

# Finding the Gaps in LLM Translation with *CoST*

Saumitra Yadav, Ananya Mukherjee, Manish Shrivastava

MT-NLP Lab, LTRC, KCIS, IIIT Hyderabad, India

saumitra.yadav@research.iiit.ac.in

ananya.mukherjee@research.iiit.ac.in

m.shrivastava@iiit.ac.in

## Abstract

Large language models (LLMs) have advanced machine translation substantially, but generic benchmarks often fail to expose where these systems struggle with complex linguistic structures and stylistic nuance. We use CoST (Complex Structures Test), a challenge suite of 1,947 English-Hindi sentences spanning eight genres, autobiography, conversation, legal, mixed, narration, play, poetry, and technical writing, previously introduced for WMT24, to evaluate 22 machine translation systems submitted to the WMT26 General Translation Shared Task for English-Hindi with reference-free, reference-based, and manual evaluation. Our results reveal a consistent genre-level pattern: systems perform well on narrative text such as autobiography and play, but degrade sharply on poetry, legal, and conversational text. We further identify a distinct failure mode in poetry translation, where systems reproduce memorized canonical Hindi verse instead of translating the given English text, and manual analysis surfaces recurring issues in named-entity handling, wrong-language output, and lexical choice, several of which persist from our prior CoST evaluation on WMT24 submissions. Our test suite is available at <https://github.com/deciphyre/CoST-WMT-26-Test-Suite-Task>.

## 1 Introduction

Neural Machine Translation (NMT) has seen substantial progress in recent years, achieving impressive quality that benefits many everyday applications. The advent of large language models (LLMs) has further enhanced translation capabilities. However, despite these advancements, there remain challenges that generic evaluation methods often fail to address. While traditional evaluations using random text samples might show overall success, they may not reveal subtle issues where MT systems struggle, such as handling complex linguistic structures, idiomatic expressions, and di-

verse text types like conversations, poetry, legal documents, and technical writing. These flaws can be obscured by average performance metrics or overlooked entirely. A more systematic method for identifying linguistic issues in translation outputs involves using test suites or challenge sets to evaluate the system’s performance on specific tasks. (Manakhimova et al., 2023). Test suites offer a standardized approach to evaluating MT systems, revealing strengths and weaknesses in handling complex text types.

In this context, we present an evaluation of state-of-the-art machine translation systems across various categories using CoST (Complex Structures Test), a test suite we previously introduced for WMT24 (Mukherjee et al., 2024). These evaluations were conducted as part of the theme “Mind the Gap: Exposing LLM Translation Blind Spots” for the 11th Conference on Machine Translation (WMT26). The test suite was used to evaluate systems submitted for the English-Hindi language pair.

CoST comprises sentences from eight categories across 16 sources, curated to assess LLM translation capabilities beyond the typical news or generic domains, which often focus on reporting or narrative writing styles. Details of the test suite are provided in Section 2.

We perform reference-free and reference-based evaluations of the Hindi translations of this test suite, produced by 22 different machine translation (MT) systems submitted to the General Translation Task at WMT26. For reference-free evaluation, we employ COMET-Kiwi (Rei et al., 2022), COMTAIL-QE (Ahsan et al., 2025), while BLEU (Papineni et al., 2002), chrF++ (Popović, 2015, 2017), MEE4 (Mukherjee et al., 2020; Mukherjee and Shrivastava, 2023), BERTScore (Zhang\* et al., 2020), COMTAIL-DA (Ahsan et al., 2025), and COMET (Rei et al., 2020) are used for reference-based evaluation. We also give subjective analy-

sis of some of the translation examples. Beyond this year’s evaluation, we compare our findings against our prior CoST evaluation of WMT24 submissions to examine which translation weaknesses persist across two years of substantially different systems, and which appear specific to this year’s models. Details of our evaluation experiments are discussed in Section 3, and our analysis, including this year-over-year comparison, is presented in Section 4.

Rather than introducing a new benchmark, we use the fixed CoST suite to longitudinally evaluate MT systems across WMT24 and WMT26, examining persistent and emerging failure modes. We also analyze metric disagreement and recurring errors, including named entities, wrong-language outputs, lexical substitutions, and canonical-verse reproduction in poetry.

## 2 CoST:Complex Structure Testsuite

Table 1 depicts the dataset categories and the distribution within our test suite. The “Original” column presents the initial count of selected sentences for each category, as gathered from the datasets. The last column, “CoST,” displays the final count of sentences included in the test suite. Our test suite is designed to evaluate translations across

- Multiple Writing Style: Prose, Conversation, Autobiography, Legal Writing, Literary Narrative and Technical Documents.
- Lexical Choice: As we are sampling test suites from various domains, there is a decent mixture of domain-specific words, e.g. Legal Text, Technical Text, etc.

In total, 1,947 English sentences were selected based on criteria such as sentence length, depth of dependency tree, combination of noun phrases, verb phrases, named entities, etc. Ensuring a test suite containing sentences with good representation from simple to complex structures.

## 3 Evaluation Strategy

To evaluate the performance of the 22 submitted MT systems, we performed automatic evaluations. And did subjective analysis on some of the translation outputs.

### 3.1 Automatic Evaluation

In automatic evaluation, we leveraged both reference-free and reference-based metrics.

| Category      | Dataset                     | Original    | CoST        |
|---------------|-----------------------------|-------------|-------------|
| poetry        | Kabir ke Dohe               | 11          | 9           |
|               | Amir Khusro                 | 9           | 9           |
| narration     | ShortStories                | 177         | 72          |
|               | Post Office                 | 440         | 10          |
|               | Glimpses of Bengal          | 101         | 64          |
|               | The Home and the World      | 236         | 183         |
|               | The gardener                | 277         | 27          |
|               | Abridged Merchant of Venice | 63          | 31          |
| legal         | Christmas Carole            | 923         | 308         |
|               | Legal Text                  | 2862        | 638         |
| mix           | IIT Bombay Jud              | 167         | 83          |
|               | IN22                        | 570         | 241         |
| conversation  | Friends                     | 77          | 53          |
| play          | King Of Dark Chamber        | 35          | 22          |
| autobiography | My Reminiscences            | 109         | 110         |
| Technical     | Technical Papers            | 185         | 87          |
| <b>Total</b>  |                             | <b>6242</b> | <b>1947</b> |

Table 1: Data Statistics of CoST.

### 3.1.1 Reference-free Evaluation

We utilize COMETKIWI (Rei et al., 2022) and COMTAIL-QE (Ahsan et al., 2025) scores, which offer quality estimation scores derived from the source sentence and MT output.

### 3.1.2 Reference-based Evaluation

With the help of professional English-to-Hindi translators, we also provide one gold reference translation for each source sentence in the test suite. We evaluate the machine translation outputs against these references using BLEU (Papineni et al., 2002), chrF++ (Popović, 2015, 2017), MEE4<sup>1</sup> (Mukherjee et al., 2020; Mukherjee and Shrivastava, 2023), BERTScore (Zhang\* et al., 2020), COMTAIL-DA (Ahsan et al., 2025) and COMET (Rei et al., 2020).

## 3.2 Manual Evaluation

The manual analysis was done by professional native speakers. They were instructed to identify mis-translations and hallucinations and make note of other translation errors like wrong post positions to get more nuanced information regarding the system performances.

## 4 Results and Analysis

### 4.1 Quantitative Analysis

The results of the automatic evaluation are reported in Table 2. Ranks are shown in parentheses for each metric, where (1) is the highest

