Findings of the WMT25 Multilingual Instruction Shared Task: Persistent Hurdles in Reasoning, Generation, and Evaluation

Tom Kocmi
Cohere

Ekaterina Artemova
Toloka AI

Eleftherios Avramidis
DFKI

Eleftheria Briakou
Google

Pinzhen Chen
University of Edinburgh

Marzieh Fadaee
Cohere Labs

Markus Freitag
Google

Roman Grundkiewicz
Microsoft

Yupeng Hou
UC San Diego

Philipp Koehn
JHU

Julia Kreutzer
Cohere Labs

Saab Mansour
Amazon

Stefano Perrella
Sapienza University

Lorenzo Proietti
Sapienza University

Parker Riley
Google

Eduardo Sánchez
Meta

Patrícia Schmidtová
Charles University

Mariya Shmatova
Toloka AI

Vilém Zouhar
ETH Zurich

Abstract

The WMT25 Multilingual Instruction Shared Task (MIST) introduces a benchmark to evaluate large language models (LLMs) across 30 languages. The benchmark covers five types of problems: machine translation, linguistic reasoning, open-ended generation, cross-lingual summarization, and LLM-as-a-judge. We provide automatic evaluation and collect human annotations, which highlight the limitations of automatic evaluation and allow further research into metric meta-evaluation. We run on our benchmark a diverse set of open- and closed-weight LLMs, providing a broad assessment of the multilingual capabilities of current LLMs. Results highlight substantial variation across sub-tasks and languages, revealing persistent challenges in reasoning, cross-lingual generation, and evaluation reliability. This work establishes a standardized framework for measuring future progress in multilingual LLM development.

1 Introduction

We are witnessing rapid development of multilingual large language models (LLMs). However, as pointed out by recent works (Kreutzer et al., 2025; Wu et al., 2025; Cruz Blandón et al., 2025), multilingual benchmarks lack comprehensiveness, scientific rigor, and consistent adoption across research labs, undermining their value in guiding multilingual LLM development. Among common problems are benchmark contamination (Ahuja et al., 2024), label noise (Chalamalasetti et al., 2025), reliance on non-native (machine-)translated instances (Chen et al., 2024b), and inconsistent evaluation pipelines. For instance, some leading LLM descriptions report multilinguality solely through translated MMLU. There is a mist surrounding multilingual evaluation that we aim to see through with this year's MIST shared task.

We introduce a novel multilingual evaluation benchmark that systematically assesses several key capabilities of LLMs across 30 diverse languages using the following sub-tasks:

- **Machine Translation (MT)**: A standardized, well-defined cross-lingual task.
- **Linguistic Reasoning (LR)**: Structured linguistic problem solving in multiple languages.
- **Open-Ended Generation (OEG)**: Using localized open-ended questions to assess language proficiency instead of specific capabilities.
- **Cross-lingual Summarization (XLSum)**: Synthesizing multilingual content from multiple documents written in different languages.
- **LLM-as-a-Judge**: Testing the effectiveness of LLMs in evaluating the quality of outputs in other sub-tasks that do not have definitive answers (MT, OEG, and XLSum).

We benchmark several of the most commonly used open- and closed-weight systems on our tests. These tests provide a multi-faceted evaluation framework that highlights the strengths and limi-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 108–129
November 8–9, 2025. ©2025 Association for Computational Linguistics

Linguistic Reasoning	Here are some word combinations in Hadza and their English translations: 1. chutisa zzokwanako: the giraffe's neck 2. athuitcha slimibii: the men's axe (for collecting honey) [. . .] Translate into Hadza: the male impalas' horns
Open-Ended Generation	As a news reporter, write an article about the opening of a new shopping complex, including who will enjoy it and what activities are available.
Cross-lingual Summarization	Fass bitte diese 6 Bewertungen eines Produkts auf Amazon auf Deutsch zusammen. Fleetwood Mack är som de är. Sköna att lyssna på. I am super pleased with my purchase and would order from this seller again. [...]
Machine Translation	You are a professional Czech-to-Ukrainian translator, tasked with providing translations for use in Ukraine. [...]
LLM-as-a-judge	Score the response generated by a system to a user's request in Lithuanian on a Likert scale from 1 to 7. The quality levels associated with numerical scores are provided below: [...]

Table 1: Example prompts for each sub-task.

	Langs	**Samples per lang**
Machine Translation	30	384
Linguistic Reasoning	15	90
Open Ended Generation	20	100
Cross-lingual Summarization	14	350
LLM-as-a-judge MT	16	1520
LLM-as-a-judge OEG	10	2256
LLM-as-a-judge XLSum	14	3200

Table 2: Number of languages and the number of samples (prompts) for each language or language pair in case of MT.

tations of current LLMs across diverse linguistic phenomena while drawing on the rigorous principles established within the MT evaluation research.

In addition to automatic metrics, we conduct human evaluation for all sub-tasks without definitive answers, which is then used to assess LLM-as-a-judge systems. Test sets, system outputs, and human judgments are released with a permissive license.[1]

2 Data and Methodology

In this section, we describe the datasets and preparation steps used for each of the sub-tasks in our benchmark. For every sub-task, we curated or adapted data across up to 30 languages. The high-level statistics are in Table 2. The following sections detail the sources of the data, the translation or localization processes applied, and any additional filtering or validation steps specific to each sub-task.

2.1 Linguistic Reasoning

The data for the linguistic reasoning sub-task were sourced from the 2024 International Linguistics Olympiad (IOL). In this olympiad, high school students compete in solving linguistic puzzles. Problems and solutions are released online and manually translated into the participants' languages. Previous benchmarks built from previous linguistics olympiads in English, such as Linguini (Sánchez et al., 2024) and LINGOLY (Bean et al., 2024), have shown that this type of puzzle is challenging even for the best LLMs. To evaluate multilinguality, we propose a multilingual version of this problem. This enables us to not only benchmark LLMs on challenging, unseen problems but also measure language disparities. The key to these puzzles is not retrieving acquired knowledge, but rather applying reasoning.

From problem PDFs to evaluation prompts IOL problems and solutions are published as PDFs under the CC-BY-SA 4.0 license.[2] They are typeset in the same LaTeX, format for all languages, which motivates our approach of tuning an automatic extraction for English, and then transferring it to the other languages. First, we manually extract questions and solutions from the PDFs for the five tasks of IOL (languages are Koryak, Hadza, Komnzo, Dâw, and Yanyuwa), which includes breaking tasks into sub-tasks (e.g. turning a matching task with four phrases to match across languages into four individual tasks), capturing metadata such as task authors, unifying task formulations and formats across task types. Then, we prompt an LLM to repeat this process for the other languages.[3] Last, with the help of human annotators that are proficient in the respective languages, we fix any errors, post-edit for cross-task consistency and translate task-level instructions.[4] This yields a total of

[1] github.com/wmt-conference/wmt-mist

[2] ioling.org

[3] A detailed description of this process is in Appendix A.

[4] Early attempts to translate tasks automatically rather than relying on parsing the language-specific solutions, quickly showed that automatic translation is not well-equipped for this task (at least not out-of-the-box) because the task involves disambiguating many single-word terms without much context (e.g. the word " roast" could be translated as noun or verb and in a literal or abstract sense), and special handling of

Grounding	localized	46
	generic	54
Type	brainstorming	23
	creative	35
	informational	25
	professional	17
Available Locale	ar_EG, bn_BD, cs_CZ, de_DE, el_GR, en_US, fa_IR, hi_IN, is_IS, id_ID, it_IT, ja_JP, kn_IN, ko_KR, ro_RO, ru_RU, sr_RS (both Latin and Cyrillic), uk_UA, zh_CN.	

Table 3: A breakdown of the 100 test questions in the open-ended generation sub-task.

90 prompts per language, covering five task types (classification (4), editing (1), fill-in-blanks (20), mapping (24) and translation (41)) for 15 languages (Chinese, Czech, Dutch, English, Estonian, French, German, Japanese, Korean, Persian, Portuguese, Russian, Spanish, Swedish, Ukrainian). Languages were chosen based on overlap with the 30 languages from the General MT task. For the final evaluation prompts, we add a simple instruction for context and answer format to each puzzle (e.g. English: "Solve the following linguistic puzzle with the help of the given context. The last line of your response should only contain the solution within square brackets [], nothing else."). We chose not to explicitly prompt the models for reasoning in order to avoid introducing any reasoning instruction bias and favoring models that are explicitly trained for reasoning. As a result, we can analyze how much each model tends to reason about each of these, but without having any expectations on the correctness, form, or volume of reasoning traces.

2.2 Open-Ended Generation

In this sub-task, we test multilingual language proficiency, e.g. generating native-sounding, useful, and coherent responses. Below a language's surface form are culture, values, and knowledge, so we also want to test LLMs' true ability grounded in the use of each language. The core motivation behind this is that LLMs sound native in English, but their responses in other languages are non-natural, contain English phenomena, or sound robotic (Guo et al., 2025). While some open-ended generation test sets exist, e.g. mArenaHard (Cohere Labs et al., 2024), they are often translated from English (Chen et al., 2024b) and skewed towards narrow domains like coding and math, which are not typical multilingual LLM use cases. Therefore, we focus on building a test set that asks native open-ended questions in many different domains, rather than specific tasks, e.g. writing a news article about a topic.

We prepared 100 questions manually with the help of LLMs, localized them into different languages, and asked native speakers to post-edit them to make them more natural and native. As a result, this multilingual test set contains comparable questions localized into each locale (language and country/region). The details of the process for question creation and localization are as follows.

English question creation First, we obtained a set of 100 English questions via two complementary workflows:

1. Three authors of this paper wrote a small pool of diverse questions.
2. We iteratively fed five randomly selected human-authored questions to two LLMs (GPT-4.1-mini and Command A), asking for a new question.
3. Then we manually inspected and post-edited these questions while mixing them with the original human-written questions.

To ensure each question's applicability to multiple locales, all locale-specific mentions stay as placeholders, e.g., using "`{language}`" instead of "English" in prompts like "Please suggest an idiom in `{language}`".

Localization and quality control We localized the English questions into 19 more unique language-writing script combinations, each of which is designated a country too, to better ground questions in locales. A full list of locales is available in Table 3, and the five-step process is detailed below:

1. Localization: We used four LLMs[5] to localize the questions and replace placeholders with locale-specific content, yielding four candidate variants per question.
2. Baseline: We also generated a reference translation for each question using Google Translate.
3. Sanity Check: To prevent LLMs from answering the question rather than faithfully localizing it, using the Google Translate version as a reference, we discarded any model variant that has a chrF score below 30 or exceeds the baseline

grammatical annotations such as for singular and plural.

