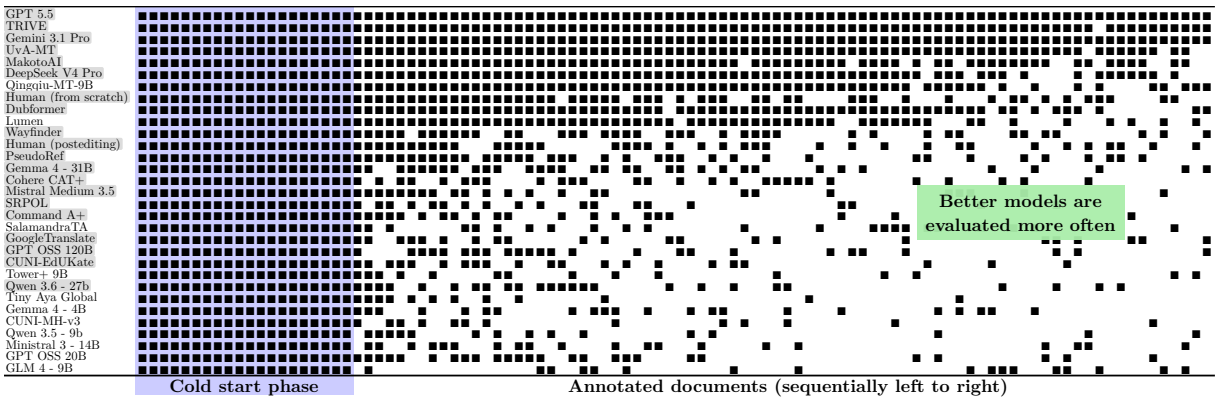


Figure 2: Evaluation coverage by model; ■ marks an evaluated document. After the cold-start phase, translations from better-performing models were evaluated on more documents. The figure shows Czech→German and is clipped for brevity.

Language pairs	Annotators' profile	#Annot.	Time/Seg.	#Errors	MAE
English→Chinese(Simpl.)	Professional paid translators.	2	89.3s	2.1	17.8
English→Chinese(Trad.)		2	92.8s	2.2	31.2
English→Japanese	Microsoft annotators: bilingual target-language native speakers, professional translators or linguists, experienced in machine translation evaluation.	49	36.5s	0.7	17.6
English→Korean		8	76.9s	1.4	19.5
English→Armenian	Native speakers and professional translators.	15	66.5s	2.8	22.0
English→Kazakh		21	52.8s	1.4	17.8
English→Belarusian		15	71.2s	3.4	18.7
Chinese→Japanese	Professional translators.	13	39.5s	1.5	17.5
English→Ladin	Unpaid academic/institutional annotators.	8	52.4s	3.3	21.6
English→Estonian	Professional translators from Luisa Language Solutions and volunteers from Sámi Giellagáldu and Sámi allaskuvla.	6	43.2s	2.2	13.3
English→Northern Sámi		4	108.2s	4.8	14.2
Czech→Vietnamese	ÚFAL Charles University paid annotators: linguists, annotators, researchers, and students who were native speakers of one of the languages and high proficiency in the other.	9	125.5s	3.3	29.4
Czech→German		11	84.7s	2.1	17.9
English→Czech		14	48.9s	2.3	19.5
Czech→Ukrainian		8	53.4s	1.8	17.2
English→Ukrainian	Toloka AI paid annotators: Bilingual native target-language speakers who were high-performing on the platform.	92	53.4s	1.7	19.9
English→Arabic		102	16.7s	0.6	5.0
English→Russian		87	41.2s	1.9	17.4
English→Icelandic	Bilingual target-language native speakers, paid translators with 3–25 years of experience in Icelandic↔English translation.	6	38.8s	2.2	21.2
English→German	Cohere annotators: professional in-house employees experienced with annotations of general LLM outputs, bilingual speakers native in German.	7	73.3s	1.8	21.7
English→Thai	Professional translators from Venga Global.	15	68.1s	2.4	16.8
English→Indonesian		10	44.1s	1.2	21.7
English→Ligurian	Volunteer native Ligurian speakers working in English, all with prior translation experience.	5	33.6s	4.8	18.1

Table 9: Overview of Annotators' profiles, Number of annotators, Time per segment, Total errors per segment, and Mean absolute error. All paid annotators were paid a fair wage in their respective countries.

5 Human Evaluation Methodology

The human evaluation is carried out through the Pearmut platform for all language pairs (Zouhar and Kocmi, 2026).

Contrastive Error Span Annotation Protocol. We use the cESA annotation protocol (Zouhar et al., 2026c) where 3 translations of the same input are presented side by side, which speeds up the annotations as the annotators do not have to reread the same source multiple times, and can more easily compare errors between the different translations. The annotators first mark the translation error spans as *Minor* and *Major* in all 3 translations and then assign an overall score between $0 - 100$ to each. The output of the cESA is a list of error spans, their severity, and an overall score from 0 to 100. A detailed description of the score categories is also provided to annotators, explaining what specific scores mean. Figure 1 shows an example of the cESA interface.

Domain-specific Instructions. In addition to the general guidelines, we provide annotators with specific detailed instructions based on the domain of the sample to align with the test set instruction following context (e.g., in the social domain, annotators are asked to translate hashtags in an appropriate natural way for social media text). See the instructions for each domain in Section A.1.

Segmentation for human evaluation. Although systems translated each test item at the document level, the human evaluation required segment-level translations to parse for the interface. We therefore parsed each document-level system output into segments before annotation. Whenever possible, we recovered segment boundaries from the test-set structure, such as HTML tags or JSON key-value pairs. When the system failed to correctly preserve them, we automatically aligned the translation with the source segments to obtain the best possible segment correspondence.

Dynamic allocation. For WMT2026, we skipped the pre-selection of systems based on automatic metrics. As a result, all systems were human-evaluated. However, we care primarily about identifying the top systems and less about distinguishing between the systems ranked 29th and 30th. To this end, we use dynamic evaluation (Zouhar et al., 2026a), where after a "cold start phase" of 20 documents, the top-performing systems are human-evaluated on more documents. The models are selected iteratively and stochastically any time an annotator wishes to annotate. Model m's probability of being chosen is proportional to $\frac{1}{\text{rank}_m}$, where rank_m is the current (not final) rank of the model. This setup concentrates the budget on separating the top-ranked systems, at the cost of a less reliable ranking among the remaining ones. In Figure 2, we show a trace of dynamic data allocation for the Czech→German evaluation. After the cold start phase, top-ranked models are evaluated on more documents.

Increasing quality of annotations. The annotators are asked to follow a short interactive tutorial in X→English at the beginning. The goal of the tutorials is to familiarize the annotators with Pearmut and the document-level cESA annotation task. An example of the 7-step tutorial is shown in Figure 6.

To ensure that annotators carefully follow the given instructions, we randomly add 3-5 "loud attention checks" shown to the annotator. For these attention checks, we know beforehand what the ranking of the 3 systems should be and if the annotator attempts to submit annotations that do not follow this ordering, a message is shown that they should pay more attention. This real-time accountability ensures annotators remain aware that their work is evaluated. Examples of attention checks are provided in Figure 5.

Lastly, to enhance the reliability of collected annotations, we filter out annotators with low reliability scores. For each annotation, we compute its likelihood given the annotations from all other annotators, and we exclude annotators with outlier annotation likelihoods (≈1% of annotators):

$$\frac{1}{N}\sum_{m_1,m_2}\exp\left(-\frac{(s_1-s_2-\mu_{m_1,m_2})^2}{2\sigma^2_{m_1,m_2}+\epsilon}\right) < 0.75$$

Where N is the total number of evaluations, s_1 and s_2 are scores assigned by the user to models m_1 and m_2, the μ_{m_1,m_2} is the population mean score difference between m_1 and m_2, and $\sigma^2_{m_1,m_2}$ is the sum of score variances for m_1 and m_2.

Analysis of annotators. We provide an overview of annotators' profiles as well as annotation volume, time (a proxy for cost), and inter-annotator disagreement (mean absolute error in scores) in Table 9. While annotation speed and the number of annotated errors vary considerably, neither is predictive of inter-annotator agreement.

6 Human Evaluation Results

6.1 Overall Ranking of Systems

Model	Avg. Rank	English→Japanese	English→Egyptian Arabic	Czech→Vietnamese	Czech→Ukrainian	English→Indonesian	Czech→German	English→Ukrainian	English→Chinese (Simp.)	English→Russian	Chinese (Simp.)→Japanese	English→Korean	English→Thai	English→Czech	English→German	English→Chinese (Trad.)	English→Kazakh	English→Estonian	English→Icelandic	English→Armenian	English→Belarusian	English→Northern Sami	English→Ligurian	English→Ladin
Grammatist	2.5																						65.9	49.4
LiLa	2.6																						66.0	48.9
Human (from scratch)	4.3			87.4			80.6											81.0	87.5			59.8	84.7	77.2
Gemini 3.1 Pro	4.8	90.3	83.7	88.5	84.1	83.5	84.0	84.3	79.3	79.4	75.7	83.5	86.2	75.5	78.2	70.1	80.0	81.1	84.9	80.7	77.6	61.3	12.4	45.6
CUNI-AR	4.8		83.2																					
TRIVE	5.1	88.3	82.7	86.6	86.9	83.5	84.3	82.7	76.8	80.7	74.2	82.5	85.0	75.0	78.5	80.5	82.1	81.4	82.5	83.4	77.6	53.3	23.8	34.6
Wayfinder	5.2	90.6	83.9	88.5	84.5	82.9	79.8	81.7	75.2	78.7	78.1	78.6	85.0	78.2	79.1	74.1	81.3	82.9	84.0	82.8	77.3	55.5	24.5	48.4
Lumen	6.8	90.6	83.1	85.9	86.6	82.6	80.2	81.5	78.6	79.8	73.7	82.5	84.0	65.5	72.0	74.0	82.2	80.3	84.5	83.4	75.1	54.5	25.4	34.8
Human (postediting)	7.4	86.0	83.4	90.6			79.8				79.6			71.0	73.0									
GPT 5.5	8.6	86.1	80.8	89.5	85.8	80.3	84.8	80.9	73.3	76.0	75.3	80.9	82.0	77.1	77.1	68.6	81.9	80.3	79.7	80.0	69.6	49.3	38.3	29.7
UvA-MT	8.9	86.1	82.1	88.1	84.4	81.8	83.5	82.5	79.5	77.6	74.9	78.5	83.1	70.8	70.9	61.6	83.1	80.1	82.2	82.8	72.0	45.8	28.0	20.2
[Anonymous]	11.0		82.2	86.1		80.9				80.7			76.6	68.3							72.0			
Yandex	12.4									75.3														
Dubformer	13.8	85.4	82.4	88.2	83.3	78.8	80.6	81.0	75.8	81.1	73.4	75.9	74.2	69.5	75.5	72.4	69.9	68.0	68.4	65.0	61.0	31.7	18.9	11.5
HW-TSC	14.0	86.2	82.5		86.2	81.8		77.6	78.4		71.5	71.7	68.3	62.0		66.7	71.0							
MakotoAI	14.5	83.4	81.3	87.5	83.4	79.7	83.1	79.5	65.4	70.8	70.2	72.4	74.7	78.2	70.1	60.8	70.6	78.5	78.3	78.6	60.8	53.5	33.8	34.8
Qingqiu-MT-9B	16.5	85.2	81.7	89.3	79.2	75.9	81.2	77.9	75.6	74.9	73.7	69.3	75.9	68.2	68.1	66.2	77.8	59.8	53.5	72.6	56.5	1.7	26.4	31.3
PseudoRef	17.4	79.6				75.1	78.4	77.4	66.0	69.2	72.2	72.8	77.0		59.7	65.3	78.4	77.1	79.1	81.0	67.5			
DeepSeek V4 Pro	17.5	83.2	80.0	77.0	82.2	78.9	81.3	80.4	75.3	66.2	72.8	74.4	71.9	56.6	66.7	64.2	74.3	71.8	71.4	67.3	63.0	43.8	20.1	22.9
Cohere CAT+	18.3	83.2	82.3	80.5	79.3	78.3	76.7	79.1	74.7	74.6	72.0	73.0	78.4	79.0	68.4	61.3	65.5	57.4	62.4	37.2	54.1	35.3	21.9	6.9
SCIR-TG-MT	20.6	90.4		81.8		79.3			58.7	36.3			75.1					68.1	63.7	68.1				
AdaptiveMT	21.5											71.5												
Krosny	21.7																				53.7			
Gemma 4 - 31B	22.0	80.5	80.5	75.7	79.3	75.7	76.9	76.3	72.3	73.7	72.2	76.2	80.0	60.1	67.8	74.3	64.4	45.8	43.6	44.0	49.9	4.7	6.6	3.6
Yandex-8B	23.9									73.0							60.8			49.7	48.4			
TartuNLP	25.9																	51.8				22.3		
Command A+	26.0	81.4	80.1	73.2	76.9	76.8	71.0	75.9	65.8	70.4	66.9	63.0	70.1	61.3	70.8	61.9	57.8	56.6	42.0	12.5	53.7	2.6	7.0	2.9
SRPOL	26.1	81.7	80.3	73.0	78.4	69.3	72.3	62.6	68.9	64.1	67.0	56.4	59.3	60.6	47.2	43.2	74.6	58.3	66.7	67.4	46.9	20.7	13.8	
RFMT	26.3	81.2									65.8													
Apertium-sme-eng	26.7																					11.7		
GoogleTranslate	26.9	69.4	72.4	68.9	74.7	70.8	64.5	65.3	69.2	66.1	40.6	63.5	70.9	60.2	57.1	51.7	58.6	62.8	66.1	52.3	54.7		47.4	
TransCore	27.3					78.7					71.4		50.8				45.3							
SalamandraTA	29.0	80.2	76.8	67.8	69.6	67.3	65.0	66.1	58.0	57.9	57.7	49.4	61.0	54.4	55.0	46.2	56.3	66.1	69.0	52.9	44.1	33.4	28.6	1.6
Mistral Medium 3.5	29.3	74.1	70.6	79.9	78.1	65.0	72.5	70.3	67.6	70.1	56.5	58.3	53.6	67.8	65.5	55.5	48.8	44.0	25.7	46.3	43.2	2.3	6.0	2.9
ReMova	29.4																				44.2			
GPT OSS 120B	32.1	78.2	77.6	66.3	63.1	68.3	62.9	58.1	61.9	56.6	64.9	54.6	45.5	52.1	58.5	58.1	48.7	47.2	46.6	32.1	30.5	2.3	4.4	5.5
AMI	32.5																		39.0					
Qwen 3.6 - 27b	35.1	68.7	65.4	63.4	59.0	59.8	58.3	44.3	63.8	54.4	55.7	55.8	42.1	61.9	51.1	59.4	43.4	39.3	33.0	21.6	27.6	13.9	5.7	3.0
Grial-SalamandraTA	35.4																		15.6				5.2	4.7
Tiny Aya Global	35.5	77.7	74.7	71.3	57.3	68.5	54.9	61.2	50.9	53.5	29.6	45.2	44.2	71.8	46.7	39.0		53.0	24.5		26.9			
Gemma 4 - 4B	35.6	77.3	78.5	74.7	63.7	70.8	54.1	56.7	64.4	59.9	61.4	60.9	62.4	32.4	49.6	53.0	25.8	15.7	16.2	30.1	18.3	0.2	1.7	0.3
CUNI-MH-v3	37.8			54.1	68.8		53.9	51.6						33.1	48.4									
GPT OSS 20B	39.8	72.4	69.9	60.6	49.7	56.0	45.4	44.6	51.7	49.5	55.1	52.0	31.3	37.7	37.5	53.3	40.8	23.3	24.1	29.0	21.6	0.8	2.4	1.1
CUNI-EdUKate	40.9				43.9		62.5	46.7							40.7									
Tower+ 9B	41.2	69.9	36.1	41.8	59.3	41.6	58.8	51.0	50.5	43.1	55.0	48.2	20.4	41.8	47.5	49.1	2.3	14.6	60.1	3.3	5.6	0.1	0.7	1.2
Ministral 3 - 14B	41.4	68.6	58.1	63.7	58.0	52.4	46.7	55.9	45.1	44.8	46.8	39.1	25.3	34.4	41.2	38.2	23.5	21.7	19.2	32.7	25.4	0.4	2.6	0.9
Qwen 3.5 - 9b	41.5	57.4	54.3	66.3	59.6	54.1	53.7	48.3	55.4	44.4	53.1	18.9	37.4	32.4	30.9	51.3	30.4	24.9	17.4	18.8	16.5	0.6	1.4	1.1
GLM 4 - 9B	43.0	65.0	65.5	53.3	49.4	42.5	45.1	46.9	57.4	32.2	50.0	26.5	8.3	40.7	35.2	47.9	5.9	13.1	7.9	3.4	10.4	1.5	2.2	2.2

Figure 3: Averaged *cursory* look at overall model ranking. Average rank is calculated as the average position over all language pairs the system participated in. For holistic and fair system comparison see the per-language results in Section 6.2.

6.2 Ranking of All Systems for each translation direction

A short gray line marks whenever a model is significantly better than the subsequent model. A longer black line marks whenever all previous models are significantly better than all subsequent models. Statistical testing with ($\alpha = 0.05$) is done using a one-sided paired or independent bootstrap test on the mean differences (100 resamples).