<sup>1</sup><https://www.kaggle.com/ananyacoder/mee4-metric-run>

| Model              | Referenceless |            | Reference-based |           |           |           |            |           |
|--------------------|---------------|------------|-----------------|-----------|-----------|-----------|------------|-----------|
|                    | COMET-Kiwi    | COMTAIL-QE | BLEU            | chrF++    | BERTScore | COMET     | COMTAIL-DA | MEE4      |
| GPT 5.5            | 82.0 (1)      | 84.3 (1)   | 40.9 (6)        | 66.2 (4)  | 87.9 (5)  | 79.8 (3)  | 79.7 (4)   | 83.6 (5)  |
| TRIVE              | 81.8 (2)      | 84.2 (4)   | 46.7 (2)        | 69.9 (1)  | 89.2 (2)  | 80.2 (2)  | 79.6 (5)   | 84.5 (2)  |
| Cohere CAT+        | 81.8 (3)      | 83.8 (6)   | 35.4 (9)        | 63.0 (10) | 86.8 (10) | 79.8 (4)  | 79.9 (2)   | 82.5 (8)  |
| GoogleTranslate    | 81.7 (4)      | 84.3 (2)   | 47.9 (1)        | 69.0 (2)  | 89.7 (1)  | 81.2 (1)  | 80.5 (1)   | 85.0 (1)  |
| Lumen              | 81.7 (5)      | 84.3 (3)   | 42.0 (4)        | 66.3 (3)  | 88.2 (4)  | 79.5 (5)  | 79.5 (6)   | 83.7 (4)  |
| DeepSeek V4 Pro    | 81.6 (6)      | 83.9 (5)   | 38.0 (8)        | 63.8 (7)  | 87.7 (8)  | 79.1 (7)  | 79.0 (7)   | 83.3 (7)  |
| Command A+         | 81.6 (7)      | 83.8 (7)   | 33.3 (12)       | 59.4 (11) | 86.3 (12) | 79.4 (6)  | 79.9 (3)   | 81.9 (11) |
| Dubformer          | 81.1 (8)      | 82.3 (8)   | 40.9 (5)        | 65.2 (5)  | 88.3 (3)  | 78.8 (8)  | 77.7 (10)  | 83.9 (3)  |
| Mistral Medium 3.5 | 80.4 (9)      | 82.2 (9)   | 33.4 (11)       | 59.3 (12) | 86.9 (9)  | 78.3 (9)  | 78.3 (8)   | 82.3 (9)  |
| GPT OSS 120B       | 80.3 (10)     | 81.5 (11)  | 24.0 (15)       | 52.9 (16) | 84.9 (13) | 77.3 (12) | 77.3 (11)  | 80.9 (12) |
| Tower 9B           | 80.2 (11)     | 82.0 (10)  | 38.7 (7)        | 63.2 (9)  | 87.8 (7)  | 78.0 (10) | 77.7 (9)   | 83.4 (6)  |
| Qwen 3.6 - 27b     | 79.4 (12)     | 80.3 (12)  | 23.9 (16)       | 53.6 (14) | 84.7 (14) | 76.8 (13) | 76.8 (13)  | 80.8 (13) |
| Gemma 4 - 31B      | 79.0 (13)     | 80.1 (13)  | 42.3 (3)        | 63.2 (8)  | 87.8 (6)  | 78.0 (11) | 77.1 (12)  | 82.2 (10) |
| Ministral 3 - 14B  | 78.4 (14)     | 78.7 (14)  | 24.7 (14)       | 53.0 (15) | 84.7 (15) | 75.7 (14) | 75.9 (14)  | 80.1 (15) |
| Gemma 4 - 4B       | 77.0 (15)     | 77.5 (15)  | 33.8 (10)       | 56.5 (13) | 86.8 (11) | 75.3 (15) | 74.5 (15)  | 80.5 (14) |
| Qingqiu-MT-9B      | 76.9 (16)     | 73.4 (18)  | 21.6 (17)       | 50.8 (17) | 83.2 (16) | 72.8 (17) | 70.2 (18)  | 79.7 (16) |
| Gemini 3.1 Pro     | 76.7 (17)     | 66.1 (20)  | 31.3 (13)       | 64.6 (6)  | 82.1 (18) | 72.9 (16) | 69.4 (19)  | 77.7 (18) |
| Qwen 3.5 - 9b      | 76.1 (18)     | 75.7 (16)  | 17.0 (19)       | 46.8 (18) | 82.4 (17) | 72.7 (18) | 72.8 (16)  | 78.1 (17) |
| GPT OSS 20B        | 75.2 (19)     | 74.3 (19)  | 18.1 (18)       | 46.1 (19) | 81.2 (19) | 71.6 (19) | 70.9 (17)  | 75.1 (19) |
| Tiny Aya Global    | 74.0 (20)     | 72.5 (19)  | 15.4 (20)       | 40.3 (20) | 80.0 (20) | 68.2 (20) | 60.6 (20)  | 74.5 (20) |
| SalamandraTA       | 65.0 (21)     | 54.0 (21)  | 9.4 (21)        | 27.8 (21) | 77.2 (21) | 59.7 (21) | 51.8 (21)  | 73.0 (21) |
| GLM 4 - 9B         | 58.9 (22)     | 42.9 (22)  | 5.3 (22)        | 23.8 (22) | 75.9 (22) | 50.3 (22) | 41.1 (22)  | 69.5 (22) |

Table 2: System-wise ranking based on reference-free and reference-based metrics.

rank. Table 2 reports system performance across eight automatic metrics, split into reference-free (COMET-Kiwi, COMTAIL-QE) and reference-based metrics, with the latter further separated into lexical (BLEU, chrF++) and embedding-based (BERTScore, COMET, COMTAIL-DA, MEE4) measures.

**Reference-Free Metrics.** On COMET-Kiwi, the top five systems, GPT 5.5, TRIVE, Cohere CAT+, GoogleTranslate, and Lumen, form a tight cluster within 0.3 points of each other (81.65 to 81.97), making it difficult to declare a single best system on this metric alone. COMTAIL-QE broadly agrees with this ordering at the top, with GPT 5.5, Lumen, GoogleTranslate, and TRIVE again occupying the top four positions, suggesting the two reference-free metrics are reasonably consistent for the strongest systems. The clearest disagreement is Gemini 3.1 Pro, which sits mid-table on COMET-Kiwi (76.70, rank 17) but drops sharply on COMTAIL-QE (66.06, rank 19), a gap of roughly 10 points that represents the largest divergence between the two reference-free metrics in the table. At the bottom, GLM 4 - 9B and SalamandraTA trail all other systems by a wide margin on both metrics, consistent with their weak performance throughout the rest of the table.

**Lexical Reference-Based Metrics.** BLEU and chrF++ show a much wider dynamic range than

any other metric, with BLEU spanning 5.27 to 47.92, which amplifies differences between systems that look far closer under embedding-based scoring. GoogleTranslate and TRIVE lead on both lexical metrics, consistent with their strong reference-free performance. Gemini 3.1 Pro is a clear outlier here: despite its comparatively weak reference-free scores, it ranks among the top six systems on both BLEU (31.26) and chrF++ (64.58), the largest positive swing between the lexical and reference-free rankings observed in the table. This hints that its outputs may align closely with reference wording even when quality-estimation metrics judge it less favorably.

**Embedding-Based Reference-Based Metrics.** BERTScore, COMET, COMTAIL-DA, and MEE4 tell a more unified story, with GoogleTranslate, TRIVE, GPT 5.5, Cohere CAT+, and Lumen consistently occupying the top ranks and GLM 4 - 9B and SalamandraTA anchoring the bottom across all four metrics. Unlike the lexical metrics, Gemini 3.1 Pro’s embedding-based scores (for example, COMET 72.90, COMTAIL-DA 69.39) fall back in line with its reference-free ranking rather than its lexical one, reinforcing that its strong BLEU and chrF++ performance is a lexical-overlap effect rather than a genuine quality advantage. Overall, the embedding-based and reference-free metrics agree far more closely with

one another than either agrees with the lexical metrics, suggesting lexical overlap and semantic or quality-estimation scores are capturing meaningfully different aspects of translation quality for at least a subset of systems.

After examining overall performance across all categories combined, we now turn to a category-wise breakdown to assess whether system rankings and the relationship between reference-free and reference-based metrics hold consistently across content types, or whether certain categories produce markedly different orderings. For each of the eight categories, autobiography, conversation, legal, mix, narration, play, poetry, and technical, we report system rankings on the reference-free metrics (COMET-Kiwi, COMTAIL-QE), followed by the reference-based metrics, further split into lexical (BLEU, chrF++) and embedding-based (BERTScore, COMET, COMTAIL, MEE4) measures, with full per-system, per-metric scores and ranks provided in Tables 3–10 in the Appendix A.

**Autobiography (Table 3).** TRIVE, GoogleTranslate, and GPT 5.5 lead COMET-Kiwi and COMTAIL-QE within 0.2 points of each other. GoogleTranslate posts the highest BLEU (39.89) and chrF++, roughly four BLEU points above TRIVE despite their close COMET-Kiwi scores, indicating closer overlap with reference wording. Embedding-based metrics again favor GoogleTranslate and TRIVE, with GLM 4 - 9B, SalamandraTA, and Tiny Aya Global anchoring the bottom across all eight metrics.

**Conversation (Table 4).** Cohere CAT+ tops COMET-Kiwi (81.62), ahead of GoogleTranslate and Command A+, a departure from its mid-table autobiography ranking. GoogleTranslate again leads BLEU (45.85) and chrF++, with Gemma 4 - 4B and Gemma 4 - 31B ranking higher on BLEU than COMET-Kiwi. GLM 4 - 9B and SalamandraTA remain weakest overall, with SalamandraTA posting the lowest COMET-Kiwi (56.95) of any system in any category.

**Legal (Table 5).** GPT 5.5 leads both reference-free metrics (COMET-Kiwi 80.40, COMTAIL-QE 79.30), narrowly ahead of TRIVE. On BLEU, GoogleTranslate and TRIVE swap to the top two, essentially tied (48.87 vs. 48.86), while GPT 5.5 drops to fifth, suggesting its output diverges more from reference phrasing despite strong quality-estimation scores. GLM 4 - 9B is a clear out-

lier here, scoring just 20.39 on COMTAIL-QE and 21.43 on COMTAIL-DA.

**Mix (Table 6).** TRIVE, GPT 5.5, and DeepSeek V4 Pro lead COMET-Kiwi within 0.05 points, the tightest race of any category. GoogleTranslate takes BLEU (52.83) and chrF++, with the lexical ranking here matching the reference-free ordering more closely than in Legal or Conversation. Embedding-based metrics again place GoogleTranslate, TRIVE, and GPT 5.5 at the top, with GLM 4 - 9B and SalamandraTA at the bottom.

**Narration (Table 7).** GoogleTranslate leads COMET-Kiwi, COMTAIL-QE, and BLEU simultaneously, the only category where one system tops all three. GPT 5.5 ranks second on COMET-Kiwi but falls to ninth on BLEU (30.60), its largest lexical-to-reference-free gap in any category. GPT OSS 20B joins GLM 4 - 9B and SalamandraTA near the bottom here, despite performing closer to mid-table elsewhere.

**Play (Table 8).** GoogleTranslate again leads COMET-Kiwi, COMTAIL-QE, and BLEU, with Cohere CAT+ and Dubformer close behind. GPT 5.5 sits further back (ninth on COMET-Kiwi, eleventh on BLEU) than in most other categories, suggesting dialogue-heavy text is comparatively harder for it. GLM 4 - 9B remains lowest throughout, and Tiny Aya Global is notably weak on BLEU (5.54, last place) despite a more moderate COMTAIL-QE ranking.

**Poetry (Table 9).** Poetry produces the most volatile rankings of any category. Command A+, GoogleTranslate, and Cohere CAT+ lead COMET-Kiwi, while TRIVE and GPT 5.5, top performers elsewhere, drop to eighteenth (69.61) and seventeenth (71.91). Gemini 3.1 Pro records its weakest result of any category here, last on COMET-Kiwi, BLEU, and COMTAIL-QE, with COMET-Kiwi as low as 58.35. GoogleTranslate again leads BLEU, while GoogleTranslate, Gemma 4 - 31B, and Cohere CAT+ top the embedding-based metrics.

**Technical (Table 10).** GoogleTranslate leads all reference-free and lexical metrics (COMET-Kiwi 84.01, BLEU 46.00) with Lumen and GPT 5.5 close behind on COMET-Kiwi. GPT 5.5 shows a lexical versus-reference-free gap, ranking third on COMET-Kiwi but eighth on BLEU (35.87), while Gemma 4 - 4B and Tower 9B rank unusually high on BLEU relative to their reference-free standing.

GoogleTranslate, Lumen, TRIVE and Tower 9B lead the embedding-based metrics, with SalamandraTA and GLM 4 - 9B trailing throughout. These results show a **consistent drop in performance across systems on the poetry, legal and conversation subsets**, while performing consistently well on autobiography, play and mixed (IN22) data.

#### Cross-category and genre-level trends.

GoogleTranslate is the most consistent top performer, leading COMET-Kiwi and BLEU simultaneously in narration, play, and technical, while GLM 4 - 9B and SalamandraTA remain the weakest systems across all eight categories, most severely in legal and conversation respectively. This points to a genre-level pattern, systems perform relatively well on narrative styles such as autobiography and play, but degrade on poetry, legal, and conversational text. One possible explanation is that models may have greater exposure to narrative prose than to the figurative language, turn-taking, and domain-specific phrasing characteristic of these genres; however, the training-data composition of the evaluated systems is not available to us, so this hypothesis cannot be directly verified. The poetry drop is not fully explained by genre difficulty alone: several source segments were English renderings of well-known Hindi *dohe* (e.g., Kabir’s couplets). Rather than translating the given wording, systems such as GPT 5.5 and Gemini 3.1 Pro appear to reproduce the canonical Hindi originals. This suggests a possible memorization-related failure mode, consistent with recent findings that LLMs can reproduce memorized content when prompted with semantically related material (Liu et al., 2026).

#### Inter-System Agreement (Difficulty Correlation)

Average pairwise Pearson correlation between each system’s per-sentence COMET score vector and all others’ (i.e., ”do the same sentences that are hard for me tend to be hard for others?”): Overall average pairwise correlation across all 22 systems:  $r = 0.47$ , moderate agreement on which sentences are intrinsically hard. Highest agreement: Lumen (0.60), GPT 5.5 (0.59), TRIVE (0.59), Cohere CAT+ (0.58), Command A+ (0.58): the strong, stable systems tend to fail (or succeed) on the same sentences as each other, consistent with those failures being genuinely sentence-level dif-

ficulty (long/complex source) rather than system-specific quirks. Lowest agreement: GLM 4-9B (0.24), SalamandraTA (0.25), Tiny Aya Global (0.26), GPT OSS 20B (0.26), Gemini 3.1 Pro (0.31), their failures are more idiosyncratic/system-specific (e.g., decoding pathologies, hallucination, script-mixing) rather than tracking universal sentence difficulty.

**Correlation between metrics** Figure 1 shows the agreement between system rankings induced by different evaluation metrics across the 22 systems. While the metrics generally exhibit high rank correlation, the agreement is not uniform. The COMET-family metrics are highly correlated with one another ( $\rho \geq 0.958$ ), whereas their agreement with lexical metrics is generally lower. In particular, BLEU shows its lowest correlations with COMTAIL-DA ( $\rho = 0.801$ ) and COMTAIL-QE ( $\rho = 0.831$ ). Thus, although the metrics broadly agree on system quality, they can differ substantially in the relative ranking of systems, highlighting the importance of cross-metric analysis when evaluating LLM-based translation.

#### 4.2 Qualitative Analysis

The automatic metrics reveal substantial differences across genres, but they do not fully capture the nature of the errors made by individual systems. We therefore manually analyze representative outputs from the CoST suite, focusing on failure modes that are difficult to characterize through aggregate scores alone. We observe recurring problems in named-entity preservation and lexical selection, as well as more severe failures involving target-language control, memorized-text reproduction, meta-responses, and output degeneration. Examples are given in Appendix (Table 11).

**Named-entity handling.** Named entities are generally expected to remain invariant in translation, apart from language-appropriate transliteration. However, several systems alter or distort proper names. For example, in the sentence “Labanya said to her sister in soothing tones: ...”, multiple systems produce variations of *Labanya*. A similar pattern occurs for *Phoebe*, which is rendered as variants such as *Phobey*, *Phoyeb*, and *Phoyebe* by several systems. These errors are notable because the surrounding sentence can remain fluent while the identity of the named entity is lost.

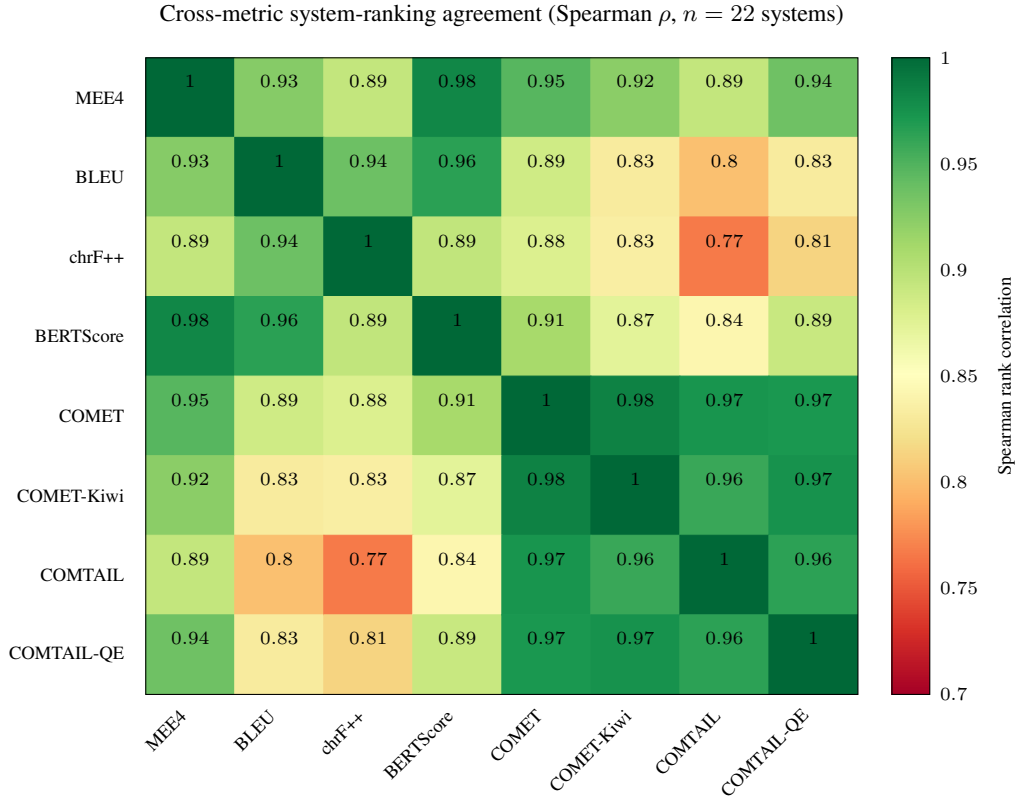


Figure 1: Spearman rank correlations between system rankings induced by the eight evaluation metrics across the 22 systems. The strong correlations among COMET-family metrics contrast with their weaker agreement with lexical metrics such as BLEU and chrF++.

**Lexical selection.** We also observe errors in fine-grained lexical choice even when the surrounding sentence is otherwise well formed. In the sentence “Chandler, Monica, and Phoebe are baby-sitting the triplets”, multiple systems select inappropriate equivalents for *triplets*, while SalamandraTA reproduces the English word unchanged. Similar lexical disagreements were observed in the WMT24 evaluation, including substitutions involving *well* and *pulling up*. The recurrence of such errors across evaluation years suggests that fine-grained lexical selection remains a persistent weakness rather than a model-specific artifact.

**Target-language failures.** More severe failures involve loss of the target-language constraint. Tiny Aya Global produces Marathi instead of Hindi for some inputs, while SalamandraTA produces outputs that are predominantly English, with occasional Mandarin segments. These failures are qualitatively different from ordinary lexical or grammatical errors: although the generated text may be fluent, the system fails to satisfy the fundamental requirement that the output be in the requested tar-

get language.

**Memorization-driven canonical-text reproduction.** Poetry exposes a distinct failure mode in which systems appear to reproduce a memorized canonical text rather than translate the supplied source. For example, given the English rendering “Tomorrow’s work do today, today’s work now; if the moment is lost, the work be done how”, GPT-5.5 produces the canonical Hindi Kabir couplet *Kal kare so āj kar, āj kare so ab; pal mem pralay hoegī, bahuri karegā kab*. Similarly, for “In anguish everyone prays to Him, in joy does none; To One who prays in happiness, how sorrow can come”, the system produces the corresponding canonical Hindi doha rather than translating the given wording. The outputs are fluent and semantically related to the inputs, but they violate source fidelity by replacing the supplied text with a memorized formulation. This suggests that the poetry degradation is not explained by genre difficulty alone and may instead reflect a memorization-driven override.

**Meta-responses and output degeneration.** We further observe more severe generation failures. In some cases, systems produce a meta-response, such as a translation label or instruction, rather than the requested translation. In others, generation degenerates into repetitive or numerically corrupted sequences, with outputs containing patterns such as *29-500000000000....* These cases represent a more fundamental failure of task adherence: the system does not merely mistranslate the source but loses the expected translation behavior altogether.

**Persistence across evaluation years.** The longitudinal comparison allows us to distinguish persistent weaknesses from failures newly observed in the current evaluation. Named-entity mishandling was prominent in both years: the WMT24 evaluation reported distorted renderings of *Labanya* and *Phoebe* by six and seven systems, respectively, compared with fourteen and seven systems in the current evaluation. Lexical selection errors likewise recur, with WMT24 showing substitutions such as *well* → *alright/okay* and disagreement over *pulling up* versus *removing*, while the current evaluation shows comparable errors for *triplets*. These recurring patterns suggest that named-entity preservation and fine-grained lexical selection remain persistent weaknesses across successive generations of MT systems.

In contrast, wrong-language outputs and memorization-driven reproduction of canonical poetry were not reported in the WMT24 qualitative analysis. The available evidence does not establish whether these represent genuinely new failure modes or failures that were simply not observed in the earlier manual analysis. Nevertheless, their presence in the current evaluation highlights failure behaviors that may be missed by aggregate metrics and motivates continued qualitative and longitudinal analysis of LLM-based translation.

## 5 Conclusion

This paper evaluates 22 machine translation systems submitted to the WMT26 General Translation Shared Task using CoST, a 1,947-sentence, eight-genre benchmark designed to expose weaknesses that generic, narrative-heavy evaluation data tends to obscure. Combining reference-free, reference-based, and manual evaluation, we find a consistent genre-level pattern: systems perform well on narrative-style text such as autobiography and play,

but degrade markedly on poetry, legal, and conversational content, plausibly reflecting a training-data skew toward narrative prose. Qualitative analysis further reveals a distinct poetry-specific failure mode, systems reproducing memorized canonical Hindi verse instead of translating the given English input, alongside recurring issues in named-entity handling, wrong-language output, and lexical word choice. Comparing these findings to a prior CoST evaluation on WMT24 submissions shows that named-entity mishandling and lexical selection errors persist across two years of substantially different systems, suggesting structural rather than model-specific limitations, while wrong-language output and memorization-driven poetry reproduction appear specific to this year’s submissions. canonical-verse reproduction were observed in our WMT26 analysis but were not reported in the WMT24 analysis.

Together, these results show that despite continued advances in general translation quality, LLM-based MT systems retain systematic blind spots in stylistically and structurally complex text, underscoring the need for more genre-aware evaluation and broader, more stylistically diverse training data.

## References

- Arafat Ahsan, Vandan Mujadia, Pruthwik Mishra, Yash Bhaskar, and Dipti Misra Sharma. 2025. [Crosslingual optimized metric for translation assessment of indian languages](#).
- Xinyue Liu, Niloofar Mireshghallah, Jane C. Ginsburg, and Tuhin Chakrabarty. 2026. [Alignment whack-a-mole : Finetuning activates verbatim recall of copyrighted books in large language models](#).
- Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian MÅller. 2023. [Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can chatgpt outperform nmt?](#) In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.
- Ananya Mukherjee, Hema Ala, Manish Shrivastava, and Dipti Misra Sharma. 2020. Mee: An automatic metric for evaluation using embeddings for machine translation. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 292–299. IEEE.
- Ananya Mukherjee and Manish Shrivastava. 2023. [Mee4 and xlsim : Iit hyd’s submissions’ for wmt23 metrics shared task](#). In *Proceedings of the Eighth*

*Conference on Machine Translation*, pages 798–803, Singapore. Association for Computational Linguistics.

Ananya Mukherjee, Saumitra Yadav, and Manish Shrivastava. 2024. [CoST of breaking the LLMs](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 299–306, Miami, Florida, USA. Association for Computational Linguistics.

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

Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

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.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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*.

its rank within that category and metric, shown in parentheses (rank 1 indicating the highest score). Systems are ordered by COMET-Kiwi in descending order within each table. Cell shading reflects the relative score within each metric column for that category, from lightest (lowest) to darkest (highest).