[5] DeepSeek V3, Gemini 2.5 Pro, Command A, GPT 4.1

length by more than 50% per NLLB-200 tokenization.

4. Selection: From the remaining variants, we discarded the lowest-scoring chrF candidate, then randomly selected a variant translation for inclusion. If no variant passed filtering, we defaulted to the Google Translate baseline.
5. Human Inspection: we conducted a review and applied post edits if necessary for all languages to minimize non-nativeness and translationese.

Nature of the questions In Table 3, we present a breakdown of the types of test questions and expected responses. By counting placeholders in the seed English questions, we find that 46 questions explicitly mention a language/country-specific entity (i.e., locale-grounded), and 54 questions are more generic. Using Gemini 2.5 Pro followed by human inspection, we classified the nature of the expected responses into one of "brainstorming", "creative", "informational", or "professional". It is worth noting that while we assigned only one label to each question, the labels are not strictly mutually exclusive.

2.3 Cross-lingual Summarization

Our cross-lingual summarization dataset combines multilingual review data from two complementary sources: Amazon product reviews and Google Maps restaurant reviews. The dataset construction process involved systematic sampling, language balancing, and content filtering to ensure high-quality cross-lingual evaluation data.

Data collection We integrated data from two distinct domains to maximize linguistic diversity: Amazon product reviews for consumer products, and Google Maps restaurant reviews for restaurants This resulted in an initial scraped dataset of 12,040 reviews spanning 853 products and restaurants. Each data item was paired with a product or restaurant-specific summarization prompt in 14 target languages: Arabic (Egyptian), Czech, Chinese (Simplified), French, German, Hindi, Indonesian, Italian, Japanese, Korean, Russian, Spanish, Swedish, and Turkish. The summarization prompt instructions were created by translating the original English summarization prompt into all target languages, with all translations checked by proficient speakers of each respective language to ensure linguistic accuracy and cultural appropriateness.

Content filtering and quality control We applied comprehensive filtering criteria to ensure high-quality multilingual content suitable for cross-lingual evaluation:

- Language-based filtering: Using language identification[6], we omitted reviews in languages not covered by the sub-task and retained only products/venues with reviews in more than one languages.
- Content length filtering: Reviews shorter than 50 characters (Amazon) or 20 characters (Google Maps) were removed as non-informative. We applied IQR-based outlier removal per language to eliminate excessively long individual reviews, while enforcing a 1,500-character limit on the final merged multi-document input for manageable human evaluation.
- Language pair balancing: We removed over-represented language combinations to maintain dataset balance and promote multilingual scenarios. We implemented a mixed-content counting algorithm that handles both alphabetic and logographic writing systems appropriately.

Balanced sampling To ensure equal representation of each target language while maximizing data diversity, we implemented a two-stage sampling approach, which first maximizes coverage across unique data items, then achieves exactly 350 examples per target language (4,900 total examples). We prioritized examples without English input to promote true cross-lingual scenarios for less-explored languages.

Data characteristics The final dataset contains 1.1M words across all examples, with an average of 230 words per example. it exhibits strong cross-lingual properties: 86.3% of examples require summarization in a target language different from any of the input languages, and 46.8% contain no English in the source reviews. The dataset comprises 66.0% Google Maps restaurant reviews (3,232 examples) and 34.0% Amazon product reviews (1,668 examples).

2.4 Machine Translation

The MT sub-task adopts the WMT25 General MT test set; full details on data sourcing, difficulty sampling, and human references collection are documented in Kocmi et al. (2025a).

[6] github.com/saffsd/langid.py

Sources and domains Source documents were collected across six domains (news, social, speech, literary, educational, dialogue) and three source languages (Czech, English, Japanese). Speech includes source audio with ASR transcripts, and social includes thread screenshots, with the objective of looking at some of the impacts of multimodal translation. The focus is on the most recent data possible to minimize potential overlap with the pretraining and fine-tuning data of the models under evaluation. All source texts were originally authored in the source language. 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). To balance domains and source languages, for each domain and source language combination, we targeted ∼9k words and 60–100 segments, with an average segment length of ∼100 words. This design enables the micro-averaging of results across languages and domains without any single category disproportionately influencing the final scores. However, there are some exceptions, as keeping these variables fixed was impractical. For example, the average segment length for the English and Japanese Speech data is 145.27 and 180.59 words, respectively, which is higher than the 100-word objective. Similarly, the dialogue domain's segments have an average length of 178.8 and 147.3 words, respectively. Comprehensive domain-specific collection procedures and final test set statistics are detailed in Kocmi et al. (2025a).

Translation instructions There is no standardized prompt instructions for WMT machine translation evaluation, various are used, from simple 'Translate into {target_lang}:' to more complex instructions adding additional instructions such as 'Your goal is to accurately convey the meaning and nuances of the original {source_lang} text while adhering to {target_lang} grammar, vocabulary, and cultural sensitivities.' (Deutsch et al., 2025).

For our use case, we extend the instruction to cover more details that human translators are asked for. Furthermore, we modify the instructions for each domain. Detailed prompt instructions are in Table 18.

2.5 LLM-as-a-judge for OEG and XLSum

LLM-as-a-judge has recently emerged as an automated solution to open-ended generation evaluation (Zheng et al., 2023b; Verga et al., 2024). It achieves high correlation with human judgment, but its efficacy for languages other than English remains little known (Son et al., 2024). To evaluate the capabilities of models to perform quality assessment of other LLM outputs, we set up the sub-tasks of LLM-as-a-judge for open-ended generation, cross-lingual summarization, and machine translation, where participating systems run evaluation on system outputs from those sub-tasks.

The LLM judges are given the same instructions provided to human annotators, and are assessed by computing their judgments' correlation to human judgment. To evaluate LLM-as-a-judge for the OEG and XLSum sub-tasks, we take all samples that are evaluated with human annotators and use a prompt instruction to judge the system output on a Likert scale of 1–7. For each system output, we run LLM-as-a-judge separately on different evaluation criteria, guided by a rubric each. Specifically:

- OEG: instruction following, naturalness, and coherence
- XLSum: faithfulness, coverage, naturalness, and coherence

The exact prompt instructions are provided in Appendix B. As human evaluation was available for only a subset of languages and systems, LLM-as-a-judge was tested on the same set of data.

2.6 LLM-as-a-judge for MT

Automatic machine translation evaluation is the catalyst of progress in translation technologies, offering a quick, low-cost signal of quality. Early metrics were string-matching against a reference, such as BLEU or ChrF (Papineni et al., 2002; Popović, 2015), which were replaced by trained metrics, such as COMET or MetricX (Rei et al., 2020; Juraska et al., 2023), and finally LLM-as-a-judge (Kocmi and Federmann, 2023). Even though each replacement increased the correlation with human judgment of translation quality, new concerns have emerged regarding language bias, robustness (Moghe et al., 2025; Zouhar et al., 2024a,b), and self-bias for evaluation (Wataoka et al., 2024; Zheng et al., 2023a; Stureborg et al., 2024). This meta-evaluation of automated metrics is usually handled by the WMT Metrics Shared Task (Lavie et al., 2025; Freitag et al., 2024).

In order to test the capabilities of models to perform as LLM-as-a-judge to judge machine translation, we adjust the GEMBA-DA (Kocmi and Fe-

Model and size	Technical report
AyaExpanse 8B	Cohere Labs et al. (2024)
Command R 7B	Cohere et al. (2025)
EuroLLM (9B)	Martins et al. (2025)
Gemma 3 (12B)	DeepMind et al. (2025)
Llama 3.1 (8B)	Grattafiori et al. (2024)
Mistral (7B)	Mistral et al. (2023)
Qwen 2.5 (7B)	Alibaba et al. (2024)
TowerPlus (9B)	Rei et al. (2025)
AyaExpanse 32B	Cohere Labs et al. (2024)
Claude 4 Sonnet	
Command A (111B)	Cohere et al. (2025)
DeepSeek V3 (671B)	DeepSeek et al. (2024)
EuroLLM (22B)	Martins et al. (2025)
Gemini 2.5 Pro	Google et al. (2025)
Gemma 3 (27B)	DeepMind et al. (2025)
GPT 4.1	
Llama 4 Maverick (400B)	
Mistral Medium	
Qwen3 (235B)	Alibaba et al. (2025)
TowerPlus (72B)	Rei et al. (2025)

Table 4: List of all LLMs evaluated in this work. Unshaded models represent "constrained" models, which are smaller and open weights in contrast to "unconstrained" which do not have any limits on being public or size.

dermann, 2023) prompt with the latest WMT25 human evaluation instruction. The exact prompt instruction is in the Appendix B.

3 Benchmarked Models

For this shared task, we defined two categories for model participation: constrained with several restrictions on model size and licensing; and unconstrained without any limitations. The same way as the General Machine Translation task (Kocmi et al., 2025a). Specifically, the constrained category is restricted to models with fewer than 20B parameters and requires that models be shared as open weights.

Unfortunately, our shared task did not obtain any (valid) participating systems. However, we collected and benchmarked outputs of popular models. The selection process was to identify the strongest-performing system per category for each of the popular model families. This approach ensured that both constrained and unconstrained models were consistently represented; the resulting model list thus reflects a broad yet balanced selection of models, enabling multilingual assessment of the current LLM landscape across languages and problems.

The list of all systems is in Table 4. During the output collection, we ran into budget and API throttling restrictions and thus could not collect some of the systems' outputs for all sub-tasks. When collecting outputs, we set the temperature to 0 and used a unified script.[7]

Model	LR	MT	OEG	XLSum
Gemini 2.5 Pro	100%	95%	94%	100%
GPT 4.1	85%	90%	100%	94%
DeepSeek V3	90%	80%	88%	71%
Claude 4	95%	78%	81%	88%
Mistral Medium	70%	75%	75%	82%
Llama 4 Maverick	80%	61%	50%	47%
Qwen3 235B	65%	63%	56%	41%
CommandA	75%	56%	62%	59%
Gemma 3 27B	60%	53%	69%	65%
Gemma 3 12B	55%	42%	44%	76%
AyaExpanse 32B	50%	31%	38%	53%
AyaExpanse 8B	30%	20%	31%	29%
Llama 3.1 8B	40%	17%	25%	18%
CommandR7B	20%	14%	19%	-
Qwen2.5 7B	35%	8%	12%	12%
Mistral 7B	5%	5%	6%	6%
TowerPlus 72B	45%	37%	-	35%
TowerPlus 9B	25%	27%	-	24%
EuroLLM 22B	15%	25%	-	-
EuroLLM 9B	10%	19%	-	-

Table 5: Aggregate results across four sub-tasks, converted into percentile ranking (100%=first).