Rank	Czech→Vietnamese	Score	Docs
1-6	Human (postediting)	90.6	17
1-5	GPT 5.5	89.5	17
1-11	Qingqiu-MT-9B	89.3	17
1-8	Gemini 3.1 Pro	88.5	19
1-12	Wayfinder	88.5	20
3-13	Dubformer	88.2	19
2-12	UvA-MT	88.1	15
1-13	MakotoAI	87.5	14
5-12	Human (from scratch)	87.4	19
3-12	TRIVE	86.6	20
3-13	[Anonymous]	86.1	20
4-13	Lumen	85.9	17
11-16	SCIR-TG-MT	81.8	18
13-18	Cohere CAT+	80.5	20
13-18	Mistral Medium 3.5	79.9	20
13-21	DeepSeek V4 Pro	77.0	19
14-21	Gemma 4 - 31B	75.7	18
14-23	Gemma 4 - 4B	74.7	17
16-20	Command A+	73.2	20
16-25	SRPOL	73.0	20
20-25	Tiny Aya Global	71.3	18
18-28	GoogleTranslate	68.9	20
18-28	SalamandraTA	67.8	16
20-28	GPT OSS 120B	66.3	20
19-28	Qwen 3.5 - 9b	66.3	13
22-28	Ministral 3 - 14B	63.7	19
21-28	Qwen 3.6 - 27b	63.4	19
22-29	GPT OSS 20B	60.6	20
28-30	CUNI-MH-v3	54.1	17
29-30	GLM 4 - 9B	53.3	20
31-31	Tower+ 9B	41.8	19

Rank	Czech→German	Score	Docs
1-3	GPT 5.5	84.8	237
1-5	TRIVE	84.3	192
1-5	Gemini 3.1 Pro	84.0	244
2-5	UvA-MT	83.5	190
2-7	MakotoAI	83.1	132
5-9	DeepSeek V4 Pro	81.3	121
5-12	Qingqiu-MT-9B	81.2	114
6-12	Human (from scratch)	80.6	90
6-12	Dubformer	80.6	124
7-10	Lumen	80.2	97
11-15	Wayfinder	79.8	74
4-15	Human (postediting)	79.8	71
11-13	PseudoRef	78.4	70
14-15	Gemma 4 - 31B	76.9	67
11-15	Cohere CAT+	76.7	67
16-18	Mistral Medium 3.5	72.5	61
15-18	SRPOL	72.3	49
16-18	Command A+	71.0	55
19-23	SalamandraTA	65.0	53
19-23	GoogleTranslate	64.5	54
19-24	GPT OSS 120B	62.9	55
19-24	CUNI-EdUKate	62.5	49
19-26	Tower+ 9B	58.8	41
21-24	Qwen 3.6 - 27b	58.3	60
25-25	Tiny Aya Global	54.9	38
26-27	Gemma 4 - 4B	54.1	51
24-28	CUNI-MH-v3	53.9	36
27-28	Qwen 3.5 - 9b	53.7	45
29-30	Ministral 3 - 14B	46.7	50
29-30	GPT OSS 20B	45.4	55
29-31	GLM 4 - 9B	45.1	48

Rank	English→Czech	Score	Docs
1-4	Cohere CAT+	79.0	93
1-4	MakotoAI	78.2	93
1-3	Wayfinder	78.2	56
4-4	GPT 5.5	77.1	85
5-6	Gemini 3.1 Pro	75.5	96
4-6	TRIVE	75.0	69
7-14	Tiny Aya Global	71.8	34
7-11	Human (postediting)	71.0	64
5-11	UvA-MT	70.8	69
7-10	Dubformer	69.5	39
11-11	[Anonymous]	68.3	45
12-16	Qingqiu-MT-9B	68.2	30
12-14	Mistral Medium 3.5	67.8	40
10-14	Lumen	65.5	41
15-15	HW-TSC	62.0	40
16-21	Qwen 3.6 - 27b	61.9	33
15-18	Command A+	61.3	43
15-19	SRPOL	60.6	52
18-21	GoogleTranslate	60.2	28
19-23	Gemma 4 - 31B	60.1	27
18-23	DeepSeek V4 Pro	56.6	32
20-23	SalamandraTA	54.4	30
20-23	GPT OSS 120B	52.1	30
24-24	Tower+ 9B	41.8	30
25-28	GLM 4 - 9B	40.7	25
25-29	GPT OSS 20B	37.7	26
24-28	Ministral 3 - 14B	34.4	34
25-29	CUNI-MH-v3	33.1	35
28-29	Gemma 4 - 4B	32.4	28
28-30	Qwen 3.5 - 9b	32.4	24

Rank	English→German	Score	Docs
1-4	Wayfinder	79.1	198
1-4	TRIVE	78.5	198
1-4	Gemini 3.1 Pro	78.2	198
1-4	GPT 5.5	77.1	131
5-5	Dubformer	75.5	132
6-10	Human (postediting)	73.0	123
6-8	Lumen	72.0	190
6-8	UvA-MT	70.9	116
9-13	Command A+	70.8	73
9-11	MakotoAI	70.1	78
8-12	Cohere CAT+	68.4	114
11-14	Qingqiu-MT-9B	68.1	83
12-15	Gemma 4 - 31B	67.8	107
12-15	DeepSeek V4 Pro	66.7	82
13-15	Mistral Medium 3.5	65.5	71
16-16	PseudoRef	59.7	68
17-19	GPT OSS 120B	58.5	59
16-19	GoogleTranslate	57.1	59
17-19	SalamandraTA	55.0	54
20-25	Qwen 3.6 - 27b	51.1	51
20-25	Gemma 4 - 4B	49.6	48
20-26	CUNI-MH-v3	48.4	53
20-25	Tower+ 9B	47.5	58
20-25	SRPOL	47.2	46
20-25	Tiny Aya Global	46.7	38
26-28	Ministral 3 - 14B	41.2	41
26-29	CUNI-EdUKate	40.7	54
26-28	GPT OSS 20B	37.5	46
29-29	GLM 4 - 9B	35.2	39
29-30	Qwen 3.5 - 9b	30.9	44

Rank	English→Icelandic	Score	Docs
1-1	Human (from scratch)	87.5	198
2-4	Gemini 3.1 Pro	84.9	198
2-4	Lumen	84.5	198
2-4	Wayfinder	84.0	198
5-6	TRIVE	82.5	198
5-6	UvA-MT	82.2	198
7-9	GPT 5.5	79.7	198
7-9	PseudoRef	79.1	192
7-9	MakotoAI	78.3	142
10-10	DeepSeek V4 Pro	71.4	108
11-11	SalamandraTA	69.0	124
12-14	Dubformer	68.4	104
11-15	SRPOL	66.7	84
11-14	GoogleTranslate	66.1	102
15-16	SCIR-TG-MT	63.7	77
14-17	Cohere CAT+	62.4	67
16-17	Tower+ 9B	60.1	71
18-18	Qingqiu-MT-9B	53.5	69
19-21	GPT OSS 120B	46.6	61
19-22	Gemma 4 - 31B	43.6	65
19-22	Command A+	42.0	50
20-22	AMI	39.0	63
23-23	Qwen 3.6 - 27b	33.0	69
24-25	Mistral Medium 3.5	25.7	71
24-26	Tiny Aya Global	24.5	50
25-26	GPT OSS 20B	24.1	53
27-27	Ministral 3 - 14B	19.2	52
28-29	Qwen 3.5 - 9b	17.4	50
28-30	Gemma 4 - 4B	16.2	47
29-30	Grial-SalamandraTA	15.6	48
31-31	GLM 4 - 9B	7.9	46

Rank	English→Chinese (Trad.)	Docs
1-1	TRIVE	80.5	50
2-5	Gemma 4 - 31B	74.3	42
2-5	Wayfinder	74.1	48
2-5	Lumen	74.0	41
2-8	Dubformer	72.4	39
5-8	Gemini 3.1 Pro	70.1	44
5-8	GPT 5.5	68.6	36
5-10	HW-TSC	66.7	37
8-11	Qingqiu-MT-9B	66.2	28
7-16	PseudoRef	65.3	28
9-17	DeepSeek V4 Pro	64.2	33
10-17	Command A+	61.9	27
8-13	UvA-MT	61.6	29
14-18	Cohere CAT+	61.3	27
10-17	MakotoAI	60.8	38
10-19	Qwen 3.6 - 27b	59.4	30
14-21	GPT OSS 120B	58.1	30
16-22	Mistral Medium 3.5	55.5	23
16-20	GPT OSS 20B	53.3	29
17-23	Gemma 4 - 4B	53.0	26
20-25	GoogleTranslate	51.7	25
20-25	Qwen 3.5 - 9b	51.3	24
20-25	Tower+ 9B	49.1	27
21-25	GLM 4 - 9B	47.9	23
21-25	SalamandraTA	46.2	27
26-27	SRPOL	43.2	25
26-27	Tiny Aya Global	39.0	23
26-28	Ministral 3 - 14B	38.2	27

Rank	English→Egyptian Arabic		Docs
1-2	Wayfinder	83.9	188
1-5	Gemini 3.1 Pro	83.7	191
2-5	Human (postediting)	83.4	186
1-5	CUNI-AR	83.2	187
2-6	Lumen	83.1	176
5-10	TRIVE	82.7	184
6-11	HW-TSC	82.5	184
6-11	Dubformer	82.4	169
6-11	Cohere CAT+	82.3	187
6-13	[Anonymous]	82.2	177
7-12	UvA-MT	82.1	176
10-14	Qingqiu-MT-9B	81.7	190
12-14	MakotoAI	81.3	172
12-17	GPT 5.5	80.8	179
14-18	Gemma 4 - 31B	80.5	186
14-18	SRPOL	80.3	191
14-18	Command A+	80.1	188
15-18	DeepSeek V4 Pro	80.0	187
19-20	Gemma 4 - 4B	78.5	187
19-21	GPT OSS 120B	77.6	191
20-21	SalamandraTA	76.8	191
22-22	Tiny Aya Global	74.7	191
23-24	GoogleTranslate	72.4	158
23-25	Mistral Medium 3.5	70.6	189
24-25	GPT OSS 20B	69.9	162
26-27	GLM 4 - 9B	65.5	180
26-27	Qwen 3.6 - 27b	65.4	165
28-28	Ministral 3 - 14B	58.1	155
29-29	Qwen 3.5 - 9b	54.3	173
30-30	Tower+ 9B	36.1	167

Rank	English→Russian	Score	Docs
1-5	Dubformer	81.1	198
1-5	[Anonymous]	80.7	198
1-4	TRIVE	80.7	198
1-6	Lumen	79.8	198
4-6	Gemini 3.1 Pro	79.4	198
4-7	Wayfinder	78.7	198
6-8	UvA-MT	77.6	198
7-11	GPT 5.5	76.0	193
8-11	Yandex	75.3	197
8-13	Qingqiu-MT-9B	74.9	190
8-13	Cohere CAT+	74.6	195
10-13	Gemma 4 - 31B	73.7	195
10-13	Yandex-8B	73.0	193
14-17	MakotoAI	70.8	188
14-17	Command A+	70.4	191
14-17	Mistral Medium 3.5	70.1	187
14-17	PseudoRef	69.2	172
18-19	DeepSeek V4 Pro	66.2	164
17-19	GoogleTranslate	66.1	151
20-20	SRPOL	64.1	135
21-21	Gemma 4 - 4B	59.9	116
22-24	SalamandraTA	57.9	125
22-23	GPT OSS 120B	56.6	92
24-26	Qwen 3.6 - 27b	54.4	103
24-26	Tiny Aya Global	53.5	98
24-26	GPT OSS 20B	49.5	106
27-29	Ministral 3 - 14B	44.8	67
27-29	Qwen 3.5 - 9b	44.4	92
27-29	Tower+ 9B	43.1	76
30-30	SCIR-TG-MT	36.3	79
30-31	GLM 4 - 9B	32.2	70

Rank	English→Belarusian	Score	Docs
1-3	TRIVE	77.6	198
1-3	Gemini 3.1 Pro	77.6	198
1-3	Wayfinder	77.3	198
4-4	Lumen	75.1	198
5-6	[Anonymous]	72.0	164
5-6	UvA-MT	72.0	169
7-8	GPT 5.5	69.6	111
7-8	PseudoRef	67.5	136
9-11	DeepSeek V4 Pro	63.0	102
9-10	Dubformer	61.0	104
11-13	MakotoAI	60.8	74
11-16	Qingqiu-MT-9B	56.5	78
11-17	GoogleTranslate	54.7	76
12-14	Cohere CAT+	54.1	72
15-18	Krosny	53.7	60
12-16	Command A+	53.7	94
17-17	Gemma 4 - 31B	49.9	53
18-20	Yandex-8B	48.4	56
17-20	SRPOL	46.9	58
18-22	ReMova	44.2	60
20-22	SalamandraTA	44.1	61
20-22	Mistral Medium 3.5	43.2	50
23-23	GPT OSS 120B	30.5	49
24-26	Qwen 3.6 - 27b	27.6	51
24-27	Tiny Aya Global	26.9	53
24-27	Ministral 3 - 14B	25.4	49
25-27	GPT OSS 20B	21.6	43
28-28	Gemma 4 - 4B	18.3	37
29-29	Qwen 3.5 - 9b	16.5	41
30-30	GLM 4 - 9B	10.4	47
31-31	Tower+ 9B	5.6	47

Rank	English→Ladin	Score	Docs
1-1	Human (from scratch)	77.2	198
2-4	Grammatist	49.4	198
2-4	LiLa	48.9	177
2-5	Wayfinder	48.4	132
4-5	Gemini 3.1 Pro	45.6	186
6-8	Lumen	34.8	110
6-7	MakotoAI	34.8	108
8-8	TRIVE	34.6	92
9-10	Qingqiu-MT-9B	31.3	87
9-10	GPT 5.5	29.7	80
11-12	DeepSeek V4 Pro	22.9	58
11-12	UvA-MT	20.2	72
13-13	Dubformer	11.5	61
14-15	Cohere CAT+	6.9	64
14-17	GPT OSS 120B	5.5	47
15-19	Grial-SalamandraTA	4.7	59
15-17	Gemma 4 - 31B	3.6	53
18-22	Qwen 3.6 - 27b	3.0	54
16-21	Command A+	2.9	51
17-22	Mistral Medium 3.5	2.9	49
17-21	GLM 4 - 9B	2.2	49
22-23	SalamandraTA	1.6	50
21-23	Tower+ 9B	1.2	49
24-26	Qwen 3.5 - 9b	1.1	49
23-26	GPT OSS 20B	1.1	39
20-26	Ministral 3 - 14B	0.9	43
25-27	Gemma 4 - 4B	0.3	40

Rank	English→Northern Sami		Docs
1-3	Gemini 3.1 Pro	61.3	43
1-3	Human (from scratch)	59.8	44
1-6	Wayfinder	55.5	39
3-7	Lumen	54.5	38
3-6	MakotoAI	53.5	31
3-6	TRIVE	53.3	39
7-7	GPT 5.5	49.3	35
8-8	UvA-MT	45.8	33
9-9	DeepSeek V4 Pro	43.8	27
10-10	Cohere CAT+	35.3	32
11-12	SalamandraTA	33.4	26
11-12	Dubformer	31.7	27
13-14	TartuNLP	22.3	30
13-14	SRPOL	20.7	22
15-15	Qwen 3.6 - 27b	13.9	29
16-16	Apertium-sme-eng	11.7	21
17-18	Gemma 4 - 31B	4.7	25
17-22	Command A+	2.6	23
18-20	Mistral Medium 3.5	2.3	27
18-20	GPT OSS 120B	2.3	22
21-23	Qingqiu-MT-9B	1.7	23
20-23	GLM 4 - 9B	1.5	21
21-23	GPT OSS 20B	0.8	26
24-25	Qwen 3.5 - 9b	0.6	24
24-26	Ministral 3 - 14B	0.4	22
25-26	Gemma 4 - 4B	0.2	23
26-27	Tower+ 9B	0.1	25

Rank	English→Estonian	Score	Docs
1-1	Wayfinder	82.9	198
2-4	TRIVE	81.4	198
2-6	Gemini 3.1 Pro	81.1	198
2-7	Human (from scratch)	81.0	198
3-7	Lumen	80.3	198
3-7	GPT 5.5	80.3	198
4-7	UvA-MT	80.1	198
8-9	MakotoAI	78.5	177
8-9	PseudoRef	77.1	172
10-10	DeepSeek V4 Pro	71.8	168
11-13	SCIR-TG-MT	68.1	121
11-12	Dubformer	68.0	124
13-13	SalamandraTA	66.1	115
14-15	GoogleTranslate	62.8	85
14-18	Qingqiu-MT-9B	59.8	96
15-18	SRPOL	58.3	82
15-17	Cohere CAT+	57.4	100
18-18	Command A+	56.6	85
19-19	Tiny Aya Global	53.0	89
20-20	TartuNLP	51.8	67
21-22	GPT OSS 120B	47.2	74
21-23	Gemma 4 - 31B	45.8	64
22-23	Mistral Medium 3.5	44.0	71
24-24	Qwen 3.6 - 27b	39.3	57
25-26	Qwen 3.5 - 9b	24.9	62
25-27	GPT OSS 20B	23.3	59
26-27	Ministral 3 - 14B	21.7	67
28-28	Gemma 4 - 4B	15.7	69
29-29	Tower+ 9B	14.6	50
28-30	GLM 4 - 9B	13.1	59

Rank	English→Japanese	Score	Docs
1-4	Lumen	90.6	191
1-4	Wayfinder	90.6	192
1-4	SCIR-TG-MT	90.4	189
1-4	Gemini 3.1 Pro	90.3	189
5-5	TRIVE	88.3	185
6-11	HW-TSC	86.2	181
6-11	GPT 5.5	86.1	184
6-11	UvA-MT	86.1	182
6-11	Human (postediting)	86.0	187
6-11	Dubformer	85.4	177
6-11	Qingqiu-MT-9B	85.2	182
12-15	MakotoAI	83.4	178
12-16	DeepSeek V4 Pro	83.2	168
12-16	Cohere CAT+	83.2	171
12-19	SRPOL	81.7	178
13-18	Command A+	81.4	177
15-20	RFMT	81.2	185
15-20	Gemma 4 - 31B	80.5	172
17-21	SalamandraTA	80.2	168
16-22	PseudoRef	79.6	171
19-23	GPT OSS 120B	78.2	172
20-23	Tiny Aya Global	77.7	177
21-23	Gemma 4 - 4B	77.3	176
24-25	Mistral Medium 3.5	74.1	177
24-25	GPT OSS 20B	72.4	188
26-29	Tower+ 9B	69.9	174
26-29	GoogleTranslate	69.4	189
26-29	Qwen 3.6 - 27b	68.7	181
26-29	Ministral 3 - 14B	68.6	172
30-30	GLM 4 - 9B	65.0	178
31-31	Qwen 3.5 - 9b	57.4	177

Rank	English→Kazakh	Score	Docs
1-4	UvA-MT	83.1	198
1-5	Lumen	82.2	198
1-5	TRIVE	82.1	198
1-5	GPT 5.5	81.9	198
2-6	Wayfinder	81.3	198
5-7	Gemini 3.1 Pro	80.0	198
6-8	PseudoRef	78.4	181
7-8	Qingqiu-MT-9B	77.8	190
9-10	SRPOL	74.6	133
9-10	DeepSeek V4 Pro	74.3	123
11-13	HW-TSC	71.0	97
11-12	MakotoAI	70.6	117
13-13	Dubformer	69.9	96
14-14	Cohere CAT+	65.5	96
15-15	Gemma 4 - 31B	64.4	102
16-18	Yandex-8B	60.8	69
16-19	GoogleTranslate	58.6	77
16-19	Command A+	57.8	73
17-19	SalamandraTA	56.3	74
20-21	Mistral Medium 3.5	48.8	71
20-21	GPT OSS 120B	48.7	60
22-23	TransCore	45.3	50
22-23	Qwen 3.6 - 27b	43.4	67
24-24	GPT OSS 20B	40.8	55
25-25	Qwen 3.5 - 9b	30.4	67
26-27	Gemma 4 - 4B	25.8	62
26-27	Ministral 3 - 14B	23.5	55
28-28	GLM 4 - 9B	5.9	43
29-29	Tower+ 9B	2.3	56

Rank	English→Indonesian	Score	Docs
1-4	Gemini 3.1 Pro	83.5	198
1-4	TRIVE	83.5	198
1-5	Wayfinder	82.9	190
1-6	Lumen	82.6	198
3-8	UvA-MT	81.8	193
4-8	HW-TSC	81.8	196
5-10	[Anonymous]	80.9	187
5-10	GPT 5.5	80.3	189
7-14	MakotoAI	79.7	192
7-14	SCIR-TG-MT	79.3	192
9-14	DeepSeek V4 Pro	78.9	190
9-14	Dubformer	78.8	188
9-15	TransCore	78.7	190
9-17	Cohere CAT+	78.3	190
13-17	Command A+	76.8	193
14-17	Qingqiu-MT-9B	75.9	190
14-18	Gemma 4 - 31B	75.7	193
17-18	PseudoRef	75.1	192
19-20	GoogleTranslate	70.8	189
19-20	Gemma 4 - 4B	70.8	191
21-23	SRPOL	69.3	195
21-24	Tiny Aya Global	68.5	191
19-24	GPT OSS 120B	68.3	193
22-25	SalamandraTA	67.3	193
24-25	Mistral Medium 3.5	65.0	197
26-26	Qwen 3.6 - 27b	59.8	193
27-28	GPT OSS 20B	56.0	193
27-28	Qwen 3.5 - 9b	54.1	191
29-29	Ministral 3 - 14B	52.4	195
30-30	GLM 4 - 9B	42.5	192
30-31	Tower+ 9B	41.6	193

Rank	Chinese (Simp.)→Japanese	Docs
1-1	Human (postediting)	79.6	190
2-2	Wayfinder	78.1	190
3-6	Gemini 3.1 Pro	75.7	190
3-5	GPT 5.5	75.3	190
3-9	UvA-MT	74.9	190
5-9	TRIVE	74.2	190
5-11	Qingqiu-MT-9B	73.7	190
5-11	Lumen	73.7	190
4-13	Dubformer	73.4	190
7-15	DeepSeek V4 Pro	72.8	190
7-15	PseudoRef	72.2	190
9-15	Gemma 4 - 31B	72.2	190
9-15	Cohere CAT+	72.0	190
10-16	HW-TSC	71.5	190
10-16	TransCore	71.4	190
14-16	MakotoAI	70.2	190
17-19	SRPOL	67.0	190
17-19	Command A+	66.9	190
17-20	RFMT	65.8	190
19-20	GPT OSS 120B	64.9	190
21-21	Gemma 4 - 4B	61.4	190
22-23	SalamandraTA	57.7	190
22-26	Mistral Medium 3.5	56.5	190
23-27	Qwen 3.6 - 27b	55.7	190
23-27	GPT OSS 20B	55.1	190
23-27	Tower+ 9B	55.0	190
24-27	Qwen 3.5 - 9b	53.1	190
28-28	GLM 4 - 9B	50.0	190
29-29	Ministral 3 - 14B	46.8	190
30-30	GoogleTranslate	40.6	190
31-31	Tiny Aya Global	29.6	190

Rank	English→Korean	Score	Docs
1-3	Gemini 3.1 Pro	83.5	105
1-3	Lumen	82.5	95
1-4	TRIVE	82.5	71
3-5	GPT 5.5	80.9	83
4-7	Wayfinder	78.6	62
5-6	UvA-MT	78.5	85
7-10	Gemma 4 - 31B	76.2	44
7-9	Dubformer	75.9	50
7-10	DeepSeek V4 Pro	74.4	58
9-15	Cohere CAT+	73.0	38
10-15	PseudoRef	72.8	44
10-15	MakotoAI	72.4	35
10-15	HW-TSC	71.7	40
9-15	AdaptiveMT	71.5	40
9-15	Qingqiu-MT-9B	69.3	38
16-18	GoogleTranslate	63.5	32
16-19	Command A+	63.0	30
16-18	Gemma 4 - 4B	60.9	38
19-19	Mistral Medium 3.5	58.3	31
20-24	SRPOL	56.4	26
19-23	Qwen 3.6 - 27b	55.8	32
20-24	GPT OSS 120B	54.6	32
20-24	GPT OSS 20B	52.0	30
22-24	SalamandraTA	49.4	26
25-27	Tower+ 9B	48.2	35
23-26	Tiny Aya Global	45.2	31
27-27	Ministral 3 - 14B	39.1	23
28-28	GLM 4 - 9B	26.5	28
29-29	Qwen 3.5 - 9b	18.9	23

Rank	English→Thai	Score	Docs
1-3	Gemini 3.1 Pro	86.2	198
1-4	Wayfinder	85.0	198
1-4	TRIVE	85.0	198
2-5	Lumen	84.0	198
4-6	UvA-MT	83.1	198
5-6	GPT 5.5	82.0	198
7-8	Gemma 4 - 31B	80.0	198
7-8	Cohere CAT+	78.4	197
9-11	PseudoRef	77.0	198
9-11	[Anonymous]	76.6	198
9-12	Qingqiu-MT-9B	75.9	197
11-14	SCIR-TG-MT	75.1	198
12-14	MakotoAI	74.7	198
12-14	Dubformer	74.2	197
15-16	DeepSeek V4 Pro	71.9	198
15-17	GoogleTranslate	70.9	198
16-18	Command A+	70.1	197
17-18	HW-TSC	68.3	198
19-20	Gemma 4 - 4B	62.4	198
19-21	SalamandraTA	61.0	197
20-21	SRPOL	59.3	198
22-23	Mistral Medium 3.5	53.6	198
22-23	TransCore	50.8	198
24-25	GPT OSS 120B	45.5	198
24-25	Tiny Aya Global	44.2	198
26-26	Qwen 3.6 - 27b	42.1	198
27-27	Qwen 3.5 - 9b	37.4	197
28-28	GPT OSS 20B	31.3	198
29-29	Ministral 3 - 14B	25.3	198
30-30	Tower+ 9B	20.4	198
31-31	GLM 4 - 9B	8.3	198

Rank	English→Ligurian	Score	Docs
1-1	Human (from scratch)	84.7	198
2-3	LiLa	66.0	198
2-3	Grammatist	65.9	198
4-4	GoogleTranslate	47.4	198
5-5	GPT 5.5	38.3	198
6-6	MakotoAI	33.8	198
7-8	SalamandraTA	28.6	178
7-9	UvA-MT	28.0	198
8-11	Qingqiu-MT-9B	26.4	198
9-12	Lumen	25.4	197
9-11	Wayfinder	24.5	163
12-12	TRIVE	23.8	164
13-14	Cohere CAT+	21.9	142
13-15	DeepSeek V4 Pro	20.1	135
14-15	Dubformer	18.9	91
16-17	SRPOL	13.8	116
16-17	Gemini 3.1 Pro	12.4	94
18-22	Command A+	7.0	80
18-23	Gemma 4 - 31B	6.6	77
18-23	Mistral Medium 3.5	6.0	83
18-23	Qwen 3.6 - 27b	5.7	100
18-23	Grial-SalamandraTA	5.2	81
19-23	GPT OSS 120B	4.4	69
24-28	Ministral 3 - 14B	2.6	76
24-28	GPT OSS 20B	2.4	57
24-26	GLM 4 - 9B	2.2	72
27-28	Gemma 4 - 4B	1.7	72
24-28	Qwen 3.5 - 9b	1.4	48
27-29	Tower+ 9B	0.7	68

Rank	Czech→Ukrainian	Score	Docs
1-4	TRIVE	86.9	173
1-2	Lumen	86.6	183
3-4	HW-TSC	86.2	109
1-4	GPT 5.5	85.8	138
5-9	Wayfinder	84.5	105
5-7	UvA-MT	84.4	103
5-8	Gemini 3.1 Pro	84.1	88
7-10	MakotoAI	83.4	74
8-10	Dubformer	83.3	68
8-11	DeepSeek V4 Pro	82.2	50
10-17	Gemma 4 - 31B	79.3	68
11-12	Cohere CAT+	79.3	57
13-13	Qingqiu-MT-9B	79.2	31
14-14	SRPOL	78.4	54
15-17	Mistral Medium 3.5	78.1	45
15-17	Command A+	76.9	41
15-17	GoogleTranslate	74.7	44
18-19	SalamandraTA	69.6	47
18-21	CUNI-MH-v3	68.8	39
19-24	Gemma 4 - 4B	63.7	32
19-23	GPT OSS 120B	63.1	48
20-26	Qwen 3.5 - 9b	59.6	35
20-24	Tower+ 9B	59.3	32
22-26	Qwen 3.6 - 27b	59.0	34
24-27	Ministral 3 - 14B	58.0	37
22-26	Tiny Aya Global	57.3	43
27-28	GPT OSS 20B	49.7	34
27-28	GLM 4 - 9B	49.4	38
28-29	CUNI-EdUKate	43.9	34

Rank	English→Ukrainian	Score	Docs
1-1	Gemini 3.1 Pro	84.3	186
2-5	TRIVE	82.7	180
2-6	UvA-MT	82.5	185
2-5	Wayfinder	81.7	186
2-8	Lumen	81.5	191
5-7	Dubformer	81.0	186
4-8	GPT 5.5	80.9	190
7-10	DeepSeek V4 Pro	80.4	187
8-11	MakotoAI	79.5	187
8-13	Cohere CAT+	79.1	183
9-14	Qingqiu-MT-9B	77.9	190
10-14	HW-TSC	77.6	191
10-14	PseudoRef	77.4	194
11-15	Gemma 4 - 31B	76.3	191
14-15	Command A+	75.9	193
16-16	Mistral Medium 3.5	70.3	192
17-18	SalamandraTA	66.1	191
17-18	GoogleTranslate	65.3	187
19-19	SRPOL	62.6	186
20-20	Tiny Aya Global	61.2	181
21-23	GPT OSS 120B	58.1	136
21-23	Gemma 4 - 4B	56.7	112
21-23	Ministral 3 - 14B	55.9	134
24-24	CUNI-MH-v3	51.6	131
25-26	Tower+ 9B	51.0	107
24-26	Qwen 3.5 - 9b	48.3	90
27-29	GLM 4 - 9B	46.9	116
26-29	CUNI-EdUKate	46.7	73
26-29	GPT OSS 20B	44.6	95
26-30	Qwen 3.6 - 27b	44.3	95

Rank	English→Armenian	Score	Docs
1-4	TRIVE	83.4	198
1-4	Lumen	83.4	198
1-4	UvA-MT	82.8	198
1-4	Wayfinder	82.8	198
5-7	PseudoRef	81.0	198
5-7	Gemini 3.1 Pro	80.7	198
5-7	GPT 5.5	80.0	198
8-8	MakotoAI	78.6	110
9-9	Qingqiu-MT-9B	72.6	144
10-10	SCIR-TG-MT	68.1	116
11-13	SRPOL	67.4	99
10-13	DeepSeek V4 Pro	67.3	131
10-13	Dubformer	65.0	88
14-16	SalamandraTA	52.9	60
14-15	GoogleTranslate	52.3	85
16-16	Yandex-8B	49.7	75
17-18	Mistral Medium 3.5	46.3	70
17-18	Gemma 4 - 31B	44.0	72
19-20	Cohere CAT+	37.2	73
19-20	Ministral 3 - 14B	32.7	65
21-22	GPT OSS 120B	32.1	61
20-22	Gemma 4 - 4B	30.1	60
23-23	GPT OSS 20B	29.0	57
24-25	Qwen 3.6 - 27b	21.6	55
24-25	Qwen 3.5 - 9b	18.8	61
26-26	Command A+	12.5	52
27-27	GLM 4 - 9B	3.4	48
27-28	Tower+ 9B	3.3	54

Rank	English→Chinese (Simp.)	Docs
1-1	UvA-MT	79.5	50
2-3	Gemini 3.1 Pro	79.3	51
2-9	Lumen	78.6	35
3-9	HW-TSC	78.4	43
2-9	TRIVE	76.8	38
2-12	Dubformer	75.8	33
3-11	Qingqiu-MT-9B	75.6	33
3-12	DeepSeek V4 Pro	75.3	32
3-12	Wayfinder	75.2	31
6-12	Cohere CAT+	74.7	25
6-12	GPT 5.5	73.3	32
8-13	Gemma 4 - 31B	72.3	28
12-18	GoogleTranslate	69.2	24
13-18	SRPOL	68.9	22
13-18	Mistral Medium 3.5	67.6	26
13-21	PseudoRef	66.0	25
13-21	Command A+	65.8	26
13-18	MakotoAI	65.4	27
19-23	Gemma 4 - 4B	64.4	26
13-25	Qwen 3.6 - 27b	63.8	22
14-23	GPT OSS 120B	61.9	24
19-28	SCIR-TG-MT	58.7	20
19-25	SalamandraTA	58.0	26
22-26	GLM 4 - 9B	57.4	22
22-28	Qwen 3.5 - 9b	55.4	22
24-28	GPT OSS 20B	51.7	25
25-28	Tiny Aya Global	50.9	24
25-28	Tower+ 9B	50.5	22
29-29	Ministral 3 - 14B	45.1	23

6.3 Ranking of Constrained Systems for each translation direction

In this section, we present results for only the constrained systems against one another as systems from constrained track do not compete against unconstrained systems.

A short gray line marks whenever a model is significantly better than the subsequent model. A longer black line marks whenever all previous models are significantly better than all subsequent models. Statistical testing with ($\alpha = 0.05$) is done using a one-sided paired or independent bootstrap test on the mean differences (100 resamples).

Rank	Czech→Vietnamese	Score	Docs
1-1	Qingqiu-MT-9B	89.3	17
2-2	Lumen	85.9	17
3-4	Gemma 4 - 4B	74.7	17
3-6	Tiny Aya Global	71.3	18
4-8	SalamandraTA	67.8	16
3-8	Qwen 3.5 - 9b	66.3	13
5-8	Ministral 3 - 14B	63.7	19
5-9	GPT OSS 20B	60.6	20
8-10	CUNI-MH-v3	54.1	17
9-10	GLM 4 - 9B	53.3	20
10-11	Tower+ 9B	41.8	19

Rank	Czech→German	Score	Docs
1-2	Qingqiu-MT-9B	81.2	114
1-2	Lumen	80.2	97
3-3	SalamandraTA	65.0	53
4-4	Tower+ 9B	58.8	41
5-5	Tiny Aya Global	54.9	38
6-7	Gemma 4 - 4B	54.1	51
4-8	CUNI-MH-v3	53.9	36
7-9	Qwen 3.5 - 9b	53.7	45
8-10	Ministral 3 - 14B	46.7	50
9-10	GPT OSS 20B	45.4	55
9-11	GLM 4 - 9B	45.1	48

Rank	English→Czech	Score	Docs
1-2	Tiny Aya Global	71.8	34
1-3	Qingqiu-MT-9B	68.2	30
2-3	Lumen	65.5	41
4-4	SalamandraTA	54.4	30
5-5	Tower+ 9B	41.8	30
6-9	GLM 4 - 9B	40.7	25
6-10	GPT OSS 20B	37.7	26
5-9	Ministral 3 - 14B	34.4	34
6-10	CUNI-MH-v3	33.1	35
9-10	Gemma 4 - 4B	32.4	28
9-11	Qwen 3.5 - 9b	32.4	24

Rank	English→German	Score	Docs
1-1	Lumen	72.0	190
2-2	Qingqiu-MT-9B	68.1	83
3-3	SalamandraTA	55.0	54
4-7	Gemma 4 - 4B	49.6	48
4-7	CUNI-MH-v3	48.4	53
4-7	Tower+ 9B	47.5	58
4-8	Tiny Aya Global	46.7	38
7-9	Ministral 3 - 14B	41.2	41
8-9	GPT OSS 20B	37.5	46
10-10	GLM 4 - 9B	35.2	39
10-11	Qwen 3.5 - 9b	30.9	44

Rank	English→Icelandic	Score	Docs
1-1	Lumen	84.5	198
2-2	SalamandraTA	69.0	124
3-3	Tower+ 9B	60.1	71
4-4	Qingqiu-MT-9B	53.5	69
5-5	AMI	39.0	63
6-7	Tiny Aya Global	24.5	50
6-7	GPT OSS 20B	24.1	53
8-10	Ministral 3 - 14B	19.2	52
8-11	Qwen 3.5 - 9b	17.4	50
8-11	Gemma 4 - 4B	16.2	47
9-11	Grial-SalamandraTA	15.6	48
12-12	GLM 4 - 9B	7.9	46

Rank	English→Chinese (Trad.)	Docs
1-1	Lumen	74.0	41
2-2	Qingqiu-MT-9B	66.2	28
3-4	GPT OSS 20B	53.3	29
3-8	Gemma 4 - 4B	53.0	26
4-8	Qwen 3.5 - 9b	51.3	24
4-8	Tower+ 9B	49.1	27
4-8	GLM 4 - 9B	47.9	23
4-8	SalamandraTA	46.2	27
9-9	Tiny Aya Global	39.0	23
9-10	Ministral 3 - 14B	38.2	27

Rank	English→Egyptian Arabic	Docs
1-1	Lumen	83.1	176
2-2	Qingqiu-MT-9B	81.7	190
3-3	Gemma 4 - 4B	78.5	187
4-4	SalamandraTA	76.8	191
5-5	Tiny Aya Global	74.7	191
6-6	GPT OSS 20B	69.9	162
7-7	GLM 4 - 9B	65.5	180
8-8	Ministral 3 - 14B	58.1	155
9-9	Qwen 3.5 - 9b	54.3	173
10-10	Tower+ 9B	36.1	167

Rank	English→Russian	Score	Docs
1-1	Lumen	79.8	198
2-3	Qingqiu-MT-9B	74.9	190
2-3	Yandex-8B	73.0	193
4-4	Gemma 4 - 4B	59.9	116
5-5	SalamandraTA	57.9	125
6-6	Tiny Aya Global	53.5	98
7-7	GPT OSS 20B	49.5	106
8-10	Ministral 3 - 14B	44.8	67
8-10	Qwen 3.5 - 9b	44.4	92
8-10	Tower+ 9B	43.1	76
11-11	GLM 4 - 9B	32.2	70

Rank	English→Belarusian	Score	Docs
1-1	Lumen	75.1	198
2-2	Qingqiu-MT-9B	56.5	78
3-3	Krosny	53.7	60
4-4	Yandex-8B	48.4	56
5-5	ReMova	44.2	60
6-6	SalamandraTA	44.1	61
7-7	Tiny Aya Global	26.9	53
8-9	Ministral 3 - 14B	25.4	49
8-9	GPT OSS 20B	21.6	43
10-10	Gemma 4 - 4B	18.3	37
11-11	Qwen 3.5 - 9b	16.5	41
12-12	GLM 4 - 9B	10.4	47
13-13	Tower+ 9B	5.6	47

Rank	English→Ladin	Score	Docs
1-1	LiLa	48.9	177
2-2	Lumen	34.8	110
3-3	Qingqiu-MT-9B	31.3	87
4-4	Grial-SalamandraTA	4.7	59
5-5	GLM 4 - 9B	2.2	49
6-7	SalamandraTA	1.6	50
5-7	Tower+ 9B	1.2	49
8-10	Qwen 3.5 - 9b	1.1	49
7-10	GPT OSS 20B	1.1	39
5-10	Ministral 3 - 14B	0.9	43
10-11	Gemma 4 - 4B	0.3	40

