



WMT26 by the Numbers: 46 Systems, 23 Directions, and the 29.30 BLEU Ceiling on Low-Resource Translation

What does machine translation actually deliver for a language your vendor does not care about? WMT26 put 46 systems through human annotators, and the best English-to-Khasi score came in at 29.30 BLEU.

THE SHORT VERSION

The WMT26 General MT findings, published October 29, 2026, evaluated 46 systems across 23 directions with human error-span annotation. In the parallel low-resource Indic task, peak English-to-Indic BLEU was 29.30 for Khasi, 21.11 for Mizo and 19.38 for Assamese. Those are the best results the field produced under its own benchmark, and they are nowhere near production numbers.

The headline number is 29.30, and it needs unpacking. BLEU is a surface overlap score



against a reference translation, so reading it is a judgment call rather than an exact measurement, but a mid-to-high-20s BLEU in a high-resource pair like English-German is roughly “fluent with occasional errors you can post-edit.” In Khasi, a language from one of the three Northeast Indian language families covered by the [WMT26 Low-resource Indic task](#), 29.30 is the ceiling of the entire competitive field across Healthcare, Political, Travel, Sports and Entertainment domains. Mizo tops out at 21.11. Assamese, a language with tens of millions of speakers, sits at 19.38. The task covered 10 English-Indic pairs, evaluated in both directions.

At 29.30 BLEU the human in the loop is still doing the translating

My read: at those scores the output works as a draft substrate for a bilingual editor and as a gist tool for internal search or triage. It does not work for patient instructions, contract text or customer-facing copy without a full human rewrite. The organisers evaluated Healthcare as one of five domains, which is exactly where I would be most careful about extrapolating from an aggregate score.

The field is also still arguing about what “best” means. One WMT26 system paper reports a pruned, fine-tuned M2M100 at BLEU 42.04 against OPUS-MT-en-ar’s 44.59, while beating the same baseline on COMET at 0.8730 versus 0.7911. Two metrics, two different winners, one pair of systems.

MEASUREMENT	PRUNED FINE-TUNED M2M100	OPUS-MT-EN-AR
BLEU	42.04	44.59
COMET	0.8730	0.7911
Source	WMT26 proceedings, 2026-10	WMT26 proceedings, 2026-10

That disagreement matters commercially. If you are procuring translation capability and your acceptance criteria name a single automatic metric, you have handed the vendor the choice of which number to optimise. I have written before about how [benchmark selection quietly decides the ranking](#), and this is the same mechanism operating inside a single shared task.



A 26-point gap that comes from the test set, not the model

The Chinese–Southeast Asian task covered 14 directions between Chinese and seven Southeast Asian languages. One system reported macro-average BLEU of 43.70 and 49.03 on a 6,859-pair internal test set, and 17.52 and 26.03 on FLORES-200. Same system, same weights, roughly 26 points of difference from where the sentences came from. Unless a vendor names the test set, the domain mix and the metric, a BLEU figure tells you close to nothing. Ask for FLORES-200 or an equivalent public set alongside whatever internal set produced the marketing figure.

The same paper reports throughput of 237.02 and 241.96 sentences per second on an RTX 5090, median of three steady-state measurements. I find that the more actionable of the two numbers, because it scales linearly into a cost model and does not depend on which reference corpus someone picked.

Contrastive error span annotation replaced the old scoring setup

WMT26 moved to contrastive Error Span Annotation, which shows an annotator three translations of the same source item at once, asks them to mark error spans, then asks for a holistic score. Test sets moved to document level using HTML or JSON structure rather than mandatory segment alignment, and dynamic sampling sent more annotation budget to the higher-performing systems. The findings paper is titled "[Contrastive Dynamic Human Evaluation at Scale](#)", which is an accurate description of what changed. The 46 systems break down into 30 participant submissions plus 16 translations from LLMs and industry translation providers.

Two details I would weigh before treating this year's scores as directly comparable to last year's. Some WMT26 reference translations were created by post-editing open-weight model outputs rather than written from scratch, which makes references cheaper to produce and plausibly closer in style to model output. I cannot size that bias from the published numbers, but it is a direction of bias worth naming. And dynamic sampling means



WMT26 by the Numbers: 46 Systems, 23 Directions, and the 29.30 BLEU Ceiling on Low-Resource Translation

not every system got equal annotation volume, so confidence is uneven across the table by design.

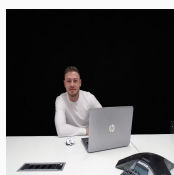
WMT26 also merged the old metrics and quality-estimation tracks into a single Automated Translation Quality Evaluation Systems task, with three segment-level views: error span detection, continuous quality-score prediction, and identification of error-free translations. That task used the human cESA judgments from General MT and covered 21 translation directions. The third view, finding the segments that are actually clean, is the one I would watch, because it maps onto the only automation decision most teams need: which outputs skip human review.

WHERE I LAND

Treat WMT as the floor estimate rather than the pessimistic case. These are 30 participant teams plus 16 LLM and industry systems competing on curated test sets with a known domain mix, and the top low-resource scores still sit between 19 and 30 BLEU. Your production distribution is messier than the test set. So if a vendor quotes you a number above the WMT26 ceiling for one of these directions, the first question is which test set, and the second is whether the references were post-edited model output.

Participation scale says something about where the effort went: UvA-MT and BSC each covered all 23 directions, and MakotoAI processed 4,897 documents across all 23 pairs. Coverage is there. Quality in the long tail is not, and the organisers' own numbers say so.

If you are planning a localisation roadmap that includes Assamese, Mizo, Khasi or similar, budget human translation capacity as the primary cost line and treat MT as a throughput multiplier on top of it. The 19.38 for Assamese is the best the field produced in October 2026, and nothing in the WMT26 proceedings points to a near-term path to a different order of magnitude.





WMT26 by the Numbers: 46 Systems, 23 Directions, and the 29.30 BLEU Ceiling on Low-Resource Translation

WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →