

Preliminary Ranking of WMT25 General Machine Translation Systems

Tom Kocmi Eleftherios Avramidis Rachel Bawden Ondřej Bojar Konstantin Dranch
Anton Dvorkovich Sergey Dukanov Natalia Fedorova Mark Fishel Markus Freitag
Thamme Gowda Roman Grundkiewicz Barry Haddow Marzena Karpinska
Philipp Koehn Howard Lakounga Jessica Lundin Kenton Murray Masaaki Nagata
Stefano Perrella Lorenzo Proietti Martin Popel Maja Popović Parker Riley
Mariya Shmatova Steinþór Steingrímsson Lisa Yankovskaya Vilém Zouhar

Introduction

We present the *preliminary* ranking of the WMT25 General Machine Translation Shared Task, in which MT systems have been evaluated using automatic metrics. As this ranking is based on automatic evaluations, it may be biased in favor of systems that employ re-ranking techniques, such as Quality Estimation re-ranking or Minimum Bayes Risk decoding. The official WMT25 ranking will be based on human evaluation, which is more reliable and will supersede the automatic ranking.

The purpose of this report is not to present the final findings of the General MT task, but rather to share preliminary results with task participants, which may be useful when preparing their system submission papers.

Types of Systems

We distinguish two types of MT systems participating in the shared task:

- **Constrained systems** are those using only publicly available training data and models. The maximum size of their parameter counts is 20B and participants are required to release their weights under an open license.
- **Unconstrained systems** (marked with gray) are all the remaining systems, with no limitations on their training data, model sizes and no requirements to publish their model weights. Systems where the relevant information is not publicly known also fall into the unconstrained category.

Evaluated Systems

Details of all systems are going to be available in the upcoming WMT25 findings. In addition to participating systems, we also collect open-weight and proprietary LLMs together with three popular commercial MT systems. For each provider, we selected their largest/best performing model for each of the subtracks (when applicable).

Constrained systems: AyaExpanse-8B, CommandR7B, EuroLLM-9B, Gemma-3-12B, Llama-3.1-8B, Mistral-7B, NLLB, Qwen2.5-7B, TowerPlus-9B

Unconstrained systems: AyaExpanse-32B, Claude-4, CommandA, DeepSeek-V3, EuroLLM-22B, Gemma-3-27B, Gemini-2.5-Pro, GPT-4.1, Llama-4-Maverick, Mistral-Medium, ONLINE-B, ONLINE-G, ONLINE-W, Qwen3-235B, TowerPlus-72B

We used a zero shot instruction following approach, translating data on a document-level whenever possible, having a paragraph-level backup for failed translations. The instructions were provided in the blindset. This may hurt some of the systems trained for specific MT instructions such as TowerLLM or EuroLLM, we mark them with [M].

We turned off the reasoning for Qwen3-235B, however, we didn't set the reasoning budget for Gemini-2.5-Pro which increased output tokens count 6.6 times making it the most expensive model in the evaluation.

The code for collecting translations is available at github.com/wmt-conference/wmt-collect-translations and we marked all systems collected by us with ▲.

Evaluated Data

We evaluated 32 language pairs: half of them will be evaluated by humans, while the other half belong to the multilingual subtrack and will rely solely on automatic ranking.

Most language pairs are in the English-to-X direction and contain approximately 37k words. Each segment contains about 100 words, which typically corresponds to paragraphs, although sometimes natural paragraphs had to be further segmented to meet this constraint. The data are aggregated into documents. The test sets combine material from four domains:

- **News commentary**
- **Social** (texts from social networks, collected with screenshots)
- **Speech** (transcripts of speeches obtained automatically)
- **Literary** (two documents of roughly 5,000 words each)

Participants could use image and video modalities (when available) to improve their translations; however, their use was not required. Humans providing reference translations were offered these extra modalities, too. Language pairs with a non-English source have a similar distribution but differ slightly in domains and sizes.

We do not provide sentence splitting; consequently, many segments contain multiple sentences.

We release all data, including references, system outputs, automatic segment scores, or latex sources of this document at: github.com/wmt-conference/wmt25-general-mt.

Automatic Ranking

We call the automatic ranking described in the rest of this section “AutoRank”. Compared to last year, both the set of automatic metrics and the aggregation procedure changed slightly.

Metrics used. For each language pair (except where noted below), we combine three families of evaluation methods:

- **LLM-as-a-Judge (reference-less).** GEMBA-ESA (Kocmi and Federmann, 2023) with two independent judges: GPT-4.1 (OpenAI, 2025) and Command A (Cohere Team, 2025), both used in a reference-less setting.

- **Trained reference-based metrics.** Two reference-based supervised metrics explicitly trained to approximate human judgments of translation quality: MetricX-24-Hybrid-XL (Juraska et al., 2024) and XCOMET-XL (Guerreiro et al., 2024).

- **Trained Quality Estimation (QE).** One QE metric trained to mimic human judgments without a reference: CometKiwi-XL (Rei et al., 2023).

Including both reference-based and reference-less (or QE) methods balances complementary failure modes: reference-based metrics typically achieve higher correlation with human judgments when references are high-quality, whereas reference-less methods reduce susceptibility to reference bias when references are suboptimal (Freitag et al., 2023). A known pitfall for multilingual QE is that it can be fooled by fluent output in the wrong target language; in contrast, the GEMBA-ESA prompt explicitly specify the target language, which should mitigate this issue.

The use of LLM-as-a-judge metrics (GEMBA-ESA) is intended to mitigate biases by models employing re-ranking or similar techniques during training or inference. Nevertheless, some systems incorporated GEMBA directly as their reward model.

For each metric and language pair, the system-level score of an MT system is computed as the average of the metric’s paragraph-level (segment-level) scores over all translations the system produced on the test set for that language pair. For language pairs without human references, we exclude CometKiwi-XL from the corresponding AutoRank computation, since MetricX-24-Hybrid-XL and XCOMET-XL are hybrid metrics and can be run in reference-less (QE) mode, thus already providing the QE signal from trained metrics for those pairs.

Low-resource exception. For the two most low-resource target languages, i.e., Bhojpuri and Maasai, we rely solely on chrF++ (Popović, 2017) because it is not known whether the above metrics are reliable in these settings (Falcão et al., 2024; Singh et al., 2024; Wang et al., 2024; Sindhuja et al., 2025) and human references are available. We compute chrF++ using sacrebleu (Post, 2018).

From system-level scores to AutoRank To combine the metrics into a single score, we first nor-

malize them using median-interpercentile scaling to address differences in scale and reduce the influence of low-performing outliers. We then compute the average using equal weights. Finally, we linearly rescale the results to the range from 1 to N systems. A detailed description is provided below:

Let S be the set of submitted systems for a given language pair, $|S| = N$, and let M be the set of automatic metrics used for that language pair (for Bhojpuri and Maasai, $|M| = 1$). For each metric $m \in M$ and system $s \in S$, we compute a system-level score $x_s^{(m)}$ as the average of that metric over all available test segments. To combine scores across metrics, we first map them to a common scale; however, classical min-max normalization is highly sensitive to outliers. To downweight extremes without discarding any system, we apply a *median-interpercentile* scaling to each metric m :

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

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

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

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

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

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

mitigation provided by the robust scaling. We refer to the resulting value as AutoRank in the various tables.

Human Evaluation

This year, we received 36 unique teams,¹ the highest number of participants ever. We are not able to evaluate them all with human annotators. Therefore, we select a subset of about 18 systems per language pair (some language pairs have this system count higher) which will be evaluated by humans with the Error Span Annotation protocol (Kocmi et al., 2024). For the remaining systems, AutoRank is going to be the official final ranking.

When selecting systems for human evaluation, we prioritize constrained systems over unconstrained systems. Therefore, we select the systems for human evaluation based on the following two rules:

1. We select top eight constrained systems ignoring unconstrained systems.
2. Then, we take the top performing systems until we have total of 18 systems selected for human evaluation.

Limitations

A key limitation of our evaluation is that some models have been optimized for the very metrics we employ in AutoRank, either during training or at inference time (Freitag et al., 2022a; Finkelstein and Freitag, 2024). This can result in artificially inflated scores that do not accurately reflect a model’s true capabilities (Kovacs et al., 2024). To mitigate this issue, we aggregate the assessments from multiple learned metrics and LLM-as-a-judge approaches. However, even this strategy has shortcomings. First, scores from different learned metrics often exhibit high correlation among themselves. Second, LLM-as-a-judge approaches, including the Gemba-ESA we use, may also have been utilized to optimize machine translation models.

Another limitation is that we use automatic metrics to evaluate entire paragraphs, whereas their reliability is typically established at the sentence level. Additionally, learned metrics struggle when evaluating translation directions involving low-resource languages, such as English-to-Bhojpuri

¹We received 43 different teams, however, 7 of them have withdrawn or been disqualified

and English-to-Maasai. Therefore, we evaluate these language pairs using chrF++. However, chrF++ is a surface-level metric that, like BLEU, has been repeatedly shown to correlate poorly with human judgments (Kocmi et al., 2021; Freitag et al., 2022b, 2023).

Furthermore, our automatic evaluation is conducted at the paragraph level, without incorporating document-level context. This may lead to inflated scores for systems that translate the dataset paragraph by paragraph, disregarding dependencies and coherence across paragraphs.

The LLM-as-a-judge approach also depends on the language performance of the underlying LLMs. For our evaluation, we selected two top-performing multilingual systems: GPT-4.1 and Command A. Command A officially supports only 23 languages (Cohere Team, 2025), while the set of languages supported by GPT-4.1 is not publicly documented. Nevertheless, as both metrics correlate well across all languages and show strong agreement with other evaluation metrics, we retained them as judges for all 30 language pairs.

Finally, using automatically generated speech recognition transcripts as source text in the speech domain introduces additional noise, as the evaluation metrics are unlikely to be robust to ASR errors. Consequently, systems that handle the speech domain well may receive lower scores if their outputs diverge from the ASR transcript, even when their translations are correct.

Given these issues, along with the well-documented biases and limitations of automatic metrics (Karpinska et al., 2022; Moghe et al., 2024), human evaluation remains indispensable. Therefore, the results from human assessments will supersede the automatic rankings presented here.

Acknowledgement

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

| English-Egyptian Arabic | | | | | | | | | |
|-------------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.658 | 76.3 | 75.0 | -5.7 | 0.388 |
| Wenyii | ✓ | 14 | ✓ | 2.5 | 0.65 | 79.2 | 73.3 | -6.4 | 0.337 |
| Algharb | ✓ | 14 | ✓ | 2.6 | 0.645 | 80.0 | 73.9 | -6.5 | 0.328 |
| GemTrans | ✓ | 27 | ✓ | 3.4 | 0.644 | 73.0 | 69.6 | -6.0 | 0.345 |
| CommandA-WMT | ✓ | 111 | ✓ | 4.0 | 0.621 | 77.8 | 75.4 | -7.0 | 0.311 |
| UvA-MT | ✓ | 12 | ✓ | 4.1 | 0.637 | 74.4 | 73.4 | -7.1 | 0.325 |
| Yolu | ✓ | 14 | ✓ | 5.4 | 0.658 | 67.8 | 63.9 | -6.6 | 0.323 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 5.6 | 0.552 | 79.5 | 84.5 | -7.6 | 0.267 |
| ▲ ONLINE-B | ✓ | ? | ✓ | 6.4 | 0.627 | 70.4 | 67.4 | -7.1 | 0.288 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 6.5 | 0.534 | 78.4 | 84.1 | -7.8 | 0.265 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 6.9 | 0.573 | 74.2 | 75.7 | -7.7 | 0.273 |
| ▲ Mistral-Medium | ✓ | ? | ✓ | 7.5 | 0.586 | 71.7 | 71.0 | -7.8 | 0.274 |
| ▲ Claude-4 | ✓ | ? | ✓ | 7.6 | 0.552 | 76.5 | 80.0 | -8.5 | 0.246 |
| SRPOL | ✗ | 12 | ✓ | 7.9 | 0.641 | 65.7 | 61.7 | -7.8 | 0.286 |
| ▲ CommandA | ✓ | 111 | ✓ | 8.3 | 0.533 | 75.8 | 80.0 | -8.5 | 0.238 |
| ▲ AyaExpanse-32B | ✓ | 32 | | 8.4 | 0.585 | 70.7 | 68.8 | -8.1 | 0.261 |
| ▲ ONLINE-W | ? | ? | | 9.0 | 0.607 | 67.7 | 64.0 | -8.2 | 0.258 |
| ▲ AyaExpanse-8B | ✓ | 8 | ✓ | 9.7 | 0.596 | 66.1 | 61.6 | -8.2 | 0.259 |
| ▲ Qwen3-235B | ✓ | 235 | | 10.7 | 0.571 | 66.1 | 64.1 | -8.7 | 0.247 |
| ▲ Gemma-3-27B | ✓ | 27 | | 10.7 | 0.549 | 64.8 | 63.3 | -8.6 | 0.281 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | | 10.7 | 0.592 | 64.0 | 60.5 | -8.5 | 0.246 |
| IRB-MT | ✓ | 12 | ✓ | 10.8 | 0.532 | 69.0 | 67.5 | -8.5 | 0.236 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 11.1 | 0.526 | 67.9 | 70.0 | -8.8 | 0.234 |
| IR-MultiagentMT | ✗ | ? | | 11.3 | 0.543 | 66.0 | 64.2 | -8.7 | 0.247 |
| ▲ CommandR7B | ✓ | 7 | ✓ | 11.3 | 0.588 | 62.7 | 59.0 | -8.8 | 0.248 |
| ▲ Gemma-3-12B | ✓ | 12 | | 11.7 | 0.529 | 67.9 | 67.6 | -9.0 | 0.22 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | | 14.0 | 0.548 | 58.7 | 54.5 | -9.3 | 0.233 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | | 15.5 | 0.534 | 58.2 | 54.0 | -10.5 | 0.224 |
| TranssionTranslate | ? | ? | | 15.8 | 0.501 | 59.0 | 57.4 | -9.9 | 0.2 |
| TranssionMT | ✓ | 1 | | 16.9 | 0.488 | 58.5 | 56.1 | -10.4 | 0.194 |
| ▲ NLLB | ✓ | 1 | | 18.0 | 0.499 | 53.9 | 51.2 | -10.8 | 0.201 |
| SalamandraTA | ✓ | 8 | | 20.1 | 0.492 | 50.0 | 44.4 | -11.4 | 0.195 |
| ▲ ONLINE-G | ✓ | ? | | 22.6 | 0.445 | 53.5 | 48.3 | -13.5 | 0.152 |
| ▲ Llama-3.1-8B | ✗ | 8 | | 22.8 | 0.458 | 45.5 | 41.8 | -12.3 | 0.18 |
| ▲ Qwen2.5-7B | ✓ | 7 | | 24.0 | 0.436 | 44.5 | 39.3 | -12.6 | 0.176 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | | 31.9 | 0.337 | 31.1 | 26.9 | -15.2 | 0.162 |
| ▲ Mistral-7B | ✗ | 7 | | 37.0 | 0.262 | 27.9 | 23.2 | -18.4 | 0.157 |

| English-Bhojpuri | | | | | |
|-----------------------|--------------|-------------|----------|------------|----------|
| System Name | LP Supported | Params. (B) | Humeval? | AutoRank ↓ | chrF++ ↑ |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 1.0 | 40.6 |
| Wenyii | ✓ | 14 | ✓ | 2.5 | 38.9 |
| Algharb | ✓ | 14 | ✓ | 2.8 | 38.6 |
| ▲ ONLINE-B | ✓ | ? | ✓ | 4.1 | 37.1 |
| TranssionTranslate | ? | ? | ✓ | 4.4 | 36.9 |
| ▲ Claude-4 | ? | ? | ✓ | 4.5 | 36.7 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 5.1 | 36.0 |
| ▲ GPT-4.1 | ? | ? | ✓ | 5.5 | 35.6 |
| Yolu | ✓ | 14 | ✓ | 5.6 | 35.4 |
| TranssionMT | ✓ | 1 | ✓ | 6.2 | 34.8 |
| ▲ Llama-4-Maverick | ✓ | 400 | ✓ | 6.5 | 34.4 |
| ▲ CommandA | ✗ | 111 | ✓ | 6.5 | 34.4 |
| ▲ NLLB | ✓ | 1 | ✓ | 6.6 | 34.3 |
| ▲ Gemma-3-27B | ? | 27 | ✓ | 8.3 | 32.4 |
| CommandA-WMT | ✗ | 111 | | 8.8 | 31.8 |
| COILD-BHO | ✓ | 7 | ✓ | 8.9 | 31.8 |
| ▲ Mistral-Medium | ? | ? | | 9.0 | 31.6 |
| ▲ Qwen3-235B | ✗ | 235 | | 11.1 | 29.2 |
| IRB-MT | ✓ | 12 | ✓ | 11.4 | 28.9 |
| ▲ AyaExpans-32B | ✗ | 32 | | 11.4 | 28.9 |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 11.5 | 28.8 |
| GemTrans | ✓ | 27 | | 11.9 | 28.3 |
| SalamandraTA | ✓ | 8 | ✓ | 12.1 | 28.2 |
| ▲ Gemma-3-12B | ? | 12 | | 12.3 | 27.9 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | | 12.7 | 27.4 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | | 12.8 | 27.3 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | | 13.6 | 26.4 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | | 14.7 | 25.2 |
| IR-MultiagentMT | ✗ | ? | | 15.9 | 23.9 |
| ▲ CommandR7B | ✗ | 7 | | 16.7 | 22.9 |
| ▲ AyaExpans-8B | ✗ | 8 | | 16.7 | 22.9 |
| ▲ Qwen2.5-7B | ? | 7 | | 17.7 | 21.8 |
| ▲ Mistral-7B | ✗ | 7 | | 20.9 | 18.2 |
| UvA-MT | ✓ | 12 | | 28.4 | 9.7 |
| ▲ Llama-3.1-8B | ✗ | 8 | | 35.0 | 2.3 |

| English-Czech | | | | | | | | | |
|-----------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.658 | 83.7 | 89.4 | -5.5 | 0.639 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 3.4 | 0.633 | 83.8 | 91.5 | -6.2 | 0.574 |
| CommandA-WMT | ✓ | 111 | ✓ | 3.5 | 0.645 | 81.3 | 86.2 | -6.0 | 0.594 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 3.9 | 0.63 | 84.2 | 89.7 | -6.6 | 0.576 |
| Wenyiil | ✓ | 14 | ✓ | 4.4 | 0.645 | 79.4 | 86.3 | -6.4 | 0.586 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 5.0 | 0.628 | 81.4 | 87.0 | -6.5 | 0.565 |
| GemTrans | ✓ | 27 | ✓ | 5.0 | 0.636 | 76.6 | 81.8 | -5.8 | 0.596 |
| Algharb | ✓ | 14 | ✓ | 6.2 | 0.627 | 79.4 | 85.0 | -6.9 | 0.552 |
| Yolu | ✓ | 14 | ✓ | 6.3 | 0.651 | 74.6 | 78.6 | -6.5 | 0.582 |
| UvA-MT | ✓ | 12 | ✓ | 6.4 | 0.637 | 77.3 | 82.9 | -6.9 | 0.562 |
| ▲ Mistral-Medium | ? | ? | ✓ | 7.0 | 0.621 | 78.4 | 84.4 | -7.1 | 0.547 |
| SRPOL | ✓ | 12 | ✓ | 8.6 | 0.641 | 72.9 | 76.2 | -7.3 | 0.552 |
| ▲ CommandA | ✓ | 111 | ✓ | 8.6 | 0.609 | 78.2 | 82.5 | -7.6 | 0.524 |
| Lanigo | ✓ | 9 | ✓ | 8.6 | 0.643 | 67.3 | 69.9 | -6.5 | 0.608 |
| ▲ Claude-4 | ? | ? | ✓ | 8.8 | 0.606 | 78.6 | 83.0 | -7.9 | 0.522 |
| ▲ Gemma-3-27B | ✓ | 27 | ✓ | 9.0 | 0.606 | 76.9 | 81.5 | -7.5 | 0.523 |
| ▲ ONLINE-B | ✓ | ? | | 10.2 | 0.612 | 73.1 | 77.0 | -7.4 | 0.513 |
| ▲ AyaExpanse-32B | ✓ | 32 | | 10.2 | 0.604 | 74.2 | 78.9 | -7.8 | 0.519 |
| SalamandraTA | ✓ | 8 | ✓ | 10.3 | 0.624 | 70.1 | 74.5 | -7.3 | 0.528 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 11.1 | 0.595 | 75.3 | 79.7 | -8.3 | 0.494 |
| ▲ ONLINE-W | ? | ? | | 11.2 | 0.602 | 74.5 | 77.9 | -8.3 | 0.495 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | ✓ | 11.4 | 0.605 | 72.0 | 75.8 | -7.9 | 0.505 |
| ▲ Qwen3-235B | ✓ | 235 | | 11.5 | 0.599 | 71.8 | 76.0 | -7.8 | 0.505 |
| CUNI-MH-v2 | ✓ | 9 | ✓ | 11.9 | 0.609 | 69.2 | 73.4 | -7.9 | 0.517 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | | 12.5 | 0.593 | 72.2 | 75.0 | -8.4 | 0.488 |
| IRB-MT | ✓ | 12 | | 12.6 | 0.591 | 71.0 | 73.6 | -7.8 | 0.484 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | | 12.9 | 0.592 | 70.8 | 74.9 | -8.4 | 0.485 |
| TranssionTranslate | ? | ? | | 13.2 | 0.597 | 68.5 | 72.0 | -7.8 | 0.48 |
| ▲ Gemma-3-12B | ✓ | 12 | | 13.4 | 0.583 | 71.6 | 74.1 | -8.5 | 0.48 |
| CUNI-SFT | ✓ | 9 | | 15.9 | 0.575 | 66.9 | 68.2 | -8.9 | 0.468 |
| ▲ AyaExpanse-8B | ✓ | 8 | | 16.0 | 0.572 | 67.1 | 67.9 | -8.7 | 0.457 |
| CUNI-DocTransformer | ✓ | <1 | | 17.5 | 0.558 | 68.7 | 71.1 | -10.0 | 0.425 |
| IR-MultiagentMT | ✗ | ? | | 17.7 | 0.546 | 66.5 | 68.7 | -9.3 | 0.442 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | | 18.9 | 0.527 | 63.1 | 63.7 | -9.0 | 0.466 |
| ▲ NLLB | ✓ | 1 | | 25.5 | 0.485 | 55.7 | 57.3 | -10.8 | 0.392 |
| ▲ CommandR7B | ✓ | 7 | | 28.0 | 0.457 | 58.5 | 51.3 | -11.6 | 0.369 |
| ▲ ONLINE-G | ✓ | ? | | 28.7 | 0.472 | 58.1 | 58.0 | -12.8 | 0.313 |
| ▲ Llama-3.1-8B | ✗ | 8 | | 28.9 | 0.48 | 55.7 | 52.0 | -12.1 | 0.317 |
| ▲ Qwen2.5-7B | ? | 7 | | 37.5 | 0.41 | 46.2 | 43.8 | -14.2 | 0.239 |
| ▲ Mistral-7B | ✗ | 7 | | 40.8 | 0.374 | 45.8 | 41.2 | -15.5 | 0.207 |
| ctpc_nlp | ? | ? | | 41.1 | 0.369 | 43.3 | 39.8 | -14.8 | 0.207 |
| TranssionMT | ✓ | 1 | | 42.0 | 0.364 | 45.4 | 45.4 | -16.7 | 0.196 |

| English-Estonian | | | | | | | | | |
|-----------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.72 | 78.8 | 87.8 | -7.3 | 0.628 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 2.5 | 0.7 | 74.1 | 90.7 | -8.0 | 0.59 |
| Wenyiil | ✓ | 14 | ✓ | 2.6 | 0.708 | 74.4 | 86.0 | -8.0 | 0.599 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 3.0 | 0.695 | 75.2 | 87.9 | -8.6 | 0.577 |
| Yolu | ✓ | 14 | ✓ | 3.7 | 0.72 | 72.1 | 77.4 | -8.3 | 0.587 |
| Algharb | ✓ | 14 | ✓ | 3.8 | 0.692 | 73.6 | 84.1 | -8.7 | 0.558 |
| GemTrans | ✓ | 27 | ✓ | 4.9 | 0.689 | 70.8 | 74.3 | -8.3 | 0.558 |
| Laniqo | ✓ | 9 | ✓ | 5.1 | 0.711 | 67.2 | 68.1 | -8.2 | 0.602 |
| SRPOL | ✓ | 12 | ✓ | 5.5 | 0.705 | 70.5 | 74.2 | -9.7 | 0.538 |
| UvA-MT | ✓ | 12 | ✓ | 5.8 | 0.696 | 71.9 | 72.6 | -10.0 | 0.531 |
| ▲ ONLINE-B | ✓ | ? | ✓ | 5.8 | 0.678 | 69.9 | 76.5 | -9.2 | 0.521 |
| CommandA-WMT | ✗ | 111 | ✓ | 5.9 | 0.689 | 71.6 | 71.8 | -9.7 | 0.527 |
| SalamandraTA | ✓ | 8 | ✓ | 6.1 | 0.695 | 68.4 | 71.5 | -9.3 | 0.532 |
| ▲ Claude-4 | ? | ? | ✓ | 6.3 | 0.673 | 71.4 | 77.3 | -10.6 | 0.505 |
| TranssionTranslate | ? | ? | ✓ | 7.2 | 0.669 | 66.1 | 73.2 | -9.5 | 0.501 |
| ▲ Gemma-3-27B | ✓ | 27 | ✓ | 7.4 | 0.662 | 70.2 | 71.8 | -10.8 | 0.491 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | ✓ | 7.9 | 0.654 | 68.6 | 72.2 | -10.8 | 0.479 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 8.0 | 0.655 | 69.0 | 71.9 | -11.1 | 0.474 |
| ▲ ONLINE-W | ? | ? | | 8.6 | 0.654 | 67.9 | 70.3 | -11.6 | 0.471 |
| ▲ DeepSeek-V3 | ? | 671 | | 10.1 | 0.613 | 64.0 | 66.5 | -11.4 | 0.468 |
| IRB-MT | ✓ | 12 | ✓ | 11.1 | 0.609 | 65.6 | 60.5 | -11.8 | 0.413 |
| IR-MultiagentMT | ✗ | ? | | 11.1 | 0.605 | 64.5 | 62.7 | -11.9 | 0.423 |
| ▲ Gemma-3-12B | ✓ | 12 | | 12.1 | 0.597 | 65.6 | 59.4 | -13.0 | 0.387 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | | 13.5 | 0.522 | 57.3 | 55.0 | -11.0 | 0.463 |
| ▲ Mistral-Medium | ? | ? | | 13.9 | 0.574 | 59.9 | 54.8 | -13.5 | 0.4 |
| ▲ Qwen3-235B | ✓ | 235 | | 14.1 | 0.576 | 62.6 | 54.1 | -13.7 | 0.349 |
| ▲ CommandA | ✗ | 111 | | 15.9 | 0.546 | 63.9 | 48.4 | -15.4 | 0.316 |
| ▲ NLLB | ✓ | 1 | | 16.1 | 0.528 | 56.6 | 53.4 | -14.2 | 0.35 |
| ▲ ONLINE-G | ✓ | ? | | 16.9 | 0.532 | 57.2 | 55.3 | -15.6 | 0.297 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | | 20.2 | 0.491 | 54.7 | 40.3 | -17.0 | 0.254 |
| TranssionMT | ✓ | 1 | | 23.7 | 0.436 | 46.6 | 43.1 | -19.1 | 0.176 |
| ▲ Llama-3.1-8B | ✗ | 8 | | 24.5 | 0.424 | 47.8 | 33.7 | -18.7 | 0.166 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | | 27.2 | 0.403 | 42.1 | 13.2 | -19.0 | 0.19 |
| ▲ AyaExpanse-32B | ✗ | 32 | | 32.3 | 0.284 | 33.4 | 20.2 | -23.2 | 0.135 |
| ▲ Qwen2.5-7B | ? | 7 | | 33.6 | 0.273 | 27.6 | 17.8 | -23.6 | 0.144 |
| ▲ CommandR7B | ✗ | 7 | | 35.8 | 0.169 | 23.4 | 9.2 | -22.6 | 0.193 |
| ▲ Mistral-7B | ✗ | 7 | | 37.4 | 0.182 | 18.1 | 11.4 | -24.5 | 0.151 |
| ▲ AyaExpanse-8B | ✗ | 8 | | 38.0 | 0.151 | 17.4 | 10.1 | -24.7 | 0.171 |

| English-Icelandic | | | | | | | | | |
|-----------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuanyan-MT | ✓ | 7 | ✓ | 1.0 | 0.663 | 71.6 | 83.9 | -7.5 | 0.543 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 1.8 | 0.647 | 69.2 | 87.6 | -7.7 | 0.512 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 1.9 | 0.653 | 70.2 | 84.5 | -8.3 | 0.516 |
| Erlendur | ✓ | ? | ✓ | 2.2 | 0.646 | 69.5 | 85.1 | -8.2 | 0.506 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | ✓ | 3.9 | 0.64 | 67.1 | 76.3 | -8.8 | 0.471 |
| ▲ ONLINE-B | ✓ | ? | ✓ | 4.4 | 0.636 | 66.1 | 73.5 | -8.8 | 0.464 |
| ▲ Claude-4 | ? | ? | ✓ | 5.2 | 0.628 | 67.5 | 73.8 | -10.6 | 0.43 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | ✓ | 5.7 | 0.621 | 66.7 | 67.7 | -10.1 | 0.435 |
| TranssionTranslate | ? | ? | ✓ | 5.8 | 0.625 | 63.2 | 68.9 | -9.1 | 0.43 |
| UvA-MT | ✓ | 12 | ✓ | 6.8 | 0.627 | 68.1 | 59.1 | -11.6 | 0.402 |
| CommandA-WMT | ✗ | 111 | ✓ | 6.8 | 0.619 | 68.0 | 57.4 | -11.1 | 0.404 |
| GemTrans | ✓ | 27 | ✓ | 7.0 | 0.609 | 65.0 | 59.1 | -9.7 | 0.401 |
| AMI | ✓ | 3 | ✓ | 7.4 | 0.627 | 59.6 | 58.1 | -9.7 | 0.426 |
| SalamandraTA | ✓ | 8 | ✓ | 8.6 | 0.605 | 61.6 | 53.9 | -11.0 | 0.386 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 8.8 | 0.587 | 64.7 | 58.8 | -12.3 | 0.357 |
| ▲ Mistral-Medium | ? | ? | | 9.7 | 0.583 | 65.3 | 51.5 | -13.0 | 0.337 |
| ▲ Gemma-3-27B | ✓ | 27 | | 9.7 | 0.572 | 62.2 | 54.9 | -12.4 | 0.364 |
| ▲ DeepSeek-V3 | ? | 671 | | 10.5 | 0.547 | 58.0 | 56.6 | -12.1 | 0.378 |
| IRB-MT | ✓ | 12 | ✓ | 11.9 | 0.542 | 61.2 | 47.2 | -13.6 | 0.306 |
| IR-MultiagentMT | ✗ | ? | | 12.1 | 0.53 | 60.0 | 51.3 | -13.7 | 0.31 |
| ▲ Qwen3-235B | ✗ | 235 | | 13.5 | 0.525 | 60.5 | 41.5 | -15.0 | 0.275 |
| ▲ Gemma-3-12B | ✓ | 12 | ✓ | 13.8 | 0.517 | 60.3 | 42.1 | -15.4 | 0.268 |
| ▲ NLLB | ✓ | 1 | ✓ | 15.2 | 0.477 | 53.0 | 48.2 | -15.0 | 0.27 |
| ▲ ONLINE-G | ✓ | ? | | 15.8 | 0.477 | 53.4 | 49.2 | -16.1 | 0.243 |
| ▲ CommandA | ✗ | 111 | | 16.2 | 0.475 | 59.0 | 37.4 | -17.0 | 0.221 |
| ▲ Llama-3.1-8B | ✗ | 8 | ✓ | 24.8 | 0.323 | 42.7 | 24.6 | -21.3 | 0.133 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | | 25.5 | 0.303 | 32.9 | 9.2 | -17.4 | 0.237 |
| ▲ AyaExpanse-32B | ✗ | 32 | | 28.0 | 0.275 | 35.2 | 18.4 | -23.3 | 0.145 |
| ▲ CommandR7B | ✗ | 7 | | 30.3 | 0.2 | 23.4 | 9.1 | -20.9 | 0.216 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | | 30.8 | 0.206 | 26.5 | 13.7 | -23.7 | 0.171 |
| ▲ Mistral-7B | ✗ | 7 | | 31.8 | 0.177 | 25.2 | 14.3 | -24.3 | 0.17 |
| ▲ Qwen2.5-7B | ? | 7 | | 31.8 | 0.186 | 24.1 | 13.1 | -24.3 | 0.174 |
| ▲ AyaExpanse-8B | ✗ | 8 | | 33.0 | 0.153 | 21.7 | 11.3 | -24.6 | 0.177 |

| English-Italian | | | | | | | | |
|-----------------------|-------------------|----------------|----------|---------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuanyuan-MT | ✓ | 7 | ✓ | 1.0 | 84.6 | 88.7 | -4.7 | 0.62 |
| CommandA-WMT | ✓ | 111 | ✓ | 2.6 | 83.4 | 88.0 | -4.8 | 0.59 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 4.4 | 85.5 | 90.5 | -5.6 | 0.537 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 4.5 | 85.0 | 89.8 | -5.8 | 0.553 |
| GemTrans | ✓ | 27 | ✓ | 5.2 | 78.2 | 83.5 | -4.9 | 0.581 |
| UvA-MT | ✓ | 12 | ✓ | 5.3 | 78.9 | 84.6 | -5.4 | 0.595 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 6.1 | 81.9 | 87.9 | -5.9 | 0.543 |
| ▲ Mistral-Medium | ? | ? | ✓ | 7.1 | 79.9 | 86.4 | -6.0 | 0.544 |
| ▲ Qwen3-235B | ✓ | 235 | ✓ | 7.2 | 80.1 | 84.9 | -5.8 | 0.541 |
| Laniqo | ✓ | 9 | ✓ | 7.6 | 70.5 | 75.3 | -4.9 | 0.63 |
| ▲ Claude-4 | ✓ | ? | ✓ | 8.4 | 81.7 | 85.2 | -6.4 | 0.52 |
| ▲ CommandA | ✓ | 111 | ✓ | 8.5 | 79.4 | 83.7 | -6.2 | 0.537 |
| ▲ ONLINE-B | ✓ | ? | | 9.4 | 76.7 | 78.6 | -5.6 | 0.53 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | | 9.4 | 76.2 | 81.8 | -6.1 | 0.539 |
| ▲ AyaExpanse-32B | ✓ | 32 | | 10.1 | 75.7 | 80.9 | -6.1 | 0.527 |
| ▲ ONLINE-W | ? | ? | | 10.1 | 74.6 | 81.1 | -6.0 | 0.531 |
| IRB-MT | ✓ | 12 | ✓ | 10.2 | 73.8 | 79.8 | -5.7 | 0.523 |
| SalamandraTA | ✓ | 8 | ✓ | 10.3 | 71.9 | 76.9 | -5.8 | 0.561 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | | 11.0 | 74.2 | 79.8 | -6.4 | 0.53 |
| TranssionTranslate | ? | ? | | 11.0 | 72.9 | 77.1 | -5.7 | 0.523 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | ✓ | 11.3 | 73.5 | 78.6 | -6.2 | 0.526 |
| ▲ Gemma-3-27B | ✓ | 27 | | 12.6 | 73.4 | 78.3 | -6.7 | 0.513 |
| IR-MultiagentMT | ✗ | ? | | 13.6 | 73.0 | 77.0 | -6.8 | 0.499 |
| ▲ AyaExpanse-8B | ✓ | 8 | ✓ | 14.9 | 69.5 | 73.9 | -6.7 | 0.502 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | ✓ | 15.2 | 68.3 | 73.5 | -6.8 | 0.509 |
| ▲ Gemma-3-12B | ✓ | 12 | ✓ | 15.5 | 69.7 | 74.7 | -7.1 | 0.494 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 18.0 | 67.0 | 71.9 | -7.5 | 0.479 |
| ▲ CommandR7B | ✓ | 7 | | 18.0 | 67.3 | 69.4 | -7.4 | 0.486 |
| ▲ Llama-3.1-8B | ✓ | 8 | | 22.8 | 61.8 | 64.1 | -8.1 | 0.449 |
| ▲ Qwen2.5-7B | ✓ | 7 | | 23.5 | 60.8 | 61.5 | -7.8 | 0.44 |
| ▲ NLLB | ✓ | 1 | | 27.1 | 58.5 | 61.6 | -9.3 | 0.421 |
| ▲ ONLINE-G | ✓ | ? | | 30.0 | 58.7 | 60.6 | -9.9 | 0.368 |
| ▲ Mistral-7B | ✗ | 7 | | 33.0 | 53.5 | 52.1 | -9.8 | 0.363 |