Rank	English→Northern Sami	Docs
1-1	Lumen	54.5	38
2-2	SalamandraTA	33.4	26
3-3	TartuNLP	22.3	30
4-6	Qingqiu-MT-9B	1.7	23
4-6	GLM 4 - 9B	1.5	21
4-7	GPT OSS 20B	0.8	26
6-8	Qwen 3.5 - 9b	0.6	24
7-8	Ministral 3 - 14B	0.4	22
9-9	Gemma 4 - 4B	0.2	23
9-10	Tower+ 9B	0.1	25

Rank	English→Estonian	Score	Docs
1-1	Lumen	80.3	198
2-2	SalamandraTA	66.1	115
3-3	Qingqiu-MT-9B	59.8	96
4-4	Tiny Aya Global	53.0	89
5-5	TartuNLP	51.8	67
6-7	Qwen 3.5 - 9b	24.9	62
6-8	GPT OSS 20B	23.3	59
7-8	Ministral 3 - 14B	21.7	67
9-9	Gemma 4 - 4B	15.7	69
10-10	Tower+ 9B	14.6	50
10-11	GLM 4 - 9B	13.1	59

Rank	English→Japanese	Score	Docs
1-1	Lumen	90.6	191
2-2	Qingqiu-MT-9B	85.2	182
3-4	RFMT	81.2	185
3-4	SalamandraTA	80.2	168
5-6	Tiny Aya Global	77.7	177
5-6	Gemma 4 - 4B	77.3	176
7-7	GPT OSS 20B	72.4	188
8-9	Tower+ 9B	69.9	174
8-9	Ministral 3 - 14B	68.6	172
10-10	GLM 4 - 9B	65.0	178
11-11	Qwen 3.5 - 9b	57.4	177

Rank	English→Kazakh	Score	Docs
1-1	Lumen	82.2	198
2-2	Qingqiu-MT-9B	77.8	190
3-3	Yandex-8B	60.8	69
4-4	SalamandraTA	56.3	74
5-5	TransCore	45.3	50
6-6	GPT OSS 20B	40.8	55
7-7	Qwen 3.5 - 9b	30.4	67
8-9	Gemma 4 - 4B	25.8	62
8-9	Ministral 3 - 14B	23.5	55
10-10	GLM 4 - 9B	5.9	43
11-11	Tower+ 9B	2.3	56

Rank	English→Indonesian	Score	Docs
1-1	Lumen	82.6	198
2-2	TransCore	78.7	190
3-3	Qingqiu-MT-9B	75.9	190
4-4	Gemma 4 - 4B	70.8	191
5-6	Tiny Aya Global	68.5	191
5-6	SalamandraTA	67.3	193
7-8	GPT OSS 20B	56.0	193
7-9	Qwen 3.5 - 9b	54.1	191
8-9	Ministral 3 - 14B	52.4	195
10-10	GLM 4 - 9B	42.5	192
10-11	Tower+ 9B	41.6	193

Rank	Chinese (Simp.)→Japanese	Docs
1-2	Qingqiu-MT-9B	73.7	190
1-2	Lumen	73.7	190
3-3	TransCore	71.4	190
4-4	RFMT	65.8	190
5-5	Gemma 4 - 4B	61.4	190
6-6	SalamandraTA	57.7	190
7-9	GPT OSS 20B	55.1	190
7-9	Tower+ 9B	55.0	190
7-9	Qwen 3.5 - 9b	53.1	190
10-10	GLM 4 - 9B	50.0	190
11-11	Ministral 3 - 14B	46.8	190
12-12	Tiny Aya Global	29.6	190

Rank	English→Korean	Score	Docs
1-1	Lumen	82.5	95
2-2	Qingqiu-MT-9B	69.3	38
3-3	Gemma 4 - 4B	60.9	38
4-5	GPT OSS 20B	52.0	30
4-5	SalamandraTA	49.4	26
6-8	Tower+ 9B	48.2	35
4-7	Tiny Aya Global	45.2	31
8-8	Ministral 3 - 14B	39.1	23
9-9	GLM 4 - 9B	26.5	28
10-10	Qwen 3.5 - 9b	18.9	23

Rank	English→Thai	Score	Docs
1-1	Lumen	84.0	198
2-2	Qingqiu-MT-9B	75.9	197
3-4	Gemma 4 - 4B	62.4	198
3-4	SalamandraTA	61.0	197
5-5	TransCore	50.8	198
6-6	Tiny Aya Global	44.2	198
7-7	Qwen 3.5 - 9b	37.4	197
8-8	GPT OSS 20B	31.3	198
9-9	Ministral 3 - 14B	25.3	198
10-10	Tower+ 9B	20.4	198
11-11	GLM 4 - 9B	8.3	198

Rank	English→Ligurian	Score	Docs
1-1	LiLa	66.0	198
2-2	SalamandraTA	28.6	178
3-4	Qingqiu-MT-9B	26.4	198
3-4	Lumen	25.4	197
5-5	Grial-SalamandraTA	5.2	81
6-6	Ministral 3 - 14B	2.6	76
7-10	GPT OSS 20B	2.4	57
6-8	GLM 4 - 9B	2.2	72
9-10	Gemma 4 - 4B	1.7	72
7-10	Qwen 3.5 - 9b	1.4	48
9-11	Tower+ 9B	0.7	68

Rank	Czech→Ukrainian	Score	Docs
1-1	Lumen	86.6	183
2-2	Qingqiu-MT-9B	79.2	31
3-4	SalamandraTA	69.6	47
3-4	CUNI-MH-v3	68.8	39
5-7	Gemma 4 - 4B	63.7	32
5-9	Qwen 3.5 - 9b	59.6	35
5-7	Tower+ 9B	59.3	32
8-9	Ministral 3 - 14B	58.0	37
6-9	Tiny Aya Global	57.3	43
10-10	GPT OSS 20B	49.7	34
10-11	GLM 4 - 9B	49.4	38

Rank	English→Ukrainian	Score	Docs
1-1	Lumen	81.5	191
2-2	Qingqiu-MT-9B	77.9	190
3-3	SalamandraTA	66.1	191
4-4	Tiny Aya Global	61.2	181
5-5	Gemma 4 - 4B	56.7	112
6-6	Ministral 3 - 14B	55.9	134
7-7	CUNI-MH-v3	51.6	131
8-10	Tower+ 9B	51.0	107
8-9	Qwen 3.5 - 9b	48.3	90
10-10	GLM 4 - 9B	46.9	116
9-11	GPT OSS 20B	44.6	95

Rank	English→Armenian	Score	Docs
1-1	Lumen	83.4	198
2-2	Qingqiu-MT-9B	72.6	144
3-4	SalamandraTA	52.9	60
3-4	Yandex-8B	49.7	75
5-6	Ministral 3 - 14B	32.7	65
5-6	Gemma 4 - 4B	30.1	60
7-7	GPT OSS 20B	29.0	57
8-8	Qwen 3.5 - 9b	18.8	61
9-9	GLM 4 - 9B	3.4	48
9-10	Tower+ 9B	3.3	54

Rank	English→Chinese (Simp.)	Docs
1-2	Lumen	78.6	35
1-2	Qingqiu-MT-9B	75.6	33
3-4	Gemma 4 - 4B	64.4	26
3-7	SalamandraTA	58.0	26
4-6	GLM 4 - 9B	57.4	22
4-9	Qwen 3.5 - 9b	55.4	22
6-9	GPT OSS 20B	51.7	25
6-9	Tiny Aya Global	50.9	24
6-9	Tower+ 9B	50.5	22
7-10	Ministral 3 - 14B	45.1	23

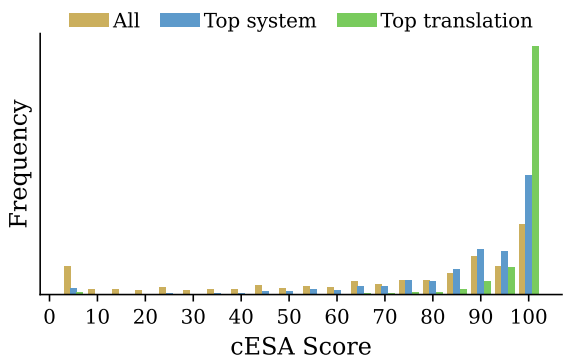


Figure 4: Distribution of final human segment-level scores (cESA with step of 5) of all systems, top-system in each language, and top translation for each segment. Results include all languages.

	Factcheck	Edu	Software	News	Speech	Social	Avg.
En.→Japanese			96.7	93.7	88.2	87.2	91.4
Cz.→Vietnamese	98.0	96.8		92.5	88.9	92.2	91.2
En.→Korean			97.3	88.5	80.8	84.6	87.8
Cz.→Ukrainian	90.3	87.9		88.3	83.0	90.7	87.4
En.→Ukrainian			93.2	87.3	78.4	87.0	86.4
En.→Icelandic			93.0	92.1	84.1	76.0	86.3
En.→Thai			94.9	91.1	81.1	77.8	86.2
En.→Armenian			94.2	87.2	84.5	78.0	86.0
En.→Indonesian			95.3	90.9	80.5	74.6	85.3
En.→Czech			100.0	90.9	80.8	68.5	85.1
En.→Kazakh			91.2	86.5	82.8	79.0	84.9
En.→Arabic (Egy.)			89.4	85.4	85.0	77.7	84.4
En.→Chinese (Simp.)			91.5	81.3	81.2	82.0	84.0
Cz.→German	90.2	90.2		85.4	78.3	87.7	83.8
En.→Estonian			91.6	87.7	79.7	72.9	83.0
En.→Russian			93.1	84.6	74.9	78.1	82.7
En.→Chinese (Trad.)			98.2	88.4	75.9	66.9	82.4
Zh.→Japanese				92.1	82.6	67.3	80.7
En.→Belarusian			90.4	86.4	69.3	72.4	79.6
En.→German			96.2	87.0	67.6	67.3	79.5
En.→Ligurian			76.5	76.3	67.2	54.4	68.6
En.→Northern Sami			85.5	64.3	67.7	43.1	65.2
En.→Ladin			72.2	47.6	51.7	46.2	54.4

Table 10: Human evaluation average scores for the top-performing system per language, by domain. Social and speech domains show most headroom for improvement.

6.4 Discussion

System ranks and scores. The tables in Section 6.2 show that only some systems achieved average scores in the "Very good" range (85–100, representing near-perfect transfer), and only for a small number of language pairs. However, the maximum scores vary notably across the language pairs, from 61 (English→Northern Sami) to 90 (English→Japanese and Czech→Vietnamese). The lowest scores also vary across the languages, from 0.1 for English→Northern Sami to 57.4 for English→Japanese.

Among the constrained systems in Section 6.3, Lumen emerges as the clear winner, ranking in the top cluster for 20 of the 23 language pairs. It is followed by Qingqiu-MT-9B, which ranks second place.

Wayfinder was the best system overall when unconstrained systems are included (only skipping human reference), appearing in the top cluster for 17 of 23 language pairs, followed by Gemini 3.1 Pro (15 times), TRIVE (14 times), and the constrained system Lumen (13 times). It is especially encouraging that Lumen, a small constrained system, reached the top cluster while competing against unconstrained systems.

Three systems targeting only one or two translation directions reached the top-performing participant cluster in every direction they entered: Grammatist and LiLa for English→Ladin and English→Ligurian, and CUNI-AR for English→Egyptian Arabic. This highlights the potential of language-specific specialization, although other targeted submissions did not show the same pattern.

Finally, human translations, often treated as a "gold standard," place in the top cluster for only eight language pairs out of twelve. This finding highlights the inherent difficulty of translation, though it could also reflect the stylistic or lexical preferences of the annotators.

Domain difficulty. In Table 10, we present the scores of the top-performing system for each translation direction and each domain. Focusing on the top-performing system mitigates the effect of low-performing outliers. While absolute scores are not directly comparable across language pairs (due to different sets of systems, annotators, and sources), certain patterns can be observed: the social domain received the lowest scores (between 43 and 92, over 90 only for two language pairs and 5 between 80 and 90), followed by the speech domain (between 51 and 88, over 80 for 13 language pairs). News received higher scores (between 47.6 and 93.7, seven language pairs over 90 and thirteen between 80 and 90). The software domain, on the other hand, received the highest scores (between 72 and 100, with 15 language pairs over 90), possibly due to the specific vocabulary and consistent style.

Results for low-resource languages vary considerably. For some low-resource target languages,

systems approach or exceed human baselines, while Ladin and Ligurian show much larger gaps: human translation from scratch scores 77.2 and 84.7 ESA, respectively, compared with 49.4 and 66.0 for the best systems. Qualitative inspection of these two languages suggests challenges beyond data scarcity alone.

For Ladin, annotators reported that systems mix forms from different contemporary varieties rather than consistently producing the requested Val Badia variety. One possible contributing factor is the loss of variety metadata when datasets are pooled under ISO 639-3 language codes, collapsing distinctions present in the original resources. For Ligurian, annotators reported that several systems produced unnatural or archaic forms. We hypothesize that this reflects exposure to transcriptions of medieval manuscripts on Ligurian Wikisource, and learner-written or template-generated content from Ligurian Wikipedia, where linguistic quality issues have been previously documented (Lusito and Maillard, 2021). These two cases suggest that low-resource translation is not simply a problem of insufficient data: careful curation – particularly with respect to quality and linguistic variation – can also play an important role.

7 Test Suites Task: "Mind the Gap - Exposing LLM Translation Blind Spots"

This is the eighth consecutive year that test suites have been part of WMT's General MT task and the third year that they have focused on translations produced by LLMs. Automated or human-curated (Zouhar et al., 2026b) challenge sets are increasingly needed to expose critical blind spots that cannot be detected by the generic evaluation setup, and help answer what the nature of the failures are, beyond mere aggregate scores.

7.1 Setup of the Test Suites Sub-task

Test suites extend the standard test sets to target specific aspects of machine translation (MT) output. As in previous editions (Bojar et al., 2018), evaluation was carried out in a decentralised manner. Test suite providers were invited to submit their customised test sets that are integrated into the standard test sets by the organisers of the General MT Shared Task. After system outputs are generated by the system participants of the main General MT task, they are returned to the providers for evaluation according to their respective methodologies. Detailed findings are presented in the individual test suite description papers, and a consolidated overview is provided below.

7.2 Submissions

Nine participants took part in the test suite task this year (Table 11), covering several language pairs, translation phenomena and domains. Descriptions of each submission, along with their main findings, can be found below.

AlphaMWE (Han et al., 2026). LLMs' performance on machine translation (MT) tasks is often dependent on the data availability in the specific domains and language pairs that they are trained on. To examine whether multiword expressions (MWEs) still present a bottleneck for LLMs regarding language understanding and translation, the participants measure the performance of 31 WMT2026 MT systems using portions of the publicly available multilingual parallel corpus AlphaMWE as test suites. The analysis focuses on English to Chinese (zh), Polish (pl), German (de), and Arabic (ar), including Modern Standard Arabic (MSA) and two dialectal varieties, Egyptian and Tunisian Arabic. The participants carried out automatic evaluations using BLEU, chrF++, and BERTScore to select the top three systems per language pair, followed by human evaluation of selected systems.

The findings show that figurative and MWE-related phenomena remain challenging, automatic metrics sometimes disagree, and human evaluation uncovers language-specific errors hidden by aggregate scores.

ASC: AMI Syntactic Constructions (Sigurðsson et al., 2026). The English→Icelandic syntactic divergence test suite is intended to study how well systems do when translating syntactic constructions found in English that do not have a direct equivalent in Icelandic. Manual evaluation of the results reveals a large gap between the best-performing models and the lower-scoring ones. The results also demonstrate that while the best-performing models achieve high accuracy in various categories of syntactic constructions, they still struggle with others.

CoST: Complex Structures Test (Yadav et al., 2026). LLMs have advanced machine translation substantially, but generic benchmarks often

Test suite	Focus	Language pairs	Segments	Tokens
ASC	syntactic constructions	en→is	690	6,002
AlphaMWE	multi-word expressions	en→{aeb, ar, arz, de, pl, zh}	2,541	54,914
CoST	complex structures	en→hi	1,947	172,619
EURAC	legal terminology	it→de	3,833	104,563
GENDER1PERSON	author gender form	en→{cs, hr, ru}	3,000	569,754
KUEST	game localization	en→ko	1,796	20,842
LLM-TR	hallucination, over-generation	en→{cs, de, es, et, ru}	4,585	67,975
Lost-in-Morphology	morphological phenomena	ko→en	3,536	37,596
TR-Gender-Bias	occupational gender bias	tr→en	200	2,234

Table 11: Overview of the participating test suites.

fail to expose where these systems struggle with complex linguistic structures and stylistic nuance. CoST, a challenge suite of 1,947 English→Hindi sentences spanning eight genres—autobiography, conversation, legal, mixed, narration, play, poetry, and technical writing—previously introduced for WMT24, is used to evaluate 22 systems submitted to the WMT26 General Translation Shared Task for English→Hindi with reference-free, reference-based, and manual evaluation.

The results reveal a consistent genre-level pattern: systems perform well on narrative text such as autobiography and play but degrade sharply on poetry, legal, and conversational text. The authors further identify a distinct failure mode in poetry translation, where systems reproduce memorised canonical Hindi verse instead of translating the given English text, and manual analysis surfaces recurring issues in named-entity handling, wrong-language output, and lexical choice, several of which persist from their prior evaluation in WMT24.

EURAC: Legal Terminology Challenges in German Language Varieties (Di Natale et al., 2026). Language varieties and dialects remain an open challenge for machine translation (MT). Moreover, existing benchmarks rarely combine the evaluation of specific varieties with domain-specific knowledge. EURAC introduces a termbase-derived MT benchmark for legal terminology across four German varieties: Germany, Austria, Switzerland, and South Tyrol (a low-resource German variety spoken in Italy). In the *inter-variety* setting, models translate Italian source sentences by selecting the correct South Tyrolean German term among semantically equivalent terms from other German varieties. In the *intra-variety* settings, they test sense disambiguation and ontological relation capabilities using homographic and ontologically related distractors for all four language varieties.

The authors find that distinguishing language varieties remains difficult. High performance on low-resourced South Tyrolean terminology is partly driven by indirect exposure to terms shared with better-resourced varieties; when this overlap is removed, even the best-performing models collapse toward small models. Additionally, there is mild evidence of dominant-variety interference, although stronger systems more often err toward the geographically and culturally closer Austrian variety. Generally, locale-oriented adaptation appears more challenging than linguistic and semantic reasoning, where the ontological relation task is largely saturated and no generalized bias is shown against a specific variety. The participants conclude that model scaling mainly reinforces dominant language varieties and that LLM-based MT shifts part of the performance bottleneck from individual term rarity to the underrepresentation of domain-variety intersections across training and alignment stages.