4 Results

In this section, we present the results and key insights for each sub-task and benchmarked model. Although automatic evaluation was applied to all prompts and outputs, human evaluation was not conducted for all tasks, systems, or languages, due to budget constraints and annotator availability. Nonetheless, human evaluation often proved more reliable than automatic metrics, so we release all annotations for future work on meta-evaluation.

4.1 Linguistic Reasoning

In order to evaluate linguistic reasoning, we choose to break them as much as possible into tasks so that we can grade LLM answers as precisely as possible (which distinguishes this work from previous linguistic reasoning benchmarks). Depending on the task type, we choose either exact match (classification, mapping, fill-in-blanks, editing) or ChrF (translation) as a metric. The scores have to be taken with a grain of salt because ChrF is likely not perfectly expressing the degradations between useless and perfect translation. We assume it rather overestimates translation quality compared to IOL judges. Each task comes with a number of points ([0.5, 1.0, 1.5, 2.0, 2.5]), summing to 20

[7] github.com/wmt-conference/wmt-collect-translations

Model	**Average**	Spanish	Portuguese	English	French	German	Dutch	Average	Russian	Swedish	Japanese	Ukrainian	Korean	Czech	Estonian	Persian	Chinese
Gemini 2.5 Pro	36.3	40.3	38.1	38.3	39.9	37.3	37.5	36.3	39.5	35.5	35.4	33.9	40.9	36.8	32.5	29.4	28.8
Claude 4	29.7	33.8	35.5	24.6	32.9	33.8	27.9	29.7	30.8	28.4	29.8	32.1	26.3	29.3	27.0	26.3	26.2
DeepSeek V3	23.6	28.5	27.9	23.2	27.9	28.2	24.1	23.6	22.4	21.8	23.1	22.9	20.4	21.5	24.5	20.6	17.4
GPT 4.1	23.4	29.0	27.9	20.5	27.4	24.3	27.1	23.4	24.1	22.7	21.6	23.0	15.3	24.6	29.4	21.8	12.7
Llama 4 Maverick	22.9	30.5	27.2	22.4	26.5	25.9	23.2	22.9	24.6	20.8	20.2	24.5	20.2	22.1	19.4	21.3	14.2
CommandA	19.8	22.0	21.1	17.8	20.6	18.8	23.2	19.8	18.8	17.8	21.3	20.8	21.8	20.1	18.8	18.2	15.7
Mistral Medium	19.8	25.8	23.5	20.4	21.2	24.9	20.8	19.8	21.8	23.2	22.9	16.8	15.2	15.1	14.6	15.7	14.8
Qwen3 235B	17.6	19.9	22.6	22.0	19.1	20.9	21.1	17.6	14.6	21.5	16.0	18.6	17.8	13.0	13.4	14.3	9.6
Gemma 3 27B	17.0	17.1	17.1	18.4	18.3	18.1	17.8	17.0	14.2	19.9	20.0	16.7	18.2	12.5	17.0	14.8	15.0
Gemma 3 12B	16.5	15.6	18.7	21.1	17.3	17.2	18.9	16.5	12.3	15.8	15.9	17.6	12.8	13.3	17.9	16.6	16.0
AyaExpanse 32B	15.3	16.7	17.2	18.9	17.7	14.8	18.7	15.3	19.0	12.3	18.1	10.7	15.7	15.0	3.5	15.1	15.7
TowerPlus 72B	13.4	17.9	17.6	17.6	14.8	11.1	14.4	13.4	14.2	14.6	16.5	13.2	15.1	9.2	13.1	7.9	3.2
Llama 3.1 8B	10.8	14.0	15.5	14.7	16.1	13.1	13.2	10.8	11.1	11.3	10.2	6.0	6.6	6.7	6.4	7.1	10.1
Qwen2.5 7B	10.7	12.6	11.5	13.5	12.2	10.0	10.8	10.7	11.1	7.9	7.7	11.5	11.9	9.6	10.4	8.1	12.2
AyaExpanse 8B	8.7	10.4	13.2	13.5	10.7	10.9	11.8	8.7	7.1	7.4	7.4	7.0	8.4	8.8	1.8	4.4	8.1
TowerPlus 9B	8.5	13.9	6.0	13.8	8.6	13.5	9.8	8.5	7.1	6.8	6.5	3.8	8.8	5.5	9.0	8.0	6.1
CommandR7B	7.3	9.5	8.1	13.5	13.1	11.6	9.2	7.3	8.7	5.9	4.8	5.6	3.2	4.3	0.6	4.6	7.1
EuroLLM 22B	5.7	11.7	7.8	10.4	4.3	8.8	5.9	5.7	6.0	6.7	0.6	6.2	3.7	5.9	4.9	0.0	2.1
EuroLLM 9B	2.6	1.9	3.9	5.7	1.7	1.6	4.9	2.6	0.2	3.9	1.6	2.0	0.7	5.6	4.9	0.0	1.1
Mistral 7B	2.6	6.1	2.7	5.8	2.7	2.2	0.4	2.6	3.4	1.3	2.1	2.1	0.7	1.1	2.4	2.3	3.6

Table 6: Results (number of points) for the linguistic reasoning sub-task (LR) across languages.

points per task and 100 points in total. Points express difficulty, which is not the same across tasks, e.g. translation tasks typically give more points than mapping tasks. The final metric is the sum of prompt-level scores ([0–1]) multiplied by their points, such that the maximum attainable score for each language is 100. The final model ranking is determined by the average number of points across languages. The number of obtained points (out of 100) for each model and language is shown in Table 6. Below are our three key observations.

First, we note that the maximum score in a single language is 40.9 and the maximum average score is 36.3, **indicating headroom** for this kind of task overall. All models failed the majority of tasks. Due to the niche-ness of linguistic reasoning (as opposed to mathematical reasoning), it is unlikely that any of the models has seen very similar tasks during training, which lets this task measure generalization more than memorization. In the 2024 IOL, the winning participant scored 79[8] with human and not automatic scoring, but presented with the same tasks in their mother tongue. The top-scoring model here would have barely made it to a Bronze medal.

Second, the **model ranking is fairly consistent across languages in the top ranks**, with the leading model being Gemini 2.5 Pro across all languages, Claude 4 following in second place, and DeepSeek V3, GPT4.1, and Llama 4 Maverick alternating in place 3. As expected, model size also plays a major role in the ranking: closed-source (presumably large) LLMs are leading in the subtask, followed by CommandA and Qwen3 235B. Notably, Gemma 3 shows good multilingual reasoning performance, with its 27B and 12B versions outperforming TowerPlus at 72B and Aya Expanse at 32B. In the 7–9B range, Llama 3.1 8B is the best. Still, at this model size, we see a steep decline when moving from higher to lower-resource (or unsupported) languages, which is partially due to a lack of instruction following and failing to respond in the required answer format.

Third, most surprisingly, we find that **English is not the language that most models perform strongest in**, although it typically dominates reasoning tasks like math (Chen et al., 2024a). In fact, the "best" solution to the tasks was found by Gemini 2.5 Pro with Korean as the instruction language. In particular, the stronger models show surprising performance drops in English: For Claude, the top performance is 33.8 in German or Spanish, while English lags behind with 24.6 points, scoring the lowest across all languages. Overall, only Gemma 3 12B, Qwen2.5 7B, Aya Expanse 8B, CommandR7B, and EuroLLM 9B performed better in English than all other languages, and in these cases only with a small, perhaps negligible margin.

[8] ioling.org/results/2024

Model	**Average**	Naturalness	Instruction Following	Coherence
GPT 4.1	6.13	5.94	6.24	6.20
Gemini 2.5 Pro	6.09	5.80	6.25	6.22
DeepSeek V3	5.97	5.65	6.17	6.09
Claude 4	5.96	5.74	6.06	6.08
Mistral Medium	5.96	5.68	6.16	6.03
Gemma 3 27B	5.94	5.59	6.15	6.07
CommandA	5.93	5.65	6.12	6.03
Qwen3 235B	5.90	5.57	6.13	5.99
Llama 4 Maverick	5.89	5.73	6.02	5.93
Gemma 3 12B	5.87	5.57	6.10	5.95
AyaExpanse 32B	5.70	5.33	5.89	5.88
AyaExpanse 8B	5.53	5.10	5.73	5.75
Llama 3.1 8B	5.21	4.82	5.56	5.26
CommandR7B	5.20	4.77	5.47	5.38
Qwen2.5 7B	5.17	4.75	5.40	5.35
Mistral 7B	4.27	3.88	4.49	4.43

Table 7: Results in various rubrics for the open-ended generation task. The points are on a Likert-7 scale, where 7 is the best.

Model	**Average**	Naturalness	Faithfulness	Coherence	Coverage
Gemini 2.5 Pro	6.05	5.84	6.03	6.14	6.19
GPT 4.1	5.99	5.91	5.98	6.11	5.95
Claude 4	5.94	5.76	5.95	6.00	6.04
Mistral Medium	5.78	5.52	5.77	5.88	5.96
Gemma 3 12B	5.72	5.49	5.78	5.88	5.75
DeepSeek V3	5.72	5.36	5.80	5.81	5.93
Gemma 3 27B	5.72	5.55	5.77	5.81	5.75
CommandA	5.66	5.29	5.75	5.75	5.87
AyaExpanse 32B	5.63	5.59	5.61	5.81	5.53
Llama 4 Maverick	5.57	5.49	5.56	5.80	5.44
Qwen3 235B	5.56	5.26	5.55	5.69	5.75
TowerPlus 72B	5.41	5.06	5.50	5.49	5.59
AyaExpanse 8B	5.37	4.97	5.37	5.73	5.41
TowerPlus 9B	5.09	4.92	5.13	5.24	5.06
Llama 3.1 8B	4.52	4.20	4.59	4.67	4.62
Qwen2.5 7B	4.50	4.09	4.59	4.64	4.69
Mistral 7B	3.48	2.83	3.66	3.50	3.92

Table 8: Results in various rubrics for the cross-lingual summarization sub-task. The points are on Likert-7 scale where 7 is the maximum best. See per-language breakdown in Appendix Table 19.

Explanations for this could be that prompting in other languages brings up the context that is more favorable for solving linguistic reasoning tasks, or that it is just the lack of English dominance in task-relevant data that usually gives it an advantage for other tasks like math or knowledge retrieval. Another explanation could be that model uncertainty might generally be quite high, so that resampling within the same language could cause similar variance as the one we see across languages. We invite future work to dive further into these questions.

4.2 Open-Ended Generation

The open-ended generation sub-task is human-evaluated. We design a rubric to assess three aspects: instruction following, naturalness, and coherence. This rubric is given to both human evaluators and LLM judges.

We human-evaluate a subset of all OEG outputs: 16 systems, 10 languages, and the same 46 questions for all system-language combinations. This is because some questions led to an overly long response, and TowerPlus and EuroLLM models had very high failure rates.

Results are shown in Table 7, with models ranked by their average scores on naturalness, instruction following, and coherence. Three points stand out. First, proprietary models generally perform better, except for DeepSeek V3, which is a large open-source mix-of-expert model. Second, performance differences among the leading systems are narrow. Third, naturalness scores show a wider spread than instruction following or coherence, implying a larger gap between the strongest and weakest systems, and highlighting the limitations of systems to produce native sounding text that can be directly used.

4.3 Cross-lingual summarization

We perform human evaluation with rubrics in the same setup as in OEG. We specifically test for naturalness, faithfulness, coherence, and coverage.

We performed a human evaluation for all 14 target languages in the sub-task, however, Turkish and Swedish results were not yet available at the time of submission. Annotators were proficient in the target language and English but were not expected to speak any other language; therefore, we translated source reviews in all other languages to English using Gemini 2.5 Flash. The user interface allowed them to view the original phrasing of the reviews, if desired.

Three models were excluded from human evaluation for the following reasons: the two EuroLLM models frequently copied input summaries in source languages rather than summarizing them, and CommandR7B had an issue with outputting Polish rather than Czech.

Given the novelty and current lack of this type of problem in the field, we conducted the human evaluation for all 17 remaining systems that did not exhibit these evident issues. Due to budget constraints, we restricted the number of evaluated outputs to 18, resulting in 306 examples rated for each language. The samples were selected based on output diversity using BLEU as the metric. We anticipate that a diverse set of outputs with human ratings will help future efforts in validating automatic metrics for this problem.

We present our preliminary analysis based on 12 target languages in Table 8. When averaged across all target languages, closed-source models have an advantage, with Gemini 2.5 Pro being the best-performing system in the unconstrained track. However, open-weight models are not far behind, led by Gemma3-27B. Model size seems to matter with all of the constrained systems, except for Gemma 3 12B, which punches above its weight, consistently showing lower scores across all languages.

Most models performed well (average rating of 5 and higher) on German, French, Chinese, Italian, Russian, and Spanish. Japanese was the most challenging language. Egyptian Arabic was the most divisive language with 4 clusters, showing a clear advantage of Gemini 2.5 Pro, GPT 4.1, and Claude, with all other models having an average score below 5. Naturalness was the weakest aspect of the generated summaries, often suffering from the models not adhering to the specifically requested language or dialect, or containing untranslated quotes from the source documents.

4.4 Machine Translation

Automatic evaluation We evaluate the MT subtask across 31 language pairs and report AutoRank, a rank induced by automatic MT metrics where lower is better (1 is best). The AutoRank is a combination of five different metrics Kocmi et al. (2025a) from three distinct metric families:

- **LLM-as-a-Judge (reference-less).** We use GEMBA-ESA (Kocmi and Federmann, 2023) with two independent judges: GPT 4.1 [9] and CommandA (Cohere et al., 2025), both in a reference-less setting.
- **Trained Reference-based Metrics.** Two supervised metrics trained to approximate human quality judgments with references: MetricX-24-Hybrid-XL[10] (Juraska et al., 2024) and XCOMET-XL[11] (Guerreiro et al., 2024).
- **Trained Quality Estimation (QE).** The reference-less QE metric CometKiwi-XL[12] (Rei et al., 2023), which is also trained to mimic human judgments.

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

However, for the two lowest-resource languages in the test set (Bhojpuri and Maasai), we do not apply QE and instead rely solely on `chrF++` (Popović, 2017), computed with `sacrebleu` (Post, 2018). This approach was chosen because the reliability of our main metrics is unestablished for these languages (Falcão et al., 2024; Singh et al., 2024; Wang et al., 2024; Sindhujan et al., 2025), whereas human references required for `chrF++`[13] were available.

The system-level score for each language pair is the average of its paragraph-level (segment-level) scores from each metric across the test set.

Human evaluation The human evaluation is done by Kocmi et al. (2025a) using Error Span Annotation (ESA; Kocmi et al., 2024) and for English to Korean and Japanese to Chinese it relies on Multidimensional Quality Metrics (MQM; Lommel et al., 2014).

The ESA annotators are asked to mark each translation error as well as its severity, "Minor" or "Major". In addition, the annotators are also asked to assign a score from 0 to 100 to the entire annotation segment (usually a paragraph).

In the MQM, annotators are asked to assign categories and subcategories to all error spans. Then,

[9] openai.com/index/gpt-4-1

[10] huggingface.co/google/metricx-24-hybrid-xl-v2p6

[11] huggingface.co/Unbabel/XCOMET-XL

[12] huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

[13] SacreBLEU signature: `nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.5.1`.

Model	Avg. MT AutoRank
Gemini 2.5 Pro	1.02
GPT 4.1	1.51
DeepSeek V3	2.62
Claude 4	2.86
Mistral Medium	3.10
Qwen3 235B	4.10
Llama 4 Maverick	4.34
Gemma 3 27B	4.55
CommandA	4.68
Gemma 3 12B	6.05
TowerPlus 72B	7.00
AyaExpanse 32B	7.32
TowerPlus 9B	8.31
EuroLLM 22B	9.22
AyaExpanse 8B	9.99
EuroLLM 9B	10.60
Llama 3.1 8B	11.81
CommandR7B	11.98
Qwen2.5 7B	14.61
Mistral 7B	18.57

Table 9: Average MT AUTORANK results across language pairs (lower is better). For fairness, all model averages are computed over the same 27 of 31 language pairs, matching Mistral Medium, which lacks outputs for four pairs (see Table 20).

instead of a 0 to 100 slider, the final score is calculated as a sum of error severities, where minor error equals -1 and major error equals -5.

Overall ranking Table 9 reports the average AUTORANK results across the various language pairs. **Gemini 2.5 Pro** leads with an average AUTORANK of **1.02**, followed by **GPT 4.1** (1.51), **DeepSeek V3** (2.62), and **Claude 4** (2.86). This top cluster is clearly separated from a mid-tier (4–6 average ranks; e.g., Qwen3 235B, Llama 4 Maverick, CommandA, Gemma 3 27B) and from compact open-weight models which concentrate above 7–8 on average. **Mistral Medium** remains competitive (3.10), but translating for fewer language pairs than all the other models (27 vs. 31). At the other end, small open-weight baselines (e.g., Qwen2.5-7B, Mistral-7B) cluster around ranks 15–18.

Human evaluation results are in Table 21; due to budget restrictions, not all systems have been evaluated. The overall picture highlights the AUTORANK results. However, we can already see some significant differences showing the limitation of automatic metrics: there is a significant drop in the English to Egyptian Arabic as LLMs mostly output the modern standard Arabic, and DeepSeek significantly underperforms in Serbian, which was not visible on AUTORANK.

Language-pair effects Table 20 in Appendix E reports the fine-grained AUTORANK results across the 31 language pairs. The fine-grained table reveals two consistent trends: (i) High-resource or typologically close directions (e.g., English→German, English→Italian, Japanese→Chinese) yield tight spreads among the strongest systems, often near ranks 1–3. (ii) Low-resource and/or orthography-sensitive directions are much harder. In particular, English→Maasai and English→Bhojpuri show large rank dispersion. Some leaders stay robust (e.g., Gemini 2.5 Pro), while others drop sharply on these pairs (e.g., GPT 4.1 on English→Maasai).

Open vs. closed trends Closed-weight models dominate the top cluster, but **DeepSeek V3** stands out as an open-source mix-of-expert model that competes closely with them. Among mid-sized open models, quality is uneven across language pairs and degrades most on low-resource or script-variant directions.

Relation to other tasks The qualitative picture resembles the pattern in Section 4.1: a tight group of leaders at the top, followed by a broader middle where performance varies more by condition. In MT, the key conditions are the choice of language pairs (especially low-resource and script variants), which ultimately drive the gaps we observe in AUTORANK.

4.5 LLM-as-a-judge for OEG and XLSum

Meta-evaluation of LLM-as-a-judge against humans is a research question in itself. Various correlation techniques are used, e.g., Cohen's Kappa, Kendall Tau, Pearson's, or Spearman's correlations (Liu et al., 2023; Verga et al., 2024). Meta-evaluation in machine translation highlighted many problems of common correlation metrics, such as how handling of ties affects the correlation (Deutsch et al., 2023), how critical grouping of items under Kendall Tau is (Perrella et al., 2024), or why Pearson's correlation may be misleading (Mathur et al., 2020). Thus, we build on top of the MT meta-evaluation research, following the best practices (Freitag et al., 2024).

We anticipate almost no ties in system ranking when all scores are aggregated at the system level, but a large number of ties at the instance level due to our use of rubric scores. Our preliminary data inspection also supports this. We compute two

	Pairwise Accuracy	group-by-item acc_{eq}			
		Average	Naturalness	Instruction Following	Coherence
Claude 4	0.95	0.56	0.55	0.57	0.56
GPT 4.1	0.95	0.57	0.54	0.59	0.59
CommandA	0.93	0.57	0.53	0.59	0.59
Qwen3 235B	0.91	0.57	0.53	0.59	0.59
Mistral Medium	0.88	0.55	0.54	0.55	0.56
DeepSeek V3	0.85	0.54	0.50	0.58	0.53
Llama 4 Maverick	0.83	0.53	0.51	0.52	0.56
AyaExpanse 32B	0.73	0.50	0.47	0.53	0.51
Qwen2.5 7B	0.70	0.48	0.46	0.49	0.48
Llama 3.1 8B	0.63	0.44	0.40	0.44	0.47
CommandR7B	0.62	0.49	0.44	0.52	0.51
AyaExpanse 8B	0.58	0.48	0.44	0.51	0.48
Mistral 7B	0.55	0.45	0.40	0.50	0.45

Table 10: System-level (Pairwise Accuracy) and Segment-level (group-by-item acc_{eq} by evaluation criterion) correlation between LLM-as-a-judge and human judgment for OEG.

types of correlations at the system and instance levels:

- **Pairwise Accuracy**: pairwise accuracy between system ranking and human ranking, neglecting ties.
- $\textbf{acc}_{\textbf{eq}}$: group-by-item pairwise accuracy with ties, then averaged across all items, as introduced by Deutsch et al. (2023). Without losing generality across all sub-tasks, an "item" refers to an input prompt requiring an output in a specific language. We report the results for each evaluation criterion separately, as well as an overall average.

Results for OEG LLM-as-a-judge Table 10 shows both system-level and instance-level accuracy measures between LLM judgment and human judgment. Regarding system ranking pairwise accuracy, the models are roughly split into two groups: LLMs with more than 100B parameters, including both closed-source and open-source ones, achieve high accuracy; small open-source models perform worse, with the lowest performance close to a random toss of a coin of 50%.

Instance-level acc_{eq} scores display a similar overall trend, but top-performing LLM judges are closer to each other. We see that the then top Claude 4 becomes lower than GPT 4.1, CommandA, or Qwen3 235B. The best LLM judge for each criterion also varies: Claude 4 is the best at judging naturalness, but three LLMs, GPT 4.1, CommandA, and Qwen3 235B, achieve the best individual accuracy for judging instruction following and coherence.

	Pairwise Accuracy	group-by-item acc_{eq}				
		Average	Naturalness	Faithfulness	Coverage	Coherence
CommandA	0.91	0.50	0.52	0.45	0.50	0.51
GPT 4.1	0.91	0.51	0.51	0.51	0.47	0.54
Llama 4 Maverick	0.89	0.45	0.48	0.41	0.42	0.47
Mistral Medium	0.89	0.49	0.50	0.44	0.50	0.52
Qwen3 235B	0.89	0.49	0.50	0.48	0.48	0.51
DeepSeek V3	0.87	0.47	0.49	0.46	0.46	0.47
CommandR7B	0.80	0.39	0.36	0.39	0.40	0.41
AyaExpanse 32B	0.78	0.40	0.38	0.39	0.42	0.42
Qwen2.5 7B	0.78	0.39	0.37	0.37	0.40	0.43
AyaExpanse 8B	0.76	0.38	0.35	0.39	0.38	0.41
Llama 3.1 8B	0.72	0.39	0.38	0.35	0.39	0.42
Mistral 7B	0.71	0.37	0.33	0.35	0.40	0.40

Table 11: System-level (Pairwise Accuracy) and Segment-level (group-by-item acc_{eq} by evaluation criterion) correlation between LLM-as-a-judge and human judgment for XLSum.

Results for XLSum LLM-as-a-judge Table 11 shows both system-level and instance-level accuracy measures between LLM judgment and human judgment. At the system level, pairwise accuracy follows a pattern similar to OEG: larger models (CommandA, GPT 4.1, and the 100B+ parameter models) achieve high accuracy between 0.87 and 0.91, while smaller open-source models below 10B parameters perform substantially worse, with accuracies between 0.71 and 0.80.

However, instance-level acc_{eq} scores reveal more concerning patterns. Overall correlations are lower than in OEG, with the best average scores around 0.50. GPT 4.1 demonstrates particularly severe overscoring tendencies, systematically assigning perfect scores to almost all outputs of 4–9 models across all criteria. The best-performing judge also varies considerably by criterion: CommandA achieves the highest accuracy for naturalness and coverage, while GPT 4.1 performs best on faithfulness and coherence despite its overscoring behavior. These patterns suggest that the system-level correlations may reflect spurious text properties rather than the intended evaluation criteria, raising questions about the validity of LLM-as-a-judge for this task.

Model	Avg. SPA	Avg. acc$_{eq}$
GPT 4.1	0.83	0.49
Claude 4	0.82	0.36
CommandA	0.80	0.39
DeepSeek V3	0.79	0.37
Qwen3 235B	0.78	0.38
AyaExpanse 32B	0.73	0.28
Llama 4 Maverick	0.72	0.19
Qwen2.5 7B	0.67	0.36
Llama 3.1 8B	0.66	0.28
CommandR7B	0.58	0.26
AyaExpanse 8B	0.58	0.22
Mistral 7B	0.54	0.29

Table 12: System-level (Pairwise Accuracy) and Segment-level (acc$_{eq}$) correlation between LLM-as-a-judge and human judgment for machine translation. Correlations have been averaged across translation directions. Full results are reported in Tables 22 and 23.

4.6 LLM-as-a-judge for MT

The meta-evaluation of LLM-as-a-judge was collected in the same way as this year's metric shared task Lavie et al. (2025). Correlations are computed at the system level using Pairwise Accuracy (PA, Kocmi et al., 2021) and at the segment level using Pairwise Accuracy with Tie Calibration (acc$_{eq}$, Deutsch et al., 2023).

We report the average correlations between LLM judges and human annotators in Table 12. At the system level, results resemble those reported in OEG LLM-as-a-judge (Section 4.5): the models are split into two groups, with closed-source and very large models (100+ billion parameters) achieving higher SPA scores ($\geq$ 0.78), while smaller ones range from 0.54 to 0.73. The only outlier is Llama 4 Maverick, which performs poorly compared to similar-sized models, placing in the group of smaller LLMs.

At the segment level, results align with the system-level ones, with models again splitting into the same two performance-based groups. However, two models stand out relative to their peers: GPT 4.1 achieves an acc$_{eq}$ score of 0.49, outperforming all others by a clear margin. Similarly, Qwen2.5 7B reaches 0.36 in terms of acc$_{eq}$, placing it closer to larger models than to others of comparable size.

Finally, we highlight CommandA, as a dense model with 111B parameters, surpasses several larger MOE competitors such as DeepSeek V3 and Qwen3 235B in Pariwise Ranking and ranks second in acc$_{eq}$.

5 Conclusion

We introduced the WMT25 Multilingual Instruction Shared Task, where the main contribution is a unified benchmark spanning five evaluation tasks: machine translation, linguistic reasoning, open-ended generation, cross-lingual summarization, and LLM-as-a-judge. The benchmark covers up to 30 languages evaluated both automatically and by humans, and emphasizes robust evaluation of multilingual LLM capabilities. We release all prompts, outputs, and human annotations to facilitate reproducibility and research.

- **Substantial headroom in linguistic reasoning.** Across languages, the best systems achieve well below half of the attainable LR points, indicating that current models struggle with structured, language-agnostic reasoning rather than knowledge recall.
- **English is not always the easiest instruction language.** Several leading models reach their top LR scores in non-English (e.g., Korean, German, Spanish), with noticeable drops in English, suggesting prompting language effects that merit further study.
- **Naturalness is the bottleneck for generation.** In OEG human evaluation, score spread is widest for *naturalness* compared to *instruction following* and *coherence*, echoing user reports that non-English outputs often sound robotic or translationese.
- **Closed-weight models lead, but strong open models follow closely.** Aggregate results and MT AUTORANK ranks show a top cluster of proprietary models, with large open models competitive on several tasks and language pairs.
- **MT quality varies sharply by pair and script.** High-resource or typologically close pairs exhibit tight spreads among top systems, while low-resource and script-variant directions show large gaps and instability.
- **LLM-as-a-judge correlates well at the system level, unevenly at the instance level.** Larger models achieve higher system-level accuracy in OEG/XLSum/MT, while smaller models are not suited for the task.
- **Evaluation reliability still hinges on humans.** Automatic scores enable broad coverage, but human annotations exposed language/script biases, instruction-following failures, and cases where metrics or judges disagree, underscoring the value of our released human-rated subsets.

6 Limitations

Budget-driven coverage limits and occasional model unavailability led to uneven per-task participation. Furthermore, human evaluation was performed on a subset of the samples.

While we usually report aggregate results across all languages (or language pairs), not all models are trained for all languages. This analysis inevitably penalizes them if some languages are unsupported. Practitioners can refer to the raw data for performance in individual languages of interest.

7 Acknowledgements

We thank many people for their help with annotations, reviewing the linguistic reasoning subtask, and providing or reviewing translations of instructions. They are: Özge Agca, Sweta Agrawal, Ondřej Bojar, Rawan Essam Bondok, Philipp Burlakov, Daryna Dementieva, TG Gowda, Alexandra Grieva, HyoJung Han, Katsuki Isobe, Gunny Kim, Ivan Korovin, Nalin Kumar, Mateusz Lango, Ariston Lim, Jiaming Liu, Andrei Manea, Sourabrata Mukherjee, Youssef Nafea, Ricardo Rei, Steinþór Steingrímsson, Hanka Štěříková, Dušan Variš, Gianluca Vico, and Razan Dyas Wibowo.

We thank Apify and Julian McAuley for their support in obtaining the cross-lingual summarization data.

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

Pinzhen Chen is supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546].

Patrícia Schmidtová is supported by Charles University projects GAUK 252986 and SVV 260 698.

References

Sanchit Ahuja, Varun Gumma, and Sunayana Sitaram. 2024. Contamination report for multilingual benchmarks. *arXiv preprint arXiv:2410.16186*.

Team Alibaba, An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Team Alibaba, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Andrew Michael Bean, Simeon Hellsten, Harry Mayne, Jabez Magomere, Ethan A Chi, Ryan Andrew Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. LINGOLY: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kranti Chalamalasetti, Gabriel Bernier-Colborne, Yvan Gauthier, and Sowmya Vajjala. 2025. Test set quality in multilingual LLM evaluation. *arXiv preprint arXiv:2508.02635*.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024a. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016. Association for Computational Linguistics.

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. 2024b. Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9706–9726. Association for Computational Linguistics.

Team Cohere, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, and others. 2025. Command a: An enterprise-ready large language model. *CoRR*.

Team Cohere Labs, John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

María Andrea Cruz Blandón, Jayasimha Talur, Bruno Charron, Dong Liu, Saab Mansour, and Marcello Federico. 2025. MEMERAG: A multilingual end-to-end meta-evaluation benchmark for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22577–22595. Association for Computational Linguistics.

Team DeepMind, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

DeepSeek, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and others. 2024. DeepSeek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284. Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929. Association for Computational Linguistics.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565. ELRA and ICCL.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44. Association for Computational Linguistics.

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

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628. Association for Computational Linguistics.

Team Google, Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. Do large language models have an English accent? evaluating and improving the naturalness of multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3823–3838. Association for Computational Linguistics.

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

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

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

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and others. 2025b. Preliminary ranking of WMT25 general machine translation systems. *Preprint*, arXiv:2508.14909.

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

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

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics.

Julia Kreutzer, Eleftheria Briakou, Sweta Agrawal, Marzieh Fadaee, and Kocmi Tom. 2025. Déjà vu: Multilingual LLM evaluation through the lens of machine translation evaluation. *arXiv preprint arXiv:2504.11829*.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Kanojia Diptesh, 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*, Suzhou, China. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172. European Association for Machine Translation.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–Machine parity in language translation. *Journal of Artificial Intelligence Research (JAIR)*, 67.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, and others. 2025. EuroLLM: Multilingual language models for Europe. *Procedia Computer Science*, 255:53–62.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997. Association for Computational Linguistics.

Mistral, AQ Jiang, and others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

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

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

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

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618. Association for Computational Linguistics.

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

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848. Association for Computational Linguistics.

Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André FT Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual LLMs. *arXiv preprint arXiv:2506.17080*.

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

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459. Association for Computational Linguistics.

Anushka Singh, Ananya Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. How good is zero-shot MT evaluation for low resource Indian languages? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649. Association for Computational Linguistics.

Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. 2024. MM-Eval: A multilingual meta-evaluation benchmark for LLM-as-a-judge and reward models. *arXiv preprint arXiv:2410.17578*.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *Preprint*, arXiv:2405.01724.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Linguini: A benchmark for language-agnostic linguistic reasoning. *Preprint*, arXiv:2409.12126.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123. Association for Computational Linguistics.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024. Evaluating WMT 2024 metrics shared task submissions on AfriMTE (the African challenge set). In *Proceedings of the Ninth Conference on Machine Translation*, pages 505–516. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in LLM-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *arXiv preprint arXiv:2504.15521*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023a. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024a. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024b. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500. Association for Computational Linguistics.

A LLM-in-the-loop PDF Parsing

As described in Section §2.1, the data for 14 of the 15 linguistic exams was extracted using an LLM-in-the-loop pipeline. This approach leveraged the manually parsed English data as a reference to efficiently scale the extraction process, which was followed by human verification and editing. Concretely, we prompted Gemini 2.5 Pro to structure each translated PDF's content into a JSON object by mimicking the provided English example. The LLM was given the unparsed English and translated PDFs, along with the reference JSON object from the English version as a string input.

Prompt Development and Iterations The exact prompt used is shown in Figure 13. As seen, the JSON extraction proved to be highly demanding, requiring the LLM to simultaneously parse content from PDF documents, process long-context inputs, and generate a structured output that must be syntactically valid and programmatically parsable. To improve reliability of the LLM automation process, we iterated on our approach by:

- **Adding step-by-step** instructions to the prompt.
- **Breaking the task down** to parse one problem at a time, which drastically reduced JSON validation failures. This means that for each exam we need at least 5 calls to an LLM (one for each language problem).
- **Implementing a resampling strategy** that only accepted outputs confirmed to be parsable JSON. Figure 14 shows the number of samples drawn for each task/problem language to arrive at a valid JSON. For most problems, a single API call was sufficient. However, the Dâw and Yanyuwa tasks, which employed more complex JSON structures, consistently required more attempts, with some outlier cases, such as for Persian, requiring as many as 50 calls to successfully generate a valid JSON object.

After parsing the PDFs to JSON, we imported the resulting data into spreadsheets so that native speakers could verify its correctness.

~~~~
You are given:
*   An English Linguistic Exam (PDF) with its
     solutions (PDF).
*   The JSON representation of the English
    exam (referred to as "English JSON").
*   The {language} version of the exam (PDF) and its
      solutions
    (PDF).

Objective: Generate a JSON object for the {
     no_problem}
problem ("{problem_language}") of the {language}
     exam.
This JSON should follow
the structural format of the English JSON.

Steps:
1.  Target the "{problem_language}" Problem:
    Isolate the "{problem_language}" problem data.
2.  Structural Template: Use the "{problem_language
     }"
    problem section from the English JSON as the
    structural basis for your new {language} JSON.
3.  Field Handling:
    *  Copy Directly from English JSON: For the
      "{problem_language}" problem, copy the values
           of these
      fields from the English JSON: `Identifier`, `
           Points`,
      `Work Language`, `Task Type`, `Eval Type`, and
            `Task
       Meta`.
    *  Extract from {language} PDFs: For all
         remaining fields,
       populate them with the corresponding content
            extracted
       from the "{problem_language}" section of the
            {language}
       PDFs.
    *  Content Adaptation: The English JSON models
         how PDF
       content should appear in the JSON. If its
            content isn't
       a direct PDF copy (e.g., it's formatted/
            structured),
       then similarly adapt the {language} PDF
            content to match
       this presentation style and any processing
       evident in the English JSON.

Attached are the PDFs, here is the English JSON:
```
{json_object}
```
Make sure all values in the JSON have the same
     length and
that the JSON itself is parsable with json.loads()
     in Python.
Output only the JSON object and nothing else.
~~~~

Table 13: Prompt used to extract a JSON object from a translated PDF by mimicking the structure of a manually parsed example ("json_object") from the original English PDF.

Language	Koryak	Hadza	Komnzo	Dâw	Yanyuwa
Ukrainian	1	1	1	2	4
Russian	1	1	1	4	18
Portuguese	1	1	3	11	11
Czech	1	1	1	1	11
Spanish	1	1	1	2	1
Swedish	1	1	1	1	1
Chinese	1	1	1	1	2
Korean	1	1	1	1	5
Persian	1	1	1	1	52
French	1	1	1	1	2
German	1	1	1	16	14
Japanese	1	1	1	1	4
Estonian	1	1	1	1	1
Dutch	1	1	3	3	9

Table 14: Number of API calls required to extract a parsable JSON object varied by task and language.

B LLM-as-a-judge Instructions

LLM-as-a-judge prompts for OEG, MT, and XLSum are in Tables 15 to 17 respectively.

```
Score the response generated by a system to a user's
      request in {language} on a likert scale from 1
      to 7. The quality levels associated with
     numerical scores are provided below:

7: The response fully adheres to all instructions
     that the user provided.
5: The chatbot mostly followed the instructions,
     conforming to the main points of the request
     but missing some details.
3: The chatbot followed only a small portion of the
     instructions or missed important points.
1: The chatbot entirely disregarded the instructions
     ; the response appears to have no relation to
     the user's request or is not in {language}.

Return only the score and nothing else.

Prompt: {input}
Response: {response}

Score (1-7):
```

```
Score the response generated by a system to a user's
      request in {language} on a likert scale from 1
      to 7. The quality levels associated with
     numerical scores are provided below:

7: The response represents fluent {language} text
     that might have been written by a native human
     speaker.
5: The response has some disfluencies that are
     noticeable but don't significantly hinder
     understanding the response.
3: The response is highly disfluent. There are
     several grammatical errors. Most of the meaning
      can be determined, but only with conscious
     effort.
1: The response is incomprehensible or is not in {
     language}.

Return only the score and nothing else.

Prompt: {input}
Response: {response}

Score (1-7):
```

```
Score the response generated by a system to a user's
      request in {language} on a likert scale from 1
      to 7. The quality levels associated with
     numerical scores are provided below:

7: The response is logically sound and appropriately
      structured with a clear sequence of nicely
     connected ideas and topics with no leaps in
     reasoning.
5: The response is generally well-structured and has
      a generally clear overall progression of ideas
     , but introduces a few logical gaps, or
     suddenly switches topics without an appropriate
      transition.
3: The response lacks an overall flow, and/or has
     multiple noticeable jumps between topics. It is
      possible to discern some relevant ideas, but
     the overall purpose of the response is
     incoherent.
1: The response has no overall structure, is in no
     way logically sound, and/or can be divided into
      many mostly-unrelated sections. It is
     difficult to identify any points the text is
     trying to make.

Return only the score and nothing else.

Prompt: {input}
Response: {response}

Score (1-7):
```

Table 15: Prompt instructions used in the LLM-as-a-judge for OEG sub-task.

```
Score the following translation from {source_lang}
     to {target_lang} on a scale from 0 to 100,
     where a score of 0 means a broken or poor
     translation; 33 indicates a flawed translation
     with significant issues; 66 indicates a good
     translation with only minor issues in grammar,
     fluency, or consistency; and 100 represents a
     perfect translation in both meaning and grammar
     . Answer with only a whole number representing
     the score, and nothing else.

{source_lang} source text:
{source_seg}
{target_lang} translation:
{target_seg}
```

Table 16: Prompt instructions used in the LLM-as-a-judge for MT sub-task.

```
Score the summary generated by a system based on a
     set of reviews in {language} on a likert scale
     from 1 to 7. Evaluate whether all information
     in the summary can be traced back to the
     reviews. Treat the reviews as the source of
     truth and do not consider any external
     information. The quality levels associated with
      numerical scores are provided below:
7: All of the information in the summary is fully
     supported by the reviews and no meaning was
     changed.
5: Most information is supported, but a small part
     of the summary contains information that either
      contradicts or cannot be verified by the
     reviews.
3: More than half of the information in the summary
     either contradicts or cannot be verified by the
      reviews.
1: The summary is fully made up of information that
     either contradicts or cannot be verified by the
      reviews.
Return only the score and nothing else.

Reviews: {input}
Summary: {response}

Score (1-7):
```

```
Score the summary generated by a system based on a
     set of reviews in {language} on a likert scale
     from 1 to 7. Read the reviews and identify the
     most important points, then evaluate whether
     these key points are covered by the summary.
     The quality levels associated with numerical
     scores are provided below:
7: The summary covers all key points.
5: The summary covers about two thirds of the key
     points.
3: The summary covers about a third of the key
     points.
1: The summary does not cover any of the key points
     mentioned in the reviews.
Return only the score and nothing else.

Reviews: {input}
Summary: {response}

Score (1-7):
```

```
Score the summary generated by a system based on a
     set of reviews in {language} on a likert scale
     from 1 to 7. Evaluate the degree to which the
     summary appears to be fluent, natural text in {
     language}, that is appropriate in terms of tone
      and formality. The quality levels associated
     with numerical scores are provided below:
7: The summary represents fluent {language} text
     that might have been written by a native human
     speaker.
5: The summary has some disfluencies that are
     noticeable but don't significantly hinder
     understanding the summary.
3: The summary is highly disfluent. There are
     several grammatical errors. Most of the meaning
      can be determined, but only with conscious
     effort. Alternatively, there are some words in
     a foreign language.
1: The summary is incomprehensible, or is not in {
     language}.
Return only the score and nothing else.

Reviews: {input}
Summary: {response}

Score (1-7):
```

```
Score the summary generated by a system based on a
     set of reviews in {language} on a likert scale
     from 1 to 7. Evaluate the degree to which the
     summary appears to be logically sound and
     internally consistent. The quality levels
     associated with numerical scores are provided
     below:
7: The summary is logically sound and appropriately
     structured with a clear sequence of nicely
     connected ideas and topics with no leaps in
     reasoning.
5: The summary is generally well-structured and has
     a generally clear overall progression of ideas,
      but introduces a few logical gaps, or suddenly
      switches topics without an appropriate
      transition.
3: The summary lacks an overall flow, and/or has
     multiple noticeable jumps between topics. It is
      possible to discern some relevant ideas, but
     the overall purpose of the summary is
     incoherent.
1: The summary has no overall structure, is in no
     way logically sound, and/or can be divided into
      many mostly-unrelated sections. It is
     difficult to identify any points the text is
     trying to make.
Return only the score and nothing else.

Reviews: {input}
Summary: {response}

Score (1-7):
```

Table 17: Prompt instructions used in the LLM-as-a-judge for XLSum sub-task.