| English-Japanese | | | | | | | | | |
|-----------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.687 | 82.2 | 89.6 | -5.5 | 0.592 |
| In2x | ? | 72 | ✓ | 2.3 | 0.711 | 78.4 | 86.3 | -5.9 | 0.575 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 2.4 | 0.672 | 83.2 | 91.2 | -5.7 | 0.55 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 2.9 | 0.674 | 81.8 | 89.7 | -5.9 | 0.558 |
| Wenyiil | ✓ | 14 | ✓ | 2.9 | 0.682 | 79.6 | 88.6 | -5.7 | 0.553 |
| KIKIS | ✓ | 18 | ✓ | 3.1 | 0.678 | 80.2 | 85.4 | -5.5 | 0.551 |
| Algharb | ✓ | 14 | ✓ | 3.2 | 0.678 | 80.8 | 89.2 | -5.8 | 0.541 |
| CommandA-WMT | ✓ | 111 | ✓ | 3.6 | 0.694 | 76.5 | 85.5 | -5.8 | 0.55 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 4.6 | 0.667 | 80.5 | 87.7 | -6.2 | 0.531 |
| ▲ Mistral-Medium | ? | ? | ✓ | 5.4 | 0.675 | 77.6 | 86.2 | -6.4 | 0.532 |
| GemTrans | ✓ | 27 | ✓ | 5.5 | 0.667 | 72.3 | 80.2 | -5.5 | 0.553 |
| ▲ Claude-4 | ✓ | ? | ✓ | 5.7 | 0.677 | 78.3 | 86.3 | -6.5 | 0.516 |
| Yolu | ✓ | 14 | ✓ | 5.9 | 0.697 | 69.9 | 77.9 | -5.9 | 0.541 |
| ▲ ONLINE-B | ✓ | ? | ✓ | 6.1 | 0.684 | 72.4 | 80.0 | -6.0 | 0.527 |
| UvA-MT | ✓ | 12 | ✓ | 6.4 | 0.691 | 72.5 | 82.0 | -6.3 | 0.517 |
| bb88 | ? | ? | | 7.2 | 0.674 | 74.6 | 82.6 | -6.5 | 0.498 |
| ▲ CommandA | ✓ | 111 | | 7.3 | 0.674 | 74.8 | 82.9 | -6.6 | 0.504 |
| ▲ Qwen3-235B | ✓ | 235 | | 7.3 | 0.667 | 74.3 | 83.2 | -6.4 | 0.499 |
| Systran | ✓ | 18 | ✓ | 7.3 | 0.703 | 68.7 | 77.1 | -6.5 | 0.523 |
| ▲ Gemma-3-27B | ✓ | 27 | | 7.9 | 0.666 | 74.3 | 82.2 | -6.6 | 0.497 |
| NTTSU | ✓ | 14 | ✓ | 8.0 | 0.676 | 67.7 | 74.3 | -5.6 | 0.498 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | | 8.6 | 0.671 | 71.4 | 80.6 | -6.8 | 0.499 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 9.1 | 0.661 | 71.5 | 81.1 | -6.8 | 0.487 |
| Laniqo | ✓ | 9 | ✓ | 9.3 | 0.677 | 66.1 | 70.1 | -6.3 | 0.529 |
| ▲ AyaExpanse-32B | ✓ | 32 | | 9.8 | 0.662 | 70.9 | 78.5 | -6.8 | 0.472 |
| IRB-MT | ✓ | 12 | | 10.3 | 0.643 | 70.0 | 77.9 | -6.5 | 0.474 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | | 10.4 | 0.665 | 68.7 | 76.3 | -6.9 | 0.477 |
| SRPOL | ✗ | 12 | | 10.8 | 0.683 | 66.7 | 73.9 | -7.1 | 0.472 |
| TranssionTranslate | ? | ? | | 12.0 | 0.668 | 64.6 | 71.8 | -6.9 | 0.459 |
| ▲ Gemma-3-12B | ✓ | 12 | | 13.6 | 0.623 | 64.7 | 73.8 | -7.0 | 0.461 |
| ▲ AyaExpanse-8B | ✓ | 8 | | 15.6 | 0.632 | 62.1 | 69.4 | -7.4 | 0.422 |
| ▲ ONLINE-W | ? | ? | | 16.4 | 0.611 | 61.7 | 67.5 | -7.3 | 0.432 |
| SH | ✓ | 56 | | 16.4 | 0.641 | 59.9 | 65.6 | -7.5 | 0.419 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | | 16.8 | 0.623 | 62.0 | 69.4 | -7.9 | 0.425 |
| ▲ CommandR7B | ✓ | 7 | | 19.9 | 0.62 | 59.3 | 65.1 | -8.6 | 0.379 |
| IR-MultiagentMT | ✗ | ? | | 22.9 | 0.576 | 54.6 | 62.1 | -8.6 | 0.373 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | | 23.6 | 0.561 | 53.8 | 60.5 | -8.6 | 0.391 |
| ▲ Qwen2.5-7B | ✓ | 7 | | 24.5 | 0.594 | 54.6 | 58.6 | -9.2 | 0.338 |
| ▲ Llama-3.1-8B | ✗ | 8 | | 25.5 | 0.596 | 51.4 | 54.8 | -9.0 | 0.32 |
| SalamandraTA | ✓ | 8 | | 26.4 | 0.603 | 51.8 | 53.1 | -9.4 | 0.299 |
| ▲ NLLB | ✓ | 1 | | 37.9 | 0.479 | 42.7 | 46.8 | -11.5 | 0.245 |
| ▲ ONLINE-G | ✓ | ? | | 40.7 | 0.495 | 45.0 | 45.8 | -13.2 | 0.207 |
| ▲ Mistral-7B | ✗ | 7 | | 43.0 | 0.462 | 39.2 | 40.2 | -12.6 | 0.193 |

| English-Korean | | | | | | | | | |
|-----------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuanyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.697 | 83.8 | 85.6 | -4.9 | 0.624 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 2.5 | 0.683 | 85.3 | 88.1 | -5.6 | 0.571 |
| CommandA-WMT | ✓ | 111 | ✓ | 2.8 | 0.711 | 79.6 | 82.3 | -5.6 | 0.584 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 2.8 | 0.686 | 83.6 | 86.3 | -5.7 | 0.581 |
| Wenyiil | ✓ | 14 | ✓ | 2.9 | 0.691 | 82.1 | 85.0 | -5.6 | 0.576 |
| Algharb | ✓ | 14 | ✓ | 3.0 | 0.687 | 83.2 | 85.9 | -5.7 | 0.565 |
| UvA-MT | ✓ | 12 | ✓ | 4.2 | 0.706 | 78.2 | 81.1 | -6.0 | 0.554 |
| ▲ Claude-4 | ✓ | ? | ✓ | 4.3 | 0.694 | 82.0 | 84.6 | -6.3 | 0.536 |
| GemTrans | ✓ | 27 | ✓ | 4.9 | 0.677 | 76.7 | 78.8 | -5.4 | 0.568 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 5.0 | 0.681 | 79.5 | 83.7 | -6.1 | 0.539 |
| ▲ CommandA | ✓ | 111 | ✓ | 5.8 | 0.692 | 77.8 | 80.9 | -6.6 | 0.524 |
| ▲ Mistral-Medium | ? | ? | ✓ | 6.0 | 0.684 | 77.2 | 79.4 | -6.3 | 0.53 |
| ▲ Qwen3-235B | ✓ | 235 | ✓ | 6.3 | 0.678 | 77.1 | 80.0 | -6.2 | 0.509 |
| Yolu | ✓ | 14 | ✓ | 6.8 | 0.701 | 70.1 | 73.0 | -5.9 | 0.533 |
| ▲ ONLINE-B | ✓ | ? | ✓ | 7.8 | 0.679 | 72.8 | 73.8 | -6.2 | 0.504 |
| IRB-MT | ✓ | 12 | ✓ | 8.4 | 0.657 | 74.9 | 76.3 | -6.4 | 0.489 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | | 8.5 | 0.684 | 72.2 | 75.2 | -6.8 | 0.487 |
| ▲ AyaExpanse-32B | ✓ | 32 | | 8.6 | 0.673 | 72.1 | 75.7 | -6.7 | 0.493 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 8.8 | 0.665 | 73.6 | 75.0 | -6.7 | 0.487 |
| Laniqo | ✓ | 9 | ✓ | 8.9 | 0.689 | 64.9 | 66.2 | -6.1 | 0.54 |
| ▲ Gemma-3-12B | ✓ | 12 | ✓ | 9.0 | 0.667 | 73.6 | 77.0 | -7.0 | 0.474 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | ✓ | 9.8 | 0.678 | 70.2 | 73.5 | -7.2 | 0.472 |
| ▲ ONLINE-W | ? | ? | | 10.4 | 0.674 | 67.7 | 69.4 | -6.8 | 0.467 |
| TranssionTranslate | ? | ? | | 12.0 | 0.675 | 64.0 | 65.0 | -6.9 | 0.439 |
| ▲ AyaExpanse-8B | ✓ | 8 | | 12.7 | 0.657 | 64.6 | 67.7 | -7.3 | 0.434 |
| ▲ Gemma-3-27B | ✓ | 27 | | 12.9 | 0.626 | 67.1 | 67.9 | -7.4 | 0.477 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | | 13.0 | 0.654 | 66.4 | 68.9 | -7.7 | 0.422 |
| IR-MultiagentMT | ✗ | ? | | 16.3 | 0.614 | 61.2 | 64.2 | -8.1 | 0.41 |
| ▲ CommandR7B | ✓ | 7 | | 18.2 | 0.619 | 59.8 | 61.3 | -8.8 | 0.364 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | | 19.1 | 0.594 | 56.8 | 57.7 | -8.3 | 0.39 |
| SalamandraTA | ✓ | 8 | | 22.8 | 0.624 | 50.6 | 50.3 | -9.7 | 0.29 |
| ▲ Llama-3.1-8B | ✗ | 8 | | 24.8 | 0.586 | 50.7 | 50.8 | -10.2 | 0.278 |
| ▲ Qwen2.5-7B | ✓ | 7 | | 25.0 | 0.568 | 47.9 | 48.7 | -9.4 | 0.291 |
| ▲ NLLB | ✓ | 1 | | 28.1 | 0.549 | 42.8 | 44.3 | -10.4 | 0.286 |
| ▲ ONLINE-G | ✓ | ? | | 32.4 | 0.532 | 44.8 | 44.3 | -12.8 | 0.187 |
| ▲ Mistral-7B | ✗ | 7 | | 36.0 | 0.478 | 37.8 | 39.3 | -12.7 | 0.174 |

| English-Maasai | | | | | |
|-----------------------|--------------|-------------|----------|------------|----------|
| System Name | LP Supported | Params. (B) | Humeval? | AutoRank ↓ | chrF++ ↑ |
| Shy-hunyuan-MT | ✗ | 7 | ✓ | 1.0 | 27.7 |
| ▲ Claude-4 | ? | ? | ✓ | 2.6 | 26.1 |
| ▲ Qwen3-235B | ✗ | 235 | ✓ | 3.0 | 25.6 |
| ▲ Llama-4-Maverick | ✗ | 400 | ✓ | 3.2 | 25.4 |
| ▲ CommandR7B | ✗ | 7 | ✓ | 4.3 | 24.3 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | ✓ | 5.3 | 23.2 |
| TranssionMT | ✓ | 1 | ✓ | 5.9 | 22.6 |
| ▲ Gemini-2.5-Pro | ? | ? | ✓ | 6.1 | 22.5 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 6.2 | 22.4 |
| CommandA-WMT | ✗ | 111 | ✓ | 6.4 | 22.2 |
| ▲ AyaExpanse-32B | ✗ | 32 | ✓ | 7.1 | 21.4 |
| ▲ CommandA | ✗ | 111 | ✓ | 7.9 | 20.6 |
| ▲ Llama-3.1-8B | ✗ | 8 | ✓ | 8.1 | 20.4 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | ✓ | 8.2 | 20.3 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | ✓ | 8.2 | 20.3 |
| ▲ AyaExpanse-8B | ✗ | 8 | ✓ | 8.2 | 20.2 |
| ▲ Qwen2.5-7B | ? | 7 | ✓ | 8.6 | 19.9 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | | 8.8 | 19.7 |
| ▲ Gemma-3-12B | ? | 12 | ✓ | 8.8 | 19.6 |
| IR-MultiagentMT | ✗ | ? | | 9.0 | 19.5 |
| IRB-MT | ✓ | 12 | | 9.7 | 18.7 |
| ▲ Mistral-7B | ✗ | 7 | | 11.3 | 17.1 |
| ▲ Gemma-3-27B | ? | 27 | | 13.3 | 15.1 |
| UvA-MT | ✓ | 12 | | 14.7 | 13.6 |
| ▲ GPT-4.1 | ? | ? | | 14.9 | 13.4 |
| GemTrans | ✗ | 27 | | 16.7 | 11.6 |
| ▲ NLLB | ✗ | 1 | | 27.0 | 0.9 |

| English-Russian | | | | | | | | | |
|-----------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.657 | 84.3 | 85.9 | -4.9 | 0.652 |
| CommandA-WMT | ✓ | 111 | ✓ | 4.2 | 0.656 | 81.3 | 80.5 | -5.8 | 0.607 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 4.3 | 0.634 | 85.9 | 87.8 | -6.1 | 0.575 |
| Yandex | ✓ | ? | ✓ | 4.4 | 0.638 | 81.2 | 80.6 | -5.3 | 0.617 |
| UvA-MT | ✓ | 12 | ✓ | 4.5 | 0.662 | 78.6 | 80.5 | -6.1 | 0.611 |
| Wenyiil | ✓ | 14 | ✓ | 4.7 | 0.644 | 82.5 | 84.1 | -6.1 | 0.588 |
| GemTrans | ✓ | 27 | ✓ | 5.1 | 0.639 | 77.8 | 79.5 | -5.3 | 0.617 |
| Algharb | ✓ | 14 | ✓ | 5.1 | 0.637 | 84.4 | 85.5 | -6.4 | 0.573 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 5.3 | 0.631 | 84.6 | 85.8 | -6.5 | 0.577 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 5.6 | 0.632 | 84.2 | 84.7 | -6.4 | 0.57 |
| Yolu | ✓ | 14 | ✓ | 6.9 | 0.658 | 73.1 | 73.6 | -6.0 | 0.596 |
| ▲ Claude-4 | ✓ | ? | ✓ | 8.5 | 0.619 | 82.0 | 81.6 | -7.5 | 0.548 |
| ▲ Qwen3-235B | ✓ | 235 | ✓ | 8.7 | 0.625 | 78.2 | 79.9 | -6.9 | 0.543 |
| ▲ Gemma-3-27B | ✓ | 27 | ✓ | 8.7 | 0.626 | 78.9 | 79.6 | -7.3 | 0.551 |
| Laniqo | ✓ | 9 | ✓ | 8.7 | 0.649 | 67.9 | 67.0 | -6.0 | 0.622 |
| RuZh | ? | 9 | ✓ | 9.5 | 0.633 | 74.5 | 74.7 | -7.0 | 0.558 |
| IRB-MT | ✓ | 12 | ✓ | 9.9 | 0.616 | 75.8 | 76.5 | -6.7 | 0.541 |
| SRPOL | ✓ | 12 | ✓ | 10.5 | 0.647 | 71.8 | 71.6 | -7.7 | 0.549 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | | 10.5 | 0.624 | 75.7 | 75.2 | -7.6 | 0.543 |
| ▲ CommandA | ✓ | 111 | | 11.0 | 0.618 | 76.9 | 76.3 | -8.1 | 0.536 |
| ▲ ONLINE-W | ? | ? | | 11.5 | 0.624 | 73.7 | 73.6 | -7.9 | 0.534 |
| DLUT_GTCOM | ✓ | 27 | | 11.6 | 0.626 | 71.1 | 71.2 | -7.3 | 0.537 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | | 11.8 | 0.617 | 73.2 | 72.7 | -7.5 | 0.533 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 12.3 | 0.616 | 75.3 | 75.5 | -8.5 | 0.513 |
| SalamandraTA | ✓ | 8 | | 12.3 | 0.632 | 69.6 | 68.2 | -7.6 | 0.534 |
| ▲ ONLINE-B | ✓ | ? | | 12.7 | 0.616 | 73.8 | 72.8 | -8.2 | 0.517 |
| TranssionTranslate | ? | ? | | 12.7 | 0.618 | 70.7 | 71.0 | -7.5 | 0.52 |
| ▲ AyaExpanse-32B | ✓ | 32 | | 13.5 | 0.603 | 71.8 | 72.1 | -7.9 | 0.517 |
| ▲ ONLINE-G | ✓ | ? | | 14.2 | 0.613 | 67.8 | 66.6 | -7.6 | 0.522 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | | 14.4 | 0.606 | 70.9 | 71.3 | -8.4 | 0.502 |
| ▲ Gemma-3-12B | ✓ | 12 | | 14.7 | 0.589 | 72.6 | 73.2 | -8.4 | 0.503 |
| ▲ AyaExpanse-8B | ✓ | 8 | | 17.4 | 0.589 | 67.1 | 66.2 | -8.7 | 0.48 |
| IR-MultiagentMT | ✗ | ? | | 19.4 | 0.564 | 65.7 | 65.7 | -8.9 | 0.467 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | | 21.5 | 0.547 | 63.6 | 63.1 | -9.5 | 0.471 |
| ▲ Qwen2.5-7B | ✓ | 7 | | 24.8 | 0.546 | 60.2 | 57.5 | -10.2 | 0.411 |
| ▲ Llama-3.1-8B | ✗ | 8 | | 29.7 | 0.521 | 58.6 | 55.1 | -12.6 | 0.372 |
| ▲ NLLB | ✓ | 1 | | 31.5 | 0.483 | 54.3 | 53.3 | -11.7 | 0.389 |
| TranssionMT | ✓ | 1 | | 34.2 | 0.483 | 54.3 | 54.9 | -13.7 | 0.332 |
| ▲ Mistral-7B | ✗ | 7 | | 39.0 | 0.45 | 52.4 | 46.3 | -14.4 | 0.288 |
| ▲ CommandR7B | ✓ | 7 | | 40.0 | 0.41 | 52.1 | 39.9 | -13.6 | 0.347 |

| English-Serbian (Cyrilics) | | | | | | | | | |
|----------------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.687 | 76.6 | 83.3 | -4.2 | 0.64 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 3.0 | 0.663 | 74.6 | 87.2 | -5.1 | 0.566 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 3.4 | 0.655 | 74.4 | 83.4 | -5.2 | 0.582 |
| GemTrans | ✓ | 27 | ✓ | 4.6 | 0.663 | 71.6 | 74.5 | -4.9 | 0.554 |
| UvA-MT | ✓ | 12 | ✓ | 5.8 | 0.658 | 71.4 | 70.2 | -4.5 | 0.46 |
| ▲ ONLINE-B | ✓ | ? | ✓ | 6.1 | 0.644 | 71.0 | 75.2 | -5.7 | 0.517 |
| ▲ Claude-4 | ? | ? | ✓ | 6.8 | 0.628 | 72.6 | 77.4 | -6.6 | 0.503 |
| CommandA-WMT | ✗ | 111 | ✓ | 7.0 | 0.641 | 71.8 | 67.3 | -6.0 | 0.512 |
| TranssionTranslate | ? | ? | ✓ | 8.0 | 0.631 | 67.3 | 70.9 | -6.0 | 0.484 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 8.6 | 0.603 | 68.0 | 72.0 | -6.6 | 0.501 |
| SalamandraTA | ✓ | 8 | ✓ | 8.8 | 0.635 | 66.2 | 65.1 | -6.2 | 0.48 |
| DLUT_GTCOM | ✓ | 27 | ✓ | 9.3 | 0.618 | 66.9 | 68.1 | -6.6 | 0.463 |
| IRB-MT | ✓ | 12 | ✓ | 9.9 | 0.604 | 67.9 | 64.2 | -6.5 | 0.435 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 10.0 | 0.603 | 68.2 | 68.9 | -7.3 | 0.444 |
| ▲ Qwen3-235B | ✓ | 235 | | 11.9 | 0.591 | 65.8 | 60.1 | -7.6 | 0.425 |
| ▲ Gemma-3-12B | ✓ | 12 | ✓ | 12.1 | 0.583 | 65.7 | 61.8 | -7.4 | 0.394 |
| ▲ Gemma-3-27B | ✓ | 27 | | 12.2 | 0.583 | 61.7 | 62.1 | -7.4 | 0.444 |
| CUNI-SFT | ✓ | 9 | ✓ | 13.5 | 0.569 | 61.1 | 52.4 | -5.8 | 0.328 |
| IR-MultiagentMT | ✗ | ? | | 14.1 | 0.548 | 63.3 | 59.2 | -8.1 | 0.386 |
| ▲ ONLINE-G | ✓ | ? | | 14.5 | 0.566 | 58.8 | 56.2 | -7.7 | 0.383 |
| ▲ CommandA | ✗ | 111 | | 17.6 | 0.527 | 62.9 | 50.8 | -10.0 | 0.323 |
| ▲ Llama-3.1-8B | ✗ | 8 | ✓ | 19.4 | 0.489 | 53.9 | 44.2 | -7.5 | 0.233 |
| ▲ NLLB | ✓ | 1 | ✓ | 19.8 | 0.468 | 53.5 | 50.3 | -9.4 | 0.33 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | | 20.6 | 0.469 | 53.6 | 41.4 | -8.6 | 0.269 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | ✓ | 22.4 | 0.454 | 51.5 | 37.4 | -9.4 | 0.265 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | | 26.0 | 0.424 | 51.6 | 36.9 | -12.4 | 0.203 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | | 26.7 | 0.368 | 43.7 | 29.2 | -9.1 | 0.182 |
| ▲ Mistral-7B | ✗ | 7 | | 27.0 | 0.414 | 49.2 | 38.3 | -13.0 | 0.207 |
| ▲ AyaExpanse-8B | ✗ | 8 | | 29.7 | 0.306 | 40.9 | 27.3 | -10.1 | 0.157 |
| ▲ CommandR7B | ✗ | 7 | | 31.3 | 0.307 | 38.0 | 26.0 | -11.6 | 0.171 |
| ▲ AyaExpanse-32B | ✗ | 32 | | 31.8 | 0.354 | 46.4 | 29.6 | -15.2 | 0.142 |
| ▲ Qwen2.5-7B | ? | 7 | | 32.0 | 0.306 | 37.0 | 27.2 | -11.9 | 0.144 |