GENDER1PERSON (Popović and Lapshinova-Koltunski, 2025). This test suite is used for measuring the ability of systems to use an inclusive gender form when translating from English into three Slavic languages: Czech, Russian, and Croatian. The test suite consists of 1 000 Amazon product reviews, uniformly distributed over 10 different product categories. The few-shot prompt was formulated to ask for inclusive forms consisting of both masculine and feminine variants separated by the symbol "/". For each system, a score is calculated as a ratio of words in the translation containing the "/" symbol and the number of the first-person pronoun "I" in the original English text.

The results show that for each language pair, there is a group of systems with a clear tendency to generate inclusive forms: Gemini 3.1 Pro, DeepSeek v4 Pro, TRIVE, and GPT OSS 120B. On the other hand, Google Translate and SRPOL never generate such forms, while SalamandraTA and CommandA+ only rarely.

KUEST: Korean Game Usage Evaluation Suite for LLM Translation (Choi et al., 2026). Despite the growing demand for multilingual game localization, systematic evaluation frameworks for game translation constraints remain largely unexplored. As part of the WMT26 Test Suite Track, the authors propose KUEST, a test suite comprising 1,796 English→Korean segments across four core categories, and evaluate 27 submitted systems.

Experimental results demonstrate that standard metrics obscure true model capabilities (e.g., the overall 1st-place system dropping to 15th in creative translation). Furthermore, we reveal residual capability fragmentation—category-specific variance that survives the shared general-competence factor and inverts top-tier rankings—and observe a "prompt gap" where context metadata degrades performance. As such, KUEST serves not only as a test suite but also as an evaluation framework that multi-dimensionally dissects domain-specific translation blind spots hidden behind single composite scores. The dataset is available at huggingface.co/datasets/JudyChoi/KUEST.

LLM Translation Challenge: "I'm Sorry, I Can't Translate That" (Vasileva, 2026). LLMs have significantly raised the bar for translation quality in recent years. However, they are still prone to novel failure modes, including hallucinations and overgenerations, where the generated translation is partially or fully detached from the source text and contains information that is not present in the input. To facilitate the analysis of these behaviors and assess the robustness of LLM-based translation systems, the participants introduce a challenge set specifically designed to probe scenarios that they have observed to be particularly prone to overgeneration, hallucination, and other forms of semantic detachment from the source.

The test suite results show that general-purpose LLMs are far more prone to failures than dedicated MT systems: many baseline LLMs have more than 50% of their segments flagged (and in some cases more than 60%), while participant submissions mostly remain below 5%, likely due to added QE/filtering stages in their pipelines. A few general-purpose models (Mistral Medium 3.5, Gemma 4-4B/31B, GPT-5.5) are an exception to this trend with sub-10% rates, showing robustness isn't strictly tied to model type. Failures cluster around specific phenomena, including sensitive content (safety mechanisms preventing models from translating and resulting in refusal to translate), truncated inputs (incomplete inputs result in overgeneration, where outputs contain content not present in the source), and instructions (misread as commands to follow rather than as input to translate). Translation-specific systems submitted by participants substantially outperform general-purpose LLMs on DNT (Do Not Translate), punctuation, and emoji challenge categories, but truncated inputs and inputs that contain instructions remain hard for both system types.

Lost in Morphology (Lee et al., 2026). The participants present a 3,536-item Korean→English challenge suite spanning nine phenomena and evaluated on 23 systems submitted to WMT26. The suite targets ellipsis punctuation, stacked particles, wordplay, topic/subject marking, evidentiality, scrambling, auxiliary chains, connective endings, and ideophones. They match each phenomenon to its observable evidence: deterministic rules for visible or lexically constrained contrasts, structured LLM observations for semantic features, joint scoring for minimal pairs, distributional metrics for multi-variant frames, and a strategy-aware panel for wordplay. Learned judges and classifiers must pass subset-specific admission tests before their outputs contribute to a system score.

Results reveal failures hidden by aggregate MT quality. Ellipsis is usually preserved, but one model family deletes the marked utterance in most of its failures. Particle fidelity falls from 0.95 at stack depth 1 to 0.70 at depth 4, while distinct scalar particles collapse into bare English even. Evidential and auxiliary meanings are often neutralised although the core proposition survives, and wordplay remains the hardest subset: the best system scores 0.385 against 0.546 for the curated references. Judge validation is itself diagnostic: the only conflict-free wordplay judge fails through ceiling saturation, whereas three judges that are also evaluated systems pass without detectable family preference. The suite therefore measures both translation failures and the reliability limits of the instruments used to identify them.

TR-Gender-Bias: Occupational Gender Bias and Male Default in Turkish (Öztop, 2026) Turkish has no grammatical gender: its third-person pronoun "o" refers to a man, a woman, or an inanimate object alike, so a system translating into English must supply a gender the source never

states. The participants introduce tr_gender_bias, a reference-free Turkish→English test suite of 200 templated items, each built around a single human referent, and use it to measure how 24 machine translation systems and large language models assign gender when the source withholds it. The suite has three parts: ambiguous items whose only signal is an occupational stereotype grounded in labor-force statistics, items whose gender is fixed by a kinship term or a name against that stereotype, and neutral controls. Every output is scored under a six-way, abstention-aware taxonomy that treats neutralization, hedging, and malformed output as distinct outcomes.

Under ambiguity, stereotype-driven gender assignment is universal, significant for all 22 testable systems after correction, and uniformly male-skewed, with no system ever favoring the feminine. Gender that the context states, by contrast, is now resolved almost perfectly, a marked change from earlier evaluations in which stereotype overrode context. The systems split into committers and neutralizers, and neutralizing with singular they removes both stereotype bias and male skew at no cost in accuracy. Gender bias in current translation is therefore less a failure to use context than an unsettled policy for what to do when the source leaves gender genuinely open.

7.3 Aggregated test suite results

To facilitate a cross-test-suite comparison of system performance, we juxtapose the rankings obtained from all available test suites for each language direction. Table 12 summarizes the results for language directions covered by at least two test suites. The systems are arranged to highlight overall agreement and disagreement across evaluations. For reference, the human-based rankings from the official General MT test set are included in the first column. The primary ordering criterion is the human ranking on the General MT test set, while ties are resolved using the mean ranking across the remaining test suites. Table 13 presents an aggregate ranking across all test suites, computed as the weighted average of ranks across test suites and language pairs, with weights proportional to the respective number of test items.

These orderings are intended to improve readability and support visual inspection of ranking patterns. It must be noted that they originate from heterogeneous evaluation settings, including different test collections, varying dataset sizes, and distinct ranking methodologies. Uncertainty estimates are not consistently available or directly compatible across evaluation methods, and therefore confidence intervals have not been incorporated into the analysis.

Overall, the specialized test suites show a positive but clearly imperfect relationship with the WMT26 human evaluation. Systems that perform well in the human evaluation tend to perform better in the test suites, but the test suites do not follow the human ranking closely. English→Russian shows the strongest and most consistent alignment. Both GENDER1PERSON and LLM-TR reach approximately $\rho = 0.586$. English→German shows moderate alignment as AlphaMWE reaches $\rho = 0.519$, while LLM-TR reaches $\rho = 0.480$. English→Czech is less consistent as LLM-TR retains a moderate correlation of $\rho = 0.403$, but GENDER1PERSON falls to only $\rho = 0.268$. Pooling the language-specific LLM-TR observations after normalizing the ranks within each language gives an overall correlation of approximately $\rho = 0.49$ across 80 system-language observations. Thus, LLM-TR captures a meaningful cross-language quality signal, but around half of the ordering variation remains unrelated to the WMT26 human ordering. This is also consistent with the intended role of test suites.

There are few striking differences between the human evaluation ranks and the test suite scores. Gemini 3.1 Pro is highly ranked by humans but performs poorly when scored by LLM-TR in all three language pairs and AlphaMWE in English→German. This suggests that strong overall human-rated output does not necessarily imply resistance to the hallucination and over-generation phenomena. It is also noteworthy that Tiny Aya Global falls from human rank 7 to ranks 21 and 26 on two English→Czech test suites; GPT 5.5 has a consistently worse LLM-TR ranking than its human ranking across all three languages (most notably, human rank 1 versus LLM-TR rank 14 in English→German); and Dubformer shows a specific gender-handling weakness in English→Russian, ranking 1st in human evaluation and 2nd in LLM-TR but 14th in GENDER1PERSON.

Focusing on the reverse pattern, for English→Russian, SCIR-TG-MT is ranked 30th by the humans but tied at 4th on LLM-TR, a

English→Czech	WMT26 human	gender1person	LLM-TR
Wayfinder	1	7	3
Cohere CAT+	1	17	9
MakotoAI	1	–	–
GPT 5.5	4	12	14
TRIVE	5	6	5
Gemini 3.1 Pro	5	4	25
Human (postediting)	6	–	–
Dubformer	7	11	1
UvA-MT	7	15	12
Tiny Aya Global	7	21	26
[Anonymous]	7	–	–
Lumen	11	10	2
Qingqiu-MT-9B	12	8	3
Mistral Medium 3.5	12	2	13
HW-TSC	15	9	7
SRPOL	15	27	8
Command A+	15	23	18
Qwen 3.6 - 27b	16	18	23
DeepSeek V4 Pro	18	5	17
GoogleTranslate	18	25	15
Gemma 4 - 31B	19	13	10
SalamandraTA	19	24	10
GPT OSS 120B	21	3	27
Tower+ 9B	24	20	22
Ministral 3 - 14B	25	1	21
CUNI-MH-v3	25	26	6
GLM 4 - 9B	25	22	19
GPT OSS 20B	25	19	24
Qwen 3.5 - 9b	28	16	20
Gemma 4 - 4B	29	14	16

English→Russian	WMT26 human	gender1person	LLM-TR
TRIVE	1	3	4
Lumen	1	8	7
Dubformer	1	14	2
[Anonymous]	1	15	1
Wayfinder	4	6	3
Gemini 3.1 Pro	4	1	21
UvA-MT	6	11	16
GPT 5.5	7	9	13
Qingqiu-MT-9B	8	5	4
Yandex	8	–	–
Cohere CAT+	9	12	8
Gemma 4 - 31B	10	10	9
Yandex-8B	11	–	–
MakotoAI	13	–	–
Mistral Medium 3.5	14	7	14
Command A+	14	25	17
PseudoRef	14	–	–
DeepSeek V4 Pro	18	2	18
GoogleTranslate	18	26	12
SRPOL	20	27	10
Gemma 4 - 4B	21	13	11
GPT OSS 120B	21	4	26
SalamandraTA	22	23	15
Qwen 3.6 - 27b	24	17	24
Tiny Aya Global	24	24	27
GPT OSS 20B	25	20	25
Ministral 3 - 14B	27	16	20
Qwen 3.5 - 9b	27	19	22
Tower+ 9B	27	21	23
SCIR-TG-MT	30	18	4
GLM 4 - 9B	30	22	19

English→German	WMT26 human	alphaMWE	LLM-TR
TRIVE	1	3	3
Wayfinder	1	7	2
GPT 5.5	1	6	14
Gemini 3.1 Pro	1	28	22
Dubformer	5	5	1
Lumen	6	11	5
UvA-MT	6	15	16
Human (postediting)	6	–	–
Cohere CAT+	8	19	9
Command A+	9	24	20
MakotoAI	9	–	–
Qingqiu-MT-9B	11	8	5
Gemma 4 - 31B	12	4	12
DeepSeek V4 Pro	12	2	18
Mistral Medium 3.5	12	14	17
GoogleTranslate	16	9	14
PseudoRef	16	–	–
SalamandraTA	17	10	10
GPT OSS 120B	17	21	28
CUNI-MH-v3	20	16	7
Gemma 4 - 4B	20	12	13
SRPOL	20	18	11
Tower+ 9B	20	13	25
Tiny Aya Global	20	22	21
Qwen 3.6 - 27b	20	20	24
Ministral 3 - 14B	26	25	27
GPT OSS 20B	26	27	26
CUNI-EdUKate	27	17	8
GLM 4 - 9B	29	23	19
Qwen 3.5 - 9b	29	26	23

Table 12: System ranking for English→{Czech, Russian, German} according to human evaluation and test suites.

26-position difference. This is the largest absolute rank discrepancy in the three datasets. A similar, though smaller, pattern occurs for systems such as CUNI-EdUKate in English→German, which is human-ranked 27th but LLM-TR-ranked 8th, and CUNI-MH-v3, which is human-ranked 20th but LLM-TR-ranked 7th.

7.4 Test Suite Findings Summary

The evaluation of multiple test suites across diverse language pairs and domains provides a qualitatively oriented but empirically justified perspective to the performance of MT systems, revealing systematic issues that are not obvious in the generic scores.

Figurative and idiomatic multiword expressions remain difficult, with automatic metrics often disagreeing and human evaluation exposing errors aggregate scores hide. For English→Icelandic, even top systems struggle with select syntactic constructions, with a wide performance gap between the strongest and weakest models. For Korean→English, particle and evidential meaning degrade under structural complexity; wordplay is the hardest phenomenon overall. General-purpose LLMs are far more prone to hallucination and semantic detachment than dedicated MT systems, especially with sensitive content, truncated input, and embedded instructions.

For English→Hindi, narrative and dialogue text is handled well, but quality drops sharply for poetry, legal, and conversational registers, with a tendency to reproduce memorized canonical translations instead of the source. For German legal terminology, apparent competence in a low-resource variety mostly reflects term overlap with better-resourced varieties, not genuine mastery. For English→Korean game localization, aggregate scores mask large category-specific variance, and added context can counterintuitively reduce quality.

For inclusive 1st-person gender forms in Slavic

System	W/R	H/R	AST	alphamwe						CoS	EURAC				KU	gender1person			LLM-TR					LoM	TR-G
			en-is	en-aeb	en-ar	en-arz	en-de	en-pl	en-zh	en-hi	it-deAT	it-deCH	it-deDE	it-deIT	en-ko	en-cs	en-hr	en-ru	en-cs	en-de	en-es	en-et	en-ru	ko-en	tr-en
CUNI-AR	2.7	4.8	–	2	4	2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
TRIVE	3.6	5.1	2	1	3	5	3	2	12	2	3	3	1	3	2	6	3	3	5	3	6	6	4	7	5
GPT 5.5	4.2	8.6	1	3	9	9	6	4	6	3	1	2	1	2	1	12	10	9	14	14	10	11	13	2	4
[Anonymous]	4.9	11.0	–	–	13	7	1	–	–	–	–	–	–	–	–	–	–	15	–	3	2	–	1	–	1
Wayfinder	4.9	5.2	4	–	10	3	7	7	14	–	–	–	–	–	3	7	–	6	3	2	3	4	3	5	3
Dubformer	5.2	13.8	13	5	12	11	5	3	4	8	5	6	5	6	5	11	7	14	1	1	1	2	2	1	–
SCIR-TG-MT	6.3	20.6	5	–	–	–	–	–	1	–	–	–	–	–	–	–	–	18	–	–	–	1	4	–	–
DeepSeek V4 Pro	6.4	17.5	11	4	5	6	2	1	11	7	4	4	4	4	9	5	5	2	17	18	14	17	18	3	14
Lumen	6.6	6.8	7	–	8	8	11	10	15	5	–	–	–	–	6	10	–	8	2	5	4	3	7	6	7
Gemma 4 - 31B	8.2	22.0	17	7	6	1	4	9	10	11	8	7	7	7	11	13	9	10	10	12	9	8	9	4	12
UvA-MT	8.7	8.9	6	–	–	10	15	–	22	–	6	5	6	5	14	15	–	11	12	16	–	11	16	–	–
Gemini 3.1 Pro	9.3	4.8	3	21	26	26	28	23	27	16	2	1	3	1	27	4	2	1	25	22	17	20	21	9	1
Cohere CAT+	9.9	18.3	16	16	14	16	19	17	23	4	7	8	8	11	13	17	16	12	9	9	7	9	8	7	15
Mistral Medium 3.5	11.1	29.3	21	9	7	12	14	5	7	9	9	10	9	8	20	2	6	7	13	17	13	14	14	17	11
Qingqiu-MT-9B	11.6	16.5	15	6	21	4	8	19	13	17	15	13	12	13	4	8	13	5	3	5	4	5	4	16	20
GPT OSS 120B	13.5	32.1	20	12	15	13	21	12	19	12	10	9	11	10	21	3	1	4	27	28	23	27	26	14	20
SRPOL	14.0	26.1	10	–	–	18	18	–	20	–	–	–	–	–	7	27	–	27	8	11	–	7	10	–	–
HW-TSC	14.0	14.0	–	–	11	–	–	–	25	–	–	–	–	–	16	9	–	–	7	–	–	–	–	–	–
Qwen 3.6 - 27b	14.1	35.1	19	13	17	16	20	20	17	13	11	11	10	12	18	18	12	17	23	24	20	19	24	11	10
Command A+	14.7	26.0	18	17	16	23	24	18	16	6	14	12	19	9	15	23	19	25	18	20	15	18	17	10	23
Gemma 4 - 4B	15.5	35.6	26	10	2	19	12	8	5	15	17	19	13	15	17	14	14	13	16	13	12	15	11	20	16
SalamandraTA	16.1	29.0	14	8	22	14	10	16	9	21	16	16	15	17	8	24	20	23	10	10	8	10	15	23	22
Ministral 3 - 14B	16.4	41.4	22	20	24	24	25	14	21	14	12	14	17	16	24	1	4	16	21	27	18	21	20	19	6
GoogleTranslate	16.7	26.9	9	11	1	20	9	6	3	1	22	22	22	22	10	25	21	26	15	14	11	16	12	13	8
Tower+ 9B	17.0	41.2	8	19	25	25	13	11	2	10	19	15	14	18	19	20	17	21	22	25	22	25	23	18	13
GPT OSS 20B	17.4	39.8	23	14	23	15	27	21	24	19	13	16	16	14	23	19	11	20	24	26	21	24	25	12	19
Qwen 3.5 - 9b	18.7	41.5	24	15	19	21	26	22	26	18	18	18	18	19	22	16	8	19	20	23	19	22	22	15	17
GLM 4 - 9B	20.4	43.0	25	18	20	22	23	15	8	22	20	20	21	20	26	22	18	22	19	19	16	23	19	21	18
CUNI-MH-v3	21.3	37.8	–	–	–	–	16	–	–	–	23	23	23	23	–	26	–	–	6	7	–	–	–	–	–
Tiny Aya Global	21.5	35.5	27	–	18	27	22	13	18	20	21	21	20	21	25	21	15	24	26	21	24	26	27	22	9
CUNI-EdUKate	22.8	40.9	–	–	–	–	17	–	–	–	24	24	24	24	–	–	–	–	–	8	–	–	–	–	–

Table 13: Aggregated ranking from all test suites. W/R: weighted-average test suite rank; H/R: GenMT human rank

languages, systems split sharply between those that consistently generate inclusive forms and those that rarely do. For Turkish→English occupational gender bias, all systems default to stereotype-driven, male-skewed assignments under ambiguity —except those using gender-neutral phrasing, which avoid the bias with no accuracy loss.

