Mind the Gap: Exposing LLM Translation Blind Spots Using the AlphaMWE Multilingual Parallel Corpus

Lifeng Han[1,2], **Jiahui Liang**[1], **Anna Latusek**[3], **Karim El Haff**[4], **Amal Haddad Haddad**[5], **Josua Höfgen**[6], **Kilian Evang**[7], **Min Ma**[8], **Maryia Zhyrko**[1]
[1]Leiden University [2]Leiden University Medical Centre
[3]Institute of Computer Science PAS, Warsaw
[4]Faculty of Computer and Information Science, University of Ljubljana
[5]University of Granada [6] Technical University of Munich
[7]Heinrich Heine University Düsseldorf [8]Google DeepMind
j.h.l.jiahui@hum.leidenuniv.nl | l.han@lumc.nl

Abstract

LLMs' performance on machine translation (MT) tasks is often dependent on the data availability in the specific domains and language pairs that they are trained upon. To examine if multiword expressions (MWEs) still present a bottleneck for LLMs regarding language understanding and translation, we report the performance of systems from the WMT 2026 Test Suites shared task, using portions of the publicly available multilingual parallel corpus AlphaMWE as test suites. We received 31 MT systems' outputs covering English to Chinese (zh), Polish (pl), German (de), and Arabic (ar) including Modern Standard Arabic (MSA) and two dialectal ones (Egyptian and Tunisian Arabic). We carried out automatic evaluations using BLEU, ChrF, and BERT-score to select the top 3 systems per language pair, followed up with human evaluations on the selected systems. Our findings show that figurative and MWE-related phenomena remain challenging for contemporary MT systems, automatic metrics sometimes disagree in system ranking, and human evaluation uncovers language-specific errors that remain hidden by aggregate scores. Inter-annotator analysis further reveals challenges in consistently identifying and calibrating linguistically subtle translation errors, highlighting the need for explicit evaluation guidelines and careful annotator calibration.

1 Introduction

Large Language Models (LLMs) have substantially advanced machine translation (MT), achieving remarkable fluency and strong automatic evaluation scores across many language pairs. Nevertheless, recent studies suggest that high overall translation quality does not imply that all linguistic phenomena are equally well handled. In particular, multiword expressions (MWEs) such as idioms, metaphorical language, and other non-compositional constructions remain challenging because their meaning cannot always be inferred compositionally from individual words. Correct translation therefore requires not only lexical knowledge but also contextual reasoning, semantic interpretation, and language-specific linguistic competence.

Recent work has begun to investigate these remaining blind spots. Cheng and Amiri (2025) argue that LLMs exhibit systematic linguistic blind spots despite their strong overall performance. Laureano De Leon et al. (2025) demonstrate that multilingual and code-switched MWEs remain difficult for current LLMs, while Mi et al. (2025) show that idiomatic expressions are frequently mistranslated when successful interpretation depends on contextual understanding rather than surface word associations. More recently, Liu et al. (2026); Liang and Han (2026) reported that verbal MWEs and metaphors continue to pose substantial challenges for machine translation systems. Together, these studies suggest that automatic MT quality improvements have not eliminated linguistically difficult translation phenomena.

However, most previous work evaluates LLMs in controlled experimental settings or focuses on individual linguistic phenomena. Less is known about how a diverse collection of contemporary MT systems behaves on a shared multilingual benchmark specifically designed to expose these challenging constructions. Furthermore, automatic evaluation metrics such as BLEU or BERTScore often provide only an overall quality estimate and may overlook linguistically important translation errors, particularly those involving idiomaticity, lexical ambiguity, or figurative meaning.

To investigate these issues, we participated in the WMT 2026 Test Suites Shared Task by submitting AlphaMWE, a multilingual parallel corpus containing manually annotated multiword expressions, as a diagnostic test suite. AlphaMWE covers English–Chinese, English–German, English–Polish, and English–Arabic (Modern Standard Ara-

Proceedings of the Eleventh Conference on Machine Translation, pages 1533–1554
October 28-29, 2026 ©2026 Association for Computational Linguistics

bic, Egyptian Arabic, and Tunisian Arabic), and contains texts from multiple genres including technical documentation, news, and literature. The shared task attracted 31 MT systems, providing a broad comparison of contemporary translation models.

Our evaluation follows a three-stage pipeline. We first construct the WMT-compatible test suites from AlphaMWE, then evaluate all submitted systems using SacreBLEU, chrF++, and BERTScore, and finally perform HOPE-based human evaluation on the top-ranked systems for each language pair. This combination enables us to compare automatic and human assessment while analysing translation behaviour across multiple languages and linguistic phenomena.

It is worth noting that, at this stage of our work, we decided to conduct the analysis and comparisons at the level of complete sentences. We therefore assessed the meaning of each sentence as a whole, rather than focusing solely on the translations of the MWEs. Thus, our primary focus was on the overall meaning conveyed by sentences containing MWEs, as well as on identifying which types of errors resulting from the use of MWEs occur in different types of texts.

Our analysis shows that despite strong automatic evaluation scores, idioms, verbal MWEs, metaphorical expressions, lexical ambiguity, and language-specific collocations remain important sources of translation errors. These shortcomings are especially evident in literary and expression-rich texts, whereas technical texts are translated much more reliably. We further observe that automatic evaluation metrics do not always agree on system ranking, and that human evaluation reveals language-specific translation issues which remain largely invisible to aggregate automatic scores.

Our main contributions are:

- We introduce AlphaMWE as a multilingual diagnostic test suite in the WMT 2026 Test Suites Shared Task, covering six target-language variants and outputs from 31 participating MT systems.
- We combine three complementary automatic evaluation metrics with HOPE-based human evaluation to investigate both system ranking and linguistically motivated translation quality.
- We provide a cross-language analysis showing that MWEs, idiomatic expressions, figurative language, lexical ambiguity, and language-specific grammatical phenomena remain persistent blind spots for contemporary MT systems despite their overall high translation quality.

2 WMT2026 Test Suites: AlphaMWE

We present the overall methodology of this shared task in Figure 1 including three stages: Stage I: Test Suite Construction; Stage II: Automatic Evaluation; Stage III: Human Evaluation & Analysis.

For Stage I, AlphaMWE (Han et al., 2026) is a multilingual parallel corpus featuring multiword expression (MWE) annotations, machine translation post-editing, and human-in-the-loop quality controls. It builds on data from sources such as the PARSEME shared task (Savary et al., 2017) to evaluate how translation engines handle complex phrasal and verbal expressions. It was originally designed by Han et al. (2020) for English–German, English–Polish, and English–Chinese language pairs, which was later extended to the Arabic edition by Hadj Mohamed et al. (2023). AlphaMWE has a mix of text genres, including IT document instructions, news/politics, and literature texts.

The original AlphaMWE corpus contained 5 files of equal size, with around 150 segments per file, for a total of 747 segments. However, their annotation was not complete for English–Arabic and English–Polish, resulting in 150 segments for English–MSA (modern standard Arabic), English–AEB (Tunisian Arabic), English–ARZ (Egyptian Arabic), and 597 segments for English–Polish[1].

For Stage II, we deployed SacreBLEU (Post, 2018), chrF++ (Popović, 2017), and BERTScore (Zhang et al., 2020) for precision-oriented n-gram matching, robust character-level matching, and contextual semantic similarity.

For system selection, we apply average rank instead of average score. Average rank is preferable because BLEU, chrF++, and BERTScore have different numerical scales. We rank the system per language pair using each metric, then average their ranking on all tested metrics.

For Stage III, we adopt the human-centric post-editing based metric HOPE (Gladkoff and Han, 2022) with its original eight error types and sever-

[1]Following ISO 639-3 language identifiers.