| English-Ukrainian | | | | | | | | | |
|-----------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.65 | 84.1 | 85.3 | -5.0 | 0.662 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 3.3 | 0.625 | 84.6 | 89.8 | -6.3 | 0.59 |
| Wenyiil | ✓ | 14 | ✓ | 3.4 | 0.635 | 83.7 | 85.4 | -6.2 | 0.597 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 3.4 | 0.626 | 82.8 | 87.0 | -6.2 | 0.611 |
| CommandA-WMT | ✓ | 111 | ✓ | 3.8 | 0.641 | 80.4 | 82.4 | -6.0 | 0.599 |
| Algharb | ✓ | 14 | ✓ | 4.1 | 0.625 | 83.2 | 86.0 | -6.5 | 0.586 |
| UvA-MT | ✓ | 12 | ✓ | 4.3 | 0.641 | 78.7 | 81.5 | -6.3 | 0.6 |
| GemTrans | ✓ | 27 | ✓ | 4.5 | 0.628 | 78.0 | 80.1 | -5.7 | 0.606 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 4.9 | 0.619 | 81.7 | 84.0 | -6.5 | 0.574 |
| Yolu | ✓ | 14 | ✓ | 5.9 | 0.643 | 73.4 | 74.4 | -6.2 | 0.589 |
| ▲ Mistral-Medium | ? | ? | ✓ | 5.9 | 0.617 | 79.8 | 82.1 | -6.9 | 0.566 |
| ▲ Claude-4 | ? | ? | ✓ | 6.9 | 0.604 | 81.1 | 82.6 | -7.6 | 0.544 |
| ▲ CommandA | ✓ | 111 | ✓ | 7.3 | 0.61 | 78.1 | 79.8 | -7.4 | 0.546 |
| Laniqo | ✓ | 9 | ✓ | 7.5 | 0.638 | 67.3 | 66.3 | -6.3 | 0.613 |
| IRB-MT | ✓ | 12 | ✓ | 8.0 | 0.604 | 74.8 | 76.9 | -6.9 | 0.539 |
| SRPOL | ✓ | 12 | ✓ | 8.2 | 0.631 | 70.9 | 72.9 | -7.3 | 0.548 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | ✓ | 8.6 | 0.603 | 73.4 | 75.2 | -7.2 | 0.541 |
| ▲ Llama-4-Maverick | ✓ | 400 | ✓ | 8.6 | 0.603 | 76.3 | 78.1 | -7.8 | 0.519 |
| CGFOKUS | ✓ | 235 | | 8.7 | 0.597 | 75.7 | 78.1 | -7.4 | 0.513 |
| ▲ ONLINE-B | ✓ | ? | | 8.8 | 0.609 | 73.2 | 73.3 | -7.3 | 0.531 |
| ▲ AyaExpanse-32B | ✓ | 32 | | 9.1 | 0.6 | 73.9 | 75.0 | -7.5 | 0.528 |
| ▲ ONLINE-W | ? | ? | | 9.1 | 0.605 | 72.8 | 75.0 | -7.5 | 0.527 |
| ▲ Qwen3-235B | ✓ | 235 | | 9.3 | 0.596 | 73.8 | 75.7 | -7.5 | 0.515 |
| SalamandraTA | ✓ | 8 | | 9.9 | 0.613 | 68.3 | 68.6 | -7.2 | 0.528 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | | 10.2 | 0.592 | 72.4 | 73.8 | -7.9 | 0.514 |
| TranssionTranslate | ? | ? | | 10.7 | 0.594 | 69.1 | 71.3 | -7.5 | 0.505 |
| DLUT_GTCOM | ✓ | 27 | | 11.0 | 0.592 | 69.8 | 71.4 | -7.9 | 0.498 |
| ▲ Gemma-3-27B | ✓ | 27 | | 11.9 | 0.575 | 68.1 | 71.0 | -8.1 | 0.51 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | | 12.5 | 0.577 | 68.9 | 69.4 | -8.6 | 0.492 |
| ▲ AyaExpanse-8B | ✓ | 8 | | 13.1 | 0.576 | 66.5 | 67.8 | -8.4 | 0.477 |
| CUNI-SFT | ✓ | 9 | | 13.3 | 0.579 | 66.1 | 65.5 | -8.5 | 0.484 |
| ▲ ONLINE-G | ✓ | ? | | 13.7 | 0.575 | 64.2 | 65.0 | -8.4 | 0.479 |
| IR-MultiagentMT | ✗ | ? | | 14.0 | 0.555 | 67.3 | 67.3 | -8.5 | 0.467 |
| ▲ Gemma-3-12B | ✓ | 12 | | 14.4 | 0.559 | 64.9 | 65.8 | -8.6 | 0.473 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | | 17.0 | 0.518 | 63.1 | 61.8 | -9.0 | 0.459 |
| ▲ NLLB | ✓ | 1 | | 24.0 | 0.467 | 53.2 | 53.6 | -11.2 | 0.368 |
| ▲ Llama-3.1-8B | ✗ | 8 | | 24.3 | 0.488 | 55.5 | 51.0 | -11.9 | 0.331 |
| TranssionMT | ✓ | 1 | | 28.1 | 0.441 | 51.8 | 52.2 | -13.5 | 0.286 |
| ▲ CommandR7B | ✓ | 7 | | 29.0 | 0.411 | 54.6 | 43.6 | -13.2 | 0.323 |
| ▲ Mistral-7B | ✗ | 7 | | 29.3 | 0.428 | 52.2 | 46.2 | -13.4 | 0.277 |
| ▲ Qwen2.5-7B | ? | 7 | | 36.6 | 0.362 | 41.8 | 36.0 | -15.2 | 0.2 |
| KYUoM | ? | <1 | | 42.0 | 0.265 | 35.9 | 34.7 | -16.6 | 0.201 |

| English-Simplified Chinese | | | | | | | | | |
|----------------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.67 | 87.2 | 88.3 | -4.0 | 0.576 |
| Wenyi1 | ✓ | 14 | ✓ | 3.9 | 0.663 | 84.2 | 87.7 | -5.0 | 0.52 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 4.0 | 0.657 | 85.2 | 88.7 | -4.9 | 0.512 |
| Algharb | ✓ | 14 | ✓ | 4.1 | 0.66 | 84.7 | 87.8 | -5.0 | 0.515 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 4.6 | 0.652 | 84.9 | 86.8 | -5.0 | 0.512 |
| ▲ Qwen3-235B | ✓ | 235 | ✓ | 4.8 | 0.661 | 82.7 | 85.0 | -5.0 | 0.513 |
| Yolu | ✓ | 14 | ✓ | 4.8 | 0.687 | 74.9 | 77.1 | -4.6 | 0.542 |
| GemTrans | ✓ | 27 | ✓ | 4.9 | 0.658 | 77.0 | 80.2 | -4.3 | 0.546 |
| ▲ Mistral-Medium | ? | ? | ✓ | 4.9 | 0.658 | 82.4 | 84.9 | -5.0 | 0.514 |
| CommandA-WMT | ✓ | 111 | ✓ | 5.6 | 0.665 | 78.9 | 81.5 | -5.0 | 0.508 |
| UvA-MT | ✓ | 12 | ✓ | 6.3 | 0.671 | 76.8 | 81.0 | -5.4 | 0.499 |
| ▲ Claude-4 | ✓ | ? | ✓ | 7.0 | 0.649 | 80.4 | 82.8 | -5.6 | 0.487 |
| ▲ DeepSeek-V3 | ✓ | 671 | ✓ | 7.1 | 0.618 | 84.9 | 85.1 | -5.2 | 0.473 |
| ▲ Llama-4-Maverick | ✓ | 400 | ✓ | 8.0 | 0.65 | 74.9 | 79.4 | -5.5 | 0.489 |
| ▲ ONLINE-B | ✓ | ? | | 8.2 | 0.656 | 73.0 | 74.7 | -5.2 | 0.492 |
| ▲ Gemma-3-27B | ✓ | 27 | | 9.0 | 0.638 | 75.5 | 78.7 | -5.8 | 0.475 |
| Lanigo | ✓ | 9 | ✓ | 9.1 | 0.665 | 65.6 | 67.4 | -4.9 | 0.513 |
| IRB-MT | ✓ | 12 | ✓ | 9.3 | 0.633 | 73.7 | 77.5 | -5.3 | 0.467 |
| ▲ CommandA | ✓ | 111 | | 9.4 | 0.645 | 76.5 | 76.8 | -6.1 | 0.464 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | | 9.8 | 0.645 | 73.3 | 76.6 | -6.1 | 0.466 |
| SRPOL | ✗ | 12 | ✓ | 10.3 | 0.666 | 68.2 | 71.1 | -6.0 | 0.461 |
| RuZh | ? | 9 | ✓ | 10.4 | 0.648 | 71.2 | 74.2 | -5.9 | 0.454 |
| ▲ Gemma-3-12B | ✓ | 12 | | 10.6 | 0.636 | 73.4 | 76.6 | -6.1 | 0.446 |
| ▲ Qwen2.5-7B | ✓ | 7 | | 11.5 | 0.625 | 70.6 | 73.6 | -5.9 | 0.451 |
| ▲ AyaExpanse-32B | ✓ | 32 | | 11.6 | 0.631 | 70.9 | 74.6 | -6.3 | 0.444 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | | 11.9 | 0.634 | 69.9 | 71.7 | -6.2 | 0.446 |
| TranssionTranslate | ? | ? | | 12.7 | 0.638 | 66.9 | 70.1 | -6.4 | 0.438 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | | 12.8 | 0.627 | 68.9 | 71.5 | -6.4 | 0.43 |
| ▲ ONLINE-W | ? | ? | | 13.4 | 0.627 | 66.4 | 69.2 | -6.5 | 0.437 |
| ▲ AyaExpanse-8B | ✓ | 8 | | 15.0 | 0.615 | 65.1 | 68.6 | -6.7 | 0.403 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | | 16.4 | 0.604 | 63.6 | 66.6 | -6.9 | 0.394 |
| IR-MultiagentMT | ✗ | ? | | 17.2 | 0.575 | 64.0 | 66.0 | -6.6 | 0.399 |
| SalamandraTA | ✓ | 8 | | 17.9 | 0.618 | 59.5 | 59.2 | -7.1 | 0.376 |
| ▲ Llama-3.1-8B | ✗ | 8 | | 18.4 | 0.594 | 61.6 | 62.8 | -7.4 | 0.379 |
| ▲ CommandR7B | ✓ | 7 | | 18.5 | 0.595 | 63.1 | 65.0 | -7.9 | 0.376 |
| ▲ ONLINE-G | ✓ | ? | | 31.2 | 0.508 | 52.2 | 51.7 | -11.1 | 0.256 |
| ▲ Mistral-7B | ✗ | 7 | | 32.0 | 0.5 | 47.6 | 46.7 | -10.4 | 0.257 |
| ▲ NLLB | ✓ | 1 | | 38.0 | 0.441 | 44.4 | 45.6 | -12.8 | 0.238 |

| Czech-Ukrainian | | | | | | | | | |
|-----------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.601 | 79.1 | 85.3 | -5.0 | 0.681 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 1.0 | 0.582 | 81.4 | 89.5 | -5.1 | 0.671 |
| CommandA-WMT | ✓ | 111 | ✓ | 1.3 | 0.593 | 80.3 | 84.3 | -4.8 | 0.664 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 1.3 | 0.592 | 80.4 | 89.0 | -5.3 | 0.666 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 3.2 | 0.578 | 79.3 | 84.3 | -5.5 | 0.654 |
| ▲ Claude-4 | ? | ? | ✓ | 3.6 | 0.587 | 78.8 | 85.6 | -6.0 | 0.645 |
| ▲ Mistral-Medium | ? | ? | ✓ | 4.1 | 0.58 | 77.9 | 83.5 | -5.8 | 0.642 |
| GemTrans | ✓ | 27 | ✓ | 4.3 | 0.58 | 75.6 | 79.2 | -5.2 | 0.645 |
| ▲ CommandA | ✓ | 111 | ✓ | 4.5 | 0.582 | 78.9 | 81.7 | -6.0 | 0.637 |
| ▲ Gemma-3-27B | ✓ | 27 | ✓ | 4.9 | 0.581 | 77.3 | 81.7 | -6.0 | 0.63 |
| UvA-MT | ✓ | 12 | ✓ | 5.0 | 0.597 | 74.7 | 79.1 | -6.0 | 0.64 |
| Wenyil | ✓ | 14 | ✓ | 5.3 | 0.585 | 75.6 | 79.1 | -5.9 | 0.635 |
| Yolu | ✓ | 14 | ✓ | 5.9 | 0.606 | 72.1 | 73.8 | -5.9 | 0.634 |
| Algharb | ✓ | 14 | ✓ | 7.1 | 0.572 | 74.0 | 79.5 | -6.4 | 0.619 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 7.3 | 0.574 | 75.5 | 80.3 | -6.7 | 0.601 |
| ▲ AyaExpanse-32B | ✓ | 32 | | 7.4 | 0.57 | 73.4 | 76.1 | -6.1 | 0.618 |
| Laniqo | ✓ | 9 | ✓ | 7.5 | 0.596 | 68.1 | 68.6 | -5.9 | 0.645 |
| SRPOL | ✓ | 12 | ✓ | 7.6 | 0.6 | 71.4 | 73.3 | -6.6 | 0.618 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | ✓ | 7.7 | 0.57 | 74.0 | 76.7 | -6.4 | 0.608 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | | 8.7 | 0.567 | 72.7 | 75.7 | -6.7 | 0.602 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | | 8.7 | 0.566 | 72.8 | 74.7 | -6.7 | 0.606 |
| IRB-MT | ✓ | 12 | ✓ | 8.9 | 0.559 | 72.4 | 74.8 | -6.4 | 0.598 |
| ▲ Gemma-3-12B | ✓ | 12 | | 9.7 | 0.559 | 73.0 | 75.9 | -6.9 | 0.583 |
| ▲ Qwen3-235B | ✓ | 235 | | 10.4 | 0.557 | 71.5 | 73.6 | -6.9 | 0.582 |
| IR-MultiagentMT | ✗ | ? | | 11.2 | 0.544 | 70.1 | 71.8 | -6.7 | 0.579 |
| ▲ ONLINE-B | ✓ | ? | | 11.5 | 0.542 | 69.1 | 69.8 | -6.5 | 0.578 |
| SalamandraTA | ✓ | 8 | | 11.7 | 0.562 | 66.9 | 66.5 | -6.7 | 0.583 |
| CUNI-EdUKate-v1 | ✓ | 9 | | 12.5 | 0.555 | 67.2 | 67.4 | -7.1 | 0.573 |
| ▲ AyaExpanse-8B | ✓ | 8 | | 13.3 | 0.54 | 66.9 | 66.7 | -6.9 | 0.565 |
| ▲ ONLINE-W | ? | ? | | 13.4 | 0.534 | 66.6 | 68.0 | -6.9 | 0.564 |
| TranssionTranslate | ? | ? | | 14.6 | 0.521 | 66.5 | 67.2 | -7.0 | 0.541 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | | 14.8 | 0.533 | 66.6 | 67.3 | -7.6 | 0.545 |
| DLUT_GTCOM | ✓ | 27 | | 15.0 | 0.523 | 65.9 | 67.3 | -7.3 | 0.54 |
| CUNI-SFT | ✓ | 9 | | 15.2 | 0.528 | 63.7 | 64.8 | -7.2 | 0.552 |
| ▲ ONLINE-G | ✓ | ? | | 24.0 | 0.471 | 58.0 | 55.3 | -8.8 | 0.458 |
| ▲ Llama-3.1-8B | ✗ | 8 | | 25.3 | 0.493 | 58.2 | 53.8 | -10.0 | 0.432 |
| CUNI-Transformer | ✓ | <1 | | 25.7 | 0.47 | 58.0 | 56.3 | -10.1 | 0.449 |
| ▲ CommandR7B | ✓ | 7 | | 25.8 | 0.481 | 57.1 | 52.2 | -10.2 | 0.467 |
| ▲ NLLB | ✓ | 1 | | 28.5 | 0.46 | 51.7 | 50.8 | -10.3 | 0.439 |
| TranssionMT | ✓ | 1 | | 33.0 | 0.444 | 52.1 | 49.7 | -12.1 | 0.371 |
| ▲ Mistral-7B | ✗ | 7 | | 33.5 | 0.443 | 52.7 | 47.2 | -12.0 | 0.359 |
| ▲ Qwen2.5-7B | ? | 7 | | 42.0 | 0.382 | 45.6 | 40.2 | -13.8 | 0.287 |

| Czech-German | | | | | | | | | |
|-----------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.596 | 78.4 | 88.3 | -3.6 | 0.653 |
| CommandA-WMT | ✓ | 111 | ✓ | 2.1 | 0.582 | 77.9 | 87.5 | -3.2 | 0.634 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 2.3 | 0.58 | 79.4 | 91.0 | -3.7 | 0.634 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 2.5 | 0.577 | 79.1 | 90.8 | -3.6 | 0.633 |
| ▲ DeepSeek-V3 | ? | 671 | ✓ | 3.5 | 0.577 | 79.5 | 88.6 | -3.8 | 0.624 |
| ▲ Mistral-Medium | ✓ | ? | ✓ | 4.1 | 0.577 | 77.6 | 86.9 | -3.8 | 0.627 |
| ▲ CommandA | ✓ | 111 | ✓ | 4.7 | 0.579 | 77.8 | 85.5 | -4.0 | 0.624 |
| ▲ Claude-4 | ✓ | ? | ✓ | 4.7 | 0.577 | 77.7 | 87.1 | -3.9 | 0.618 |
| GemTrans | ✓ | 27 | ✓ | 6.2 | 0.569 | 75.0 | 82.1 | -3.7 | 0.619 |
| UvA-MT | ✓ | 12 | ✓ | 6.8 | 0.584 | 74.2 | 82.3 | -4.3 | 0.617 |
| ▲ Gemma-3-27B | ✓ | 27 | ✓ | 7.1 | 0.572 | 74.9 | 82.5 | -4.1 | 0.612 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 7.5 | 0.569 | 74.7 | 84.7 | -4.2 | 0.604 |
| ▲ AyaExpanse-32B | ✓ | 32 | | 8.0 | 0.568 | 74.3 | 80.5 | -4.1 | 0.606 |
| Yolu | ✓ | 14 | ✓ | 9.0 | 0.589 | 70.0 | 75.2 | -4.4 | 0.613 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | | 9.3 | 0.572 | 73.1 | 78.5 | -4.4 | 0.6 |
| ▲ Qwen3-235B | ✓ | 235 | | 9.3 | 0.565 | 73.1 | 80.9 | -4.2 | 0.594 |
| Laniqo | ✓ | 9 | ✓ | 10.1 | 0.587 | 67.5 | 70.3 | -4.2 | 0.619 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | ✓ | 10.1 | 0.568 | 71.7 | 77.4 | -4.4 | 0.599 |
| Wenyiil | ✓ | 14 | ✓ | 10.7 | 0.559 | 71.1 | 77.9 | -4.3 | 0.597 |
| SRPOL | ✓ | 12 | ✓ | 10.8 | 0.593 | 69.2 | 73.2 | -4.7 | 0.591 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | | 11.0 | 0.567 | 70.7 | 77.4 | -4.6 | 0.596 |
| ▲ Gemma-3-12B | ✓ | 12 | ✓ | 11.2 | 0.561 | 71.9 | 77.5 | -4.6 | 0.592 |
| ▲ ONLINE-B | ✓ | ? | | 11.7 | 0.555 | 69.3 | 74.4 | -4.1 | 0.597 |
| IRB-MT | ✓ | 12 | ✓ | 12.1 | 0.557 | 70.6 | 75.4 | -4.5 | 0.588 |
| Algharb | ✓ | 14 | ✓ | 12.9 | 0.551 | 70.8 | 77.1 | -4.7 | 0.58 |
| IR-MultiagentMT | ✗ | ? | | 13.0 | 0.559 | 68.0 | 75.3 | -4.7 | 0.592 |
| CUNI-MH-v2 | ✓ | 9 | ✓ | 13.8 | 0.562 | 68.2 | 72.5 | -4.7 | 0.577 |
| SalamandraTA | ✓ | 8 | | 15.3 | 0.554 | 65.8 | 69.5 | -4.6 | 0.574 |
| ▲ AyaExpanse-8B | ✓ | 8 | | 15.4 | 0.555 | 66.4 | 70.9 | -4.7 | 0.564 |
| TranssionTranslate | ? | ? | | 16.6 | 0.538 | 67.0 | 71.1 | -4.7 | 0.56 |
| ▲ ONLINE-W | ? | ? | | 16.7 | 0.542 | 67.0 | 71.1 | -4.9 | 0.56 |
| DLUT_GTCOM | ✓ | 27 | | 17.4 | 0.537 | 66.6 | 70.5 | -4.8 | 0.553 |
| ▲ CommandR7B | ✓ | 7 | | 17.9 | 0.545 | 65.8 | 68.9 | -5.1 | 0.556 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | | 22.4 | 0.531 | 57.1 | 61.1 | -5.6 | 0.579 |
| ▲ Llama-3.1-8B | ✓ | 8 | | 25.3 | 0.524 | 59.6 | 61.6 | -5.8 | 0.508 |
| ▲ ONLINE-G | ✓ | ? | | 32.1 | 0.492 | 58.1 | 58.4 | -6.9 | 0.47 |
| ▲ Qwen2.5-7B | ✓ | 7 | | 32.1 | 0.503 | 54.0 | 54.5 | -6.6 | 0.476 |
| ▲ NLLB | ✓ | 1 | | 33.4 | 0.5 | 52.5 | 54.2 | -6.9 | 0.479 |
| ▲ Mistral-7B | ✗ | 7 | | 36.4 | 0.492 | 53.3 | 51.1 | -7.1 | 0.434 |
| TranssionMT | ✓ | 1 | | 40.0 | 0.473 | 51.2 | 52.4 | -7.9 | 0.425 |

| Japanese-Simplified Chinese | | | | | | | | | |
|-----------------------------|----------------------|----------------|----------|---------------|------------------|-----------------------------|-------------------------------|------------------------------------|-----------------|
| System Name | LP Sup- ported | Params. (B) | Humeval? | AutoRank ↓ | CometKiw XL ↑ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | ✓ | 1.0 | 0.577 | 85.1 | 85.5 | -4.2 | 0.629 |
| In2x | ? | 72 | ✓ | 3.0 | 0.624 | 77.0 | 77.7 | -4.7 | 0.618 |
| ▲ Gemini-2.5-Pro | ✓ | ? | ✓ | 3.2 | 0.549 | 84.8 | 84.8 | -4.6 | 0.596 |
| Kaze-MT | ✓ | 72 | ✓ | 3.8 | 0.569 | 81.5 | 81.8 | -4.8 | 0.605 |
| Algharb | ✓ | 14 | ✓ | 4.2 | 0.547 | 83.5 | 84.1 | -4.8 | 0.583 |
| ▲ GPT-4.1 | ✓ | ? | ✓ | 4.4 | 0.549 | 83.8 | 84.7 | -5.1 | 0.582 |
| Wenyiil | ✓ | 14 | ✓ | 4.5 | 0.555 | 81.4 | 81.9 | -4.8 | 0.591 |
| CommandA-WMT | ✓ | 111 | ✓ | 5.1 | 0.558 | 80.2 | 79.7 | -4.7 | 0.575 |
| NTTSU | ✓ | 14 | ✓ | 5.8 | 0.563 | 77.5 | 74.8 | -4.6 | 0.577 |
| bb88 | ? | ? | | 6.1 | 0.551 | 80.1 | 78.9 | -5.2 | 0.573 |
| ▲ Claude-4 | ✓ | ? | ✓ | 6.2 | 0.545 | 82.9 | 83.7 | -5.6 | 0.556 |
| ▲ DeepSeek-V3 | ✓ | 671 | ✓ | 6.3 | 0.534 | 82.9 | 80.9 | -5.1 | 0.552 |
| ▲ Mistral-Medium | ? | ? | ✓ | 6.4 | 0.546 | 81.1 | 81.1 | -5.4 | 0.558 |
| GemTrans | ✓ | 27 | ✓ | 6.5 | 0.556 | 76.0 | 74.9 | -4.8 | 0.579 |
| Yolu | ✓ | 14 | ✓ | 6.9 | 0.578 | 74.6 | 73.6 | -5.0 | 0.565 |
| ▲ Qwen3-235B | ✓ | 235 | ✓ | 7.5 | 0.549 | 78.4 | 77.0 | -5.4 | 0.555 |
| ▲ CommandA | ✓ | 111 | | 7.6 | 0.54 | 79.4 | 77.6 | -5.5 | 0.556 |
| UvA-MT | ✓ | 12 | | 8.3 | 0.564 | 73.9 | 75.2 | -5.6 | 0.561 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | | 9.7 | 0.537 | 76.5 | 75.0 | -5.9 | 0.536 |
| ▲ AyaExpanse-32B | ✓ | 32 | | 10.7 | 0.537 | 73.2 | 72.0 | -5.8 | 0.521 |
| Lanigo | ✓ | 9 | ✓ | 11.1 | 0.579 | 63.1 | 62.1 | -5.4 | 0.557 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | ✓ | 11.2 | 0.535 | 71.9 | 69.8 | -5.8 | 0.523 |
| IRB-MT | ✓ | 12 | ✓ | 12.1 | 0.521 | 72.2 | 70.4 | -6.0 | 0.509 |
| ▲ Gemma-3-27B | ✓ | 27 | | 12.8 | 0.526 | 70.4 | 70.2 | -6.2 | 0.503 |
| ▲ Llama-4-Maverick | ✓ | 400 | | 13.1 | 0.524 | 71.5 | 66.1 | -6.3 | 0.518 |
| ▲ Qwen2.5-7B | ✓ | 7 | | 13.6 | 0.524 | 68.9 | 67.4 | -6.3 | 0.502 |
| IR-MultiagentMT | ✗ | ? | | 13.7 | 0.523 | 67.8 | 68.5 | -6.2 | 0.492 |
| SRPOL | ✗ | 12 | | 13.8 | 0.56 | 63.8 | 62.5 | -6.4 | 0.522 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | | 14.7 | 0.521 | 66.4 | 66.2 | -6.3 | 0.486 |
| ▲ AyaExpanse-8B | ✓ | 8 | | 15.5 | 0.518 | 65.6 | 64.4 | -6.4 | 0.472 |
| ▲ ONLINE-B | ✓ | ? | | 16.2 | 0.499 | 63.7 | 63.2 | -6.2 | 0.472 |
| ▲ Gemma-3-12B | ✓ | 12 | | 17.1 | 0.509 | 65.0 | 64.1 | -7.1 | 0.465 |
| ▲ CommandR7B | ✓ | 7 | | 18.4 | 0.496 | 59.8 | 58.5 | -6.9 | 0.486 |
| TranssionTranslate | ? | ? | | 18.8 | 0.488 | 59.9 | 60.6 | -6.7 | 0.45 |
| ▲ Llama-3.1-8B | ✗ | 8 | | 20.2 | 0.507 | 58.8 | 57.3 | -7.2 | 0.423 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | | 20.8 | 0.479 | 59.4 | 57.2 | -7.6 | 0.461 |
| ▲ ONLINE-W | ? | ? | | 25.2 | 0.456 | 52.3 | 52.9 | -7.9 | 0.387 |
| ▲ Mistral-7B | ✗ | 7 | | 32.8 | 0.445 | 42.9 | 43.4 | -9.8 | 0.317 |
| SalamandraTA | ✓ | 8 | | 33.1 | 0.426 | 36.5 | 38.0 | -8.6 | 0.328 |
| ▲ ONLINE-G | ✓ | ? | | 40.8 | 0.352 | 39.5 | 39.8 | -12.1 | 0.28 |
| ▲ NLLB | ✓ | 1 | | 41.0 | 0.371 | 35.5 | 35.8 | -12.1 | 0.303 |

| English-Bengali | | | | | | | |
|-----------------------|-----------------|----------------|---------------|--------------------------|----------------------------|------------------------------------|-----------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✗ | 7 | 1.0 | 67.9 | 83.2 | -4.8 | 0.449 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 2.2 | 66.5 | 86.6 | -5.2 | 0.382 |
| ▲ GPT-4.1 | ✓ | ? | 3.2 | 66.9 | 81.6 | -5.9 | 0.373 |
| GemTrans | ✓ | 27 | 3.5 | 64.3 | 75.1 | -5.0 | 0.374 |
| ▲ Mistral-Medium | ? | ? | 3.9 | 65.5 | 78.0 | -6.0 | 0.366 |
| ▲ Claude-4 | ✓ | ? | 4.0 | 65.6 | 80.6 | -6.1 | 0.348 |
| UvA-MT | ? | 12 | 4.2 | 64.1 | 75.0 | -6.1 | 0.381 |
| ▲ DeepSeek-V3 | ? | 671 | 4.3 | 63.7 | 77.7 | -6.1 | 0.364 |
| IRB-MT | ✓ | 12 | 5.1 | 62.9 | 72.7 | -6.0 | 0.34 |
| CommandA-WMT | ✗ | 111 | 5.3 | 63.5 | 69.4 | -6.2 | 0.345 |
| ▲ Llama-4-Maverick | ✓ | 400 | 5.5 | 63.9 | 73.9 | -6.3 | 0.315 |
| ▲ Qwen3-235B | ✓ | 235 | 6.0 | 62.7 | 71.2 | -6.4 | 0.313 |
| ▲ ONLINE-B | ✓ | ? | 7.1 | 59.5 | 65.9 | -6.4 | 0.304 |
| TranssionTranslate | ? | ? | 7.3 | 59.4 | 63.9 | -6.4 | 0.301 |
| ▲ Gemma-3-12B | ✓ | 12 | 7.6 | 59.8 | 65.9 | -7.4 | 0.316 |
| ▲ Gemma-3-27B | ✓ | 27 | 7.8 | 55.7 | 65.6 | -7.1 | 0.335 |
| ▲ CommandA | ✗ | 111 | 9.2 | 60.8 | 59.2 | -8.0 | 0.254 |
| ▲ NLLB | ✓ | 1 | 11.4 | 53.9 | 55.5 | -8.6 | 0.235 |
| IR-MultiagentMT | ✗ | ? | 11.5 | 53.7 | 55.5 | -8.6 | 0.238 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | 13.5 | 55.0 | 47.0 | -9.9 | 0.189 |
| ▲ Llama-3.1-8B | ✗ | 8 | 14.2 | 50.7 | 46.1 | -9.5 | 0.176 |
| ▲ ONLINE-G | ✓ | ? | 15.8 | 48.3 | 48.1 | -10.9 | 0.151 |
| ▲ AyaExpanse-32B | ✗ | 32 | 17.9 | 46.2 | 36.1 | -11.7 | 0.143 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | 20.3 | 27.9 | 9.6 | -9.0 | 0.228 |
| ▲ Qwen2.5-7B | ? | 7 | 21.1 | 36.6 | 30.8 | -12.8 | 0.122 |
| ▲ CommandR7B | ✗ | 7 | 22.7 | 30.6 | 22.4 | -13.8 | 0.181 |
| ▲ AyaExpanse-8B | ✗ | 8 | 25.1 | 27.7 | 21.5 | -16.1 | 0.16 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | 27.5 | 15.5 | 6.2 | -15.2 | 0.189 |
| ▲ Mistral-7B | ✗ | 7 | 28.6 | 19.3 | 14.4 | -18.6 | 0.175 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | 30.0 | 16.5 | 12.7 | -19.5 | 0.171 |

| English-German | | | | | | | |
|-----------------------|-----------------|----------------|---------------|--------------------------|----------------------------|------------------------------------|-----------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | 1.0 | 84.3 | 90.6 | -3.1 | 0.703 |
| CommandA-WMT | ✓ | 111 | 2.2 | 82.8 | 89.0 | -3.0 | 0.686 |
| ▲ GPT-4.1 | ✓ | ? | 3.2 | 84.6 | 91.4 | -3.6 | 0.671 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 3.4 | 84.0 | 91.7 | -3.5 | 0.665 |
| ▲ DeepSeek-V3 | ? | 671 | 3.6 | 84.0 | 90.0 | -3.6 | 0.671 |
| ▲ Mistral-Medium | ✓ | ? | 3.8 | 83.6 | 88.2 | -3.6 | 0.676 |
| GemTrans | ✓ | 27 | 4.7 | 78.7 | 84.8 | -3.1 | 0.672 |
| ▲ CommandA | ✓ | 111 | 4.8 | 82.3 | 87.1 | -3.8 | 0.672 |
| ▲ ONLINE-B | ✓ | ? | 5.4 | 77.7 | 83.4 | -3.3 | 0.678 |
| ▲ Claude-4 | ✓ | ? | 5.5 | 81.4 | 86.9 | -3.9 | 0.669 |
| UvA-MT | ✓ | 12 | 5.9 | 77.5 | 83.3 | -3.6 | 0.679 |
| ▲ Qwen3-235B | ✓ | 235 | 6.2 | 80.0 | 85.4 | -3.7 | 0.659 |
| ▲ AyaExpanse-32B | ✓ | 32 | 6.5 | 78.2 | 83.8 | -3.8 | 0.669 |
| ▲ Llama-4-Maverick | ✓ | 400 | 7.0 | 79.0 | 83.5 | -3.9 | 0.663 |
| ▲ ONLINE-W | ? | ? | 8.0 | 76.4 | 80.9 | -3.9 | 0.664 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | 8.2 | 76.0 | 80.0 | -3.9 | 0.667 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | 8.4 | 76.0 | 80.2 | -4.0 | 0.665 |
| TranssionTranslate | ? | ? | 8.7 | 73.5 | 78.1 | -3.4 | 0.653 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | 8.7 | 75.4 | 79.0 | -4.1 | 0.669 |
| SalamandraTA | ✓ | 8 | 9.8 | 72.4 | 75.4 | -3.8 | 0.663 |
| IRB-MT | ✓ | 12 | 9.8 | 74.8 | 79.0 | -3.7 | 0.63 |
| ▲ Gemma-3-12B | ✓ | 12 | 12.2 | 73.0 | 76.2 | -4.4 | 0.633 |
| ▲ AyaExpanse-8B | ✓ | 8 | 12.2 | 70.1 | 75.3 | -4.3 | 0.644 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | 12.4 | 70.5 | 73.6 | -4.5 | 0.654 |
| IR-MultiagentMT | ✗ | ? | 12.9 | 71.9 | 77.2 | -4.7 | 0.63 |
| ▲ CommandR7B | ✓ | 7 | 15.7 | 67.8 | 68.8 | -4.8 | 0.628 |
| ▲ Gemma-3-27B | ✓ | 27 | 17.6 | 67.4 | 71.7 | -5.1 | 0.589 |
| ▲ ONLINE-G | ✓ | ? | 17.9 | 66.5 | 67.7 | -5.2 | 0.609 |
| ▲ Llama-3.1-8B | ✓ | 8 | 20.9 | 64.3 | 62.6 | -5.5 | 0.588 |
| ▲ Qwen2.5-7B | ✓ | 7 | 23.1 | 60.0 | 59.1 | -5.5 | 0.575 |
| ▲ NLLB | ✓ | 1 | 26.1 | 58.1 | 59.3 | -6.7 | 0.573 |
| ▲ Mistral-7B | ✗ | 7 | 32.0 | 54.9 | 50.2 | -7.0 | 0.51 |