## A Appendix

### A.1 Category wise Scores

Tables 3-10 present category-wise system rankings across all eight evaluation metrics, split into referenceless (COMET-Kiwi, COMTAIL-QE) and reference-based (BLEU, chrF++, BERTScore, COMET, COMTAIL-DA, MEE4) groups. Each table corresponds to one of the eight test-suite categories (autobiography, conversation, legal, mix, narration, play, poetry, and technical) and reports, for every system, the raw metric score along with

| Model              | Referenceless |            | Reference-based |           |           |           |            |           |
|--------------------|---------------|------------|-----------------|-----------|-----------|-----------|------------|-----------|
|                    | COMET-Kiwi    | COMTAIL-QE | BLEU            | chrF++    | BERTScore | COMET     | COMTAIL-DA | MEE4      |
| TRIVE              | 83.2 (1)      | 88.1 (1)   | 35.6 (2)        | 60.2 (2)  | 88.8 (2)  | 80.5 (2)  | 82.5 (3)   | 83.8 (2)  |
| GoogleTranslate    | 83.0 (2)      | 88.0 (2)   | 39.9 (1)        | 62.4 (1)  | 89.5 (1)  | 81.5 (1)  | 83.1 (1)   | 84.4 (1)  |
| GPT 5.5            | 83.0 (3)      | 87.7 (3)   | 31.4 (7)        | 57.6 (5)  | 87.4 (8)  | 80.0 (6)  | 82.3 (5)   | 82.8 (6)  |
| Lumen              | 83.0 (4)      | 87.7 (4)   | 33.0 (5)        | 58.1 (4)  | 88.0 (6)  | 80.1 (5)  | 82.3 (4)   | 83.2 (4)  |
| DeepSeek V4 Pro    | 82.9 (5)      | 87.5 (6)   | 31.0 (8)        | 56.3 (7)  | 87.8 (7)  | 79.7 (7)  | 81.9 (8)   | 82.6 (8)  |
| Gemma 4 - 31B      | 82.8 (6)      | 87.4 (7)   | 35.0 (3)        | 59.1 (3)  | 88.3 (3)  | 80.2 (4)  | 82.2 (6)   | 83.3 (3)  |
| Cohere CAT+        | 82.7 (7)      | 87.7 (5)   | 28.4 (10)       | 55.2 (9)  | 86.9 (9)  | 80.4 (3)  | 82.8 (2)   | 82.1 (10) |
| Command A+         | 82.2 (8)      | 86.8 (8)   | 24.4 (12)       | 50.3 (13) | 85.5 (12) | 78.8 (9)  | 82.0 (7)   | 80.9 (12) |
| Tower 9B           | 81.7 (9)      | 85.7 (10)  | 30.2 (9)        | 55.0 (10) | 86.8 (10) | 77.7 (12) | 80.2 (12)  | 82.2 (9)  |
| Dubformer          | 81.5 (10)     | 86.0 (9)   | 33.1 (4)        | 56.6 (6)  | 88.1 (4)  | 79.1 (8)  | 80.8 (10)  | 82.9 (5)  |
| Gemma 4 - 4B       | 81.4 (11)     | 85.2 (12)  | 31.8 (6)        | 56.1 (8)  | 88.1 (5)  | 77.9 (11) | 80.3 (11)  | 82.7 (7)  |
| Mistral Medium 3.5 | 81.1 (12)     | 85.6 (11)  | 28.0 (11)       | 51.2 (12) | 86.5 (11) | 78.1 (10) | 80.8 (9)   | 81.3 (11) |
| GPT OSS 120B       | 80.9 (13)     | 84.2 (14)  | 20.6 (14)       | 48.0 (15) | 84.4 (15) | 76.4 (14) | 79.2 (14)  | 80.2 (14) |
| Qwen 3.6 - 27b     | 80.7 (14)     | 84.6 (13)  | 21.1 (13)       | 48.6 (14) | 84.9 (13) | 76.8 (13) | 79.9 (13)  | 80.2 (13) |
| Ministral 3 - 14B  | 79.2 (15)     | 82.1 (15)  | 20.5 (15)       | 47.4 (16) | 84.8 (14) | 75.5 (15) | 78.5 (15)  | 79.7 (15) |
| Qingqiu-MT-9B      | 78.4 (16)     | 80.0 (16)  | 18.3 (17)       | 44.6 (17) | 83.8 (16) | 73.5 (16) | 76.0 (16)  | 79.1 (16) |
| Qwen 3.5 - 9b      | 77.8 (17)     | 79.5 (17)  | 17.8 (19)       | 43.7 (19) | 83.5 (17) | 73.0 (17) | 75.3 (17)  | 78.5 (17) |
| GPT OSS 20B        | 77.1 (18)     | 79.5 (18)  | 18.7 (16)       | 44.3 (18) | 81.8 (18) | 71.5 (18) | 73.7 (18)  | 75.8 (18) |
| Gemini 3.1 Pro     | 76.5 (19)     | 64.6 (21)  | 18.0 (18)       | 51.9 (11) | 80.4 (19) | 70.2 (19) | 68.4 (19)  | 75.6 (19) |
| Tiny Aya Global    | 76.4 (20)     | 78.8 (19)  | 13.9 (20)       | 37.7 (20) | 80.2 (20) | 68.7 (20) | 63.5 (21)  | 74.1 (20) |
| SalamandraTA       | 72.6 (21)     | 70.1 (20)  | 11.0 (21)       | 35.0 (21) | 79.7 (21) | 67.6 (21) | 66.5 (20)  | 73.8 (21) |
| GLM 4 - 9B         | 63.5 (22)     | 54.3 (22)  | 9.7 (22)        | 33.8 (22) | 78.9 (22) | 55.3 (22) | 50.3 (22)  | 71.0 (22) |

Table 3: System-wise ranking for the *Autobiography* category based on reference-free and reference-based metrics.

| Model              | Referenceless |            | Reference-based |           |           |           |            |           |
|--------------------|---------------|------------|-----------------|-----------|-----------|-----------|------------|-----------|
|                    | COMET-Kiwi    | COMTAIL-QE | BLEU            | chrF++    | BERTScore | COMET     | COMTAIL-DA | MEE4      |
| Cohere CAT+        | 81.6 (1)      | 85.6 (1)   | 32.5 (11)       | 54.6 (11) | 86.7 (11) | 77.7 (3)  | 77.1 (2)   | 80.0 (11) |
| GoogleTranslate    | 81.2 (2)      | 85.1 (2)   | 45.9 (1)        | 61.6 (1)  | 89.0 (1)  | 78.7 (1)  | 78.1 (1)   | 82.4 (3)  |
| Command A+         | 80.9 (3)      | 84.1 (4)   | 30.8 (13)       | 53.3 (13) | 86.2 (13) | 76.9 (5)  | 76.6 (3)   | 79.4 (13) |
| Lumen              | 80.2 (4)      | 83.6 (5)   | 41.0 (6)        | 59.4 (6)  | 88.2 (7)  | 77.2 (4)  | 75.6 (5)   | 82.4 (2)  |
| Gemma 4 - 31B      | 80.0 (5)      | 84.2 (3)   | 42.6 (3)        | 60.0 (3)  | 88.8 (3)  | 77.9 (2)  | 75.9 (4)   | 82.0 (6)  |
| GPT 5.5            | 79.8 (6)      | 82.9 (6)   | 41.2 (5)        | 60.0 (4)  | 88.4 (6)  | 76.2 (7)  | 73.8 (11)  | 82.2 (4)  |
| DeepSeek V4 Pro    | 79.6 (7)      | 82.7 (9)   | 39.7 (8)        | 58.2 (8)  | 88.1 (8)  | 76.0 (9)  | 74.6 (7)   | 81.4 (8)  |
| TRIVE              | 79.5 (8)      | 82.8 (7)   | 42.1 (4)        | 61.3 (2)  | 88.6 (4)  | 76.5 (6)  | 74.0 (9)   | 82.1 (5)  |
| Dubformer          | 79.3 (9)      | 81.2 (13)  | 40.5 (7)        | 59.3 (7)  | 88.5 (5)  | 76.2 (8)  | 73.2 (12)  | 81.9 (7)  |
| Mistral Medium 3.5 | 79.3 (10)     | 82.8 (8)   | 36.6 (10)       | 55.7 (10) | 87.3 (10) | 75.2 (11) | 74.7 (6)   | 80.4 (10) |
| GPT OSS 120B       | 79.1 (11)     | 82.2 (11)  | 31.6 (12)       | 53.7 (12) | 86.4 (12) | 74.4 (14) | 73.9 (10)  | 79.4 (12) |
| Qwen 3.6 - 27b     | 78.3 (12)     | 80.9 (14)  | 28.6 (14)       | 50.3 (15) | 85.2 (14) | 74.5 (13) | 73.1 (13)  | 79.2 (14) |
| Gemma 4 - 4B       | 78.2 (13)     | 82.3 (10)  | 43.0 (2)        | 59.5 (5)  | 89.0 (2)  | 74.7 (12) | 72.5 (14)  | 82.5 (1)  |
| Tower 9B           | 78.1 (14)     | 81.7 (12)  | 38.9 (9)        | 58.0 (9)  | 88.0 (9)  | 75.3 (10) | 74.5 (8)   | 80.7 (9)  |
| Gemini 3.1 Pro     | 77.1 (15)     | 72.9 (19)  | 23.1 (19)       | 53.2 (14) | 82.8 (18) | 70.6 (18) | 65.5 (18)  | 76.1 (18) |
| Ministral 3 - 14B  | 76.7 (16)     | 79.5 (15)  | 24.9 (16)       | 48.4 (16) | 84.6 (15) | 71.9 (15) | 72.4 (15)  | 77.3 (16) |
| Qwen 3.5 - 9b      | 75.0 (17)     | 76.5 (17)  | 23.8 (18)       | 45.7 (19) | 84.3 (17) | 71.0 (16) | 70.0 (16)  | 76.5 (17) |
| Qingqiu-MT-9B      | 74.5 (18)     | 72.1 (20)  | 25.0 (15)       | 45.7 (18) | 84.3 (16) | 70.6 (17) | 64.3 (19)  | 77.4 (15) |
| GPT OSS 20B        | 74.1 (19)     | 74.2 (18)  | 24.9 (17)       | 47.9 (17) | 81.9 (19) | 68.9 (19) | 66.4 (17)  | 72.5 (20) |
| Tiny Aya Global    | 72.3 (20)     | 77.0 (16)  | 7.2 (22)        | 23.7 (21) | 78.9 (21) | 64.0 (20) | 57.7 (20)  | 70.0 (21) |
| GLM 4 - 9B         | 62.5 (21)     | 57.0 (21)  | 18.6 (20)       | 38.0 (20) | 81.8 (20) | 58.9 (21) | 53.7 (21)  | 73.0 (19) |
| SalamandraTA       | 57.0 (22)     | 53.5 (22)  | 11.9 (21)       | 22.8 (22) | 76.5 (22) | 55.0 (22) | 43.4 (22)  | 69.2 (22) |