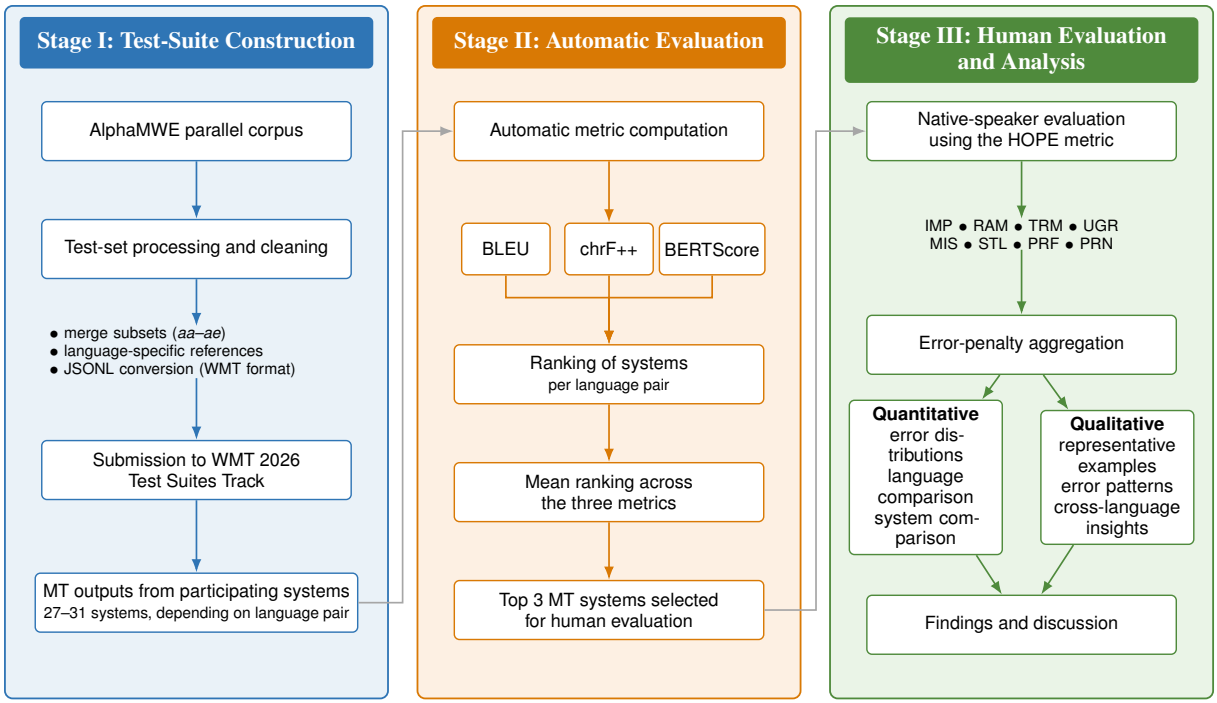

Figure 1: Overall methodology for AlphaMWE test-suite construction, automatic evaluation of WMT 2026 system outputs, top 3 system selection per language pair, and HOPE-based human evaluation followed by quantitative and qualitative analyses.

ity levels (0, 1, 2, 4, 8, 16) for (minor, medium, major, severe, and critical). The eight error types are: Impact (IMP), Required Adaptation Missing (RAM), Terminology (TRM), Ungrammatical (UGR), Mistranslation (MIS), Style (STL), Proofreading (PRF), and Proper Name (PRN).

3 Systems Received

We list the statistics of MT system outputs we received from WMT in Table 1. The varying numbers of participating systems reflect the fact that not all WMT 2026 Test Suites participants submitted translations for every language pair.

4 Automatic Evaluation

The top 3 systems per translation language pair are displayed in Table 2, where we presented the metrics of BLEU, chrF++, BERT, average rank (Avg. Rank), rank standard deviation (Rank SD), and the selection rank. Regarding the rank standard deviation (Rank-SD), a low standard deviation indicates that all three automatic metrics agree on the system's quality, while a high standard deviation highlights disagreement between metrics.[2]

The top 3 systems for each language pair were selected based on the mean of the three automatic metric rankings. The rank standard deviation (SD) is reported to indicate the level of agreement among the automatic metrics but was not used as a selection criterion. Nevertheless, we can see from the table that while half of the selected systems across language pairs have Rank-SD value 0 reflecting their stable performances across three metrics, there are another half of the systems having different level of Rank-SD, with the extremist from GPT 5.5 and Gemma 4-31B on the en→arz language pair having SD values of 3.559 and 1.886, very unbalanced ranking across metric. Appendix Section F shows the three metric-specific ranking tables for transparency and reproducibility.

For Chinese, the competition-ranking tie is preserved: SCIR-TG-MT: aggregate rank 1; Google-Translate: aggregate rank 1; Tower 9B: aggregate rank 3. Here, the SD immediately explains the tie for en→zh: both the first two systems have the same mean rank (1.667). SCIR-TG-MT has a lower SD (0.471), meaning the three metrics agree more closely. Google Translate has a higher SD (0.943), indicating more disagreement because BERTScore

[2] Score 0 → all three metrics ranked the system identically (very high agreement). Small SD (0.4–1.0) → often used for strong agreement. Large SD (>2) → the metrics disagree substantially, suggesting that the system's quality depends on the evaluation metric.

Table 1: Coverage of MT systems that returned translations for the AlphaMWE test suites submitted to the WMT 2026 Test Suites Shared Task. Numbers in parentheses indicate the number of evaluation segments in each language pair. English–Polish contains 597 segments because the *ad* subset was unavailable, while the Arabic test suites each contain the 150-segment *aa* subset only.

System	**en→aeb (150)**	**en→ar (150)**	**en→arz (150)**	**en→de (747)**	**en→pl (597)**	**en→zh (747)**
CUNI-AR	✓	✓	✓			
CUNI-EdUKate				✓		
CUNI-MH-v3				✓		
Cohere CAT+	✓	✓	✓	✓	✓	✓
Command A+	✓	✓	✓	✓	✓	✓
DeepSeek V4 Pro	✓	✓	✓	✓	✓	✓
Dubformer	✓	✓	✓	✓	✓	✓
GLM 4–9B	✓	✓	✓	✓	✓	✓
GPT 5.5	✓	✓	✓	✓	✓	✓
GPT OSS 120B	✓	✓	✓	✓	✓	✓
GPT OSS 20B	✓	✓	✓	✓	✓	✓
Gemini 3.1 Pro	✓	✓	✓	✓	✓	✓
Gemma 4–31B	✓	✓	✓	✓	✓	✓
Gemma 4–4B	✓	✓	✓	✓	✓	✓
Google Translate	✓	✓	✓	✓	✓	✓
HW-TSC		✓				✓
Lumen		✓	✓	✓	✓	✓
Ministral 3–14B	✓	✓	✓	✓	✓	✓
Mistral Medium 3.5	✓	✓	✓	✓	✓	✓
Qingqiu-MT-9B	✓	✓	✓	✓	✓	✓
Qwen 3.5–9B	✓	✓	✓	✓	✓	✓
Qwen 3.6–27B	✓	✓	✓	✓	✓	✓
SCIR-TG-MT						✓
SRPOL			✓	✓		✓
SalamandraTA	✓	✓	✓	✓	✓	✓
TRIVE	✓	✓	✓	✓	✓	✓
Tiny Aya Global		✓	✓	✓	✓	✓
Tower 9B	✓	✓	✓	✓	✓	✓
UvA-MT			✓	✓		✓
VoxNexus_V1		✓	✓	✓		
Wayfinder		✓	✓	✓	✓	✓
# Systems	**24**	**28**	**29**	**31**	**27**	**27**

ranked it lower than BLEU and chrF++. Even though the official competition ranking remains 1, 1, 3, the SD provides valuable context for interpreting that tie.

5 Human Evaluation

In Table 3, we list the summed HOPE error penalties for the top 3 systems per language pair over 150 segments. Because every language pair contains exactly 150 segments, these raw totals are directly comparable in terms of corpus size. However, possible differences in annotator severity across target languages should still be acknowledged when making cross-language interpretations. Figure 2 visualises the HOPE-based human evaluation results. Notably, the human-evaluation ranking does not fully reproduce the automatic-evaluation ranking of the selected top-three systems: the ordering changes for English–Arabic, English–German, and English–Polish, while human evaluation resolves the automatic-ranking tie between SCIR-TG-MT and Google Translate for English–Chinese.

5.1 English–Arabic

For Arabic, error annotations across the evaluated sentences revealed substantial differences in translation quality among the three machine translation systems. TRIVE achieved the strongest overall performance, with errors identified in fewer sentences and a cumulative severity score of 57 on all segments. Google Translate exhibited an intermediate profile, with errors present in more sentences and a cumulative severity score of 84 on all 150 segments, mostly due to MIS error type over TRIVE. Gemma 4-4B showed the weakest performance, with errors identified in even more sentences and a cumulative severity score of 284, more than the sum of Google Translate and TRIVE, also introducing many errors in Impact category, followed

Table 2: Top 3 MT systems selected for human evaluation on each AlphaMWE language pair. Systems were selected according to the average of their SacreBLEU, chrF++, and BERTScore F1 rankings. The rank standard deviation (Rank SD) measures the agreement among the three automatic metrics; lower values indicate greater agreement but were *not* used as a selection criterion. Competition ranking was adopted for the final selection rank.

Language Pair	System	BLEU	chrF++	BERT	Avg. Rank	Rank SD	Selection Rank
en→aeb	TRIVE	1	1	1	1.000	0.000	1
	CUNI-AR	2	2	2	2.000	0.000	2
	DeepSeek V4 Pro	4	4	4	4.000	0.000	3
en→ar	GoogleTranslate	1	1	1	1.000	0.000	1
	Gemma 4–4B	2	2	2	2.000	0.000	2
	TRIVE	3	3	3	3.000	0.000	3
en→arz	Wayfinder	2	1	3	2.000	0.816	1
	Gemma 4–31B	5	5	1	3.667	1.886	2
	GPT 5.5	1	2	9	4.000	3.559	3
en→de	VoxNexus_V1	1	1	1	1.000	0.000	1
	DeepSeek V4 Pro	2	2	2	2.000	0.000	2
	TRIVE	3	3	3	3.000	0.000	3
en→pl	DeepSeek V4 Pro	1	1	1	1.000	0.000	1
	TRIVE	3	2	2	2.333	0.471	2
	Dubformer	2	3	3	2.667	0.471	3
en→zh	SCIR-TG-MT	2	2	1	1.667	0.471	1
	GoogleTranslate	1	1	3	1.667	0.943	1
	Tower 9B	3	4	2	3.000	0.816	3

by Terminology, Style, and Proofreading.