C MT Prompt instructions

```
You are a professional {source_language}-to-{
     target_language} translator, tasked with
     providing translations suitable for use in {
     target_region} ({tgt_language_code}). Your goal
      is to accurately convey the meaning and
     nuances of the original {source_language} text
     while adhering to {target_language} grammar,
     vocabulary, and cultural sensitivities. The
     original {source_language} text {
     domain_description}. {domain_instruction}
     Produce only the {target_language} translation,
      without any additional explanations or
     commentary. Retain the paragraph breaks (double
      new lines) from the input text. Please
     translate the following {source_language} text
     into {target_language} ({tgt_language_code}):\n
     \n{input_text}
```

```
news: Ensure the translation is formal, objective,
     and clear. Maintain a neutral and informative
     tone consistent with journalistic standards.

social: Ensure you do not reproduce spelling
     mistakes, abbreviations or marks of
     expressivity. Platform-specific elements such
     as hashtags or userids should be translated as-
     is.

literary: Aim to maintain the original tone and
     register, retaining the emotional depth of the
     story. Dialogues should sound natural and
     follow the conventions of the target language.

speech: Pay attention to errors that mimic speech
     transcription errors and fix as necessary.
     Maintain the flow and colloquial style of the
     speaker in the translation.

edu: Preserve the line breaks. Use precise
     terminology and a tone appropriate for academic
      or instructional materials.

dialogue: Maintain dialog turn structure and speaker
      indicators (X, Y). Ensure natural flow,
     consistent tone (feminine/masculine, polite/
     familiar), and preserve any HTML tags (e.g.,
     italics).
```

Table 18: Prompt instruction used in the machine translation sub-task together with domain information.

D Cross-lingual Summarization Results by Language

Table 19 details the cross-lingual summarization performance in each language. In most target languages, the differences between models are relatively small: Italian only had one cluster. Egyptian Arabic had 4 clusters and Hindi had 3. The remaining languages had two clusters: Mistral-7B (worse performing) and all remaining models.

E Machine Translation Fine-Grained Results

Table 20 reports the fine-grained MT AUTORANK scores for all models by language pair.

Model	Egyptian Arabic	Spanish	German	Japanese	Russian	Italian	Czech	Indonesian	Simplified Chinese	Korean	French	Hindi
Gemini 2.5 Pro	5.57	5.79	6.51	5.44	6.42	6.11	6.11	5.81	6.49	5.74	6.42	6.22
GPT 4.1	5.56	5.68	6.31	5.71	6.54	6.06	6.00	5.79	6.33	5.58	6.38	5.93
Claude 4	5.31	5.86	6.54	5.96	6.32	6.06	5.60	5.44	6.47	5.51	6.33	5.88
Mistral Medium	4.94	5.71	6.36	5.32	6.18	5.88	5.58	5.56	6.38	5.49	6.06	5.93
Gemma 3 12B	4.69	5.50	6.32	5.56	6.21	6.17	5.36	5.64	6.22	5.32	5.88	5.83
DeepSeek V3	4.90	5.57	6.25	5.24	6.03	5.92	5.47	5.40	6.38	5.46	6.00	6.07
Gemma 3 27B	4.82	5.69	6.28	5.64	6.18	5.99	5.35	5.01	6.22	5.29	6.42	5.76
CommandA	4.82	5.60	6.15	5.53	5.92	5.89	5.07	5.36	6.08	5.57	6.31	5.65
AyaExpanse 32B	4.89	5.61	6.44	5.51	6.08	5.53	5.31	5.88	5.78	5.12	5.86	5.60
Llama 4 Maverick	4.71	5.54	5.86	5.26	5.67	5.46	5.39	5.28	5.99	5.57	6.25	5.90
Qwen3 235B	4.12	5.72	6.39	5.15	6.17	5.94	4.46	5.42	6.47	5.12	5.92	5.88
TowerPlus 72B	2.49	5.58	5.60	5.22	5.81	5.79	5.36	5.68	6.18	5.21	6.31	5.69
AyaExpanse 8B	4.49	5.28	5.68	4.93	5.24	5.97	5.31	5.11	5.51	4.97	6.22	5.72
TowerPlus 9B	2.28	5.62	5.54	4.82	5.26	5.78	5.29	4.12	5.94	5.00	5.74	5.64
Llama 3.1 8B	2.71	5.04	4.92	3.58	5.14	5.21	3.57	4.51	5.15	3.97	5.68	4.76
Qwen2.5 7B	2.65	5.43	5.21	4.40	4.65	4.58	3.21	4.69	5.26	4.50	5.76	3.65
Mistral 7B	1.57	3.89	3.43	2.68	4.65	4.88	2.35	3.69	3.83	3.18	4.76	2.79

Table 19: Human evaluation results for cross-lingual summarization by language. The scores are averaged across the evaluated criteria.

Model	Czech→German	Czech→Ukrainian	English→Arabic	English→Bhojpuri	English→Bengali	English→Czech	English→German	English→Greek	English→Estonian	English→Persian	English→Hindi	English→Indonesian	English→Icelandic	English→Italian	English→Japanese	English→Kannada	English→Korean	English→Lithuanian	English→Masai	English→Marathi	English→Romanian	English→Russian	English→Serbian Cyr.	English→Serbian Lat.	English→Swedish	English→Thai	English→Turkish	English→Ukrainian	English→Vietnamese	English→Chinese	Japanese→Chinese
Gemini 2.5 Pro	1.1	1.0	1.0	1.0	1.0	1.0	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.2	1.0	1.0	1.0	1.0	6.2	1.0	1.0	1.0	1.0	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
GPT 4.1	1.0	1.1	1.6	3.5	1.7	1.3	1.0	1.3	1.3	1.4	1.5	1.5	1.1	1.0	1.1	3.8	1.2	1.4	19.0	2.4	1.1	1.4	1.2	1.1	1.5	1.7	1.2	1.0	1.1	1.4	1.6
DeepSeek V3	1.7	1.9	1.2	3.3	2.5	1.8	1.1	3.6	5.1	2.1	1.7	1.9	6.6	2.0	2.0	4.4	2.4	5.7	6.3	3.0	2.0	1.6	4.5	1.0	2.0	2.0	1.4	1.9	1.4	3.2	3.2
Claude 4	2.3	2.1	2.3	3.0	2.4	3.7	2.4	2.5	3.1	2.8	3.0	3.3	3.4	3.7	2.3	3.2	2.0	3.2	1.0	3.0	3.4	3.2	3.4	2.9	2.8	2.7	2.9	3.1	2.6	3.0	2.8
Mistral Medium	1.9	2.3	1.4	5.5	2.2	2.7	1.2	2.8	7.2	2.4	4.3	2.5	6.5	2.5	2.2	4.3	2.9	6.4	–	4.0	2.6	–	–	–	2.3	2.4	2.6	2.5	1.7	1.5	2.8
Qwen3 235B	4.8	5.2	3.3	6.7	3.9	4.9	2.8	4.3	7.3	5.1	4.3	1.9	8.9	2.5	3.1	4.5	3.1	4.5	1.7	5.0	3.1	3.2	6.6	4.3	4.6	1.9	4.2	4.4	1.4	1.4	3.3
Llama 4 Maverick	3.8	3.8	4.1	4.1	3.6	4.9	3.2	3.8	4.0	3.3	4.3	4.1	5.8	9.6	3.9	5.6	4.5	4.0	2.0	4.7	3.5	5.1	5.4	4.2	3.4	3.0	4.0	4.1	3.7	3.5	7.0
CommandA	2.2	2.5	2.8	4.1	6.5	3.6	1.8	2.4	8.4	2.9	3.6	4.3	10.7	3.4	3.0	10.1	2.8	6.6	8.8	8.4	2.7	4.4	10.1	5.5	5.1	7.4	4.5	3.3	5.1	4.6	3.6
Gemma 3 27B	3.5	2.7	3.0	5.1	4.8	3.8	10.4	7.0	3.7	2.4	3.2	2.8	6.3	6.1	3.3	5.2	6.8	3.9	16.6	3.2	2.8	3.2	6.7	7.0	2.6	2.4	5.1	5.8	5.9	4.3	6.8
Gemma 3 12B	5.8	4.9	4.5	7.3	4.9	6.0	6.6	5.7	6.3	3.7	4.4	3.6	9.2	8.0	6.2	9.1	4.6	6.1	10.2	7.8	4.4	6.2	6.7	5.6	6.7	5.9	4.9	7.2	4.7	5.4	9.7
TowerPlus 72B	4.6	4.4	6.3	7.6	9.7	5.7	3.9	13.9	10.7	11.2	6.0	5.3	3.7	3.8	3.6	17.1	4.3	14.7	10.1	11.3	5.1	4.1	15.4	10.3	4.2	4.3	8.1	4.9	5.0	4.8	4.8
AyaExpanse 32B	4.0	3.8	1.9	6.8	12.8	4.3	2.7	3.1	17.1	3.7	5.3	4.4	17.5	4.3	4.2	16.3	4.3	13.9	7.7	11.8	3.4	5.6	19.0	11.5	12.9	14.7	5.7	4.2	3.0	6.1	5.4
TowerPlus 9B	5.0	4.0	16.4	7.6	13.3	4.8	3.7	16.4	14.4	13.6	4.8	12.0	2.4	5.0	4.5	11.9	5.0	16.4	5.1	4.8	3.0	4.7	15.7	15.1	3.2	12.0	13.9	3.9	10.6	6.1	5.7
EuroLLM 22B	5.5	4.4	3.1	8.1	20.0	5.5	4.0	3.6	4.0	19.0	6.9	16.9	18.9	4.8	7.5	19.0	6.9	4.8	9.3	11.5	4.4	6.1	12.0	6.8	4.7	19.3	5.0	6.2	20.0	6.8	8.0
AyaExpanse 8B	8.0	6.5	2.4	9.8	17.1	7.2	6.3	4.5	20.0	5.8	7.0	5.4	20.0	7.3	7.0	18.8	6.6	18.3	9.3	13.4	5.3	7.6	17.5	18.4	20.0	17.7	7.8	6.5	4.4	8.3	8.5
EuroLLM 9B	11.5	7.2	5.1	8.7	18.2	8.5	6.3	4.9	6.8	20.0	7.2	20.0	15.6	7.5	10.9	18.8	10.2	5.3	9.2	8.7	5.6	9.6	13.0	8.0	5.6	20.0	6.0	8.6	17.2	9.3	12.5
Llama 3.1 8B	13.6	12.1	10.8	20.0	10.2	13.8	12.3	11.7	13.0	9.2	9.1	8.4	15.8	12.8	11.5	13.6	13.5	14.9	9.0	11.2	10.7	14.1	11.3	11.2	8.7	8.6	11.1	13.0	7.2	10.6	11.4
CommandR7B	9.4	12.3	3.4	9.8	15.3	13.2	8.6	9.0	18.8	8.4	10.5	10.4	18.3	9.5	8.9	16.9	9.8	16.5	3.6	14.0	9.9	19.0	18.6	19.0	17.2	17.4	10.9	15.6	8.0	10.7	10.6
Qwen2.5 7B	17.4	20.0	11.6	10.4	14.9	18.2	13.6	18.7	17.8	14.6	16.5	8.5	19.4	13.2	11.0	20.0	13.6	18.6	9.8	18.4	20.0	11.4	19.0	18.2	16.1	7.9	14.5	20.0	6.5	5.9	7.3
Mistral 7B	20.0	15.9	20.0	12.1	19.1	20.0	20.0	20.0	19.7	19.3	20.0	17.1	19.4	20.0	20.0	19.6	20.0	20.0	13.7	20.0	17.5	18.9	15.9	15.5	12.9	16.6	20.0	15.9	16.2	20.0	20.0

Table 20: Machine translation AUTORANK results across language pairs (lower is better).

Model	Czech→German		Czech→Ukrainian		English→Arabic		English→Bhojpuri		English→Chinese		English→Czech		English→Estonian		English→Icelandic		English→Italian		English→Japanese		English→Korean		English→Maasai		English→Russian		English→Serbian Cyrillic		English→Ukrainian		Japanese→Chinese	
Gemini 2.5 Pro	91	1	93	1	61	1	95	1	84	1	89	1	79	1	78	1	79	1	86	1	-3	1	10	6	83	1	94	1	90	1	-4	1
GPT 4.1	89	1	92	1	77	2	83	4	84	1	81	1	72	1	68	1	79	1	84	1	-3	1		19	76	1	92	1	88	1	-6	2
Claude 4	89	2	89	2	56	2	83	3	87	3	80	4	53	3	48	3	72	4	79	2	-3	2	8	1	76	3	90	3	86	3	-6	3
DeepSeek V3	88	2	89	2	57	1	77	3	85	3	85	2		5		7	72	2	79	2	-4	2	3	6	74	2	79	4	86	2	-8	3
Mistral Medium	87	2	89	2	36	1		6	80	2	80	3		7		6	74	2	85	2	-5	3							85	2	-10	3
CommandA	87	2	86	2	74	3	73	4		5	76	4		8		11	73	3		3	-5	3	1	9		4		10	84	3		4
TowerPlus 9B	80	5	85	4		16		8		6	66	5		14	57	2	61	5		4	-7	5	1	5		5		16	84	4	-13	6
Gemma 3 27B	82	4	89	3		3	56	5		4	76	4	46	4		6		6		3		7		17	62	3		7		6		7
Llama 4 Maverick		4		4		4	76	4	81	4		5		4		6		10		4		4	5	2		5		5	86	4		7
Qwen3 235B		5		5		3		7	84	1		5		7		9	67	2		3		3	3	2	68	3		7		4		3
Gemma 3 12B	77	6		5		4		7		5		6		6	16	9	54	8		6	-6	5	3	10		6	75	7		7		10
EuroLLM 9B		12		7		5		9		9		8		7		16	57	8		11		10	1	9		10	42	13		9		12
Llama 3.1 8B		14		12		11		20		11		14		13	11	16		13		12		14	3	9		14	58	11		13		11
AyaExpanse 8B		8		6	2	2		10		8		7		20		20	57	7		7		7	6	9		8		18		6		8
EuroLLM 22B		6		4		3		8		7		6	47	4		19		5		8		7	0	9		6		12		6		8
TowerPlus 72B		5		4		6		8		5		6		11	46	4		4		4		4		10		4		15		5		5
CommandR7B		9		12	4	3		10		11		13		19		18		10		9		10	2	4		19		19		16		11
AyaExpanse 32B		4		4		2		7		6		4		17		18		4		4		4	3	8		6		19		4		5
Qwen2.5 7B		17		20		12		10		6		18		18		19		13		11		14	3	10		11		19		20		7
Mistral 7B		20		16		20		12		20		20		20		19		20		20		20		14		19		16		16		20

Table 21: Combination of human evaluation of LLMs for machine translation and AUTORANK (Kocmi et al., 2025a,b). The left column shows human evaluation (higher is better) either with ESA or MQM annotation protocol (Kocmi et al., 2024; Freitag et al., 2021) and the right column for each language shows the AUTORANK (lower is better) based on Table 20.

	Czech → German	Czech → Ukrainian	English → Arabic	English → Bhojpuri	English → Czech	English → Estonian	English → Icelandic	English → Italian	English → Japanese	English → Korean	English → Maasai	English → Russian	English → Serbian	English → Ukrainian	English → Chinese	Japanese → Chinese
GPT 4.1	0.90	0.85	0.87	0.78	0.87	0.84	0.93	0.84	0.76	0.88	0.61	0.82	0.89	0.78	0.80	0.94
Claude 4	0.88	0.84	0.73	0.77	0.88	0.79	0.91	0.84	0.75	0.92	0.62	0.77	0.88	0.78	0.85	0.93
CommandA	0.86	0.83	0.74	0.69	0.87	0.78	0.81	0.83	0.70	0.89	0.59	0.81	0.83	0.80	0.83	0.91
DeepSeek V3	0.86	0.85	0.55	0.64	0.85	0.85	0.86	0.83	0.67	0.90	0.61	0.76	0.85	0.81	0.84	0.88
Qwen3 235B	0.84	0.84	0.53	0.61	0.88	0.81	0.84	0.83	0.70	0.89	0.59	0.80	0.84	0.81	0.86	0.81
AyaExpanse 32B	0.79	0.83	0.54	0.64	0.83	0.63	0.65	0.82	0.63	0.90	0.58	0.74	0.69	0.77	0.82	0.79
Llama 4 Maverick	0.82	0.82	0.59	0.66	0.66	0.75	0.78	0.74	0.65	0.82	0.60	0.65	0.74	0.65	0.80	0.84
Qwen2.5 7B	0.72	0.66	0.34	0.54	0.78	0.54	0.66	0.82	0.59	0.80	0.57	0.77	0.74	0.64	0.79	0.78
Llama 3.1 8B	0.75	0.70	0.20	0.58	0.74	0.63	0.74	0.78	0.59	0.80	0.55	0.71	0.77	0.62	0.81	0.63
CommandR7B	0.77	0.71	0.17	0.55	0.68	0.47	0.44	0.80	0.61	0.76	0.57	0.57	0.37	0.50	0.68	0.67
AyaExpanse 8B	0.76	0.66	0.34	0.49	0.70	0.46	0.44	0.72	0.56	0.73	0.50	0.63	0.42	0.51	0.67	0.70
Mistral 7B	0.64	0.56	0.21	0.48	0.62	0.49	0.46	0.63	0.44	0.60	0.56	0.53	0.61	0.55	0.62	0.57

Table 22: System-level Soft Pairwise Accuracy (SPA) computed between the LLM judges and human annotators in the task of machine translation evaluation.

	Czech → German	Czech → Ukrainian	English → Arabic	English → Bhojpuri	English → Czech	English → Estonian	English → Icelandic	English → Italian	English → Japanese	English → Korean	English → Maasai	English → Russian	English → Serbian	English → Ukrainian	English → Chinese	Japanese → Chinese
GPT 4.1	0.46	0.40	0.53	0.56	0.47	0.52	0.66	0.45	0.45	0.50	0.54	0.46	0.51	0.43	0.41	0.46
CommandA	0.41	0.35	0.44	0.34	0.38	0.35	0.43	0.40	0.33	0.47	0.49	0.36	0.43	0.34	0.34	0.38
Qwen3 235B	0.39	0.33	0.37	0.29	0.35	0.33	0.42	0.38	0.34	0.47	0.49	0.36	0.41	0.33	0.36	0.40
DeepSeek V3	0.36	0.33	0.37	0.38	0.31	0.33	0.47	0.34	0.28	0.48	0.49	0.33	0.45	0.31	0.35	0.39
Claude 4	0.35	0.29	0.42	0.35	0.28	0.30	0.52	0.28	0.30	0.48	0.56	0.28	0.42	0.25	0.30	0.39
Qwen2.5 7B	0.36	0.32	0.37	0.31	0.31	0.29	0.38	0.37	0.28	0.47	0.49	0.36	0.37	0.34	0.33	0.36
Mistral 7B	0.28	0.28	0.37	0.22	0.23	0.22	0.24	0.26	0.25	0.46	0.49	0.25	0.27	0.23	0.23	0.30
AyaExpanse 32B	0.28	0.29	0.37	0.27	0.22	0.15	0.21	0.28	0.23	0.46	0.49	0.23	0.27	0.22	0.24	0.33
Llama 3.1 8B	0.27	0.25	0.37	0.20	0.22	0.25	0.22	0.24	0.23	0.46	0.49	0.25	0.26	0.20	0.28	0.31
CommandR7B	0.22	0.21	0.37	0.18	0.19	0.20	0.19	0.26	0.25	0.47	0.49	0.24	0.17	0.22	0.26	0.32
AyaExpanse 8B	0.19	0.20	0.37	0.20	0.12	0.11	0.11	0.18	0.13	0.46	0.49	0.13	0.18	0.12	0.18	0.29
Llama 4 Maverick	0.14	0.17	0.37	0.19	0.05	0.06	0.24	0.11	0.10	0.46	0.49	0.07	0.18	0.06	0.12	0.26

Table 23: Segment-level Pairwise Accuracy with Tie Calibration (acc_{eq} computed between the LLM judges and human annotators in the task of machine translation evaluation.