Table 4: System-wise ranking for the *Conversation* category based on reference-free and reference-based metrics.

| Model              | Referenceless |            | Reference-based |           |           |           |            |           |
|--------------------|---------------|------------|-----------------|-----------|-----------|-----------|------------|-----------|
|                    | COMET-Kiwi    | COMTAIL-QE | BLEU            | chrF++    | BERTScore | COMET     | COMTAIL-DA | MEE4      |
| GPT 5.5            | 80.4 (1)      | 79.3 (1)   | 42.6 (5)        | 68.3 (5)  | 88.0 (4)  | 80.6 (3)  | 77.8 (3)   | 85.3 (4)  |
| TRIVE              | 80.0 (2)      | 79.1 (3)   | 48.9 (2)        | 72.0 (1)  | 89.4 (1)  | 80.5 (4)  | 77.2 (6)   | 86.2 (1)  |
| Cohere CAT+        | 79.9 (3)      | 78.1 (7)   | 36.7 (10)       | 65.0 (8)  | 86.5 (9)  | 80.7 (2)  | 77.8 (2)   | 84.3 (8)  |
| Lumen              | 79.9 (4)      | 79.2 (2)   | 43.9 (3)        | 68.4 (4)  | 88.3 (3)  | 80.1 (6)  | 77.2 (5)   | 85.4 (3)  |
| Command A+         | 79.7 (5)      | 78.4 (5)   | 34.7 (11)       | 60.8 (11) | 85.8 (11) | 80.1 (5)  | 77.9 (1)   | 83.6 (10) |
| DeepSeek V4 Pro    | 79.7 (6)      | 78.4 (4)   | 39.0 (8)        | 65.2 (7)  | 87.1 (7)  | 79.2 (8)  | 76.4 (8)   | 84.6 (7)  |
| GoogleTranslate    | 79.3 (7)      | 78.2 (6)   | 48.9 (1)        | 70.1 (2)  | 89.3 (2)  | 81.0 (1)  | 77.7 (4)   | 86.1 (2)  |
| Mistral Medium 3.5 | 79.2 (8)      | 77.3 (8)   | 33.9 (12)       | 60.4 (12) | 86.1 (10) | 79.3 (7)  | 76.7 (7)   | 83.7 (9)  |
| Dubformer          | 78.9 (9)      | 76.0 (9)   | 42.1 (6)        | 66.8 (6)  | 88.0 (5)  | 78.6 (9)  | 73.8 (12)  | 85.3 (5)  |
| GPT OSS 120B       | 78.4 (10)     | 76.0 (10)  | 23.2 (16)       | 52.9 (16) | 83.2 (15) | 77.7 (10) | 75.0 (9)   | 81.5 (14) |
| Tower 9B           | 77.8 (11)     | 75.7 (11)  | 39.8 (7)        | 64.5 (9)  | 87.3 (6)  | 77.7 (11) | 74.5 (10)  | 84.8 (6)  |
| Qwen 3.6 - 27b     | 77.0 (12)     | 73.5 (12)  | 23.2 (15)       | 53.8 (14) | 82.8 (16) | 76.9 (12) | 73.7 (13)  | 81.6 (12) |
| Ministral 3 - 14B  | 76.9 (13)     | 73.4 (13)  | 24.5 (14)       | 53.5 (15) | 83.2 (14) | 76.6 (13) | 74.2 (11)  | 81.1 (16) |
| Gemini 3.1 Pro     | 75.5 (14)     | 64.2 (19)  | 38.2 (9)        | 68.6 (3)  | 84.5 (12) | 76.5 (14) | 71.5 (15)  | 82.1 (11) |
| Gemma 4 - 31B      | 74.7 (15)     | 71.1 (14)  | 42.9 (4)        | 63.8 (10) | 86.6 (8)  | 76.3 (15) | 72.6 (14)  | 81.6 (13) |
| Qingqiu-MT-9B      | 74.7 (16)     | 65.4 (18)  | 21.7 (17)       | 51.8 (17) | 81.9 (17) | 73.0 (16) | 65.5 (19)  | 81.4 (15) |
| Qwen 3.5 - 9b      | 73.5 (17)     | 68.8 (15)  | 16.3 (20)       | 47.1 (18) | 80.2 (18) | 72.9 (17) | 69.5 (16)  | 78.8 (17) |
| GPT OSS 20B        | 72.1 (18)     | 67.7 (16)  | 17.1 (19)       | 45.7 (19) | 79.3 (20) | 71.5 (19) | 68.5 (17)  | 75.2 (20) |
| Gemma 4 - 4B       | 71.2 (19)     | 66.3 (17)  | 32.4 (13)       | 55.5 (13) | 83.8 (13) | 71.9 (18) | 68.0 (18)  | 77.6 (18) |
| Tiny Aya Global    | 70.0 (20)     | 63.5 (20)  | 17.2 (18)       | 41.7 (20) | 79.5 (19) | 68.2 (20) | 58.0 (20)  | 76.7 (19) |
| SalamandraTA       | 56.0 (21)     | 35.8 (21)  | 8.6 (21)        | 26.1 (21) | 73.4 (21) | 49.5 (21) | 34.9 (21)  | 72.9 (21) |
| GLM 4 - 9B         | 49.0 (22)     | 20.4 (22)  | 4.0 (22)        | 20.1 (22) | 69.2 (22) | 38.0 (22) | 21.4 (22)  | 66.6 (22) |

Table 5: System-wise ranking for the *Legal* category based on reference-free and reference-based metrics.