Across all three systems, Mistranslation constituted the single largest error category, indicating that semantic accuracy, rather than syntactic well-formedness, was the primary locus of quality degradation. Ungrammatical output was comparatively rare; the few genuine grammar violations identified involved gender agreement failures, such as a masculine verb form incorrectly applied to a feminine subject. More frequently, apparent grammatical irregularities on closer inspection reflected stylistic or collocational choices rather than true violations of Arabic grammar, underscoring the importance of distinguishing genuine ungrammaticality from register mismatch or reduced naturalness in translation evaluation. Several recurring error patterns emerged at the lexical and idiomatic level. Idiomatic expressions posed a particular challenge: an English idiom conveying collective laughter was rendered literally by two systems, producing a semantically unrelated and ultimately incoherent Arabic sentence, while the third system correctly recovered the idiomatic sense. For instance, in the following example in Figure 3, TRIVE is the only system that correctly conveyed the meaning of laughter in the expression "this broke everybody up", the other outputs convey the literal meaning of the breaking (fracturing) of 'everyone'.

Similarly, polysemous source terms were occasionally mapped to the wrong sense entirely, as when a medical usage of a common English noun was translated according to its more frequent, unrelated meaning, yielding an implausible reading within its surrounding context. Errors of this kind, in which an incorrect lexical sense generates a locally incoherent target sentence, were judged more severe than errors of nuance, since they compromise not only precision but basic interpretability for the target reader. Terminological consistency also varied across systems. In several instances, an established or institutionally fixed Arabic term used correctly elsewhere in the same output was replaced by a looser or less standard alternative, producing internal inconsistency within a single document. Related issues included inconsistent transliteration of proper names, occasional calques that preserved English syntactic structure at the expense of natural Arabic phrasing, and isolated failures to adapt source-language conventions, such as untranslated units of measurement or an uncorrected error present in the source text itself, appropriately treated as a required-adaptation failure when the target system failed to resolve it.

Taken together, these findings suggest that translation adequacy in this sample was governed less by surface grammaticality, which was generally

Table 3: Summed HOPE error penalties for the three systems selected for human evaluation on each language pair. Each value is the total penalty assigned to an error category over 150 AlphaMWE segments. Lower scores indicate fewer or less severe translation errors. SEGS is the sum of the eight error-category penalties.

Language Pair	MT System	IMP	RAM	TRM	UGR	MIS	STL	PRF	PRN	SEGS
en→zh	SCIR-TG-MT	0	0	2	1	41	2	6	0	52
	GoogleTranslate	0	0	3	1	79	0	13	2	98
	Tower 9B	0	2	13	4	113	2	27	6	167
en→pl	DeepSeek V4 Pro	27	0	24	5	40	34	6	0	136
	TRIVE	21	0	36	2	22	23	3	16	123
	Dubformer	35	0	26	5	28	30	9	0	133
en→de	VoxNexus_V1	27	0	0	5	16	23	6	0	77
	DeepSeek V4 Pro	32	4	0	16	36	36	7	0	131
	TRIVE	30	4	0	12	32	17	2	0	97
en→ar	GoogleTranslate	10	2	9	5	39	14	5	0	84
	Gemma 4–4B	48	4	33	37	114	18	17	13	284
	TRIVE	10	1	12	1	12	13	8	0	57

well preserved across all systems, than by the accurate resolution of idiomatic expressions, lexical ambiguity, and terminological consistency. TRIVE demonstrated the most robust handling of these higher-order challenges, while Gemma 4-4B was disproportionately affected by mistranslation and inconsistent register, with Google Translate occupying an intermediate position.

5.2 English–Chinese

Across-system differences. The three most apparent error types are Mistranslation, Proofreading, and Terminology, especially for Tower-9B. The SCIR-TG-MT performed much better than the other two systems. Their total error penalties are (52, 98, 167) respectively for SCIR-TG-MT, GoogleMT, and Tower-9B. Even though Mistranslation is the common issue of all these three MT systems, SCIR-TG-MT has a much lower overall penalty score 41 than other two systems (79, 113) respectively for GoogleMT and Tower-9B on this error type. As one example, the source sentence "They intrigued and slandered and hated each other only on that account, – but as to effectually *lifting a little finger* – oh, no. By heavens!" (aa-127) in Figure 4, SCIR-TG-MT had the correct translation "真正动手帮一下忙" of the sentence and idiom "(effectually) lifting a little finger", while the other two systems produced wrong translation (literal) on the idiomatic expression.

Another example can be found via the vMWE "took it back" in the sentence "Afterwards I *took it back* when it was borne in upon me startlingly with what extreme nicety he had estimated the time requisite for the 'affair'." (sentence-id: aa-121), in Figure 5, where SCIR-TG-MT correctly translated it as "收回了之前的看法(withdrew my previous opinion)" while the other two systems produced "收回了那份文件– took back that file (made up)" (GoogleMT) and "把它拿了回来– took it back (literal)" (Tower-9B).

Across-system similarities. Literature text presents greater difficulty than IT document text. This exactly reflects the "gap" that LLMs still need to fill in, i.e., the text with rich expressions and idiomatic and metaphorical phrases. For future research, task-specific evaluation frameworks focusing on figurative language can be a growing topic, e.g., grounded metrics on metaphor translation from Liang and Han (2026), adapted from the general domain HOPE framework.

5.3 English–German

Overall, during annotation we noticed that TRIVE has a considerably more "literary" style. It knows and uses German idioms, cf. aa-119 (*Zahn der Zeit*), aa-105 (*kein Auge trocken*), and introduces additional words that lend the text a more literary character (cf. aa-122, aa-115, aa-109). This stylistic richness, however, sometimes comes at the cost of precision.

A translation issue we encountered repeatedly in practice concerns the word "standard" (aa-137), which, in an engineering context, should almost exclusively be translated as *Norm*. A *Norm* is a document prescribing how to design, build, test, and document a given thing or process, whereas a standard may also encompass non-binding recommendations. Accordingly, the statement "this does not comply with the standard" (*Dies entspricht nicht*

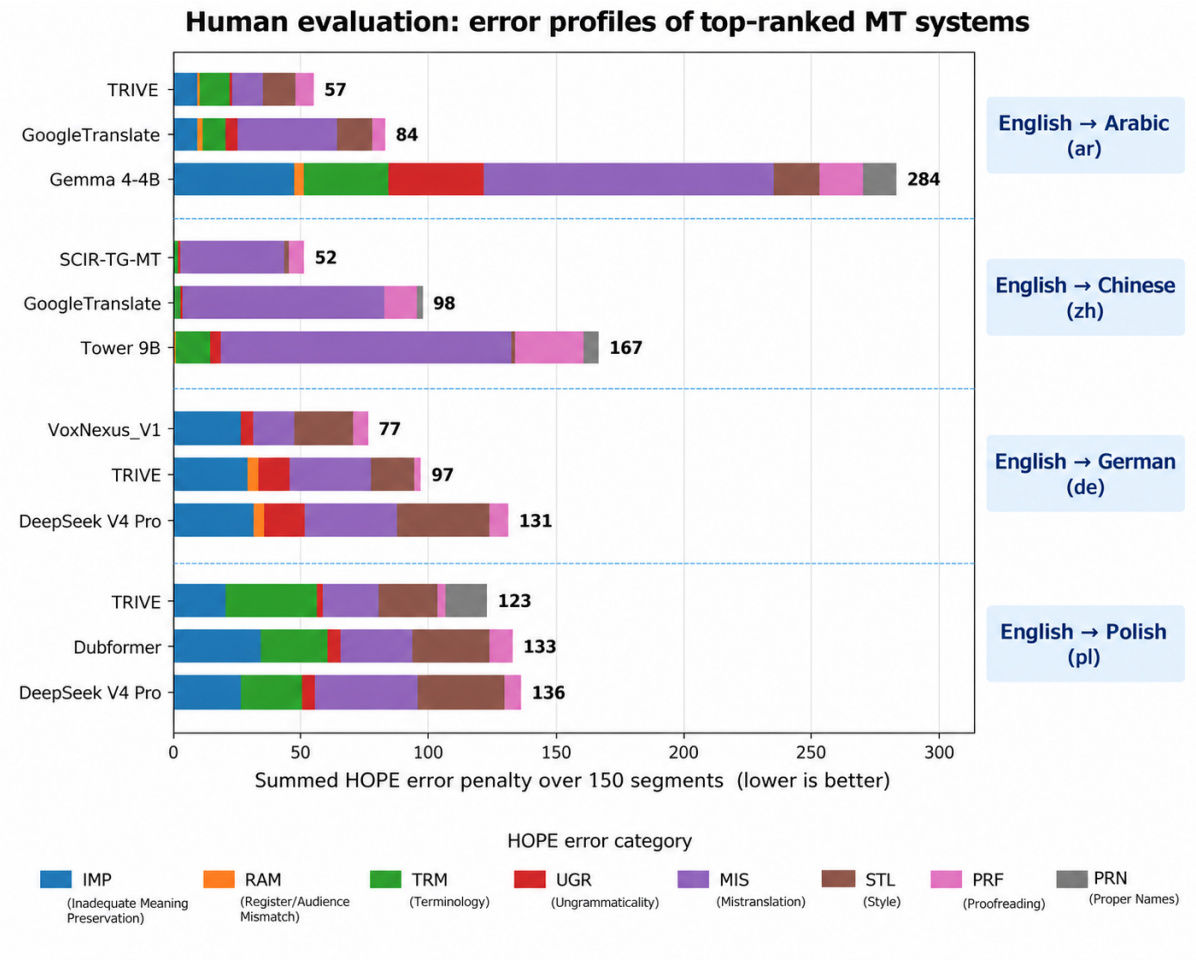

Figure 2: Visualisation of Human Evaluations on Four Language Pairs.