8 Conclusions

This year we focused on improving the quality and reproducibility of human evaluation, the final arbiter of translation quality. We switched to Pearmut annotation tool, cESA annotation protocol, and made use of dynamic evaluation rather than relying on selection of systems based on automatic metrics. The novel dynamic allocation method enabled us to include all systems in human evaluation, with better systems getting more annotations, which helped us to avoid introducing bias from automatic evaluation as in previous years.

The WMT 2026 General Machine Translation Task covered 23 language pairs and up to six domains. We prepared a more challenging test set by utilizing novel human-based difficulty sampling. In total, we evaluated 46 systems: 30 participant submissions and 16 LLMs or popular online translation services whose outputs we collected. Participation decreased slightly compared with last year but remained high, with most teams using LLMs.

We used contrastive ESA as the human evaluation protocol, which consists of marking errors and assigning scores for three systems next to each other at once.

Analysis of domains showed the social domain as the most challenging, followed by speech, while software texts were the easiest to translate.

Specialised test suites revealed remaining weaknesses in figurative and idiomatic multiword expressions, wordplay, sensitive content, truncated input, inclusive gender forms, as well as domains and genres with specific terminology and/or style such as legal texts, game localisation, poetry.

Limitations

Access to test data: Researchers who had access to the internal repository (i.e., ALL source documents) before the blindset was released to public: Archchana Sindhujan, Chi-Kiu Lo, Diptesh Kanojia, Tom Kocmi, Eleftherios Avramidis, Martin Popel, Ondrej Bojar, Lorenzo Proietti, Rachel Bawden, Samuel Larkin, Stefano Perrella.

Sergey Dukanov had access only to the speech domain source data (no access to internal repo).

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

Human evaluation: Although we use human judgments as the gold standard, we should men-

tion that human annotations are noisy (Wei and Jia, 2021) and their performance is affected by the quality of other evaluated systems (Mathur et al., 2020). Lastly, different annotators use different ranking strategies, which may affect the system ranking.

Ethical Considerations

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

Collection of human judgments using Pearmut was fully anonymous. Automatically generated accounts associated with annotation tasks with single-sign-on URLs were distributed among pools of annotators, and we do not store any personal information. We do store the mapping between which annotator (pseudoanonymized) annotated which segments. Annotators received the standard professional translator's or evaluator's wage with respect to their countries.

Acknowledgments

This report would not have been possible without the partnership with Alconost, Árni Magnússon Institute for Icelandic Studies, Charles University, Cohere, Council for Ligurian Linguistic Heritage, Dubformer, German Research Center for Artifical Intelligence (DFKI), Google, Istitut Ladin Micurà de Rü, Institute of the Estonian Language, Microsoft, NTT, Tencent, The Sámi Parliament of Norway, The Sámi University of Applied Sciences, Toloka, University of Tartu, University of Tokyo. Furthermore, we are grateful to Toshiaki Nakazawa, Konstantin Dranch, Mao Zheng, Antonina Zaleschenok, Miliana Bær, Sierge Rasmus, and Lars Ailo Bongo.

Vilém Zouhar gratefully acknowledges the support of the Google PhD Fellowship. Sara Rajaee is funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080. Ondřej Bojar acknowledges the support by the grant CZ.02.01.01/00/23_020/0008518 (Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím"). The reference translations and manual evaluations were also supported by the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO, by the grant 10131474 (Czech AI Factory, CZAI) by EuroHPC JU and by TAČR grant EdUKate (TQ01000458). Eleftherios Avramidis was partially funded by the German Federal Ministry of Research, Technology and Space (BMFTR) via the project RoGSiLT (grant number 16IW26001). Rachel Bawden's participation was supported by the French national research agency (ANR) as part of the MaTOS project (grant number: ANR-22-CE23-0033). 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 M. Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, and 17 others. 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.

Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. Machine translation in the era of large language models:a survey of historical and emerging problems. *Information*, 16(9).

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, and 2 others. 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.

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, and 2 others. 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.

Mirelle Candida Bueno, Lucas Domingues, Samuel Frontull, Jean Maillard, and Sushil Garg. 2026. LiLa at WMT 2026 general MT task: Exploring beyond NMT and learning the limits of LLM adaptation for ligurian and ladin. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. 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.

Xiayu Cao, Baohang Li, Zekun Yuan, Zekai Ye, and Xiaocheng Feng. 2026. SCIR-TG-MT's submission for the WMT 2026 general machine translation task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. 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.

Aaron Chatterji, Thomas Cunningham, David Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use ChatGPT.

Minjoo Choi, Jaejun Jung, Gunwoo Han, Junhyeop Kim, Yiji Yoon, Hyemi Kim, Yeonseo Cho, and Jinwoo Jang. 2026.

KUEST: Korean game usage evaluation suite for LLM translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. 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.

Marta R. Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, and Carleigh Wood. 2024. MuTox: Universal MUltilingual audio-based TOXicity dataset and zero-shot detector. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5725–5734, Bangkok, Thailand. Association for Computational Linguistics.

Marta Cozzini, Antoni Oliver, Gonzalo López-Sánchez, Mercè Vàzquez, Patricia Morales-Hurtado, and Sergi Alvarez-Vidal. 2026. GRIAL-TA participation in the WMT26 general translation shared task: SalamandraTA7bFFT. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Paolo Di Natale, Elena Chiocchetti, Natascia Ralli, Marlies Alber, and Egon W. Stemle. 2026. Terminology resources as (MT) benchmarks: Evaluation of legal terminology challenges in german language varieties by EURAC. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Adam Dobrowolski, Paweł Przybysz, Marcin Szymański, Paweł Przewłocki, and Artur Janicki. 2026. Relay-MT: Multi-agent LLM ensembling for machine translation – samsung r&d poland submission for WMT2026. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Sergey Dukanov, Maria Tsyurupa, and Anton Dvorkovich. 2026. Dubformer at WMT26 general MT: Translating the video, not the transcript. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Mark Fishel and Lisa Yankovskaya. 2026. The two towers for Estonian-centric and Finno-Ugric machine translation. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 515–527, Tilburg, The Netherlands. European Association for Machine Translation.

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.

Samuel Frontull, Jean Maillard, Christopher R. Haberland, Ruth Videsott, and Stefano Lusito. 2026. Grammatist at WMT 2026 general MT task: An agentic LLM framework for machine translation with explicit linguistic knowledge. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. 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, 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.

Lifeng Han, Jiahui Liang, Anna Latusek, Karim El Haff, Amal Haddad Haddad, Josua Höfgen, Kilian Evang, Min Ma, and Maryia Zhyrko. 2026. Mind the gap: Exposing LLM translation blind spots using the AlphaMWE multilingual parallel corpus. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Yu He, Xiaoqing Lan, Daimeng Wei, Jiaxin Guo, Yuanchang Luo, Hengchao Shang, Zongyao Li, Jinlong Yang, Zhanglin Wu, and Boqi Huang. 2026. HW-TSC's submission to the WMT 2026 general machine translation track. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and

Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Miroslav Hrabal, Josef Jon, and Ondřej Bojar. 2026. CUNI-MH and CUNI-EdUKate at the WMT26 general MT shared task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Josef Jon, Rawan Bondok, Miroslav Hrabal, and Ondřej Bojar. 2026. CUNI-AR team at WMT26 general translation task for egyptian arabic. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Diptesh Kanojia, Chi-kiu Lo, Archchana Sindhujan, Samuel Larkin, Greg Hanneman, and Alon Lavie. 2026. In the blind: Building pseudo-references for MT evaluation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

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

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

Tom Kocmi, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Shaun Rafael Cassini, Ona de Gibert, Aidan Gomez, Nithya Govindarajan, Shun Kiyono, Olivia Lasche, Lawrence Rogers, Kelly Marchisio, Nikita Moghe, Yash More, Camila Moran-Hidalgo, Yiyang Nan, Michael Sachs, Trisha Starostina, Daan van Stigt, and 4 others. 2026. North small translate: Advanced cost-effective translation (cohere CAT+). In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

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.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Deun Sol Lee, Dae Woong Kim, and Yun Ha Seo. 2026. Lost in morphology at WMT26: A korean-english test suite for LLM translation blind spots. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Soyoung Lee. 2026. AdaptiveMT at WMT26: A controller–executor–reducer agent harness for document-level translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. 2019. Pkuseg: A toolkit for multi-domain chinese word segmentation. *CoRR*, abs/1906.11455.

Stefano Lusito and Jean Maillard. 2021. A Universal Dependencies corpus for Ligurian. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 121–128, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Audrey Mash, Jonathan Orama Ayebakuro, Ella Paulina Bohman, Xixian Liao, Francesca De Luca Fornaciari, and Maite Melero. 2026. BSC submission for WMT26 general machine translation shared task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. 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.

Ryosuke Matsuda, Keito Kudo, Ryo Fujii, Takumi Ito, Makoto Morishita, and Jun Suzuki. 2026. RFMT at WMT 2026 general translation task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. 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.

Graham Neubig. 2011. The Kyoto free translation task.

Chris Nishihama. 2026. MakotoAI at WMT26: Per-pair model selection with deterministic output normalization for the general machine translation task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. 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, and 20 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

OpenAI Team. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Yusuf Öztop. 2026. Evaluating occupational gender bias and male default in turkish-to-english machine translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand. Association for Computational Linguistics.

Mikita Pilinka, Aliaksandr Kliuyeu, David Samuel, and Yves Scherrer. 2026. ReMova: Fine-tuning LLMs for english to belarusian translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Flammie A. Pirinen. 2026. Apertium-sme-eng for WMT2026 shared task—creating new pair by crossing dictionaries and translating rules. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Dmitrii Popov, Elizaveta Kozlova, Ekaterina Taracheva, Elizaveta Makhotina, Artem Mekhraliev, Miron Baratelia, Ekaterina Enikeeva, and Nikolay Karpachev. 2026. Yandex at WMT26: Translation adaptation at 235b and hard-label distillation to 8b. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Maja Popović and Ekaterina Lapshinova-Koltunski. 2025. GENDER1PERSON: Test suite for estimating gender bias of first-person singular forms. In *Proceedings of the Tenth Conference on Machine Translation*, pages 800–822, Suzhou, China. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24261–24285, Suzhou, China. 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, Nuno M Guerreiro, José Pombal, João Alves, Amin Farajian, Pedro Teixeirinha, and Andre Martins. 2026. TOWER+: Bridging generality and translation specialization in multilingual LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29614–29635, San Diego, California, United States. 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.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. 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.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. 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 6490–6500, Online. Association for Computational Linguistics.

Liangying Shao, Xinwei Wu, Yichong Huang, Jifang Wang, Ruoxi Xu, Ling Shi, Baosong Yang, and Linlong Xu. 2026. Wayfinder: An agentic ensemble-and-postcheck system for WMT26 GenMT. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Einar Freyr Sigurðsson, Steinþór Steingrímsson, Bjarki Ármannsson, Atli Jasonarson, Helga Svala Sigurdardottir, and Jón Símon Markússon. 2026. Translating syntactic structures: An english-icelandic test suite for syntactic divergence. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Steinþór Steingrímsson, Sveinbjörn Þórðarson, and Jón Friðrik Daðason. 2026. What a DRAG (it is being small): The AMI submission to the WMT 2026 general translation shared task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

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.

Lisa Vasileva. 2026. "i'm sorry, i can't translate that": A challenge set for refusal to translate, overgeneration and hallucination in LLM translation systems. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Oksana Volchek and Vladislav Poritski. 2026. Krosny at WMT26 general translation task: Dictionary augmentation and iterative refinement for english–belarusian translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Hao Wang, Yangyang Liu, Xiaohu Zhao, Heng Liu, Zifu Shang, Tianhao Li, Ruize Gao, Jialong Tang, Shiao Shao, Haoran Wei, Baosong Yang, Linlong Xu, Longyue Wang, and Weihua Luo. 2026. Lumen at WMT2026: Instruction-specialized translation via quality-aware training and on-policy distillation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. 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.

Di Wu, Sergey Troshin, Wafaa Mohammed, Seth Aycock, Evgeniia Tokarchuk, Vlad Niculae, and Christof Monz. 2026. UvA-MT's participation in the WMT26 general translation shared task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Tian Xia, Chao Chen, Mengpeng Yang, Jingxu Yang, Yabo Sun, and Qiang Liu. 2026. QINGQIU-MT-9b: An instruction-following multilingual translation model. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Saumitra Yadav, Ananya Mukherjee, and Manish Shrivastava. 2026. Finding the gaps in LLM translation with CoST. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Mengpeng Yang, Tian Xia, Jingxu Yang, Chao Chen, Yabo Sun, and Qiang Liu. 2026. TRIVE: A tri-fusion verified translation agent. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Hao Yu, Ziyan Chen, Hongyu Zhao, Yulong Zhang, Xiaogang Wang, Kaiyu Huang, and Degen Huang. 2026. TransCore: The NewTranx submission to the constrained track of the WMT26 general machine translation shared task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. 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).

Vilém Zouhar and Tom Kocmi. 2026. Pearmut: Human evaluation of translation made trivial. In *Proceedings of the 2026*

Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Budapest, Hungary.

Vilém Zouhar, Julia Kreutzer, Alon Lavie, Tom Kocmi, Matt Post, Ondřej Bojar, and Mrinmaya Sachan. 2026a. Dynamically allocating evaluation effort for model ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, Budapest, Hungary.

Vilém Zouhar, Niyati Bafna, Mukund Choudhary, Maike Züfle, Sara Rajaee, Pinzhen Chen, Jannis Vamvas, Sara Papi, Ona de Gibert, Bhavitvya Malik, Eliya Habba, Orfeas Menis Mastromichalakis, Patrícia Schmidtová, Michelle Wastl, Sheriff Issaka, Leshem Choshen, Stella Biderman, Antonis Anastasopoulos, Jan Niehues, and 225 others. 2026b. Last translation benchmark. *Preprint*, arXiv:2609.04173.

Vilém Zouhar, Roman Grundkiewicz, Sara Rajaee, Parker Riley, Martin Popel, Rachel Bawden, Philipp Koehn, Marine Carpuat, and Tom Kocmi. 2026c. Contrastive esa: Human evaluation of multiple translations at once. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.

A.2 Human Annotation Attention Checks

Figure 5: A screenshot of loud attention checks in the human evaluation campaigns. The annotators must assign a lower score to the lowest quality translation to pass the annotation step.

A Translator and Annotator Details

A.1 Human Annotation Domain-Specific Instructions

- **news:** Ensure the translation is formal and consistent with journalistic standards.
- **factchecking:** Ensure the translation is formal and consistent with journalistic standards.
- **speech:** Check the translation maintains the flow and colloquial style of the speaker. Translation should not include non-linguistic sounds (e.g. laughter, groans, hesitation sounds, etc.), but should include interjections. If a word is interrupted, translations should either guess the full word if possible or otherwise omit it. Foreign words should be preserved.
- **social:** Spelling mistakes should not be reproduced. Marks of expressiveness that communicate meaningful intent (e.g. enthusiasm through capitalisation or elongation) in a way that is natural should be reproduced in the translation. URLs and user handles should be copied directly rather than being translated. Hashtags should be translated as appropriate for the translation to be natural for social media text. Punctuation of the source text should be followed as best as possible. Translation should be in an informal style, like close friends talking, even if it changes the original tone.
- **software:** Translation should translate the content of the JSON, without translating keys (such as `"general.duration.years"`) and placeholders (e.g. `"{{count}}"`).
- **edu:** Translation should be accurate and maintain the instructional nature of the content. Ensure that technical terms are translated correctly and consistently.

A.4 Human Annotation cESA Instructions

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

- ***Task:*** *Read the source text and competing translations. Highlight all errors. Rate each translation (each column is an output from a single model/translator). At the end, click* ***Next*** *in the top-right corner.*
- ***Highlighting errors:***
  - ***Mark errors*** *by clicking the beginning and then clicking at the end of the erroneous text span. Then assign severity by hovering over the error span.*

A.3 Human Annotation Tutorials

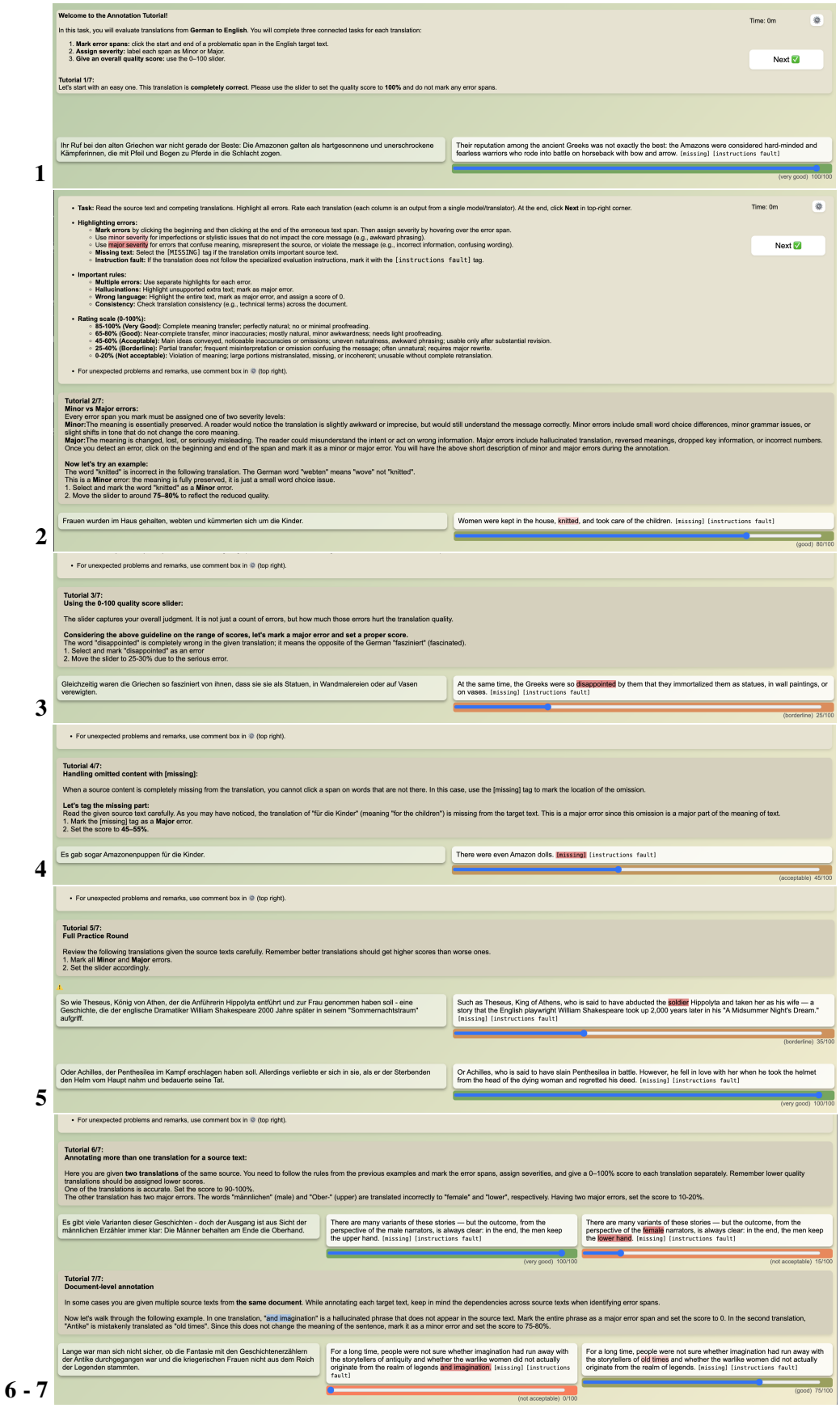


Figure 6: The screenshots of 7-step optional tutorials used at the beginning of each annotation campaign (with varying language) to familiarize the annotators with Pearmut platform and the cESA protocol.

- *Use minor severity for imperfections or stylistic issues that do not impact the core message (e.g., awkward phrasing).*
- *Use major severity for errors that confuse meaning, misrepresent the source, or violate the message (e.g., incorrect information, confusing wording).*
- ***Missing text:*** *Select the* `[MISSING]` *tag if the translation omits important source text.*

- ***Important rules:***
  - ***Multiple errors:*** *Use separate highlights for each error.*
  - ***Hallucinations:*** *Highlight unsupported extra text; mark as major error.*
  - ***Wrong language/dialect:*** *Mark the text as a major error; borrowed words are allowed if they are valid in given language.*
  - ***Consistency:*** *Check translation consistency (e.g., technical terms) across the document.*
- ***Rating scale (0–100%):***
  - ***85–100% (Very Good):*** *Complete meaning transfer; perfectly natural; no or minimal proofreading.*
  - ***65–80% (Good):*** *Near-complete transfer, minor inaccuracies; mostly natural, minor awkwardness; needs light proofreading.*
  - ***45–60% (Acceptable):*** *Main ideas conveyed, noticeable inaccuracies or omissions; uneven naturalness, awkward phrasing; usable only after substantial revision.*
  - ***25–40% (Borderline):*** *Partial transfer; frequent misinterpretation or omission confusing the message; often unnatural; requires major rewrite.*
  - ***0–20% (Not acceptable):*** *Violation of meaning; large portions mistranslated, missing, or incoherent; unusable without complete retranslation.*
- *For unexpected problems and remarks, use the comment box in the top right corner.*

A.5 Translator Brief for Post-Editing

The following instructions were given to human translators:

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

- *All translations must be produced by human post-editing.*
- *Translators must preserve paragraph boundaries. In the source files, paragraphs are always separated by empty lines, although each paragraph may itself span multiple lines. The translation must contain the same number of paragraphs as the source, with the same paragraph boundaries. Within each paragraph, translators may adjust the number of sentences as needed.*
- *In some cases, the automatic translations provided for post-editing may contain a different number of paragraphs, or different paragraph breaks, from the corresponding source documents. In these cases, please treat the source documents as authoritative for the paragraph structure. The final reference translations must follow the source documents, preserving the same number of paragraphs and the same paragraph boundaries, with paragraphs separated by empty lines.*
- *Translators should avoid inserting parenthetical explanations into the translated text and obviously avoid losing any pieces of information from the source text. We will check the translations for quality and will reject translations that contain errors.*
- *If the original data contains errors, typos, or other problems, do not change the source documents (which should not be modified in any way), instead try to prepare a correct translation as if the error had not been in the source.*
- *The data contain four or five domains (news, social, speech, software), each folder containing one domain source and each domain needing a specific handling. Maintain the formality of the source text, i.e. news texts are formal, social domain texts are likely to be informal.*

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

News Domain and Fact-Checking Domain. The texts are news articles and should be translated in a formal style, consistent with journalistic writing in the target language. The translation should accurately convey the meaning, tone, and nuances of the source text while reading naturally as a news article originally written in the target language.

Speech Domain. Your task is to translate the speech from the provided video. For this domain, each source file corresponds to one segment of video. Alongside the video, we provide a human-edited transcript of the spoken content. The transcript has already been corrected, but it should be used only as a guideline for translation from the video. The transcripts have been used to generate the automatic translations.

Preserve the flow, degree of formality, and colloquial style of the speaker(s), including informal wording, jargon, common contractions, repetitions, omissions, and unfinished phrases when they are part of natural speech. Do not make the style more formal or more polished than the source video.

If a word or phrase is unintelligible in the video and cannot be reconstructed from the context with reasonable certainty, make your best guess. If a speaker stops mid-word, either translate the intended full word if it can be inferred from the context or omit it; do not leave unfinished word fragments in the translation.

Do not include non-linguistic sounds such as laughter, groans, breathing, or hesitation sounds unless they clearly contribute to the meaning. Common interjections and filler words may be translated when they are part of the speaker's expression or the natural flow of the speech. Words spoken in a foreign language should be left as they are.

If some words are sung rather than spoken, include them only when they contribute to the meaning of the video.

Videos correspond to different domains: they differ in formality, style, topics and number of speakers. Translate into the target language using wording and a register that match the characteristics of the source video as closely as possible.

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

The texts are user-generated social media content and may contain spelling errors, slang, acronyms, informal wording, marks of expressivity, etc. The translation should sound natural as social media text in the target language, while preserving the meaning, style, familiarity, and expressive intent of the source text as much as possible. The translation must always use an informal style, even if the source sounds relatively formal.

- *Spelling mistakes should not be preserved in their translations, i.e. the translation should be spelt correctly.*
- *Marks of expressivity, such as asterisks, repeated characters, or other emphasis devices, should be preserved when they communicate meaningful intent (e.g. enthusiasm through capitalisation or elongation). The exact form does not need to be copied mechanically: for example, elongated words (e.g. gooood) should be rendered in a natural way in the target language, and the exact number of repeated characters does not need to match the source unless it makes a clear difference in quality or meaning.*
- *There will be abbreviations and acronyms (e.g. btw -> by the way, fwiw -> for what it's worth). These do not need to be translated using an abbreviation or acronyms unless an abbreviation/acronym is the best translation choice in the target language.*

- *Users (@user123) and URLs should be left as they are, i.e. not translated.*
- *Platform-specific elements such as hashtags should be translated as hashtags, but the content should be translated appropriately into the target language.*
- *Additional punctuation should be added only if not doing so would seriously alter the comprehension of the text. Otherwise, we recommend following the punctuation of the source text.*

Please always refer to the screenshots included alongside the source texts. These screenshots show the original context of the entire thread contained in the source text file and should be consulted during translation. The screenshots may also contain images attached in the thread that can provide further context for the translation.

Software Domain. The source texts in this domain come from software data and are provided in JSON format. Translate only the user-facing natural-language content, i.e. the text that appears as JSON values. Do not translate JSON keys, placeholders, and other technical content. The final output must be a valid JSON and must match the input format, preserving the overall source structure. Placeholders, such as {{website_name}}, {{count}}, or similar elements, must be copied exactly, not translated.

A.6 LLM Translation Instructions

To generate the automatic translations, we prompted LLMs with domain-specific translation instructions. The prompts used for each domain are reported below. The placeholders `{TRG}`, `{TRG_CODE}`, and `{SRC}` were instantiated with the corresponding target language, target-language code, and source language, respectively. For the Speech domain, we used the final transcripts after human correction (see Appendix A.7) of the automatic transcriptions only when providing translations for human post-editing. For both the Speech domain and the Social domain, the automatic translations were generated using a text-only prompt, without providing the accompanying videos or screenshots to the LLMs. These automatic translations were generated by translating each source document as a whole.

News Domain and Fact-checking Domain. You are a professional {TRG} translator, tasked with providing translations suitable for use in {TRG} ({TRG_CODE}). Your goal is to accurately convey the meaning and nuances of the original text while adhering to {TRG} grammar and vocabulary and ensuring that the translation is natural. The original text is a news article. Ensure the translation is formal and consistent with journalistic standards. Produce only the {TRG} translation, without any additional explanations or commentary. Maintain the HTML formatting of the original source text. Please translate the following text into {TRG} ({TRG_CODE}):

Speech Domain. You are a professional {TRG} translator, tasked with providing translations suitable for use in {TRG} ({TRG_CODE}). Your goal is to accurately convey the meaning and nuances of the original text while adhering to {TRG} grammar and vocabulary and ensuring that the translation is natural. The original text is automatically transcribed from spoken language and can contain errors. Maintain the flow and colloquial style of the speaker in the translation. Do not include non-linguistic sounds (e.g. laughter, groans, hesitation sounds, etc.), but do include interjections. If a word is interrupted, either guess the full word if possible or otherwise omit it. Keep foreign words as they are when translating. Produce only the {TRG} translation, without any additional explanations or commentary. Output the text such that each sentence is on a separate line. Please translate the following text into {TRG} ({TRG_CODE}):

Social Domain. You are a professional {TRG} translator, tasked with providing translations suitable for use in {TRG} ({TRG_CODE}). Your goal is to accurately convey the meaning and nuances of the original text while adhering to {TRG} grammar and vocabulary and ensuring that the translation is natural. The original text is user-generated content from a social media platform. Do not reproduce spelling mistakes. Reproduce marks of expressiveness that communicate meaningful intent (e.g. enthusiasm through capitalisation or elongation) in a way that is natural in {TRG}. Copy URLs and user handles directly rather than translating them. However, translate hashtags as appropriate for the translation to be natural for social media text. Follow the punctuation of the source text as best as possible. Additional punctuation should be added only if not doing so would seriously alter the comprehension of the text.

Translate text in an informal style, like close friends talking, even if it changes the original tone. Produce only the {TRG} translation, without any additional explanations or commentary. Maintain the HTML formatting of the original source text. Please translate the following text into {TRG} ({TRG_CODE}):

Software Domain. You are a professional {TRG} translator, tasked with providing translations suitable for use in {TRG} ({TRG_CODE}). Your goal is to accurately convey the meaning and nuances of the original text while adhering to {TRG} grammar and vocabulary and ensuring that the translation is natural. The original text is from software data. Translate the content of json, without translating keys and placeholders (i.e. copy them rather than translate them). Produce only the {TRG} translation, without any additional explanations or commentary. Produce a valid json output that matches the input format. Please translate the following text into {TRG} ({TRG_CODE}):

Educational Domain. You are a professional {TRG} translator, tasked with providing translations suitable for use in {TRG} ({TRG_CODE}). Your goal is to accurately convey the meaning and nuances of the original text while adhering to {TRG} grammar and vocabulary and ensuring that the translation is natural. The original {SRC} text consists of biology, chemistry, and geography exercises extracted from an educational web portal for children aged 9–16. Produce only the {TRG} translation, without any additional explanations or commentary. Maintain the HTML formatting of the original source text. Please translate the following text into {TRG} ({TRG_CODE}):

A.7 Speech Transcript Cleanup Guidelines

For the Speech domain, the automatically generated transcripts were manually corrected by human annotators before being used both to generate the automatic translations provided for possible human post-editing and as source material for translations from scratch. The following instructions were given to the annotators responsible for this transcript cleanup:

You will be given: a video and its automatically created transcription. The speech in the transcription is divided into sentences, and each sentence has start and end timings and a speaker code (explained below). Please carefully watch the video and correct the transcription.

- *Make sure that the written text fully corresponds to the spoken text. Rewrite all the words and phrases that were transcribed incorrectly, add all the missing words and phrases, delete all the words that were added by mistake. Correct spelling and punctuation.*
- *Make sure the text is correctly split into sentences.*
- *Correct the voice codes if necessary.*
- *Correct spelling and punctuation if necessary.*
- *Try to keep punctuation simple. Make sure all the crucial commas, question marks and exclamation points are in their places, but don't worry about the difference between types of dashes and quotes. Don't use bold or italics.*
- *Don't ever ever insert any comments into the text (such as "laughter" or "said he angrily"). The text must contain only the words that are meant to be spoken aloud. Remove such comments if present.*
- *The speech on the video can be unintelligible, if, for example, the environment is noisy, the speaker mumbles or has a strong accent. If you can't recognize a word or a phrase and it can't be reconstructed based on the context with a reasonable level of certainty, make your best guess.*
- *Remove laughter, groans, and other non-speech vocal sounds from the transcription.*
- *You may include in the transcription commonly used interjections (examples: wow, ouch, ah, shh) and filler words (like, I mean, well, etc.).*
- *Hesitation words (sounds used when a speaker is not sure what to say next, examples: um, er) should be omitted. They must appear in the transcription only in the rare cases when they contribute to the meaning. For example, when we need to show that the speaker is at a loss what to say, or else the long pause would appear unnatural. Or when someone says "er..." and then comments upon it. You have to rely on your own judgment in such cases.*
- *Do not correct the style. Colloquial speech, jargon, incomplete phrases, informal words and commonly used contractions (ok, gotta, coupla, yeah) may be left as they are.*

- *If the speaker stops mid-word, the word should be guessed from the context or omitted. Both variants are fine. Just don't leave the word unfinished, because it would lead to translation errors.*
- *Repetitions, omissions and unfinished phrases are fine in natural speech, you may leave them in the text.*
- *Do not translate words spoken in a foreign language, leave them as they are.*
- *Some words in the video may be not spoken but sung. Include them in the transcription if they contribute to the meaning of the video. The song in the background can usually be omitted without any loss. If sung words are part of the conversation, they certainly have to be transcribed. If your case is somewhere in-between, act according to your judgment.*
- *If at some point two or more speakers talk simultaneously, try to hear all the words, write all the phrases one after another, and give them the correct overlapping time codes.*
- *Annotate speakers assigning the same unique label for phrases of the same speaker.*
- *Timecodes have been created automatically. They are usually correct, but there may be mistakes if the speech was not recognized properly, so please keep an eye on them and correct them if they are considerably off.*
- *When splitting a phrase into several phrases or when adding a new phrase, please give your new phrases correct timecodes.*

B System Submission Summaries

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

B.1 AdaptiveMT (Lee, 2026)

We describe AdaptiveMT, our submission to the English→Korean condition of the WMT26 General Machine Translation task in the unconstrained track. AdaptiveMT is a language-general adaptive translation harness with target-language-specific profiles; this run uses the Korean profile. The system uses a controller–executor–reducer agent harness. For each document, the harness analyzes the source, generates independent candidate translations from two heterogeneous models, compares them with token-level and semantic diffs, and synthesizes a final translation with a judge model. Mandatory QA hooks and deterministic Korean-specific guards run before finalization; failures trigger bounded local repair. Multimodal documents use structured visual context extracted from images or sampled video frames before translation. We processed all 2,078 distributed blindset records under two Korean target tags; the subsequently released General MT evaluation set contains 198 English→Korean documents tagged kor_Hang. On this official subset, the harness made 1,441 successful LLM calls at an estimated list-price inference cost of $12.03, and all 198 documents finished with passing hooks. In blindset-wide diagnostics, synthesis eliminated observed candidate-level mechanical errors, and a GPT-5 judge preferred the final output over a bare GPT-4.1 baseline on 600 sampled documents (44.5% vs. 30.0%, 25.5% ties).

B.2 AMI (Steingrímsson et al., 2026)

We describe the AMI team's submission to the WMT 2026 General Translation shared task, participating in the English→Icelandic translation direction. Building on our 2025 submission, which relied on a 3B-parameter model and substantial rule-based post-processing, this year we aim to reduce that dependency and rely more on the translation model itself. We continually pre-train a Qwen3-4B checkpoint on a mixture of Icelandic, English, code, math, and structured data, then fine-tune it with LoRA on a retrieval-augmented, domain-conditioned instruction-tuning dataset covering five domains. At both training and inference time, source sentences are enriched with dictionary entries and back-translated example translations similar to the sentence being translated, retrieved by a dedicated RAG server, and served through a self-healing inference client that retries and validates generations. We report COMET and chrF++ scores comparing this system against last year's submission, a larger Llama-3.1-8B model adapted with the same recipe, and an instruction-tuned Qwen3-4B baseline, alongside ablations that isolate the contribution of each part of the retrieval-augmented prompt. The models are available on HuggingFace: Qwen3-4B-wmt26-AMI-en-is and Llama-3.1-8B-wmt26-AMI-en-is.

B.3 Apertium-sme-eng (Pirinen, 2026)

We describe the creation of a new rule-based machine translation model built from two existing models. We built Apertium system between North Sámi and English based on existing translation models between North Sámi and Finnish, and Finnish and English. There were two expert-driven processes involved in building of the system: firstly crossing and validating the bilingual dictionaries and secondly translating and modifying the grammatical and structural rules. We briefly explain why there are no preexisting systems for this language pair and what limitations and ethical issues are related to it. We will also evaluate some of the training and development corpora provided in the shared task to examine the harm that low-quality machine translation can cause to vulnerable minority languages.

B.4 Cohere CAT+ (Kocmi et al., 2026)

We present North Small Translate, an open-weight, LLM-based machine translation (MT) model with instruction-following capabilities built on the same foundation as Cohere's Command A Plus, a mixture-of-experts architecture with 25 billion active parameters out of 218 billion total parameters. North Small Translate is trained using difficulty sampling to obtain challenging documents and a five-step training protocol combining supervised fine-tuning, direct preference optimization, and online reinforcement learning. We prioritized throughput through a non-reasoning base model and supplemented with optional agentic capabilities to unlock translation quality gains. North Small Translate is trained to perform MT-related tasks, including post-editing and quality estimation, as well as related tasks such as general instruction following. The model achieves top MT performance across 50 languages in the class of models under 1T parameters, with no need to run expensive reasoning at inference time.

B.5 CUNI-AR (Jon et al., 2026)

We describe the Charles University CUNI-AR submission to the WMT26 General Translation shared task for English into Egyptian Arabic. Our pipeline is built around GEMBA (LLM-as-judge quality) estimator that scores translations independently for accuracy and dialect authenticity. We use it throughout our pipeline: to filter training data, to construct preference pairs for DPO, to select best checkpoints, and to pick the best hypothesis per segment at inference time. Our primary translation model is Gemma-4-12B-it finetuned via QLoRA, we also finetune Jais-2-8B-Chat and Aya-expanse-8B. Gemma-4-12B-it provides a high-quality Egyptian Arabic translation, so SFT yields only marginal quality gains in automatic metrics; DPO is able to improve the metrics it optimizes for, but the real gain in translation quality is to be assessed by human evaluation. The final submission selects the best-scoring translation per segment from 31 candidate checkpoints, with the unmodified base model contributing 31% of the output.

B.6 CUNI-EdUKate and CUNI-MH-v3 (Hrabal et al., 2026)

We describe our constrained CUNI-MH-v3 (based on Qwen3-14B) and unconstrained CUNI-EdUKate (based on Granite 4.1 8B) submissions for the WMT26 General MT shared task. Both submissions use a three-stage training procedure: QLoRA supervised fine-tuning on synthetic sentence pairs, continued supervised fine-tuning on a smaller mixture of paragraph-level and instruction-oriented examples, and iterative preference optimization with automatically generated and scored preference pairs. We document the data construction, training setup, checkpoint selection, and evaluation. We also share the training data and model checkpoints for our constrained system. The models are available on HuggingFace: CUNI-MH-v3-encs, CUNI-MH-v3-csvi, CUNI-MH-v3-csuk, CUNI-MH-v3-csde.