| Model              | Referenceless |            | Reference-based |           |           |           |            |           |
|--------------------|---------------|------------|-----------------|-----------|-----------|-----------|------------|-----------|
|                    | COMET-Kiwi    | COMTAIL-QE | BLEU            | chrF++    | BERTScore | COMET     | COMTAIL-DA | MEE4      |
| TRIVE              | 85.9 (1)      | 89.7 (2)   | 51.2 (2)        | 72.7 (2)  | 92.1 (2)  | 84.1 (2)  | 83.8 (1)   | 87.1 (2)  |
| GPT 5.5            | 85.9 (2)      | 89.6 (3)   | 45.9 (6)        | 69.7 (4)  | 91.2 (4)  | 83.7 (4)  | 83.6 (3)   | 86.1 (4)  |
| DeepSeek V4 Pro    | 85.9 (3)      | 89.5 (4)   | 44.1 (7)        | 69.0 (6)  | 91.2 (5)  | 83.4 (5)  | 83.3 (7)   | 85.7 (7)  |
| Lumen              | 85.8 (4)      | 89.7 (1)   | 46.6 (4)        | 69.5 (5)  | 91.0 (6)  | 83.4 (6)  | 83.7 (2)   | 86.0 (5)  |
| Gemma 4 - 31B      | 85.8 (5)      | 88.4 (8)   | 49.8 (3)        | 71.4 (3)  | 91.8 (3)  | 83.8 (3)  | 83.3 (8)   | 86.8 (3)  |
| Command A+         | 85.7 (6)      | 89.2 (5)   | 36.0 (12)       | 62.2 (13) | 88.8 (12) | 82.8 (8)  | 83.5 (4)   | 83.7 (12) |
| Cohere CAT+        | 85.6 (7)      | 88.6 (7)   | 40.1 (9)        | 66.2 (10) | 89.7 (11) | 83.1 (7)  | 83.4 (5)   | 84.8 (9)  |
| GoogleTranslate    | 85.4 (8)      | 89.1 (6)   | 52.8 (1)        | 73.3 (1)  | 92.3 (1)  | 84.1 (1)  | 83.4 (6)   | 87.1 (1)  |
| Dubformer          | 85.3 (9)      | 87.2 (13)  | 46.0 (5)        | 68.8 (7)  | 91.0 (7)  | 82.7 (9)  | 82.1 (12)  | 86.0 (6)  |
| Gemma 4 - 4B       | 85.0 (10)     | 87.8 (11)  | 43.8 (8)        | 66.9 (9)  | 90.8 (8)  | 82.0 (10) | 82.3 (11)  | 85.2 (8)  |
| GPT OSS 120B       | 84.9 (11)     | 87.4 (12)  | 31.7 (15)       | 59.6 (15) | 88.0 (13) | 81.4 (13) | 81.8 (13)  | 82.8 (15) |
| Tower 9B           | 84.8 (12)     | 87.9 (9)   | 39.6 (11)       | 64.9 (11) | 89.9 (9)  | 81.7 (12) | 82.4 (10)  | 84.6 (10) |
| Mistral Medium 3.5 | 84.8 (13)     | 87.8 (10)  | 39.8 (10)       | 64.8 (12) | 89.8 (10) | 81.9 (11) | 82.4 (9)   | 84.3 (11) |
| Qwen 3.6 - 27b     | 84.2 (14)     | 86.1 (14)  | 31.8 (14)       | 60.5 (14) | 88.0 (14) | 80.5 (14) | 81.1 (14)  | 83.1 (14) |
| GPT OSS 20B        | 84.1 (15)     | 85.4 (15)  | 26.9 (17)       | 56.3 (17) | 86.5 (17) | 79.6 (15) | 80.2 (16)  | 81.4 (18) |
| Ministral 3 - 14B  | 83.2 (16)     | 84.7 (16)  | 30.7 (16)       | 58.7 (16) | 87.6 (15) | 79.3 (16) | 80.5 (15)  | 82.1 (16) |
| Gemini 3.1 Pro     | 83.0 (17)     | 76.7 (20)  | 33.7 (13)       | 67.3 (8)  | 87.2 (16) | 79.3 (17) | 77.3 (18)  | 83.5 (13) |
| Qingqiu-MT-9B      | 81.8 (18)     | 77.1 (19)  | 25.4 (18)       | 54.5 (18) | 85.6 (18) | 76.5 (19) | 74.2 (19)  | 81.6 (17) |
| Qwen 3.5 - 9b      | 81.5 (19)     | 82.0 (18)  | 21.0 (20)       | 51.6 (19) | 85.1 (19) | 76.7 (18) | 77.5 (17)  | 79.8 (19) |
| Tiny Aya Global    | 81.2 (20)     | 82.6 (17)  | 21.9 (19)       | 48.1 (20) | 83.4 (20) | 74.3 (20) | 68.5 (20)  | 77.3 (20) |
| GLM 4 - 9B         | 70.9 (21)     | 61.3 (21)  | 14.7 (21)       | 42.1 (21) | 81.5 (21) | 63.8 (22) | 59.2 (21)  | 75.1 (21) |
| SalamandraTA       | 70.2 (22)     | 55.4 (22)  | 13.7 (22)       | 32.1 (22) | 79.7 (22) | 64.6 (21) | 56.8 (22)  | 74.9 (22) |

Table 6: System-wise ranking for the *Mix* category based on reference-free and reference-based metrics.

| Model              | Referenceless |            | Reference-based |           |           |           |            |           |
|--------------------|---------------|------------|-----------------|-----------|-----------|-----------|------------|-----------|
|                    | COMET-Kiwi    | COMTAIL-QE | BLEU            | chrF++    | BERTScore | COMET     | COMTAIL-DA | MEE4      |
| GoogleTranslate    | 82.4 (1)      | 88.1 (1)   | 41.2 (1)        | 62.0 (1)  | 89.2 (1)  | 80.3 (1)  | 82.0 (1)   | 83.4 (1)  |
| GPT 5.5            | 82.2 (2)      | 87.8 (2)   | 30.6 (9)        | 54.9 (9)  | 86.7 (9)  | 78.2 (4)  | 80.9 (4)   | 81.5 (5)  |
| TRIVE              | 82.2 (3)      | 87.7 (4)   | 36.1 (2)        | 59.4 (2)  | 88.2 (2)  | 79.1 (2)  | 81.2 (2)   | 82.7 (2)  |
| Cohere CAT+        | 82.0 (4)      | 87.8 (3)   | 28.2 (11)       | 53.3 (10) | 86.3 (11) | 78.3 (3)  | 81.1 (3)   | 80.4 (11) |
| DeepSeek V4 Pro    | 81.8 (5)      | 87.2 (6)   | 31.4 (8)        | 55.1 (8)  | 87.1 (7)  | 77.8 (7)  | 80.0 (8)   | 81.4 (6)  |
| Lumen              | 81.8 (6)      | 87.2 (5)   | 31.8 (7)        | 55.1 (7)  | 87.0 (8)  | 77.6 (8)  | 80.3 (6)   | 81.4 (7)  |
| Command A+         | 81.7 (7)      | 86.9 (7)   | 26.9 (12)       | 51.8 (11) | 86.0 (12) | 77.9 (5)  | 80.5 (5)   | 80.2 (12) |
| Dubformer          | 81.7 (8)      | 86.7 (8)   | 33.3 (5)        | 56.4 (4)  | 87.7 (3)  | 77.9 (6)  | 80.0 (7)   | 82.0 (4)  |
| Tower 9B           | 80.7 (9)      | 85.6 (9)   | 32.9 (6)        | 56.4 (5)  | 87.4 (6)  | 77.2 (10) | 79.2 (9)   | 82.1 (3)  |
| GPT OSS 120B       | 80.4 (10)     | 84.6 (11)  | 24.8 (13)       | 50.2 (13) | 85.3 (13) | 75.7 (12) | 77.9 (13)  | 79.9 (13) |
| Gemma 4 - 31B      | 80.1 (11)     | 85.2 (10)  | 35.9 (3)        | 57.3 (3)  | 87.6 (4)  | 77.4 (9)  | 79.2 (10)  | 81.1 (8)  |
| Mistral Medium 3.5 | 79.9 (12)     | 84.6 (12)  | 29.0 (10)       | 51.6 (12) | 86.5 (10) | 76.1 (11) | 78.5 (11)  | 80.5 (10) |
| Qwen 3.6 - 27b     | 79.7 (13)     | 84.2 (13)  | 23.8 (14)       | 49.3 (14) | 85.1 (14) | 75.3 (14) | 77.9 (12)  | 79.3 (14) |
| Gemma 4 - 4B       | 78.6 (14)     | 82.9 (14)  | 34.7 (4)        | 56.3 (6)  | 87.5 (5)  | 75.5 (13) | 76.9 (14)  | 80.8 (9)  |
| Ministral 3 - 14B  | 77.8 (15)     | 81.2 (15)  | 23.0 (15)       | 47.8 (16) | 84.8 (15) | 73.6 (15) | 75.9 (15)  | 78.6 (15) |
| Qingqiu-MT-9B      | 77.0 (16)     | 79.4 (17)  | 19.5 (16)       | 44.5 (17) | 83.5 (16) | 71.2 (16) | 73.2 (17)  | 77.5 (16) |
| Qwen 3.5 - 9b      | 76.3 (17)     | 79.7 (16)  | 17.6 (18)       | 43.2 (18) | 83.3 (17) | 71.0 (17) | 74.0 (16)  | 76.9 (17) |
| Gemini 3.1 Pro     | 76.0 (18)     | 65.6 (21)  | 15.3 (19)       | 49.1 (15) | 78.9 (22) | 68.5 (19) | 66.6 (19)  | 72.8 (18) |
| Tiny Aya Global    | 74.7 (19)     | 76.2 (18)  | 8.4 (22)        | 31.7 (22) | 79.4 (20) | 66.3 (21) | 60.1 (21)  | 71.8 (21) |
| GPT OSS 20B        | 74.5 (20)     | 75.9 (19)  | 18.4 (17)       | 43.1 (19) | 80.9 (18) | 68.7 (18) | 69.9 (18)  | 72.8 (19) |
| SalamandraTA       | 71.3 (21)     | 69.7 (20)  | 11.9 (20)       | 34.2 (20) | 79.5 (19) | 66.5 (20) | 65.1 (20)  | 72.5 (20) |
| GLM 4 - 9B         | 62.1 (22)     | 53.8 (22)  | 10.4 (21)       | 33.7 (21) | 79.1 (21) | 54.8 (22) | 50.1 (22)  | 69.7 (22) |