src: This broke everybody up.

ref: هذا حطم الجميع

GoogleTranslate: هذا الأمر أدى إلى تفكك الجميع

Gemma: هذا فرّق الجميع

TRIVE: أضحك هذا الجميع بشدة

Figure 3: Example on Arabic Translation Issues with vMWE (aa-105).

src: They intrigued and slandered and hated each other only on that account, – but as to **effectually lifting a little finger** – oh, no. By heavens!

ref: 他们彼此算计、诽谤、仇恨，仅仅是为了这个——但要说到**真正动手帮一下忙**——哦，不。老天爷啊！

SCIR-TG-MT: 他们彼此算计、诽谤、仇恨，仅仅是为了这个——但要说到**真正动手帮一下忙**——哦，不。老天爷啊！

GoogleTranslate: 他们就因为这件事互相陷害、诽谤和憎恨，但要说**真正动动手指**——哦，不！天哪！

TOWER 9B: 他们只是因为这个原因而互相挑拨、诽谤和憎恨——但至于**真的动一下小指头**——哦，不。天哪！

Figure 4: Example on English-Chinese Translation Issues of Metaphors (aa-127).

dem Standard) is, in an engineering context, generally less consequential than "this does not meet the Norm" (*Dies erfüllt nicht die Norm*). Failing to meet a standard may impact customer acceptance of a business, whereas failing to meet a *Norm* may entail more severe legal consequences.

5.4 English–Polish

The English–Polish translations produced by the selected MT systems were of high quality, although performance varied substantially across text genres. Technical texts were translated accurately owing to well-established terminology, whereas literary

src: Afterwards **I took it back** when it was borne in upon me startlingly with what extreme nicety he had estimated the time requisite for the 'affair.'

ref: 后来，当我猛然意识到他对那"事件"所需时间的估算精确到了何等惊人的地步时，**我收回了之前的看法** 。

SCIR-TG-MT: 他后来，当我猛然意识到他对那"事件"所需时间的估算精确到了何等惊人的地步时，**我收回了之前的看法**。

GoogleTranslate: 后来，**我收回了那份文件**，因为他以极其细致的方式估算了完成“这件事”所需的时间，这让我大吃一惊。

TOWER 9B: 后来，当他以极高的精确度估 计了“这件”事情所需的时间时，我突然意 识到这一点，于是**把它拿了回来** 。

Figure 5: Example on English-Chinese Translation Issues of Figurative Expressions (aa-121).

texts containing idiomatic expressions and MWEs remained considerably more challenging. News-style texts generally exhibited fewer semantic errors because the surrounding context was easier for the systems to infer.

The most frequent error category was stylistic errors (STL), followed by mistranslations (MIS). Stylistic errors mainly involved unnatural Polish sentence structure, awkward collocations and overly complex wording, while mistranslations often resulted from literal translation of MWEs and idiomatic expressions, sometimes changing the meaning of the entire sentence.

Examples:

- STL

  aa-005: *However, when you apply a filter by selection, the conditional filter already applied on the field is removed before the filter by selection is applied.*

  DeepSeek V4 Pro: *Jednakże, gdy zastosujesz filtr według zaznaczenia, warunkowy filtr już zastosowany na tym polu zostanie usunięty przed zastosowaniem filtru według zaznaczenia.*

  Comment: The whole sentence is just complicated and akward. It needs to be read several times to be fully understood.

- MIS

  aa-102: *Shrugging, he gives up and I turn to the **twice disagreeable chicken** and eat guiltily, my appetite spoiled.*

  TRIVE: Wzruszając ramionami, daje za wygraną, a ja wracam do **podwójnie nieapetycznego kurczaka** i jem z poczuciem winy, z popsutym apetytem.

  Comment: This sentence cannot be translated literally while preserving its meaning, and here we are dealing with a literal translation.

Linguistic observations. Several errors reflected fundamental linguistic differences between English and Polish: verbal aspect, grammatical gender, singular/plural/formal “you", idioms, and collocations.

Examples:

- verbal aspect

  aa-012: *When items in a field **are hidden** by conditional filtering, a funnel appears to the left of the drop-down arrow Field arrow.*

  DeepSeek V4 Pro: *Gdy elementy w polu **są ukrywane** przez filtrowanie warunkowe, po lewej stronie strzałki rozwijanego pola (strzałki pola) pojawia się lejek.*

  Comment: It should be a perfective verb *ukryć* used here, instead of imperfective *ukrywać*: *zostały/są ukryte*.

- grammatical gender

  aa-124: ***I heard** the name of Kurtz pronounced, then the words, “take advantage of this unfortunate accident."*

  DeepSeek V4 Pro: ***Usłyszałem** wymówione imię Kurtza, a potem słowa: “skorzystajcie z tego nieszczęśliwego wypadku".*

  Comment: Polish distinguishes grammatical genders. So it could be here either *Usłyszałem* (masculine), or *Usłyszałam* (feminine). Without a context we cannot be sure whether it is she or he speaking.

- singular/plural/formal “you":

  ab-009: *Never mind, Mrs. Bray will join **you** later.*

  DeepSeek V4 Pro: *Nieważne, pani Bray dołączy do **was** później.* → “you" in plural.

  TRIVE: *Nieważne, pani Bray dołączy do **ciebie** później.* → “you" in singular, informal.

  Dubformer: *Mniejsza z tym, pani Bray dołączy do **pana** później.* → “you" in singular, formal (he – sir).

Comment: All translations are correct. More context is needed to determine which use of "you" is appropriate here.

Annotation observations. We also noted that sentence-level evaluation without wider document context occasionally made multiple translations equally acceptable. It shows a limitation of test-suite evaluation rather than MT itself. This has also been reflected by the literature (Gladkoff et al., 2026). We list more English–Polish examples of translation issues in Section C.

Two Polish native speakers completed the annotation independently over all 150 segments; inter-annotator agreement is reported in Appendix B. Agreement on error detection was only fair ($\kappa = 0.32$) and severity calibration differed substantially between the two annotators, so the absolute penalty magnitudes reported for English–Polish should be read as annotator-dependent, even though the system ranking is stable across both annotations. Inter-annotator agreement (IAA) for English–Arabic and English–Chinese is to be reported in the future.

6 Conclusions and Future Work

In this work, we used the AlphaMWE multilingual parallel corpus as a test suite in the WMT 2026 Test Suites Shared Task to investigate whether multiword expressions and other non-compositional phenomena remain challenging for contemporary MT systems. We evaluated outputs from 31 participating systems across English–Chinese, English–German, English–Polish, Modern Standard Arabic, Egyptian Arabic, and Tunisian Arabic. We first applied SacreBLEU, chrF++, and BERTScore F1, ranked systems independently for each language pair, and selected the top-performing systems through mean rank aggregation. We then complemented the automatic evaluation with HOPE-based human evaluation by native speakers.

Across the evaluated language pairs, the human analysis indicates that high automatic metric performance does not eliminate linguistically important translation errors. In particular, idioms, verbal MWEs, metaphorical and figurative expressions, lexical ambiguity, and language-specific collocations continue to expose weaknesses in otherwise fluent MT output. These problems are especially visible in literary and expression-rich text, whereas technical and news-oriented sentences were generally handled more reliably. The observed errors also differed across languages: English–Chinese and English–Arabic showed prominent failures involving literal interpretation of idiomatic and figurative expressions, English–Polish exposed difficulties involving verbal aspect, collocation, grammatical distinctions and context-sensitive lexical choice, while the English–German analysis highlighted the tension between idiomatic or literary rendering and domain-specific terminological precision.

The comparison between automatic and human evaluation further suggests that no single automatic metric provides a complete account of translation quality. Although the three automatic metrics often agreed on the strongest systems, non-zero rank standard deviations for several systems revealed substantial metric disagreement. Human annotation subsequently identified errors – particularly those involving semantic scope, idiomaticity, terminology, and contextual interpretation – that are difficult to characterize through aggregate automatic scores alone. This difference is also reflected at the system level: HOPE-based human evaluation does not fully preserve the ranking induced by the aggregated automatic metrics, changing the relative ordering of the selected systems for several language pairs.

Our results therefore support combining complementary automatic metrics with targeted human evaluation when investigating linguistically challenging translation phenomena.

A further limitation concerns sentence-level evaluation. Several English source sentences are underspecified when removed from their wider discourse context. This is particularly important for languages that require grammatical distinctions not explicitly encoded in English, such as gender, verbal aspect, or singular/plural and formal/informal forms of *you*. In such cases, multiple target translations may be equally valid, and reference-based sentence-level evaluation can potentially penalize legitimate alternatives. The AlphaMWE resource retains contextual material originating from the PARSEME data, which provides an opportunity to address this limitation in future work through document-level or context-aware translation evaluation.

Future work will therefore extend the present analysis in three directions. First, we plan to evaluate context-aware and document-level translation of MWE-containing sentences. Second, we will investigate phenomenon-specific evaluation frameworks for idioms, metaphors, and other figurative expressions, where general-purpose MT metrics

may provide insufficient diagnostic information. Third, larger-scale human annotation with multiple annotators per language will enable analysis of inter-annotator agreement, particularly for error severity, and help separate genuine system errors from variation in linguistic preference and annotation judgement. Together, these directions can further develop AlphaMWE from a multilingual test suite into a diagnostic resource for identifying the remaining linguistic blind spots of LLM-based machine translation.

Limitations

The most important limitations are: 1) two annotators for Arabic but only for one round, 75 segments each; three annotators for Chinese, with two doing 75 segments each, then the 3rd person double-checked the 1st round annotation. Two German native speakers also completed the annotation independently over all 150 segments. Agreement is reported in Appendix D and is markedly weaker than for Polish ($\kappa = 0.10$ on error detection), with the two annotators flagging largely disjoint sets of outputs, so the English–German penalty magnitudes should likewise be treated as annotator-dependent. 2) only 150 common segments used for cross-language human comparison. 3) different systems participated in different language pairs. 4) sentence-level context limitations; and 5) automatic evaluation limited to three metrics, with more metrics considered to be included with extra time, e.g., h/LEPOR (Han et al., 2013), COMET (Rei et al., 2020), and metaphor specific MetaHOPE (Liang and Han, 2026).

Acknowledgments

LH is grateful to the HumanAI cluster fund from LIACS, Leiden University.

References

Jiali Cheng and Hadi Amiri. 2025. Linguistic blind spots of large language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–17, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Serge Gladkoff and Lifeng Han. 2022. HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13–21, Marseille, France. European Language Resources Association.

Serge Gladkoff, Angelika Vaasa, Sue Ellen Wright, Ingemar Strandvik, and Lifeng Han. 2026. Looking under the wrong lamppost: On the limitations of automated translation quality estimation. In *Proceedings of the 9th International Conference on Natural Language and Speech Processing (ICNLSP 2026)*, Trento, Italy.

Najet Hadj Mohamed, Malak Rassem, Lifeng Han, and Goran Nenadic. 2023. AlphaMWE-Arabic: Arabic edition of multilingual parallel corpora with multiword expression annotations. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 448–457, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Aaron Li-Feng Han, Derek F Wong, Lidia S Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013. Language-independent model for machine translation evaluation with reinforced factors. In *Proceedings of Machine Translation Summit XIV: Posters*.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. AlphaMWE: Construction of multilingual parallel corpora with MWE annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Lifeng Han, Najet Hadj Mohamed, Malak Rassem, Gareth JF Jones, Alan F Smeaton, and Goran Nenadic. 2026. Towards a resource for multilingual lexicons: an mt assisted and human-in-the-loop multilingual parallel corpus with multi-word expression annotation. *Language Resources and Evaluation*, 60(2):33.

Frances Adriana Laureano De Leon, Asim Abbas, Harish Tayyar Madabushi, and Mark Lee. 2025. Evaluating large language models on multiword expressions in multilingual and code-switched contexts. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 644–653, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jiahui Liang and Lifeng Han. 2026. MetaHOPE: A metaphor-oriented evaluation framework for analysing mt and llm translation errors. In *Proceedings of the 9th International Conference on Natural Language and Speech Processing (ICNLSP 2026)*. Accepted for publication. Preprint available at arXiv:2607.00848.

Linfeng Liu, Saptarshi Ghosh, and Tianyu Jiang. 2026. Evaluating the impact of verbal multiword expressions on machine translation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15291–15319, San Diego, California, United States. Association for Computational Linguistics.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity:

How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

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

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

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

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Language Identifiers

Language identifiers used in the paper are described in Table 4.

B English–Polish Cross-Annotator Agreement

Table 5 summarises the comparison between the two independent English–Polish annotations. Beyond the aggregate figures, two structural sources of disagreement are worth recording.

Category overlap. A recurring pattern is that both annotators identify the *same* defect but file it under different categories. In aa-134, the untranslated "Kong" for Hong Kong was recorded as a proper name error (PRN) by one annotator and as a missing source adaptation (RAM) by the other; in aa-104, one annotator recorded terminology (TRM) and the other mistranslation (MIS) for the same expansion of the source acronym. Per-category detection κ ranges from -0.02 (UGR) to 0.40 (TRM, PRN), and RAM was never used by one of the two annotators at all. For HOPE-style schemes we therefore recommend reporting agreement both per category and after collapsing categories that overlap by design, and supplying worked examples for the categories that are otherwise left unused.

Annotation procedure. The two annotators worked in different ways. One annotated against a Polish reference translation revised for this study, which was produced by a different person than the annotator. The other did not use the reference column, preferring to translate each source segment independently before scoring. Part of the disagreement reported above therefore reflects a difference in procedure rather than in judgement. The effect is visible in aa-104, where the original reference left the source acronym "OAS" untranslated and all three systems expanded it to *OPA* (*Organizacja Państw Amerykańskich*), whereas the passage (Amin speaking at Port Louis in July 1976) concerns the Organisation of African Unity. The annotator who worked from her own translation recorded this as a critical terminology error on all three systems, while the reference-based annotation did not flag it. A related factor is that the severity scale was not calibrated between the annotators in advance, so the same scale could be read differently across error types. We therefore recommend that IAA studies of HOPE-style schemes record the annotation procedure and a shared severity key alongside the taxonomy itself.

C Additional English–Polish Translation Examples

The following examples illustrate representative linguistic phenomena identified during the HOPE-based human evaluation of the English–Polish test suite. The source sentence, human reference, and outputs of the three selected systems are shown for each case. The discussion summarizes the Polish annotator's qualitative analysis.

Example 1: Literal translation of an idiomatic MWE (MIS) – aa-042

Source. But it did not give me the time of day.

Reference. Ale nie zwrócił na mnie uwagi.

DeepSeek V4 Pro. Ale nie dało mi pory dnia.

TRIVE. Ale nawet nie zaszczyciło mnie spojrzeniem.

Dubformer. Ale on nawet nie raczył na mnie spojrzeć.

Discussion. The DeepSeek V4 Pro output translates the idiom *give me the time of day* literally as *dało mi pory dnia*, which does not convey the intended idiomatic meaning in Polish. In contrast, TRIVE and Dubformer render the expression idiomatically. This example illustrates a severe MWE-related mistranslation, where lexical correspondence is preserved but the intended meaning is lost. It is worth noting that the reference translation differs from the correct machine translations in terms of MWE, but conveys the same meaning.

Example 2: Verbal aspect in Polish (UGR) – aa-012

Source. When items in a field are hidden by conditional filtering, a funnel appears to the left of the drop-down arrow Field arrow.

Reference. Gdy elementy w polu są ukryte przez filtrowanie warunkowe, na lewo od strzałki rozwijanej strzałki Pola pojawia się lejek.

DeepSeek V4 Pro. Gdy elementy w polu są ukrywane przez filtrowanie warunkowe, po lewej stronie strzałki rozwijanego pola (strzałki pola) pojawia się lejek.

TRIVE. Gdy elementy w polu są ukryte przez filtrowanie warunkowe, po lewej stronie strzałki rozwijania pola pojawia się lejek.

Table 4: Language identifiers used throughout this paper and in the WMT 2026 Test Suites Shared Task.

Pair	Code	Language	Description
en→ar	ar	Modern Standard Arabic (MSA)	Standard written Arabic
en→arz	arz	Egyptian Arabic	Egyptian Arabic dialect
en→aeb	aeb	Tunisian Arabic	Tunisian Arabic dialect
en→de	de	German	Standard German
en→pl	pl	Polish	Polish
en→zh	zh	Chinese	Simplified Chinese

Table 5: Cross-annotator comparison for English–Polish over 150 segments × 3 MT systems (450 MT outputs). SEGS totals are summed error penalties; lower is better. A1 and A2 denote the two independent annotators.

	A1	A2
SEGS *totals*		
DeepSeek V4 Pro	116	136
TRIVE	68	123
Dubformer	89	133
best–worst spread	48	13
Annotation profile		
error instances	180	137
mean severity	1.52	2.86
instances at 8 or 16	2	11
Agreement		
identical error profiles	265 / 450 (58.9%)	
cell-level agreement	92.9%	
detection κ	0.32	
SEGS correlation	$r = 0.70$, $\rho = 0.38$	
system ranking	identical	

Dubformer. Gdy elementy w polu są ukryte przez filtrowanie warunkowe, na lewo od strzałki rozwijania (strzałki pola) pojawia się lejek.

Discussion. The DeepSeek V4 Pro translation uses the imperfective form *ukrywane* ('being hidden repeatedly/in progress'), whereas the Polish context calls for a resultative or perfective reading, such as *są ukryte*. This reflects a structural difference between English and Polish: Polish verbal aspect is strongly lexicalized and must often be resolved explicitly during translation. The error does not prevent comprehension, but it produces a less natural grammatical interpretation.

Example 3: Multiple acceptable context-dependent translations – aa-044

Source. Uncle Vernon might still have been able to make his deal – if it hadn't been for the owl.

Reference. Wujek Vernon nadal mógłby zawrzeć umowę – gdyby nie sowa.

DeepSeek V4 Pro. Wuj Vernon mógłby jeszcze dopiąć swego – gdyby nie ta sowa.

TRIVE. Wuj Vernon być może zdołałby jeszcze ubić interes – gdyby nie sowa.

Dubformer. Wuj Vernon być może wciąż byłby w stanie dobić targu – gdyby nie sowa.

Discussion. All three system outputs are plausible Polish translations, but *dopiąć swego*, *ubić interes*, and *dobić targu* introduce slightly different semantics. The example illustrates a limitation of sentence-level evaluation: without broader discourse context, more than one target-language realization may be acceptable, and the reference should not be treated as the only valid formulation.

Example 4: Collocational naturalness (STL) – aa-106

Source. We step into the street and my friend David Shahar, whose chest is large, takes a deep breath and advises me to do the same.

Reference. Wchodzimy na ulicę i mój przyjaciel David Shahar, którego klatka piersiowa jest duża, bierze głęboki oddech i radzi mi zrobić to samo.

DeepSeek V4 Pro. Wychodzimy na ulicę, a mój przyjaciel David Shahar, którego klatka piersiowa jest duża, bierze głęboki oddech i radzi mi zrobić to samo.

TRIVE. Wychodzimy na ulicę i mój przyjaciel David Shahar, który ma szeroką klatkę piersiową, bierze głęboki oddech i radzi mi zrobić to samo.

Dubformer. Wychodzimy na ulicę, a mój przyjaciel David Shahar, który ma postawną klatkę piersiową, bierze głęboki oddech i radzi mi zro-

bić to samo.

Discussion. Dubformer uses *postawną klatkę piersiową*. In Polish, *postawny* is normally used to describe a person's build rather than a body part, making the phrase collocationally unnatural. The translation remains understandable, so the issue is primarily stylistic rather than a major semantic failure.

Example 5: Domain terminology versus literal wording (TRM) – aa-137

Source. The fiscal advantages for clean vehicles could then be introduced some two years before the new standards take effect.

Reference. Korzyści podatkowe dla ekologicznie czystych pojazdów mogłyby zostać wprowadzone na około dwa lata przed wejściem w życie nowych norm.

DeepSeek V4 Pro. Ulgi fiskalne dla pojazdów ekologicznych mogłyby wówczas zostać wprowadzone na około dwa lata przed wejściem w życie nowych norm.

TRIVE. Zachęty fiskalne dla czystych pojazdów mogłyby wówczas zostać wprowadzone na około dwa lata przed wejściem w życie nowych norm.

Dubformer. Korzyści fiskalne dla czystych pojazdów mogłyby wówczas zostać wprowadzone na około dwa lata przed wejściem w życie nowych norm.

Discussion. The literal phrase *czyste pojazdy* is understandable, but Polish domain usage normally favors established expressions such as *pojazdy ekologiczne* or *pojazdy niskoemisyjne*. DeepSeek V4 Pro therefore provides the most idiomatic terminology in this example, while the other outputs illustrate how literal lexical transfer can reduce terminological adequacy even when the overall meaning remains recoverable.

Example 6: Literal target-language construction (STL) – aa-133

Source. Greece, Italy, France, Spain and Austria, have understood that something has to be done and now have a presence in Albania.

Reference. Grecja, Włochy, Francja, Hiszpania i Austria zrozumiały, że trzeba coś zrobić i teraz są obecne w Albanii.

DeepSeek V4 Pro. Grecja, Włochy, Francja, Hiszpania i Austria zrozumiały, że trzeba coś zrobić, i teraz mają swoją obecność w Albanii.

TRIVE. Grecja, Włochy, Francja, Hiszpania i Austria zrozumiały, że coś trzeba zrobić, i są teraz obecne w Albanii.

Dubformer. Grecja, Włochy, Francja, Hiszpania i Austria zrozumiały, że trzeba coś zrobić i są teraz obecne w Albanii.

Discussion. DeepSeek V4 Pro transfers the English construction *have a presence* too directly as *mają swoją obecność*. Although the meaning is transparent, this construction is not idiomatic Polish. The reference, TRIVE, and Dubformer instead use *są obecne*, which is considerably more natural. The example demonstrates how literal translation can produce stylistic errors without necessarily causing a loss of propositional meaning.

Example 7: Scope ambiguity in MWE translation (MIS/STL) – aa-108

Source. In 1947 Copeland had been sent to Damascus ("by whom is not stated," Kedourie says) "to make unofficial contact" with Syrian leaders and "to probe for means of persuading them, on their own, to liberalize their political system."

Reference. W 1947 r. Copeland został wysłany do Damaszku ("przez kogo nie jest powiedziane", mówi Kedourie), aby "nawiązać nieoficjalny kontakt" z syryjskimi przywódcami i "poszukać sposobów przekonania ich, na własną rękę, do liberalizacji ich systemu politycznego".

DeepSeek V4 Pro. W 1947 roku Copeland został wysłany do Damaszku („przez kogo, nie podano", mówi Kedourie) „aby nawiązać nieoficjalny kontakt" z syryjskimi przywódcami oraz „aby zbadać sposoby nakłonienia ich, samodzielnie, do liberalizacji ich systemu politycznego".

TRIVE. W 1947 roku Copeland został wysłany do Damaszku („nie podano przez kogo", jak

twierdzi Kedourie), „aby nawiązać nieoficjalny kontakt” z syryjskimi przywódcami i „wybadać sposoby nakłonienia ich, by z własnej inicjatywy zliberalizowali swój system polityczny”.

Dubformer. W 1947 roku Copeland został wysłany do Damaszku („przez kogo, nie podano”, pisze Kedourie) „w celu nawiązania nieoficjalnych kontaktów” z syryjskimi przywódcami i „zbadania możliwości przekonania ich, aby sami zliberalizowali swój system polityczny”.

Discussion. The placement of DeepSeek V4 Pro's *samodzielnie* makes the scope of *on their own* difficult to recover and can suggest an unintended interpretation. TRIVE and Dubformer make the semantic relation explicit by using *z własnej inicjatywy* and *aby sami zliberalizowali*, respectively. This case shows that an MWE or modifier may be lexically translated while its syntactic scope is altered, potentially changing the interpretation of the whole sentence.

D English–German Cross-Annotator Agreement

Table 6 summarises the comparison between the two independent English–German annotations, computed exactly as for English–Polish.

Unused and disjointly used categories. Annotator A2 did not use the terminology (TRM) or proper-name (PRN) categories at any point, so the terminological analysis reported in Section 5.3 (including the *standard*/*Norm* distinction in aa-137) has no counterpart in the second annotation. More strikingly, both annotators made heavy use of the impact category, 87 times between them, yet applied it to the same MT output on only three occasions (per-category $\kappa = -0.04$); a further 63 style penalties appear in A2's annotation alone. The only category with substantial agreement is required-adaptation-missing (RAM, $\kappa = 0.57$), which both annotators applied to the uncorrected source defect in aa-134, where the source reads “Kong's civil service” for Hong Kong. These figures suggest that the divergence is driven less by differing severity thresholds than by differing readings of what the categories denote, and they reinforce the recommendation made for English–Polish in Appendix B: HOPE-style schemes need worked examples per category, and agreement should be reported per category rather than only in aggregate.

Table 6: Cross-annotator comparison for English–German over 150 segments × 3 MT systems (450 MT outputs). SEGS totals are summed error penalties; lower is better. A1 and A2 denote the two independent annotators.

	A1	**A2**
SEGS *totals*		
VoxNexus_V1	42	77
DeepSeek V4 Pro	145	131
TRIVE	140	97
best–worst spread	103	54
Annotation profile		
error instances	126	162
mean severity	2.60	1.88
instances at 8 or 16	12	3
Agreement		
identical error profiles	252 / 450 (56.0%)	
cell-level agreement	92.7%	
both found an error	38	
flagged by one only	162 (49 A1, 113 A2)	
detection κ	0.10	
SEGS correlation	$r = 0.15$, $\rho = 0.12$	
system ranking	identical	

Effect on discrimination. The two annotations separate the systems by different margins. Under A1, TRIVE and DeepSeek V4 Pro differ by 5 penalty points out of roughly 140, which is well within annotation noise, while A2 separates them by 34. Conversely, A1 places VoxNexus_V1 far ahead of both competitors (42 against 140 and 145), whereas A2's margin is narrower (77 against 97 and 131). The ordering is preserved in both cases, but the confidence one can place in the second-versus-third comparison depends on which annotation is consulted.

E Additional English–Chinese Discussion

The **annotation tasks require domain-expertise knowledge** for better understanding of the terminologies and expressions. As one example, from sent-id: aa-76:

Source. A query that uses wildcard characters in a criteria expression can produce different results under each query mode.
Reference. 在标准表达式中使用通配符的查询，在不同的查询模式下会产生不同的结果。
Tower-9B. 在每个查询模式下，使用通配符的条件表达式可能会产生不同的结果。

It is assigned 8 (major error) by our annotator (literature and translation studies background), for MIS error. However, it actually carried a similar meaning to the reference translation, but they are in very different word orders, which might have an impact to our annotators who are using/looking at the offered reference from the original AlphaMWE corpus. After discussion, we changed it to unnatural/proof-reading error with severity level 2.

Error-severity agreements. It is not an issue to agree on "whether a sentence has an error", but a problem on the error severity levels, similar findings from Liang and Han (2026). E.g., the sentence "Harry just caught sight of a pair of bright brown eyes staring at him before it closed with a snap." (Sent-id: 054). We have:

Reference. 哈利刚看到一双明亮的棕色眼睛盯着他，然后就啪的一声闭上了。
SCIR-TG.哈利只来得及瞥见一双明亮的棕色眼睛正盯着他，然后门就啪的一声关上了。
GoogleMT. (v) 哈利瞥见一双明亮的棕色眼睛正盯着他，然后那双眼睛啪嗒一声闭上了。
Tower-9B. 哈利刚瞥见一对明亮的棕色眼睛在盯着他，然后门就砰地一声关上了。

Both SCIR-TG and Tower-9B translated "door closed" while GoogleMT made a correct translation of "eyes closed". Our initial annotation gave a score of 2 for medium-level; however, after discussion, we think it should be "4 or 8" for major or severe errors.

F Detailed Automatic Scores Per System Per Language Pair

The detailed scores from SacreBLEU, ChrF++, and BERTscore on submitted systems per language pair are listed in Table 7, 8, and 9.

Their corresponding rankings are displayed in Table 10, 11, and 12.

The top 3 systems per language pair are listed in Figure 6 with their corresponding original metric scores.

Table 7: SacreBLEU scores for the submitted AlphaMWE test suites. Higher scores are better.

System	en→aeb	en→ar	en→arz	en→de	en→pl	en→zh
CUNI-AR	8.46	30.02	9.03	–	–	–
CUNI-EdUKate	–	–	–	40.52	–	–
CUNI-MH-v3	–	–	–	41.38	–	–
Cohere CAT+	3.04	23.97	3.89	39.16	29.79	32.88
Command A+	2.69	21.88	2.35	36.03	26.89	40.04
DeepSeek V4 Pro	7.35	28.75	9.90	50.00	**40.69**	37.94
Dubformer	7.01	22.80	8.72	48.30	40.52	43.89
GLM 4 - 9B	0.91	17.42	2.22	36.80	31.41	43.32
GPT 5.5	6.95	26.84	**10.32**	48.37	38.20	42.81
GPT OSS 120B	3.74	23.14	6.71	40.08	34.11	39.53
GPT OSS 20B	4.19	19.21	4.71	31.24	27.43	35.14
Gemini 3.1 Pro	1.05	5.48	1.31	7.09	6.57	7.35
Gemma 4 - 31B	5.94	27.73	9.83	47.97	36.67	42.16
Gemma 4 - 4B	7.01	31.64	2.72	44.56	37.42	44.23
GoogleTranslate	7.79	**36.46**	2.42	47.68	38.59	**47.23**
HW-TSC	–	26.84	–	–	–	35.11
Lumen	–	28.26	8.84	46.77	36.83	40.37
Ministral 3 - 14B	2.05	21.61	2.08	32.74	29.14	37.62
Mistral Medium 3.5	6.62	28.70	4.87	44.67	36.83	43.24
Qingqiu-MT-9B	6.41	20.76	8.86	46.31	28.10	40.33
Qwen 3.5 - 9b	3.90	19.13	1.68	31.58	22.99	30.87
Qwen 3.6 - 27b	4.38	22.08	4.01	38.52	27.76	36.94
SCIR-TG-MT	–	–	–	–	–	45.48
SRPOL	–	–	4.11	39.49	–	35.90
SalamandraTA	6.13	22.63	5.03	46.54	31.05	43.06
TRIVE	**8.66**	30.92	9.12	49.01	40.51	40.34
Tiny Aya Global	–	21.88	0.07	29.25	20.55	40.55
Tower 9B	0.73	8.85	0.53	44.67	33.05	44.97
UvA-MT	–	–	9.83	43.70	–	36.11
VoxNexus_V1	–	23.15	9.41	**50.04**	–	–
Wayfinder	–	26.52	10.27	47.43	37.44	39.63

Language Pair	Final Rank	System	Average Rank	Rank StdDev	SacreBLEU	SacreBLEU Rank	chrF++	chrF++ Rank	BERTScore F1	BERTScore Rank	Metrics Available
en_aeb	1	TRIVE	1.000	0.000	8.6558	1	36.8392	1	0.785693	1	3
en_aeb	2	CUNI-AR	2.000	0.000	8.4596	2	34.5010	2	0.777942	2	3
en_aeb	3	DeepSeek V4 Pro	4.000	0.000	7.3467	4	32.7572	4	0.762406	4	3
en_ar	1	GoogleTranslate	1.000	0.000	36.4647	1	65.2388	1	0.906980	1	3
en_ar	2	Gemma 4 - 4B	2.000	0.000	31.6389	2	61.3956	2	0.896778	2	3
en_ar	3	TRIVE	3.000	0.000	30.9177	3	60.3463	3	0.895420	3	3
en_arz	1	Wayfinder	2.000	0.816	10.2652	2	40.3347	1	0.784990	3	3
en_arz	2	Gemma 4 - 31B	3.667	1.886	9.8259	5	39.8351	5	0.788336	1	3
en_arz	3	GPT 5.5	4.000	3.559	10.3157	1	39.9765	2	0.779032	9	3
en_de	1	VoxNexus_V1	1.000	0.000	50.0396	1	69.9207	1	0.908204	1	3
en_de	2	DeepSeek V4 Pro	2.000	0.000	49.9973	2	69.8349	2	0.906092	2	3
en_de	3	TRIVE	3.000	0.000	49.0054	3	69.4334	3	0.905597	3	3
en_pl	1	DeepSeek V4 Pro	1.000	0.000	40.6918	1	63.3976	1	0.890576	1	3
en_pl	2	TRIVE	2.333	0.471	40.5125	3	62.8851	2	0.888687	2	3
en_pl	3	Dubformer	2.667	0.471	40.5208	2	62.7774	3	0.887188	3	3
en_zh	1	SCIR-TG-MT	1.667	0.471	45.4849	2	30.8858	2	0.868074	1	3
en_zh	1	GoogleTranslate	1.667	0.943	47.2283	1	32.1595	1	0.867518	3	3
en_zh	3	Tower 9B	3.000	0.816	44.9658	3	30.5066	4	0.867755	2	3

Figure 6: Top 3 MT systems selected for human evaluation: ranking and original scores

Table 8: chrF++ scores for the submitted AlphaMWE test suites. Higher scores are better.

System	**en→aeb**	**en→ar**	**en→arz**	**en→de**	**en→pl**	**en→zh**
CUNI-AR	34.50	59.72	39.58	–	–	–
CUNI-EdUKate	–	–	–	63.07	–	–
CUNI-MH-v3	–	–	–	63.41	–	–
Cohere CAT+	13.75	54.41	28.32	62.97	56.02	23.80
Command A+	17.68	52.65	23.78	60.54	53.47	27.17
DeepSeek V4 Pro	32.76	58.67	39.36	69.83	**63.40**	27.93
Dubformer	31.07	54.43	38.05	68.59	62.78	30.60
GLM 4 - 9B	0.85	47.84	22.94	60.60	55.24	29.19
GPT 5.5	33.63	57.29	39.98	69.13	61.84	29.96
GPT OSS 120B	26.59	54.23	33.56	63.33	58.28	26.35
GPT OSS 20B	26.02	48.94	30.34	56.88	52.80	23.96
Gemini 3.1 Pro	16.90	35.85	19.29	38.70	37.24	12.67
Gemma 4 - 31B	27.51	58.37	39.84	68.85	60.16	29.64
Gemma 4 - 4B	30.36	61.40	24.77	66.44	60.26	29.07
GoogleTranslate	31.60	**65.24**	24.92	68.62	60.97	**32.16**
HW-TSC	–	57.19	–	–	–	24.76
Lumen	–	57.68	38.69	67.56	60.38	28.06
Ministral 3 - 14B	6.11	50.15	22.15	59.97	55.71	25.35
Mistral Medium 3.5	28.04	57.86	31.06	66.28	60.19	28.73
Qingqiu-MT-9B	30.11	49.63	39.90	67.56	53.36	28.51
Qwen 3.5 - 9b	23.61	49.21	23.11	57.57	50.50	21.76
Qwen 3.6 - 27b	27.23	52.42	29.01	62.59	54.08	25.92
SCIR-TG-MT	–	–	–	–	–	30.89
SRPOL	–	–	28.29	63.63	–	26.28
SalamandraTA	28.99	50.44	29.83	67.41	54.05	28.59
TRIVE	**36.84**	60.35	39.97	69.43	62.89	28.28
Tiny Aya Global	–	52.43	0.41	57.76	54.02	28.16
Tower 9B	0.74	30.27	9.67	66.66	57.80	30.51
UvA-MT	–	–	38.62	65.41	–	25.73
VoxNexus_V1	–	54.35	39.14	**69.92**	–	–
Wayfinder	–	56.74	**40.33**	68.13	60.85	27.50

Table 9: BERTScore F1 scores for the submitted AlphaMWE test suites. Higher scores are better.

System	**en→aeb**	**en→ar**	**en→arz**	**en→de**	**en→pl**	**en→zh**
CUNI-AR	0.7779	0.8935	0.7868	–	–	–
CUNI-EdUKate	–	–	–	0.8877	–	–
CUNI-MH-v3	–	–	–	0.8906	–	–
Cohere CAT+	0.6850	0.8756	0.7349	0.8867	0.8598	0.8382
Command A+	0.6784	0.8683	0.7049	0.8757	0.8561	0.8529
DeepSeek V4 Pro	0.7624	0.8914	0.7810	0.9061	**0.8906**	0.8587
Dubformer	0.7590	0.8769	0.7745	0.9053	0.8872	0.8668
GLM 4 - 9B	0.6414	0.8579	0.7056	0.8799	0.8623	0.8623
GPT 5.5	0.7705	0.8852	0.7790	0.9036	0.8824	0.8632
GPT OSS 120B	0.7303	0.8721	0.7565	0.8841	0.8672	0.8495
GPT OSS 20B	0.7181	0.8476	0.7439	0.8587	0.8486	0.8381
Gemini 3.1 Pro	0.5947	0.7077	0.6181	0.6734	0.6927	0.6726
Gemma 4 - 31B	0.7474	0.8894	**0.7883**	0.9055	0.8780	0.8613
Gemma 4 - 4B	0.7423	0.8968	0.7141	0.8993	0.8800	0.8634
GoogleTranslate	0.7366	**0.9070**	0.7109	0.9013	0.8809	0.8675
HW-TSC	–	0.8819	–	–	–	0.8330
Lumen	–	0.8857	0.7804	0.9009	0.8776	0.8543
Ministral 3 - 14B	0.6332	0.8460	0.6886	0.8731	0.8629	0.8453
Mistral Medium 3.5	0.7437	0.8873	0.7567	0.8976	0.8816	0.8630
Qingqiu-MT-9B	0.7543	0.8562	0.7845	0.9014	0.8548	0.8557
Qwen 3.5 - 9b	0.7148	0.8612	0.7069	0.8701	0.8444	0.8305
Qwen 3.6 - 27b	0.7242	0.8665	0.7349	0.8857	0.8546	0.8506
SCIR-TG-MT	–	–	–	–	–	**0.8681**
SRPOL	–	–	0.7307	0.8874	–	0.8492
SalamandraTA	0.7460	0.8506	0.7507	0.9010	0.8613	0.8618
TRIVE	**0.7857**	0.8954	0.7840	0.9056	0.8887	0.8559
Tiny Aya Global	–	0.8645	0.5926	0.8825	0.8664	0.8497
Tower 9B	0.6379	0.7914	0.6440	0.8989	0.8685	0.8678
UvA-MT	–	–	0.7766	0.8944	–	0.8432
VoxNexus_V1	–	0.8760	0.7807	**0.9082**	–	–
Wayfinder	–	0.8830	0.7850	0.9021	0.8803	0.8554

Table 10: System rankings based on SacreBLEU for the AlphaMWE test suites. Rank 1 denotes the best-performing system.

System	en→aeb	en→ar	en→arz	en→de	en→pl	en→zh
CUNI-AR	2	4	8	–	–	–
CUNI-EdUKate	–	–	–	17	–	–
CUNI-MH-v3	–	–	–	16	–	–
Cohere CAT+	16	12	18	20	15	25
Command A+	17	19	21	23	20	15
DeepSeek V4 Pro	4	5	3	2	**1**	18
Dubformer	6	15	11	5	2	5
GLM 4 - 9B	20	24	22	22	13	6
GPT 5.5	7	9	**1**	4	5	9
GPT OSS 120B	15	14	12	18	11	17
GPT OSS 20B	13	22	15	26	19	23
Gemini 3.1 Pro	19	26	25	28	23	27
Gemma 4 - 31B	11	8	5	6	10	10
Gemma 4 - 4B	5	2	19	14	7	4
GoogleTranslate	3	**1**	20	7	4	**1**
HW-TSC	–	10	–	–	–	24
Lumen	–	7	10	9	8	12
Ministral 3 - 14B	18	20	23	24	16	19
Mistral Medium 3.5	8	6	14	12	9	7
Qingqiu-MT-9B	9	21	9	11	17	14
Qwen 3.5 - 9b	14	23	24	25	21	26
Qwen 3.6 - 27b	12	17	17	21	18	20
SCIR-TG-MT	–	–	–	–	–	2
SRPOL	–	–	16	19	–	22
SalamandraTA	10	16	13	10	14	8
TRIVE	**1**	3	7	3	3	13
Tiny Aya Global	–	18	27	27	22	11
Tower 9B	21	25	26	13	12	3
UvA-MT	–	–	4	15	–	21
VoxNexus_V1	–	13	6	**1**	–	–
Wayfinder	–	11	2	8	6	16

Table 11: System rankings based on chrF++ for the AlphaMWE test suites. Rank 1 denotes the best-performing system.

System	**en→aeb**	**en→ar**	**en→arz**	**en→de**	**en→pl**	**en→zh**
CUNI-AR	2	4	6	–	–	–
CUNI-EdUKate	–	–	–	19	–	–
CUNI-MH-v3	–	–	–	17	–	–
Cohere CAT+	18	13	17	20	13	25
Command A+	16	16	21	23	19	17
DeepSeek V4 Pro	4	5	7	2	**1**	15
Dubformer	6	12	11	7	3	3
GLM 4 - 9B	20	24	23	22	15	7
GPT 5.5	3	9	2	4	4	5
GPT OSS 120B	13	15	12	18	11	18
GPT OSS 20B	14	23	14	27	21	24
Gemini 3.1 Pro	17	25	25	28	23	27
Gemma 4 - 31B	11	6	5	5	10	6
Gemma 4 - 4B	7	2	20	13	8	8
GoogleTranslate	5	**1**	19	6	5	**1**
HW-TSC	–	10	–	–	–	23
Lumen	–	8	9	10	7	14
Ministral 3 - 14B	19	20	24	24	14	22
Mistral Medium 3.5	10	7	13	14	9	9
Qingqiu-MT-9B	8	21	4	9	20	11
Qwen 3.5 - 9b	15	22	22	26	22	26
Qwen 3.6 - 27b	12	18	16	21	16	20
SCIR-TG-MT	–	–	–	–	–	2
SRPOL	–	–	18	16	–	19
SalamandraTA	9	19	15	11	17	10
TRIVE	**1**	3	3	3	2	12
Tiny Aya Global	–	17	27	25	18	13
Tower 9B	21	26	26	12	12	4
UvA-MT	–	–	10	15	–	21
VoxNexus_V1	–	14	8	**1**	–	–
Wayfinder	–	11	**1**	8	6	16

Table 12: System rankings based on BERTScore F1 for the AlphaMWE test suites. Rank 1 denotes the best-performing system.

System	en→aeb	en→ar	en→arz	en→de	en→pl	en→zh
CUNI-AR	2	4	2	–	–	–
CUNI-EdUKate	–	–	–	17	–	–
CUNI-MH-v3	–	–	–	16	–	–
Cohere CAT+	16	14	17	19	17	23
Command A+	17	16	23	24	18	16
DeepSeek V4 Pro	4	5	6	2	**1**	11
Dubformer	5	12	11	5	3	4
GLM 4 - 9B	18	20	22	23	15	8
GPT 5.5	3	9	9	6	4	6
GPT OSS 120B	12	15	13	21	12	19
GPT OSS 20B	14	23	15	27	21	24
Gemini 3.1 Pro	21	26	26	28	23	27
Gemma 4 - 31B	7	6	**1**	4	9	10
Gemma 4 - 4B	10	2	19	12	8	5
GoogleTranslate	11	**1**	20	9	6	3
HW-TSC	–	11	–	–	–	25
Lumen	–	8	8	11	10	15
Ministral 3 - 14B	20	24	24	25	14	21
Mistral Medium 3.5	9	7	12	14	5	7
Qingqiu-MT-9B	6	21	4	8	19	13
Qwen 3.5 - 9b	15	19	21	26	22	26
Qwen 3.6 - 27b	13	17	16	20	20	17
SCIR-TG-MT	–	–	–	–	–	**1**
SRPOL	–	–	18	18	–	20
SalamandraTA	8	22	14	10	16	9
TRIVE	**1**	3	5	3	2	12
Tiny Aya Global	–	18	27	22	13	18
Tower 9B	19	25	25	13	11	2
UvA-MT	–	–	10	15	–	22
VoxNexus_V1	–	13	7	**1**	–	–
Wayfinder	–	10	3	7	7	14