B.7 Dubformer (Dukanov et al., 2026)

We describe the system Dubformer entered in the WMT26 General MT shared task, an unconstrained system built on our video-dubbing pipeline. It rests on two commitments. A document is never translated as raw text: it is first decomposed into a structure chosen for its do- main — scenes carrying speaker and audio- visual context for spoken dialogue, markup- masked segments for web documents, value- only segments for localization resources — so that non-linguistic material is protected by construction and everything needed for consistency is in front of the model at once. And no translation is taken on trust:

every segment is checked by a second model from a different family for hallucination and for drift from the meaning of the source, and a segment it rejects is re-translated with the reason attached rather than filled in from a fallback.

B.8 Grammatist (Frontull et al., 2026)

We describe Grammatist, our submission to the unconstrained track of the WMT 2026 General Machine Translation Task, developed in direct collaboration with community members of two regional languages of Italy, Ladin and Ligurian. Ladin and Ligurian represent challenging scenarios for machine translation (MT), as they have limited resources for conventional data-driven approaches. However, this scarcity of large-scale parallel data does not reflect a lack of linguistic documentation: both languages are extensively documented in dictionaries and grammar books, which contain valuable knowledge but are not readily exploitable by conventional MT approaches. Grammatist addresses this gap by making this information available to a large language model (LLM) at inference time. The framework is both agentic and model-agnostic: it dynamically consults dictionaries and grammar books during translation to retrieve the information most relevant to the input. Any LLM can serve as the underlying base model, enabling Grammatist to benefit from advances in increasingly capable models. For the WMT 2026 task, we apply Grammatist to structured document-level translation from English into Ligurian and Ladin.

B.9 Grial-SalamandraTA (Cozzini et al., 2026)

This work describes the Grial-TA team submission to the constrained track of WMT2026 General Machine Translation Task for the following language pairs: English→Icelandic, English→Ladin and English→Ligurian. Our submission leverages full fine-tuning of theSalamandraTA-7b-instruct model. Two separated models were fine-tuned: one for Icelandic and one for Ligurian and Ladin. The model is available on HuggingFace: MTUOC-recipes.

B.10 HW-TSC (He et al., 2026)

Machine translation based on large language models (LLMs) has achieved remarkable progress in recent years. However, relying on a single translation model and fixed translation instructions often limits translation quality, especially for multilingual and multi-domain scenarios. To address this issue, we propose HW-TSC-Agent, a multi-stage translation agent framework that integrates translation model fine-tuning, dual-model candidate generation, and instruction-aware translation refinement into a unified pipeline. First, we perform Supervised Fine-Tuning (SFT) and Contrastive Preference Optimization (CPO) on openPangu-Embedded-7B using high-quality bilingual data to enhance its multilingual translation capabilities. The fine-tuned openPangu-Embedded-7B and HY-MT2-7B are then used to generate candidate translations. Finally, source-language analysis, candidate translations, and task-specific instructions are incorporated into a dynamic prompt, which guides DeepSeek-V4-Flash to compare, refine, and fuse the candidate translations to produce the final output. Our system is submitted to the WMT2026 General Machine Translation Shared Task, covering 13 translation directions. On the WMT2025 English→Chinese development test set, enriching the dynamic prompt generally improves translation quality over an instruction-only DeepSeek-V4-Flash baseline.

B.11 Krosny (Volchek and Poritski, 2026)

This paper describes our submission to the WMT26 General Machine Translation shared task for the English → Belarusian direction. In lower-resourced morphologically rich languages like Belarusian, open-weight large language models (LLMs) frequently struggle with orthographic fidelity, hallucinated vocabulary, and negative transfer from related high-resource languages. To address these issues, we implemented Krosny, an iterative, resource-augmented pipeline built on top of TranslateGemma 12B. The system translates each segment in three stages: (1) dictionary-augmented generation, (2) refinement using a reference grammatical database, and (3) rule-based post-processing followed by an LLM-as-a-judge candidate selection. We find a small but consistent improvement over the baseline TranslateGemma 12B. In comparison with other constrained-track systems of WMT26, the performance of Krosny is mid-tier. We release the source code of our implementation (https://github.com/volchek/Krosny-WMT26).

B.12 LiLa (Bueno et al., 2026)

We describe LiLa, our submission to the constrained track of the WMT 2026 General Machine Translation Task. Motivated by recent advances in large language models (LLMs) and their increasing availability as open models, we investigate the adaptation of LLMs for Ligurian and Ladin, two low-resource languages included in the shared task. We evaluate several adaptation strategies for Gemma-4, including Instruction Tuning (IT-only), continuous pre-training followed by Instruction Tuning (CPT+IT), and Low-Rank Adaptation (LoRA), and compare them with a dedicated encoder-decoder pipeline based on fine-tuned No-Language Left Behind (NLLB) models. Our experiments show that, under the evaluated conditions, fine-tuned NLLB models achieve higher automatic translation scores than the investigated LLM adaptation strategies. Based on these results, our final submission adopts a fine-tuned NLLB-based translation pipeline. We discuss the main observations from the different approaches explored during development. We release the code and models produced in these experiments through a publicly accessible GitHub repository https://github.com/schtailmuel/wmt26-lila/.

B.13 Lumen (Wang et al., 2026)

We present Lumen, our submission to the WMT2026 General Machine Translation Shared Task. We focus on instruction-specialized translation, where systems must produce accurate translations while satisfying fine-grained requirements on style, terminology, and structured output. Lumen follows a four-stage pipeline. First, continued pre-training and supervised fine-tuning jointly strengthen multilingual translation competence and the ability to follow translation-specific instructions, such as terminology, style, and format constraints. Second, we extend the originally offline M^2PO preference-optimization framework into an online preference-optimization stage: the current policy samples candidate translations, an LLM judge assigns continuous direct-assessment scores, and the resulting multi-pair preferences are used to update the model. Third, on-policy distillation transfers stronger translation-instruction-following behavior from a larger teacher model to the Qwen3-14B student, improving translation-instruction adherence while retaining a 14B primary translation backbone. Finally, structure-aware inference and postprocessing preserve required formats, while selective repair addresses clear structural or decoding failures before submission. We participate in 23 translation directions. Under an absolute 1–5 LLM-judge rubric that evaluates translation-instruction following and translation quality separately, Lumen reaches an overall mean score of 4.60. Despite using a 14B translation backbone, it achieves performance comparable to leading proprietary systems in both translation quality and translation-instruction following.

The model is available on HuggingFace: Lumen_wmt26.

B.14 MakotoAI (Nishihama, 2026)

We describe MakotoAI, an unconstrained submission to the WMT26 General Machine Translation shared task covering all 23 language pairs (4,897 documents). The system performs no training or fine-tuning: it is a lightweight orchestration layer over six off-the-shelf commercial LLMs, in which each language pair is routed to the single model that won a blinded pre-competition shootout for that pair. Translation is document-level with the full task instruction provided in-prompt. Structured output (JSON and HTML domains) is repaired by a small deterministic normalization pass rather than by model-based retry: an ablation on real test data found that a post-hoc validate-and-retry instruction layer was inert on clean documents and net-negative on JSON documents, while deterministic fence and preamble stripping achieved the intended effect at zero inference cost. Quality assurance combined an exhaustive deterministic instruction-adherence scan, a two-judge LLM evaluation whose self-grading bias we measured to be directional by model family, and the organizers' official alignment checker, against which the final submission scores 1,914 of 1,914 structurally checkable documents aligned. Four defective documents (three truncations, one repetition-loop degeneration) were detected and regenerated, one requiring escalation to a different model; all four are disclosed. The entire system runs on commodity hardware; total API cost to produce and verify the submission was just under $25. We argue that for API-orchestrated MT, verification discipline and deterministic post-processing are higher-leverage than added inference-time machinery.

B.15 PseudoRef (Kanojia et al., 2026)

The WMT26 General MT task evaluates systems on 10 language pairs that have no human references (neither translated from scratch nor post-edited from MT output by humans). We describe how we built the pseudo-references for these pairs and six other language pairs (in which some forms of human references are available): seven models translate the 3,277 official documents under up to five prompt conditions, giving a total of 26 system–prompt combinations; then three reference-free quality estimation (QE) models score every candidate; and a per-document selector picks one translation, which GPT-5.5 post-edits where needed. Working without references exposed a failure mode of QE-guided selection: the metrics rank fluent output in the wrong language above correct translations. Adding a confidence-scaled language identification penalty to the score fusion drives the wrong-language count to zero, and the resulting selector still scores better on MetricX than the rank-fusion baseline it replaces. Since no references were available for these pairs while we were building them, we calibrate every selection decision on last year's WMT25 human judgments. The human evaluation, released after construction, shows the cost of getting selection wrong: our references stand with the strongest participating systems when the selector kept a frontier-model candidate, and fall up to 17 ESA points below them when it did not. We release the selection method and the provenance of every reference.

B.16 Qingqiu-MT-9B (Xia et al., 2026)

We present Qingqiu-MT-9B, an instruction-following multilingual translation model developed for the constrained track of the WMT26 General Machine Translation shared task. Our model supports translation among 20 languages and covers 22 of the 23 official language pairs. It uses a source-grounded synthesis pipeline with separate LLM-based steps for generating instance-specific translation instructions and their corresponding translations. The resulting corpus covers diverse translation requirements, task complexity, instruction formulations, domains, document lengths, and glossary constraints, reducing reliance on fixed prompt templates and supporting generalization to unseen translation instructions. Our model is based on Qwen3.5-9B and trained through two-stage full-parameter supervised fine-tuning (SFT) followed by reinforcement learning (RL). The first SFT stage performs broad adaptation on a large-scale, medium-quality corpus, while the second refines the model on a smaller, higher-quality set. The RL stage further optimizes the model with Group Relative Policy Optimization (GRPO). During development, we use an LLM-based, rubric-guided contrastive method that scores candidates by dimension and aggregates an overall score to assess translation quality. Experiments show that Qingqiu-MT-9B demonstrates strong generalization and competitiveness among similarly sized models. We release the model at Qingqiu-MT-9B.

B.17 ReMova (Pilinka et al., 2026)

This paper presents a Belarusian-specific data-cleaning pipeline and fine-tuning for English→Belarusian machine translation. Our cleaning pipeline distinguishes itself from others by employing a correction tool that addresses the issue of the two orthographies of the Belarusian language, noise in the training data, interference from other languages, and other misspelling issues common in Belarusian on the internet. A matched ablation on unfiltered training data shows substantial benefits from filtering for all fine-tuned models, with the LLM-based models gaining roughly twice as much from filtering as the dedicated encoder-decoder MT system, supporting the view that for Belarusian MT one of the primary bottlenecks is data quality. The model is available on HuggingFace: ReMova-TranslateGemma-en-be.

B.18 RFMT (Matsuda et al., 2026)

We participated in the constrained (open-weight) track of the WMT 2026 General MT Task for the English→Japanese and Simplified Chinese→Japanese directions. Our system builds on Marco-MT-Algharb, a strong open-weight translation model from the previous WMT General MT task. We further fine-tuned this model to generate more natural Japanese translations. To this end, we constructed a Context-Aware JApanese Linguistic Acceptability dataset (CAJALA). During annotation, human annotators were shown multiple paraphrased variants of a segment within its document-level context and asked to select

the most natural one. We then back-translated the CAJALA dataset to create parallel data for Direct Preference Optimization (DPO). We also trained the model to perform fill-in-the-middle (FIM) translation, in which the model reconstructs target-side segments from their surrounding document. At inference, we use Minimum Bayes Risk (MBR) decoding, followed by selective FIM post-editing of low-quality segments. The model is available on HuggingFace: RFMT.

B.19 SalamandraTA (Mash et al., 2026)

We present the BSC submission to the WMT26 General Machine Translation shared task: a 7B decoder-only model covering 23 directions, adapted from the SALAMANDRA family through vocabulary replacement, two rounds of continued pre-training (monolingual, followed by instruction-formatted parallel data), supervised fine-tuning on a translation-only instruction mixture, and preference optimisation applied selectively to six weak directions via CPO-SimPO with LoRA adapters. Submissions are decoded with Minimum Bayes Risk selection over a candidate pool filtered for truncation, degeneracy and off-target script. As official human evaluation was unavailable at the time of writing, we report a stage-wise internal evaluation on FLORES+ and WMT24++, together with automatic scores for the submitted system on the 12 test-set directions for which references have been released. Instruction tuning accounts for almost all measurable gain, but its size differs by an order of magnitude between the two evaluation sets; preference optimisation moves automatic metrics by less than the drift among untreated directions. The model is released under a Research-Only Licence and is available on request.

B.20 SCIR-TG-MT (Cao et al., 2026)

This paper presents SCIR-TG-MT, our primary submission (ID 44) to the unconstrained track of the WMT 2026 General Machine Translation shared task. The system covers ten translation directions using DeepSeek-V4-Pro and STRIDE (Structured Translation with Risk-Informed Directional Execution), an OpenCode translation skill. STRIDE converts source-side evidence and task instructions into a translation contract, an execution route, generation constraints, and selected checks. On five WMT25 development directions, STRIDE raises average SacreBLEU from 31.24 to 31.80 and COMET-22 from 83.61 to 84.17. A cumulative ablation on three directions separates the effects of routing, controls, and validation.

B.21 SRPOL (Dobrowolski et al., 2026)

We propose a method to improve machine translation quality by ensembling multiple large language models within a multi-agent framework. The method addresses the problem of ensembling models with different tokenizations by abstracting away from token-level representations and operating at the level of generated text. The approach integrates several established concepts in MT, including ensembling, quality estimation, and reranking into one translation system. The log-likelihood reranking applied in this paper significantly improves results, even without ensembling. Experimental results show improvements comparable to COMET-MBR.

B.22 TartuNLP/Tähetorn (Fishel and Yankovskaya, 2026)

We present an open-weight translation model for Estonian and its low-resource "relatives" in the Finno-Ugric language family. The training data includes 12 languages paired with Estonian as well as 23 more Finno-Ugric languages and varieties. The translation model use Unbabel Tower+ 9B as its starting point. The model is available on HuggingFace: Tahetorn9B.

B.23 TransCore (Yu et al., 2026)

This paper describes TransCore, the NewTranx submission to the WMT26 General Machine Translation Shared Task constrained track. Our system is built upon the Qwen3-8B large language model and adopts a two-stage supervised fine-tuning framework trained exclusively on publicly available data. Our approach focuses on high-quality multilingual data construction, including parallel corpus construction, knowledge-distilled corpus construction, and instruction-following data construction. For the parallel corpus, we construct a high-quality multilingual training corpus through multilingual data collection and automatic quality filtering. For knowledge distillation, we construct a KD corpus from publicly

available monolingual resources using a large language model, followed by automatic quality filtering. To support the instruction-following evaluation introduced in WMT26, we further construct instruction data covering immutable entity preservation, document-level translation, and structured translation, which are incorporated into the first-stage supervised fine-tuning. Our system participates in four language directions of the WMT26 General Machine Translation Shared Task: English→Indonesian, English→Kazakh, English→Thai, and Chinese→Japanese.

The models are available on ModelScope: TransCore-EnId, TransCore-EnKk, TransCore-EnTh, TransCore-ZhJa

B.24 TRIVE (Yang et al., 2026)

For the WMT26 General Machine Translation Task (Unconstrained Track), we propose TRIVE (Tri-Fusion Verified Translation Agent), a model-agnostic multi-agent translation framework. TRIVE organizes translation as a closed loop of Generation–Fusion–Verification–Correction: it first generates candidates in parallel under three complementary strategies—Faithful, Fluent, and Free—then synthesizes a draft that balances semantic fidelity and natural expression, and finally iteratively refines the output via rule-based validation and large language model evaluation. Because each module is implemented primarily via prompting, TRIVE can be directly overlaid onto diverse backbone models without task-specific fine-tuning. Experiments show that overlaying TRIVE on multiple backbones consistently improves instruction-following translation quality. Finally, we submitted Gemini-3.5-Flash + TRIVE as our primary system to WMT26.

B.25 UvA-MT (Wu et al., 2026)

This paper presents UvA-MT's submission to the WMT 2026 General Machine Translation shared task, competing in the unconstrained track across all 23 translation directions. This year, we fully leverage the test-time methods of Large Language Models (LLMs) for machine translation. Specifically, (1) we use sequential sampling with a refinement prompt to generate a pool of translation candidates for each source sentence, a strategy shown to outperform traditional parallel sampling; and (2) we employ LLMs as pairwise judges to select the best candidates via a round-robin voting mechanism, with offline evaluation on the WMT25 benchmark showing a very clear improvement over point-wise best-of-N selection. Lastly, we simply apply these two strategies for the submission of WMT26.

B.26 Wayfinder (Shao et al., 2026)

This paper presents Wayfinder, our submission system for the WMT26 General Machine Translation (GenMT) shared task. GenMT evaluates document-level translation under natural-language instructions, where structural preservation, formatting constraints, and output cleanliness are part of translation quality. Rather than training a new translation model, Wayfinder treats four API-based system outputs as a document-level candidate pool and produces one final hypothesis through a four-stage agentic pipeline. It aggregates source, instruction, metadata, candidates, and a DeepSeek-V4-Pro ranking-then-scoring signal; applies an anonymized pick-or-rewrite agent; and runs a conservative postcheck for severe objective defects. During development, a judge-guided loop compares outputs with a GPT-5.5 baseline, analyzes loss cases, updates prompts and repair constraints, and selectively reprocesses difficult documents. In automatic pairwise evaluation against GPT-5.5, Wayfinder obtains a pooled document-level weighted win rate of 0.6537 and remains above parity in every evaluated direction. The results suggest that candidate-pool translation, constrained agentic selection, conservative repair, and selective judge-guided iteration are practical components for instruction-conditioned shared-task MT.

B.27 Yandex and Yandex-8B (Popov et al., 2026)

We describe Yandex and Yandex-8B, our unconstrained and constrained submissions to the WMT26 General Machine Translation task. Our previous WMT25 submission, a 7B translation specialist, shows strong Fluency but comparatively weaker Accuracy. To combine its translation behavior with the capacity of a much larger model without replaying the full 7B training pipeline at 235B scale, we add two forms of supervision to the instruction-tuning data of an in-house 235B pretrained model: direct translation

targets generated by the 7B specialist and in-house translation-related tasks. The resulting Yandex model improves both Accuracy and Fluency and reduces the LLM-based error score relative to the same 235B model tuned only on general instructions. Relative to the 7B specialist, it improves Accuracy and reduces the error score, while their Fluency point estimates remain similar. Yandex then provides hard targets for Yandex-8B, which is trained for English→Russian, English→Belarusian, English→Kazakh, and English→Armenian using direct and instruction-conditioned translation data. Across four directions, Yandex-8B obtains the highest paragraph-level ChrF++ and the highest Accuracy and Fluency under our LLM-based evaluation among five public baselines. The constrained model is available on HuggingFace: Yandex-8B.