| English-Greek | | | | | | | |
|-----------------------|-----------------|----------------|---------------|--------------------------|----------------------------|------------------------------------|-----------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | 1.0 | 80.3 | 85.8 | -5.3 | 0.601 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 1.9 | 84.3 | 88.7 | -6.2 | 0.529 |
| CommandA-WMT | ✓ | 111 | 2.1 | 79.9 | 84.1 | -5.7 | 0.56 |
| ▲ GPT-4.1 | ✓ | ? | 2.4 | 82.6 | 87.1 | -6.4 | 0.528 |
| GemTrans | ✓ | 27 | 3.2 | 74.3 | 78.9 | -5.4 | 0.543 |
| UvA-MT | ? | 12 | 4.0 | 73.3 | 77.6 | -6.4 | 0.545 |
| ▲ CommandA | ✓ | 111 | 4.3 | 77.4 | 80.7 | -7.1 | 0.509 |
| ▲ Claude-4 | ? | ? | 4.3 | 79.0 | 82.2 | -7.3 | 0.496 |
| SalamandraTA | ✓ | 8 | 4.7 | 71.0 | 75.4 | -6.3 | 0.524 |
| ▲ Mistral-Medium | ? | ? | 5.1 | 73.5 | 78.2 | -7.0 | 0.498 |
| ▲ ONLINE-B | ✓ | ? | 5.5 | 71.2 | 75.0 | -6.7 | 0.495 |
| ▲ ONLINE-W | ? | ? | 5.5 | 74.0 | 77.3 | -7.4 | 0.487 |
| ▲ AyaExpanse-32B | ✓ | 32 | 5.6 | 72.1 | 75.8 | -7.2 | 0.494 |
| IRB-MT | ✓ | 12 | 5.9 | 70.3 | 73.9 | -6.9 | 0.486 |
| ▲ DeepSeek-V3 | ? | 671 | 6.4 | 69.9 | 74.7 | -7.6 | 0.48 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | 6.6 | 69.3 | 72.1 | -7.4 | 0.482 |
| ▲ Llama-4-Maverick | ✓ | 400 | 6.6 | 71.4 | 74.1 | -7.9 | 0.471 |
| TranssionTranslate | ? | ? | 6.7 | 67.7 | 71.5 | -6.9 | 0.468 |
| ▲ Qwen3-235B | ✓ | 235 | 7.6 | 67.0 | 69.8 | -7.6 | 0.455 |
| ▲ AyaExpanse-8B | ✓ | 8 | 8.0 | 65.1 | 67.7 | -7.8 | 0.46 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | 8.7 | 62.7 | 66.1 | -8.1 | 0.454 |
| IR-MultiagentMT | ✗ | ? | 9.1 | 65.6 | 67.2 | -8.6 | 0.419 |
| ▲ Gemma-3-12B | ✓ | 12 | 9.9 | 60.6 | 62.9 | -8.9 | 0.436 |
| ▲ Gemma-3-27B | ✓ | 27 | 12.0 | 54.9 | 56.9 | -9.7 | 0.411 |
| ▲ ONLINE-G | ✓ | ? | 13.2 | 58.9 | 60.1 | -10.9 | 0.333 |
| ▲ NLLB | ✓ | 1 | 13.4 | 55.1 | 57.5 | -11.1 | 0.373 |
| ▲ CommandR7B | ✓ | 7 | 17.6 | 27.9 | 17.5 | -9.9 | 0.487 |
| ▲ Llama-3.1-8B | ✗ | 8 | 19.0 | 44.8 | 41.7 | -13.2 | 0.254 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | 22.4 | 36.5 | 33.6 | -14.8 | 0.202 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | 26.5 | 26.8 | 22.9 | -16.7 | 0.148 |
| ▲ Qwen2.5-7B | ? | 7 | 29.8 | 22.1 | 20.0 | -20.0 | 0.109 |
| ▲ Mistral-7B | ✗ | 7 | 32.0 | 19.2 | 14.3 | -22.7 | 0.135 |

| English-Persian | | | | | | | |
|-----------------------|--------------|-------------|------------|------------------|--------------------|------------------------|-------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA-ESA-CMDA ↑ | GEMBA-ESA-GPT4.1 ↑ | MetricX-24-Hybrid-XL ↑ | XCOMET-XL ↑ |
| Shy-hunyuan-MT | ✗ | 7 | 1.0 | 80.4 | 84.1 | -4.6 | 0.553 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 1.7 | 82.4 | 88.4 | -5.2 | 0.476 |
| ▲ GPT-4.1 | ✓ | ? | 2.3 | 81.1 | 85.4 | -5.4 | 0.47 |
| CommandA-WMT | ✓ | 111 | 2.6 | 77.4 | 80.4 | -5.0 | 0.497 |
| GemTrans | ✓ | 27 | 2.9 | 74.1 | 77.0 | -4.5 | 0.502 |
| ▲ DeepSeek-V3 | ? | 671 | 3.3 | 78.1 | 80.6 | -5.5 | 0.456 |
| UvA-MT | ? | 12 | 3.7 | 73.2 | 76.2 | -5.3 | 0.489 |
| ▲ Gemma-3-27B | ✓ | 27 | 3.8 | 75.8 | 79.0 | -5.6 | 0.453 |
| ▲ Mistral-Medium | ? | ? | 3.9 | 75.5 | 78.7 | -5.6 | 0.453 |
| ▲ Claude-4 | ? | ? | 4.4 | 77.5 | 79.6 | -6.3 | 0.427 |
| ▲ CommandA | ✓ | 111 | 4.6 | 74.0 | 77.1 | -6.0 | 0.439 |
| ▲ ONLINE-B | ✓ | ? | 4.8 | 70.7 | 72.3 | -5.4 | 0.458 |
| IRB-MT | ✓ | 12 | 5.1 | 71.7 | 73.1 | -5.6 | 0.432 |
| ▲ Llama-4-Maverick | ✓ | 400 | 5.1 | 72.3 | 75.8 | -6.0 | 0.425 |
| TranssionTranslate | ? | ? | 5.6 | 68.5 | 69.3 | -5.5 | 0.438 |
| ▲ Gemma-3-12B | ✓ | 12 | 5.7 | 71.0 | 72.5 | -6.1 | 0.417 |
| ▲ AyaExpanse-32B | ✓ | 32 | 5.7 | 70.4 | 72.3 | -6.1 | 0.425 |
| ▲ Qwen3-235B | ✓ | 235 | 7.8 | 64.1 | 66.9 | -6.6 | 0.378 |
| IR-MultiagentMT | ✗ | ? | 8.7 | 63.8 | 63.8 | -7.1 | 0.359 |
| ▲ AyaExpanse-8B | ✓ | 8 | 8.8 | 62.1 | 62.5 | -7.0 | 0.369 |
| ▲ CommandR7B | ✓ | 7 | 12.7 | 55.1 | 49.5 | -8.9 | 0.312 |
| ▲ ONLINE-G | ✓ | ? | 13.4 | 54.6 | 53.2 | -9.3 | 0.255 |
| ▲ NLLB | ✓ | 1 | 13.8 | 52.5 | 52.4 | -9.6 | 0.27 |
| ▲ Llama-3.1-8B | ✗ | 8 | 13.8 | 51.5 | 49.2 | -8.9 | 0.261 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | 16.6 | 45.6 | 43.8 | -10.3 | 0.203 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | 20.2 | 37.7 | 32.8 | -12.0 | 0.16 |
| ▲ Qwen2.5-7B | ? | 7 | 21.6 | 32.4 | 32.0 | -12.7 | 0.134 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | 28.3 | 21.2 | 16.1 | -18.8 | 0.165 |
| ▲ Mistral-7B | ✗ | 7 | 28.6 | 21.9 | 17.6 | -19.0 | 0.131 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | 30.0 | 14.5 | 9.8 | -19.7 | 0.185 |

| English-Hindi | | | | | | | |
|-----------------------|-----------------|----------------|---------------|--------------------------|----------------------------|------------------------------------|-----------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | 1.0 | 77.0 | 82.3 | -5.1 | 0.44 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 1.9 | 78.3 | 86.3 | -5.7 | 0.376 |
| GemTrans | ✓ | 27 | 2.6 | 72.4 | 78.4 | -5.2 | 0.397 |
| ▲ GPT-4.1 | ✓ | ? | 2.7 | 75.7 | 84.5 | -5.9 | 0.372 |
| ▲ DeepSeek-V3 | ? | 671 | 3.0 | 76.2 | 82.4 | -5.9 | 0.36 |
| CommandA-WMT | ✓ | 111 | 3.2 | 73.6 | 79.0 | -5.6 | 0.375 |
| UvA-MT | ? | 12 | 4.4 | 70.8 | 77.6 | -6.0 | 0.355 |
| ▲ Claude-4 | ✓ | ? | 4.8 | 73.7 | 78.3 | -6.6 | 0.334 |
| ▲ Gemma-3-27B | ✓ | 27 | 5.2 | 71.9 | 76.6 | -6.3 | 0.319 |
| IRB-MT | ✓ | 12 | 5.3 | 69.8 | 74.3 | -6.1 | 0.33 |
| ▲ ONLINE-B | ✓ | ? | 5.6 | 68.0 | 74.5 | -6.2 | 0.331 |
| ▲ CommandA | ✓ | 111 | 5.8 | 71.0 | 74.9 | -6.6 | 0.314 |
| TranssionTranslate | ? | ? | 6.5 | 64.7 | 70.3 | -6.1 | 0.326 |
| ▲ Llama-4-Maverick | ✓ | 400 | 6.7 | 68.5 | 73.4 | -6.7 | 0.296 |
| ▲ Qwen3-235B | ✓ | 235 | 6.8 | 67.8 | 72.1 | -6.6 | 0.298 |
| ▲ Mistral-Medium | ? | ? | 6.9 | 67.2 | 71.7 | -6.9 | 0.322 |
| ▲ Gemma-3-12B | ✓ | 12 | 7.1 | 66.8 | 70.1 | -6.8 | 0.309 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | 7.5 | 67.1 | 70.8 | -7.0 | 0.287 |
| ▲ AyaExpanse-32B | ✓ | 32 | 8.1 | 65.6 | 70.4 | -7.1 | 0.27 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | 9.3 | 63.3 | 66.3 | -7.4 | 0.264 |
| IR-MultiagentMT | ✗ | ? | 10.1 | 61.9 | 62.9 | -7.6 | 0.251 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | 10.7 | 59.7 | 61.2 | -7.7 | 0.259 |
| ▲ AyaExpanse-8B | ✓ | 8 | 10.8 | 59.3 | 60.6 | -7.7 | 0.254 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | 11.6 | 53.6 | 54.4 | -7.8 | 0.3 |
| ▲ Llama-3.1-8B | ✓ | 8 | 13.8 | 54.8 | 54.2 | -8.6 | 0.195 |
| ▲ NLLB | ✓ | 1 | 14.4 | 55.2 | 55.2 | -9.4 | 0.199 |
| ▲ ONLINE-G | ✓ | ? | 15.4 | 54.5 | 51.8 | -9.6 | 0.176 |
| ▲ CommandR7B | ✓ | 7 | 15.9 | 49.6 | 49.6 | -9.3 | 0.18 |
| ▲ Qwen2.5-7B | ? | 7 | 24.8 | 30.0 | 32.5 | -12.8 | 0.107 |
| ▲ Mistral-7B | ✗ | 7 | 30.0 | 25.0 | 23.2 | -16.6 | 0.126 |

| English-Indonesian | | | | | | | |
|-----------------------|-----------------|----------------|---------------|--------------------------|----------------------------|------------------------------------|-----------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuanyuan-MT | ✓ | 7 | 1.0 | 83.2 | 87.1 | -4.4 | 0.677 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 2.8 | 83.0 | 89.3 | -5.6 | 0.576 |
| ▲ GPT-4.1 | ✓ | ? | 3.6 | 81.6 | 87.9 | -5.9 | 0.564 |
| GemTrans | ✓ | 27 | 3.7 | 76.4 | 80.8 | -4.7 | 0.622 |
| CommandA-WMT | ✓ | 111 | 4.0 | 78.2 | 83.7 | -5.5 | 0.592 |
| ▲ DeepSeek-V3 | ? | 671 | 4.1 | 81.2 | 85.1 | -5.9 | 0.558 |
| ▲ Qwen3-235B | ✓ | 235 | 4.3 | 79.8 | 84.2 | -6.0 | 0.566 |
| UvA-MT | ? | 12 | 4.4 | 78.0 | 83.2 | -5.9 | 0.584 |
| ▲ Mistral-Medium | ? | ? | 5.1 | 78.2 | 83.6 | -6.3 | 0.549 |
| ▲ Gemma-3-27B | ✓ | 27 | 5.4 | 78.3 | 83.1 | -6.4 | 0.531 |
| IRB-MT | ✓ | 12 | 5.5 | 75.8 | 80.6 | -5.8 | 0.548 |
| ▲ Claude-4 | ✓ | ? | 5.9 | 78.8 | 82.8 | -6.9 | 0.514 |
| ▲ Gemma-3-12B | ✓ | 12 | 6.6 | 75.3 | 81.1 | -6.8 | 0.515 |
| ▲ ONLINE-B | ✓ | ? | 7.0 | 72.7 | 76.7 | -6.3 | 0.528 |
| ▲ Llama-4-Maverick | ✓ | 400 | 7.3 | 74.0 | 78.5 | -6.9 | 0.507 |
| ▲ CommandA | ✓ | 111 | 7.5 | 74.9 | 77.7 | -7.1 | 0.498 |
| ▲ AyaExpanse-32B | ✓ | 32 | 7.8 | 72.7 | 76.9 | -6.9 | 0.5 |
| ▲ ONLINE-W | ? | ? | 8.2 | 69.8 | 73.9 | -6.7 | 0.522 |
| TranssionTranslate | ? | ? | 8.5 | 68.7 | 72.7 | -6.3 | 0.498 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | 9.1 | 70.2 | 74.6 | -7.5 | 0.479 |
| ▲ AyaExpanse-8B | ✓ | 8 | 9.3 | 68.6 | 72.1 | -7.2 | 0.487 |
| IR-MultiagentMT | ✗ | ? | 9.8 | 68.4 | 72.6 | -7.5 | 0.464 |
| ▲ ONLINE-G | ✓ | ? | 12.9 | 63.4 | 65.9 | -8.7 | 0.409 |
| ▲ Llama-3.1-8B | ✗ | 8 | 13.6 | 62.2 | 62.4 | -9.0 | 0.417 |
| ▲ Qwen2.5-7B | ? | 7 | 13.8 | 60.2 | 61.6 | -8.6 | 0.412 |
| ▲ CommandR7B | ✓ | 7 | 16.6 | 57.6 | 53.0 | -10.2 | 0.392 |
| ▲ NLLB | ✓ | 1 | 17.3 | 57.3 | 57.7 | -10.9 | 0.333 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | 18.7 | 52.0 | 50.2 | -10.6 | 0.339 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | 25.5 | 40.6 | 39.4 | -13.7 | 0.214 |
| ▲ Mistral-7B | ✗ | 7 | 25.6 | 43.1 | 40.1 | -14.2 | 0.197 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | 31.0 | 26.4 | 20.2 | -16.0 | 0.275 |

| English-Kannada | | | | | | | |
|-----------------------|--------------|-------------|------------|------------------|--------------------|------------------------|-------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA-ESA-CMDA ↑ | GEMBA-ESA-GPT4.1 ↑ | MetricX-24-Hybrid-XL ↑ | XCOMET-XL ↑ |
| Shy-hunyuan-MT | ✗ | 7 | 1.0 | 64.0 | 78.8 | -6.0 | 0.446 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 2.2 | 62.5 | 81.6 | -6.3 | 0.399 |
| ▲ Claude-4 | ? | ? | 5.0 | 61.3 | 76.1 | -7.6 | 0.333 |
| GemTrans | ✓ | 27 | 5.2 | 58.7 | 67.3 | -6.7 | 0.358 |
| ▲ GPT-4.1 | ✓ | ? | 5.9 | 60.2 | 71.3 | -7.9 | 0.327 |
| ▲ Mistral-Medium | ? | ? | 6.5 | 59.8 | 69.2 | -8.0 | 0.312 |
| ▲ Qwen3-235B | ✓ | 235 | 6.7 | 60.1 | 67.4 | -7.9 | 0.305 |
| ▲ DeepSeek-V3 | ? | 671 | 6.7 | 57.3 | 69.9 | -8.3 | 0.325 |
| CommandA-WMT | ✗ | 111 | 7.5 | 59.7 | 64.5 | -8.4 | 0.295 |
| ▲ ONLINE-B | ✓ | ? | 7.7 | 57.0 | 66.0 | -7.9 | 0.289 |
| ▲ Gemma-3-27B | ✓ | 27 | 7.8 | 57.2 | 66.1 | -8.3 | 0.294 |
| TranssionTranslate | ? | ? | 8.1 | 56.1 | 63.3 | -7.8 | 0.286 |
| ▲ Llama-4-Maverick | ✓ | 400 | 8.1 | 58.1 | 66.5 | -8.4 | 0.27 |
| UvA-MT | ? | 12 | 8.8 | 53.3 | 60.5 | -8.6 | 0.308 |
| IRB-MT | ✓ | 12 | 11.0 | 52.4 | 57.6 | -9.3 | 0.239 |
| ▲ NLLB | ✓ | 1 | 12.1 | 52.3 | 54.2 | -9.8 | 0.215 |
| ▲ ONLINE-G | ✓ | ? | 13.3 | 52.5 | 51.9 | -10.5 | 0.186 |
| ▲ Gemma-3-12B | ✓ | 12 | 13.4 | 46.1 | 49.4 | -10.4 | 0.244 |
| ▲ CommandA | ✗ | 111 | 14.2 | 54.3 | 48.1 | -11.6 | 0.175 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | 18.1 | 32.8 | 2.7 | -8.8 | 0.281 |
| ▲ Llama-3.1-8B | ✗ | 8 | 19.1 | 44.1 | 35.8 | -13.4 | 0.12 |
| IR-MultiagentMT | ✗ | ? | 19.5 | 40.1 | 36.4 | -13.9 | 0.149 |
| ▲ AyaExpanse-32B | ✗ | 32 | 23.9 | 34.3 | 25.5 | -18.6 | 0.157 |
| ▲ CommandR7B | ✗ | 7 | 25.6 | 17.1 | 10.6 | -16.4 | 0.222 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | 25.8 | 22.4 | 16.5 | -18.3 | 0.197 |
| ▲ AyaExpanse-8B | ✗ | 8 | 28.1 | 19.8 | 14.3 | -20.5 | 0.174 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | 28.2 | 13.1 | 2.9 | -16.9 | 0.179 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | 28.7 | 12.8 | 5.0 | -18.3 | 0.184 |
| ▲ Mistral-7B | ✗ | 7 | 29.9 | 7.4 | 4.6 | -19.2 | 0.199 |
| ▲ Qwen2.5-7B | ? | 7 | 30.0 | 15.1 | 10.2 | -21.4 | 0.163 |

| English-Lithuanian | | | | | | | |
|-----------------------|-----------------|----------------|---------------|--------------------------|----------------------------|------------------------------------|-----------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | 1.0 | 77.6 | 84.1 | -6.3 | 0.569 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 2.3 | 76.1 | 87.3 | -7.2 | 0.502 |
| ▲ GPT-4.1 | ✓ | ? | 2.9 | 75.3 | 84.8 | -7.6 | 0.5 |
| CommandA-WMT | ✗ | 111 | 4.5 | 72.6 | 72.4 | -7.8 | 0.506 |
| GemTrans | ✓ | 27 | 4.5 | 70.2 | 71.7 | -6.8 | 0.505 |
| ▲ ONLINE-B | ✓ | ? | 5.4 | 69.1 | 70.9 | -7.7 | 0.487 |
| ▲ Claude-4 | ? | ? | 5.8 | 71.8 | 75.6 | -9.3 | 0.455 |
| SalamandraTA | ✓ | 8 | 6.1 | 66.9 | 67.0 | -7.9 | 0.496 |
| ▲ ONLINE-W | ? | ? | 6.7 | 67.6 | 69.5 | -9.1 | 0.467 |
| TranssionTranslate | ? | ? | 6.7 | 66.3 | 67.6 | -7.9 | 0.454 |
| ▲ Gemma-3-27B | ✓ | 27 | 6.9 | 69.8 | 68.8 | -9.0 | 0.434 |
| ▲ Llama-4-Maverick | ✓ | 400 | 6.9 | 69.1 | 71.5 | -9.3 | 0.43 |
| UvA-MT | ? | 12 | 7.1 | 68.5 | 63.7 | -9.0 | 0.472 |
| ▲ Qwen3-235B | ✓ | 235 | 7.8 | 67.8 | 66.1 | -9.1 | 0.414 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | 8.3 | 64.7 | 66.1 | -9.7 | 0.434 |
| IRB-MT | ✓ | 12 | 8.9 | 66.0 | 61.2 | -9.5 | 0.402 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | 9.3 | 61.0 | 57.5 | -9.5 | 0.455 |
| ▲ DeepSeek-V3 | ? | 671 | 9.7 | 60.4 | 60.8 | -9.8 | 0.418 |
| ▲ Gemma-3-12B | ✓ | 12 | 10.2 | 66.0 | 58.3 | -10.7 | 0.368 |
| IR-MultiagentMT | ✗ | ? | 10.5 | 63.2 | 59.8 | -10.9 | 0.374 |
| ▲ Mistral-Medium | ? | ? | 10.8 | 64.5 | 56.7 | -10.9 | 0.362 |
| ▲ CommandA | ✗ | 111 | 11.0 | 65.1 | 55.7 | -11.4 | 0.361 |
| ▲ ONLINE-G | ✓ | ? | 15.5 | 55.7 | 52.4 | -14.2 | 0.259 |
| ▲ NLLB | ✓ | 1 | 15.9 | 53.0 | 49.3 | -14.2 | 0.283 |
| ▲ AyaExpanse-32B | ✗ | 32 | 22.3 | 46.0 | 32.7 | -17.9 | 0.143 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | 23.6 | 41.4 | 28.6 | -18.1 | 0.141 |
| ▲ Llama-3.1-8B | ✗ | 8 | 23.9 | 40.4 | 31.2 | -18.6 | 0.125 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | 26.3 | 34.3 | 18.3 | -19.8 | 0.157 |
| ▲ CommandR7B | ✗ | 7 | 27.0 | 21.5 | 4.6 | -17.7 | 0.274 |
| ▲ AyaExpanse-8B | ✓ | 8 | 29.3 | 25.8 | 17.5 | -22.7 | 0.141 |
| ▲ Qwen2.5-7B | ? | 7 | 29.7 | 24.1 | 19.0 | -22.7 | 0.116 |
| ▲ Mistral-7B | ✗ | 7 | 32.0 | 16.0 | 11.4 | -24.0 | 0.14 |

| English-Marathi | | | | | | | |
|-----------------------|--------------|-------------|------------|------------------|--------------------|------------------------|-------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA-ESA-CMDA ↑ | GEMBA-ESA-GPT4.1 ↑ | MetricX-24-Hybrid-XL ↑ | XCOMET-XL ↑ |
| Shy-hunyuan-MT | ✗ | 7 | 1.0 | 70.8 | 81.6 | -5.8 | 0.248 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 2.7 | 68.1 | 84.7 | -6.2 | 0.222 |
| GemTrans | ✓ | 27 | 4.0 | 67.3 | 65.2 | -5.8 | 0.224 |
| ▲ GPT-4.1 | ✓ | ? | 4.6 | 67.6 | 79.4 | -6.7 | 0.196 |
| UvA-MT | ? | 12 | 5.2 | 67.4 | 72.4 | -6.5 | 0.192 |
| ▲ Claude-4 | ? | ? | 5.5 | 67.2 | 76.2 | -7.2 | 0.193 |
| ▲ DeepSeek-V3 | ? | 671 | 5.5 | 67.5 | 74.7 | -7.0 | 0.19 |
| ▲ Gemma-3-27B | ✓ | 27 | 5.8 | 67.1 | 71.8 | -7.0 | 0.191 |
| ▲ Mistral-Medium | ? | ? | 6.9 | 66.8 | 70.6 | -7.3 | 0.171 |
| CommandA-WMT | ✗ | 111 | 7.2 | 66.4 | 64.7 | -7.2 | 0.178 |
| IRB-MT | ✓ | 12 | 7.3 | 64.4 | 68.1 | -7.0 | 0.175 |
| ▲ Llama-4-Maverick | ✓ | 400 | 7.8 | 64.0 | 69.4 | -7.5 | 0.169 |
| ▲ ONLINE-B | ✓ | ? | 8.0 | 62.7 | 66.0 | -7.2 | 0.172 |
| TranssionTranslate | ? | ? | 8.1 | 62.6 | 65.0 | -7.1 | 0.17 |
| ▲ Qwen3-235B | ✓ | 235 | 8.3 | 64.1 | 64.9 | -7.5 | 0.167 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | 8.9 | 63.0 | 7.7 | -7.7 | 0.277 |
| ▲ NLLB | ✓ | 1 | 12.0 | 58.3 | 55.6 | -8.7 | 0.148 |
| IR-MultiagentMT | ✗ | ? | 12.1 | 57.7 | 55.5 | -9.1 | 0.156 |
| ▲ Gemma-3-12B | ✓ | 12 | 12.4 | 52.5 | 51.8 | -9.4 | 0.189 |
| ▲ ONLINE-G | ✓ | ? | 13.2 | 57.3 | 54.2 | -9.4 | 0.138 |
| ▲ CommandA | ✗ | 111 | 13.3 | 60.9 | 49.6 | -9.7 | 0.131 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | 14.1 | 52.5 | 10.9 | -9.2 | 0.225 |
| ▲ Llama-3.1-8B | ✗ | 8 | 17.2 | 50.4 | 41.6 | -11.3 | 0.139 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | 17.8 | 49.8 | 30.5 | -12.5 | 0.175 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | 18.1 | 47.0 | 15.3 | -11.8 | 0.199 |
| ▲ AyaExpanse-32B | ✗ | 32 | 18.4 | 49.6 | 34.7 | -13.0 | 0.163 |
| ▲ AyaExpanse-8B | ✗ | 8 | 20.8 | 41.8 | 27.0 | -14.5 | 0.189 |
| ▲ CommandR7B | ✗ | 7 | 21.4 | 36.6 | 20.0 | -13.1 | 0.187 |
| ▲ Qwen2.5-7B | ? | 7 | 27.8 | 27.6 | 19.8 | -18.0 | 0.175 |
| ▲ Mistral-7B | ✗ | 7 | 30.0 | 25.0 | 12.5 | -17.9 | 0.146 |

| English-Romanian | | | | | | | |
|-----------------------|-----------------|----------------|---------------|--------------------------|----------------------------|------------------------------------|-----------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | 1.0 | 83.2 | 86.3 | -5.7 | 0.651 |
| CommandA-WMT | ✓ | 111 | 1.8 | 82.5 | 86.0 | -6.0 | 0.634 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 2.3 | 85.0 | 89.3 | -6.6 | 0.586 |
| ▲ GPT-4.1 | ✓ | ? | 2.7 | 83.5 | 88.2 | -6.8 | 0.597 |
| GemTrans | ✓ | 27 | 3.2 | 77.7 | 80.7 | -5.7 | 0.619 |
| ▲ DeepSeek-V3 | ? | 671 | 4.2 | 80.1 | 84.4 | -6.8 | 0.574 |
| UvA-MT | ? | 12 | 4.7 | 77.6 | 80.5 | -6.9 | 0.598 |
| ▲ Mistral-Medium | ? | ? | 5.1 | 77.7 | 83.2 | -7.2 | 0.568 |
| ▲ CommandA | ✓ | 111 | 5.2 | 79.4 | 82.9 | -7.4 | 0.563 |
| ▲ Gemma-3-27B | ✓ | 27 | 5.3 | 78.9 | 82.3 | -7.4 | 0.562 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | 6.0 | 74.8 | 79.9 | -7.1 | 0.566 |
| ▲ Claude-4 | ? | ? | 6.1 | 79.4 | 82.4 | -7.8 | 0.536 |
| ▲ Qwen3-235B | ✓ | 235 | 6.2 | 75.7 | 78.9 | -7.2 | 0.558 |
| ▲ AyaExpanse-32B | ✓ | 32 | 6.4 | 76.4 | 79.9 | -7.5 | 0.546 |
| IRB-MT | ✓ | 12 | 6.4 | 75.4 | 77.4 | -7.0 | 0.548 |
| SalamandraTA | ✓ | 8 | 6.5 | 70.9 | 75.2 | -6.7 | 0.589 |
| ▲ Llama-4-Maverick | ✓ | 400 | 6.5 | 76.4 | 81.1 | -7.6 | 0.541 |
| ▲ ONLINE-B | ✓ | ? | 6.7 | 73.5 | 76.9 | -7.1 | 0.556 |
| ▲ Gemma-3-12B | ✓ | 12 | 7.9 | 74.3 | 77.9 | -8.0 | 0.524 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | 8.2 | 71.6 | 76.4 | -7.8 | 0.533 |
| TranssionTranslate | ? | ? | 9.0 | 68.4 | 72.1 | -7.3 | 0.521 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | 9.3 | 70.3 | 73.2 | -8.0 | 0.512 |
| ▲ ONLINE-W | ? | ? | 9.4 | 72.2 | 75.4 | -8.7 | 0.51 |
| ▲ AyaExpanse-8B | ✓ | 8 | 9.8 | 68.3 | 71.7 | -8.0 | 0.516 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | 10.2 | 68.6 | 70.6 | -8.3 | 0.512 |
| IR-MultiagentMT | ✗ | ? | 16.5 | 57.6 | 59.3 | -10.0 | 0.425 |
| ▲ CommandR7B | ✓ | 7 | 16.9 | 59.9 | 54.3 | -10.2 | 0.434 |
| ▲ Llama-3.1-8B | ✗ | 8 | 18.0 | 56.9 | 56.7 | -10.3 | 0.38 |
| ▲ ONLINE-G | ✓ | ? | 18.7 | 59.5 | 60.6 | -11.6 | 0.359 |
| ▲ NLLB | ✓ | 1 | 19.3 | 55.0 | 57.2 | -11.6 | 0.39 |
| ▲ Mistral-7B | ✗ | 7 | 28.1 | 44.6 | 41.7 | -14.0 | 0.224 |
| ▲ Qwen2.5-7B | ? | 7 | 32.0 | 37.7 | 34.9 | -15.2 | 0.177 |

| English-Thai | | | | | | | |
|-----------------------|-----------------|----------------|---------------|--------------------------|----------------------------|------------------------------------|-----------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | 1.0 | 71.3 | 87.9 | -5.1 | 0.603 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 2.2 | 69.0 | 90.6 | -5.6 | 0.533 |
| GemTrans | ✓ | 27 | 2.7 | 67.9 | 80.4 | -5.4 | 0.558 |
| UvA-MT | ? | 12 | 3.2 | 69.5 | 79.7 | -6.0 | 0.54 |
| ▲ GPT-4.1 | ✓ | ? | 3.2 | 69.9 | 87.2 | -6.2 | 0.489 |
| ▲ Qwen3-235B | ✓ | 235 | 3.5 | 68.8 | 80.9 | -6.1 | 0.51 |
| ▲ DeepSeek-V3 | ? | 671 | 3.6 | 69.6 | 82.9 | -6.3 | 0.493 |
| ▲ Gemma-3-27B | ✓ | 27 | 4.1 | 68.3 | 82.2 | -6.5 | 0.482 |
| ▲ Mistral-Medium | ? | ? | 4.2 | 68.9 | 79.8 | -6.6 | 0.486 |
| ▲ Claude-4 | ? | ? | 4.5 | 68.5 | 80.7 | -6.8 | 0.466 |
| IRB-MT | ✓ | 12 | 4.8 | 66.2 | 77.1 | -6.4 | 0.475 |
| ▲ Llama-4-Maverick | ✓ | 400 | 5.0 | 67.0 | 76.2 | -6.6 | 0.463 |
| ▲ ONLINE-B | ✓ | ? | 5.0 | 65.1 | 72.2 | -6.1 | 0.484 |
| CommandA-WMT | ✗ | 111 | 5.8 | 66.8 | 70.0 | -6.8 | 0.449 |
| TranssionTranslate | ? | ? | 6.3 | 62.2 | 67.7 | -6.4 | 0.453 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | 6.7 | 64.1 | 70.3 | -7.2 | 0.424 |
| ▲ Gemma-3-12B | ✓ | 12 | 9.1 | 55.1 | 62.6 | -8.1 | 0.427 |
| IR-MultiagentMT | ✗ | ? | 9.5 | 56.6 | 56.5 | -7.9 | 0.404 |
| ▲ CommandA | ✗ | 111 | 11.0 | 60.6 | 54.6 | -9.4 | 0.311 |
| ▲ Qwen2.5-7B | ✓ | 7 | 11.7 | 56.4 | 51.1 | -9.2 | 0.319 |
| ▲ Llama-3.1-8B | ✓ | 8 | 12.6 | 55.6 | 51.2 | -9.7 | 0.277 |
| ▲ NLLB | ✓ | 1 | 14.8 | 51.2 | 48.4 | -11.3 | 0.247 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | 17.5 | 42.1 | 34.2 | -11.3 | 0.221 |
| ▲ ONLINE-G | ✓ | ? | 21.1 | 35.8 | 36.1 | -14.9 | 0.176 |
| ▲ AyaExpanse-32B | ✗ | 32 | 21.6 | 36.6 | 30.5 | -14.9 | 0.154 |
| ▲ Mistral-7B | ✗ | 7 | 24.6 | 31.0 | 24.9 | -17.2 | 0.132 |
| ▲ CommandR7B | ✗ | 7 | 25.8 | 27.0 | 21.6 | -18.2 | 0.156 |
| ▲ AyaExpanse-8B | ✗ | 8 | 26.4 | 24.6 | 20.1 | -18.5 | 0.167 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | 29.1 | 19.0 | 15.2 | -20.9 | 0.169 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | 30.0 | 12.8 | 7.4 | -20.2 | 0.185 |

| English-Serbian (Latin) | | | | | | | |
|-------------------------|--------------|-------------|------------|------------------|--------------------|------------------------|-------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA-ESA-CMDA ↑ | GEMBA-ESA-GPT4.1 ↑ | MetricX-24-Hybrid-XL ↑ | XCOMET-XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | 1.0 | 80.1 | 84.2 | -3.4 | 0.583 |
| Wenyii | ✓ | 14 | 2.5 | 77.8 | 84.6 | -3.8 | 0.513 |
| Algharb | ✓ | 14 | 2.8 | 77.9 | 86.5 | -4.0 | 0.493 |
| GemTrans | ✓ | 27 | 2.9 | 74.6 | 75.3 | -3.4 | 0.528 |
| ▲ GPT-4.1 | ✓ | ? | 2.9 | 78.6 | 85.3 | -4.1 | 0.501 |
| ▲ DeepSeek-V3 | ? | 671 | 2.9 | 78.5 | 80.3 | -3.9 | 0.514 |
| UvA-MT | ✓ | 12 | 2.9 | 75.0 | 75.0 | -3.7 | 0.562 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 2.9 | 77.7 | 86.9 | -4.1 | 0.488 |
| Yolu | ✓ | 14 | 3.0 | 73.0 | 73.1 | -3.4 | 0.553 |
| ▲ Claude-4 | ? | ? | 4.8 | 74.5 | 76.6 | -4.5 | 0.471 |
| SalamandraTA | ✓ | 8 | 5.0 | 68.8 | 68.3 | -3.8 | 0.491 |
| ▲ Llama-4-Maverick | ✓ | 400 | 6.1 | 71.3 | 70.4 | -4.7 | 0.448 |
| IRB-MT | ✓ | 12 | 6.3 | 69.0 | 66.7 | -4.3 | 0.441 |
| ▲ Qwen3-235B | ✓ | 235 | 6.4 | 68.8 | 65.4 | -4.3 | 0.439 |
| IR-MultiagentMT | ✗ | ? | 6.8 | 70.0 | 66.2 | -4.7 | 0.437 |
| CommandA-WMT | ✗ | 111 | 7.2 | 70.4 | 62.5 | -5.6 | 0.506 |
| ▲ ONLINE-B | ✓ | ? | 7.2 | 69.6 | 64.6 | -5.6 | 0.497 |
| ▲ Gemma-3-12B | ✓ | 12 | 7.6 | 67.8 | 63.6 | -5.0 | 0.427 |
| ▲ CommandA | ✗ | 111 | 7.6 | 67.7 | 59.9 | -4.6 | 0.41 |
| ▲ Gemma-3-27B | ✓ | 27 | 9.0 | 64.1 | 63.6 | -5.8 | 0.438 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | 9.0 | 60.6 | 54.4 | -4.9 | 0.43 |
| CUNI-SFT | ✓ | 9 | 9.4 | 61.5 | 53.2 | -4.9 | 0.392 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | 10.2 | 58.8 | 47.7 | -5.2 | 0.42 |
| TranssionTranslate | ? | ? | 10.7 | 60.8 | 59.6 | -5.9 | 0.349 |
| TranssionMT | ✓ | 1 | 11.0 | 57.3 | 52.8 | -5.4 | 0.358 |
| ▲ ONLINE-G | ✓ | ? | 12.5 | 57.8 | 52.9 | -6.9 | 0.375 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | 12.6 | 55.7 | 43.1 | -5.5 | 0.306 |
| ▲ Llama-3.1-8B | ✗ | 8 | 13.4 | 54.7 | 43.8 | -6.0 | 0.29 |
| ▲ AyaExpans-32B | ✗ | 32 | 13.8 | 52.9 | 40.4 | -5.7 | 0.259 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | 17.6 | 43.0 | 29.2 | -6.4 | 0.181 |
| ▲ Mistral-7B | ✗ | 7 | 17.6 | 49.4 | 37.0 | -7.8 | 0.213 |
| ▲ Qwen2.5-7B | ? | 7 | 20.5 | 39.3 | 29.0 | -8.1 | 0.144 |
| ▲ AyaExpans-8B | ✗ | 8 | 20.7 | 37.5 | 25.9 | -7.9 | 0.143 |
| ▲ CommandR7B | ✗ | 7 | 21.1 | 38.5 | 25.7 | -8.9 | 0.203 |
| ▲ NLLB | ✓ | 1 | 35.0 | 0.8 | 0.1 | -15.2 | 0.195 |

| English-Swedish | | | | | | | |
|-----------------------|-----------------|----------------|---------------|--------------------------|----------------------------|------------------------------------|-----------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA- ESA- CMDA ↑ | GEMBA- ESA- GPT4.1 ↑ | MetricX- 24- Hybrid- XL ↑ | XCOMET- XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | 1.0 | 84.2 | 91.0 | -4.7 | 0.685 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 2.5 | 83.1 | 92.3 | -5.4 | 0.638 |
| GemTrans | ✓ | 27 | 2.9 | 79.2 | 85.1 | -4.7 | 0.656 |
| ▲ GPT-4.1 | ✓ | ? | 3.2 | 81.5 | 91.7 | -5.9 | 0.635 |
| ▲ DeepSeek-V3 | ? | 671 | 4.1 | 81.0 | 86.8 | -5.9 | 0.621 |
| CommandA-WMT | ✗ | 111 | 4.4 | 78.2 | 81.9 | -5.3 | 0.63 |
| UvA-MT | ? | 12 | 4.5 | 79.0 | 82.9 | -5.7 | 0.636 |
| ▲ Mistral-Medium | ? | ? | 4.6 | 80.8 | 85.6 | -6.1 | 0.614 |
| ▲ Gemma-3-27B | ✓ | 27 | 5.0 | 79.5 | 84.7 | -6.1 | 0.61 |
| ▲ Claude-4 | ? | ? | 5.3 | 80.9 | 85.4 | -6.6 | 0.601 |
| IRB-MT | ✓ | 12 | 5.8 | 76.3 | 80.4 | -5.8 | 0.606 |
| ▲ TowerPlus-9B[M] | ✓ | 9 | 6.0 | 77.0 | 81.3 | -6.2 | 0.602 |
| ▲ ONLINE-B | ✓ | ? | 6.1 | 76.2 | 80.5 | -6.1 | 0.599 |
| SalamandraTA | ✓ | 8 | 6.1 | 75.0 | 78.0 | -6.0 | 0.621 |
| ▲ Llama-4-Maverick | ✓ | 400 | 6.2 | 78.4 | 81.8 | -6.6 | 0.591 |
| ▲ ONLINE-W | ? | ? | 7.2 | 75.6 | 80.7 | -7.0 | 0.591 |
| ▲ TowerPlus-72B[M] | ✓ | 72 | 7.5 | 75.2 | 77.6 | -6.8 | 0.58 |
| ▲ Qwen3-235B | ✓ | 235 | 8.2 | 73.9 | 75.3 | -6.8 | 0.571 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | 8.3 | 74.4 | 76.6 | -7.2 | 0.568 |
| IR-MultiagentMT | ✗ | ? | 8.3 | 74.6 | 76.9 | -7.2 | 0.564 |
| TranssionTranslate | ? | ? | 8.4 | 69.7 | 75.7 | -6.2 | 0.563 |
| ▲ CommandA | ✗ | 111 | 8.9 | 74.6 | 75.0 | -7.3 | 0.551 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | 9.8 | 70.8 | 72.3 | -7.3 | 0.555 |
| ▲ Gemma-3-12B | ✓ | 12 | 11.4 | 67.2 | 69.2 | -7.7 | 0.528 |
| ▲ Llama-3.1-8B | ✗ | 8 | 14.6 | 63.8 | 61.1 | -8.8 | 0.483 |
| ▲ ONLINE-G | ✓ | ? | 17.7 | 61.8 | 59.9 | -10.6 | 0.422 |
| ▲ NLLB | ✓ | 1 | 18.1 | 58.9 | 58.2 | -10.5 | 0.436 |
| ▲ Mistral-7B | ✗ | 7 | 21.0 | 56.2 | 50.3 | -11.2 | 0.374 |
| ▲ AyaExpans-32B | ✗ | 32 | 21.1 | 55.1 | 49.6 | -11.1 | 0.376 |
| ▲ Qwen2.5-7B | ? | 7 | 26.0 | 47.5 | 42.0 | -12.9 | 0.304 |
| ▲ CommandR7B | ✗ | 7 | 27.8 | 41.0 | 31.7 | -12.7 | 0.316 |
| ▲ AyaExpans-8B | ✗ | 8 | 32.0 | 40.1 | 33.7 | -15.5 | 0.211 |

| English-Turkish | | | | | | | |
|-----------------------|--------------|-------------|------------|------------------|--------------------|------------------------|-------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA-ESA-CMDA ↑ | GEMBA-ESA-GPT4.1 ↑ | MetricX-24-Hybrid-XL ↑ | XCOMET-XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | 1.0 | 81.4 | 85.2 | -7.2 | 0.542 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 2.7 | 82.7 | 87.9 | -8.4 | 0.462 |
| ▲ GPT-4.1 | ✓ | ? | 3.0 | 83.1 | 86.1 | -8.6 | 0.465 |
| CommandA-WMT | ✓ | 111 | 3.3 | 77.9 | 80.2 | -7.8 | 0.491 |
| GemTrans | ✓ | 27 | 3.3 | 74.6 | 78.8 | -7.4 | 0.506 |
| ▲ DeepSeek-V3 | ? | 671 | 3.4 | 81.2 | 84.5 | -8.6 | 0.461 |
| ▲ Mistral-Medium | ? | ? | 5.3 | 76.0 | 78.7 | -9.0 | 0.44 |
| ▲ Claude-4 | ✓ | ? | 5.5 | 77.7 | 80.1 | -9.4 | 0.424 |
| UvA-MT | ? | 12 | 5.6 | 72.6 | 76.0 | -8.8 | 0.46 |
| ▲ ONLINE-W | ? | ? | 6.2 | 73.7 | 76.1 | -9.1 | 0.431 |
| ▲ ONLINE-B | ✓ | ? | 7.1 | 72.4 | 72.5 | -9.3 | 0.414 |
| IRB-MT | ✓ | 12 | 7.2 | 69.9 | 71.8 | -8.9 | 0.415 |
| ▲ Llama-4-Maverick | ✓ | 400 | 7.3 | 72.5 | 74.7 | -9.8 | 0.409 |
| TranssionTranslate | ? | ? | 7.6 | 68.1 | 71.6 | -9.1 | 0.413 |
| ▲ Qwen3-235B | ✓ | 235 | 7.7 | 69.4 | 71.2 | -9.3 | 0.408 |
| ▲ CommandA | ✓ | 111 | 8.1 | 71.5 | 72.2 | -9.9 | 0.393 |
| ▲ Gemma-3-12B | ✓ | 12 | 8.7 | 68.5 | 69.4 | -9.8 | 0.391 |
| ▲ EuroLLM-22B-pre.[M] | ✓ | 22 | 9.0 | 66.1 | 69.1 | -9.9 | 0.397 |
| ▲ Gemma-3-27B | ✓ | 27 | 9.1 | 66.9 | 69.6 | -10.1 | 0.394 |
| ▲ AyaExpanse-32B | ✓ | 32 | 9.9 | 64.5 | 66.1 | -10.2 | 0.383 |
| ▲ EuroLLM-9B[M] | ✓ | 9 | 10.8 | 59.6 | 60.5 | -10.2 | 0.409 |
| IR-MultiagentMT | ✗ | ? | 10.9 | 64.1 | 65.3 | -10.6 | 0.351 |
| ▲ AyaExpanse-8B | ✓ | 8 | 13.0 | 58.6 | 58.7 | -11.0 | 0.325 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | 13.5 | 58.5 | 56.7 | -11.3 | 0.325 |
| ▲ ONLINE-G | ✓ | ? | 14.3 | 58.0 | 58.6 | -11.9 | 0.294 |
| ▲ NLLB | ✓ | 1 | 15.5 | 53.3 | 55.3 | -12.4 | 0.304 |
| ▲ Llama-3.1-8B | ✗ | 8 | 17.8 | 51.1 | 48.4 | -12.9 | 0.248 |
| ▲ CommandR7B | ✗ | 7 | 18.0 | 48.4 | 42.9 | -12.8 | 0.291 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | 22.1 | 43.6 | 36.9 | -14.6 | 0.192 |
| ▲ Qwen2.5-7B | ? | 7 | 22.7 | 41.2 | 38.5 | -14.9 | 0.174 |
| ▲ Mistral-7B | ✗ | 7 | 31.0 | 27.1 | 22.2 | -20.2 | 0.138 |

| English-Vietnamese | | | | | | | |
|-----------------------|--------------|-------------|------------|------------------|--------------------|------------------------|-------------|
| System Name | LP Supported | Params. (B) | AutoRank ↓ | GEMBA-ESA-CMDA ↑ | GEMBA-ESA-GPT4.1 ↑ | MetricX-24-Hybrid-XL ↑ | XCOMET-XL ↑ |
| Shy-hunyuan-MT | ✓ | 7 | 1.0 | 83.1 | 87.3 | -4.5 | 0.623 |
| ▲ Gemini-2.5-Pro | ✓ | ? | 2.7 | 82.3 | 88.6 | -5.6 | 0.539 |
| CommandA-WMT | ✓ | 111 | 2.7 | 78.4 | 83.2 | -4.9 | 0.577 |
| ▲ GPT-4.1 | ✓ | ? | 2.8 | 82.9 | 88.1 | -5.7 | 0.533 |
| ▲ DeepSeek-V3 | ? | 671 | 3.2 | 81.5 | 85.5 | -5.7 | 0.533 |
| ▲ Qwen3-235B | ✓ | 235 | 3.3 | 79.9 | 84.1 | -5.5 | 0.539 |
| GemTrans | ✓ | 27 | 3.4 | 74.8 | 80.3 | -4.8 | 0.572 |
| UvA-MT | ? | 12 | 3.7 | 77.0 | 80.8 | -5.5 | 0.559 |
| ▲ Mistral-Medium | ? | ? | 3.7 | 78.7 | 83.8 | -5.8 | 0.53 |
| ▲ Claude-4 | ? | ? | 5.0 | 78.2 | 81.5 | -6.7 | 0.494 |
| IRB-MT | ✓ | 12 | 5.1 | 74.1 | 77.7 | -5.8 | 0.506 |
| ▲ AyaExpanse-32B | ✓ | 32 | 5.8 | 72.3 | 76.9 | -6.3 | 0.498 |
| ▲ Llama-4-Maverick | ✓ | 400 | 6.6 | 72.2 | 76.4 | -6.9 | 0.47 |
| ▲ ONLINE-B | ✓ | ? | 6.6 | 70.9 | 74.4 | -6.6 | 0.478 |
| TranssionTranslate | ? | ? | 7.2 | 66.4 | 70.7 | -6.1 | 0.476 |
| ▲ AyaExpanse-8B | ✓ | 8 | 7.8 | 66.3 | 70.1 | -6.7 | 0.465 |
| ▲ Gemma-3-12B | ✓ | 12 | 8.1 | 67.2 | 70.9 | -7.2 | 0.448 |
| ▲ CommandA | ✓ | 111 | 8.7 | 66.8 | 69.2 | -7.7 | 0.442 |
| IR-MultiagentMT | ✗ | ? | 8.7 | 67.0 | 70.0 | -7.5 | 0.424 |
| ▲ TowerPlus-72B[M] | ✗ | 72 | 8.8 | 65.3 | 65.4 | -7.3 | 0.46 |
| ▲ Gemma-3-27B | ✓ | 27 | 9.9 | 62.8 | 65.5 | -7.9 | 0.42 |
| ▲ Qwen2.5-7B | ✓ | 7 | 10.8 | 61.4 | 61.2 | -8.3 | 0.41 |
| ▲ Llama-3.1-8B | ✗ | 8 | 11.8 | 59.3 | 60.7 | -8.9 | 0.385 |
| ▲ CommandR7B | ✓ | 7 | 13.1 | 55.7 | 52.0 | -9.6 | 0.406 |
| ▲ NLLB | ✓ | 1 | 15.5 | 54.0 | 53.9 | -11.4 | 0.303 |
| ▲ TowerPlus-9B[M] | ✗ | 9 | 16.8 | 46.2 | 42.1 | -10.7 | 0.319 |
| ▲ ONLINE-G | ✓ | ? | 17.4 | 52.5 | 51.0 | -12.6 | 0.238 |
| ▲ Mistral-7B | ✗ | 7 | 24.4 | 33.7 | 33.6 | -15.9 | 0.139 |
| ▲ EuroLLM-9B[M] | ✗ | 9 | 27.3 | 18.8 | 9.4 | -17.9 | 0.327 |
| ▲ EuroLLM-22B-pre.[M] | ✗ | 22 | 30.0 | 22.2 | 20.8 | -20.5 | 0.113 |

References

- Cohere Team. 2025. [Command a: An enterprise-ready large language model](#). *Preprint*, arXiv:2504.00698.
- 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, Torino, Italia. ELRA and ICCL.
- Mara Finkelstein and Markus Freitag. 2024. [MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods](#). In *The Twelfth International Conference on Learning Representations*.
- Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. [High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics](#). *Transactions of the Association for Computational Linguistics*, 10:811–825.
- Markus Freitag, Nitika Mathur, 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.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022b. [Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68. Association for Computational Linguistics.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xcomet: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- 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, Miami, Florida, USA. Association for Computational Linguistics.
- Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. [DEMETR: Diagnosing evaluation metrics for translation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561. Association for Computational Linguistics.
- 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](#). *Preprint*, arXiv:2406.11580.
- Geza Kovacs, Daniel Deutsch, and Markus Freitag. 2024. [Mitigating metric bias in minimum Bayes risk decoding](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1063–1094, Miami, Florida, USA. Association for Computational Linguistics.
- Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Senrich, and Liane Guillou. 2024. [Machine translation meta evaluation through translation accuracy challenge sets](#). *Preprint*, arXiv:2401.16313.
- OpenAI. 2025. Introducing gpt-4.1 in the api. <https://openai.com/index/gpt-4-1/>. Model announcement and documentation; accessed 2025-08-09.
- 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, José 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.
- 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anushka Singh, Ananya Sai, Raj Dabre, Ratish Pudupully, 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, Bangkok, Thailand. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenertorp. 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, Miami, Florida, USA. Association for Computational Linguistics.

A Metrics correlations

To examine how the metrics used for AutoRank correlate with each other, we calculated the Pearson correlation between paragraph-level scores for all systems, resulting in a sample size of around 14k scores per each language pair.

The results show that GEMBA-ESA on CmdA and GPT-4.1 exhibit the highest correlations for almost all languages. In contrast, the weakest correlations are generally observed between xComet and both GEMBA-ESA variants.

When examining results by language pair, Bhojpuri, Maasai, and Marathi show the lowest correlations. This is why we use chrF++ for the first two language pairs. Unfortunately, no reference translations are available for Marathi, so we must rely on QE metrics for its evaluation.

| | Kiwi G-CmdA | Kiwi G-GPT | Kiwi MetX | Kiwi xComet | G-CmdA G-GPT | G-CmdA MetX | G-CmdA xComet | G-GPT MetX | G-GPT xComet | MetX xComet |
|---------------|----------------|---------------|--------------|----------------|-----------------|----------------|------------------|---------------|-----------------|----------------|
| cs-de_DE | 0.441 | 0.484 | 0.541 | 0.709 | 0.732 | 0.583 | 0.403 | 0.636 | 0.436 | 0.560 |
| cs-uk_UA | 0.531 | 0.600 | 0.696 | 0.794 | 0.708 | 0.571 | 0.517 | 0.654 | 0.573 | 0.710 |
| en-ar_EG | 0.610 | 0.573 | 0.750 | 0.494 | 0.740 | 0.624 | 0.350 | 0.605 | 0.268 | 0.565 |
| en-bho_IN | 0.465 | 0.093 | 0.517 | 0.030 | 0.503 | 0.621 | 0.051 | 0.428 | -0.008 | 0.194 |
| en-bn_BD | 0.742 | 0.752 | 0.822 | 0.498 | 0.802 | 0.735 | 0.435 | 0.730 | 0.448 | 0.584 |
| en-cs_CZ | 0.617 | 0.696 | 0.728 | 0.747 | 0.757 | 0.642 | 0.533 | 0.682 | 0.535 | 0.712 |
| en-de_DE | 0.481 | 0.546 | 0.612 | 0.789 | 0.742 | 0.578 | 0.350 | 0.593 | 0.358 | 0.559 |
| en-el_GR | 0.736 | 0.777 | 0.787 | 0.691 | 0.863 | 0.716 | 0.542 | 0.743 | 0.544 | 0.741 |
| en-et_EE | 0.783 | 0.837 | 0.825 | 0.720 | 0.787 | 0.736 | 0.583 | 0.795 | 0.655 | 0.802 |
| en-fa_IR | 0.814 | 0.834 | 0.862 | 0.703 | 0.852 | 0.785 | 0.596 | 0.793 | 0.589 | 0.689 |
| en-hi_IN | 0.651 | 0.663 | 0.654 | 0.443 | 0.754 | 0.658 | 0.432 | 0.681 | 0.459 | 0.634 |
| en-id_ID | 0.696 | 0.777 | 0.705 | 0.680 | 0.775 | 0.633 | 0.542 | 0.653 | 0.552 | 0.775 |
| en-is_IS | 0.787 | 0.811 | 0.839 | 0.659 | 0.756 | 0.713 | 0.495 | 0.787 | 0.620 | 0.741 |
| en-it_IT | 0.549 | 0.596 | 0.691 | 0.780 | 0.735 | 0.566 | 0.470 | 0.583 | 0.456 | 0.716 |
| en-ja_JP | 0.644 | 0.668 | 0.717 | 0.691 | 0.752 | 0.626 | 0.543 | 0.637 | 0.496 | 0.715 |
| en-kn_IN | 0.796 | 0.778 | 0.826 | 0.379 | 0.790 | 0.714 | 0.324 | 0.703 | 0.375 | 0.563 |
| en-ko_KR | 0.645 | 0.667 | 0.699 | 0.680 | 0.774 | 0.643 | 0.580 | 0.648 | 0.547 | 0.738 |
| en-lt_LT | 0.798 | 0.837 | 0.858 | 0.726 | 0.828 | 0.755 | 0.556 | 0.783 | 0.601 | 0.762 |
| en-mas_KE | 0.694 | 0.325 | 0.403 | 0.124 | 0.460 | 0.406 | 0.223 | 0.096 | -0.085 | 0.533 |
| en-mr_IN | 0.738 | 0.622 | 0.785 | 0.179 | 0.610 | 0.685 | 0.124 | 0.595 | 0.034 | 0.320 |
| en-ro_RO | 0.634 | 0.707 | 0.748 | 0.796 | 0.753 | 0.619 | 0.546 | 0.648 | 0.561 | 0.762 |
| en-ru_RU | 0.580 | 0.647 | 0.677 | 0.731 | 0.707 | 0.534 | 0.499 | 0.575 | 0.500 | 0.742 |
| en-sr_Cyrl_RS | 0.699 | 0.775 | 0.714 | 0.743 | 0.737 | 0.577 | 0.577 | 0.655 | 0.664 | 0.696 |
| en-sr_Latn_RS | 0.731 | 0.789 | 0.724 | 0.691 | 0.797 | 0.672 | 0.532 | 0.661 | 0.564 | 0.610 |
| en-sv_SE | 0.662 | 0.738 | 0.777 | 0.830 | 0.780 | 0.634 | 0.573 | 0.706 | 0.641 | 0.798 |
| en-th_TH | 0.821 | 0.845 | 0.837 | 0.667 | 0.831 | 0.775 | 0.585 | 0.797 | 0.639 | 0.735 |
| en-tr_TR | 0.704 | 0.758 | 0.713 | 0.649 | 0.782 | 0.619 | 0.498 | 0.642 | 0.516 | 0.738 |
| en-uk_UA | 0.646 | 0.704 | 0.745 | 0.763 | 0.752 | 0.594 | 0.550 | 0.643 | 0.568 | 0.771 |
| en-vi_VN | 0.714 | 0.762 | 0.762 | 0.641 | 0.827 | 0.685 | 0.507 | 0.698 | 0.522 | 0.743 |
| en-zh_CN | 0.557 | 0.633 | 0.653 | 0.653 | 0.688 | 0.584 | 0.525 | 0.584 | 0.518 | 0.744 |
| ja-zh_CN | 0.508 | 0.553 | 0.658 | 0.735 | 0.779 | 0.639 | 0.532 | 0.639 | 0.545 | 0.718 |