Table 7: System-wise ranking for the *Narration* category based on reference-free and reference-based metrics.

| Model              | Referenceless |            | Reference-based |           |           |           |            |           |
|--------------------|---------------|------------|-----------------|-----------|-----------|-----------|------------|-----------|
|                    | COMET-Kiwi    | COMTAIL-QE | BLEU            | chrF++    | BERTScore | COMET     | COMTAIL-DA | MEE4      |
| GoogleTranslate    | 85.1 (1)      | 89.7 (1)   | 46.1 (1)        | 63.3 (1)  | 90.2 (1)  | 83.6 (1)  | 83.8 (1)   | 85.2 (1)  |
| Cohere CAT+        | 85.0 (2)      | 89.6 (2)   | 28.2 (8)        | 53.3 (8)  | 87.7 (7)  | 81.9 (4)  | 82.8 (4)   | 81.9 (10) |
| Dubformer          | 84.7 (3)      | 89.2 (3)   | 35.8 (3)        | 56.6 (5)  | 88.5 (4)  | 82.0 (2)  | 82.8 (3)   | 83.2 (3)  |
| DeepSeek V4 Pro    | 84.6 (4)      | 89.1 (5)   | 27.7 (10)       | 51.5 (11) | 87.2 (10) | 81.6 (6)  | 82.1 (8)   | 82.2 (7)  |
| Gemma 4 - 31B      | 84.3 (5)      | 89.0 (7)   | 33.9 (4)        | 57.8 (2)  | 89.1 (3)  | 81.8 (5)  | 82.5 (6)   | 83.2 (4)  |
| TRIVE              | 84.3 (6)      | 89.1 (6)   | 33.0 (6)        | 56.8 (4)  | 88.3 (5)  | 82.0 (3)  | 83.2 (2)   | 82.9 (5)  |
| Lumen              | 84.2 (7)      | 89.1 (4)   | 30.7 (7)        | 54.0 (7)  | 87.6 (8)  | 80.8 (10) | 82.2 (7)   | 82.1 (8)  |
| Command A+         | 84.1 (8)      | 88.2 (9)   | 28.2 (9)        | 53.1 (9)  | 86.9 (11) | 80.9 (9)  | 82.5 (5)   | 80.6 (12) |
| GPT 5.5            | 83.7 (9)      | 88.5 (8)   | 25.5 (11)       | 52.5 (10) | 87.4 (9)  | 80.7 (11) | 81.5 (10)  | 82.0 (9)  |
| Gemma 4 - 4B       | 83.4 (10)     | 86.6 (12)  | 36.8 (2)        | 57.6 (3)  | 89.5 (2)  | 81.5 (7)  | 81.9 (9)   | 84.2 (2)  |
| Qwen 3.6 - 27b     | 83.3 (11)     | 86.8 (11)  | 22.1 (14)       | 49.4 (12) | 86.1 (14) | 79.8 (14) | 80.5 (14)  | 80.5 (13) |
| Mistral Medium 3.5 | 83.2 (12)     | 87.0 (10)  | 21.4 (16)       | 47.4 (16) | 86.3 (12) | 80.1 (13) | 81.0 (13)  | 80.4 (14) |
| Tower 9B           | 83.0 (13)     | 86.5 (13)  | 33.7 (5)        | 54.5 (6)  | 88.1 (6)  | 80.5 (12) | 81.2 (11)  | 82.2 (6)  |
| GPT OSS 120B       | 82.9 (14)     | 85.6 (14)  | 23.0 (13)       | 48.9 (13) | 86.3 (13) | 81.1 (8)  | 81.1 (12)  | 81.0 (11) |
| Ministral 3 - 14B  | 81.5 (15)     | 84.9 (15)  | 24.9 (12)       | 47.6 (15) | 84.9 (15) | 78.0 (15) | 79.2 (15)  | 79.6 (15) |
| Qwen 3.5 - 9b      | 79.8 (16)     | 81.2 (17)  | 16.9 (17)       | 41.3 (18) | 82.6 (17) | 74.6 (17) | 76.0 (16)  | 76.7 (17) |
| Gemini 3.1 Pro     | 79.5 (17)     | 68.5 (21)  | 15.8 (18)       | 47.6 (14) | 78.5 (21) | 71.8 (18) | 71.8 (18)  | 73.2 (19) |
| Qingqiu-MT-9B      | 79.5 (18)     | 81.4 (16)  | 22.0 (15)       | 45.7 (17) | 84.8 (16) | 75.9 (16) | 75.5 (17)  | 79.3 (16) |
| Tiny Aya Global    | 76.2 (19)     | 77.9 (18)  | 5.5 (22)        | 19.8 (22) | 77.2 (22) | 67.8 (21) | 57.9 (21)  | 71.2 (20) |
| GPT OSS 20B        | 74.9 (20)     | 74.0 (19)  | 14.8 (19)       | 40.3 (19) | 79.1 (20) | 69.1 (20) | 69.4 (19)  | 71.1 (21) |
| SalamandraTA       | 74.5 (21)     | 71.8 (20)  | 10.5 (20)       | 34.8 (20) | 81.2 (18) | 70.0 (19) | 68.7 (20)  | 73.4 (18) |
| GLM 4 - 9B         | 64.6 (22)     | 55.1 (22)  | 10.2 (21)       | 32.8 (21) | 79.2 (19) | 54.6 (22) | 46.2 (22)  | 70.5 (22) |

Table 8: System-wise ranking for the *Play* category based on reference-free and reference-based metrics.

| Model              | Referenceless |            | Reference-based |           |           |           |            |           |
|--------------------|---------------|------------|-----------------|-----------|-----------|-----------|------------|-----------|
|                    | COMET-Kiwi    | COMTAIL-QE | BLEU            | chrF++    | BERTScore | COMET     | COMTAIL-DA | MEE4      |
| Command A+         | 80.9 (1)      | 87.4 (3)   | 20.0 (15)       | 47.3 (11) | 84.4 (13) | 80.3 (5)  | 83.8 (4)   | 77.4 (12) |
| GoogleTranslate    | 80.8 (2)      | 87.9 (2)   | 36.4 (1)        | 53.8 (1)  | 87.1 (4)  | 83.5 (1)  | 85.0 (1)   | 82.3 (2)  |
| Cohere CAT+        | 80.7 (3)      | 87.9 (1)   | 30.4 (8)        | 51.2 (4)  | 86.9 (6)  | 81.5 (2)  | 84.5 (2)   | 81.6 (5)  |
| Gemma 4 - 31B      | 79.3 (4)      | 87.2 (4)   | 35.0 (2)        | 53.2 (2)  | 87.4 (2)  | 81.2 (3)  | 83.8 (3)   | 82.9 (1)  |
| GPT OSS 120B       | 79.2 (5)      | 84.7 (6)   | 23.2 (12)       | 45.2 (13) | 84.4 (12) | 75.5 (13) | 78.4 (13)  | 76.7 (13) |
| Qwen 3.6 - 27b     | 79.2 (6)      | 86.2 (5)   | 29.3 (9)        | 50.1 (6)  | 85.8 (9)  | 80.5 (4)  | 83.2 (5)   | 79.1 (9)  |
| Tower 9B           | 78.8 (7)      | 83.9 (9)   | 26.7 (11)       | 47.7 (9)  | 84.5 (11) | 77.8 (11) | 79.3 (12)  | 77.8 (11) |
| Qingqiu-MT-9B      | 78.0 (8)      | 84.5 (8)   | 27.0 (10)       | 46.0 (12) | 85.2 (10) | 78.3 (10) | 80.9 (7)   | 78.3 (10) |
| DeepSeek V4 Pro    | 77.5 (9)      | 83.5 (10)  | 34.5 (3)        | 50.6 (5)  | 86.6 (8)  | 79.2 (9)  | 80.7 (9)   | 80.0 (8)  |
| Mistral Medium 3.5 | 77.2 (10)     | 83.4 (12)  | 33.4 (4)        | 47.5 (10) | 86.7 (7)  | 77.2 (12) | 80.8 (8)   | 81.5 (6)  |
| Dubformer          | 77.2 (11)     | 84.6 (7)   | 32.1 (6)        | 48.6 (8)  | 87.2 (3)  | 79.7 (6)  | 79.5 (10)  | 81.8 (3)  |
| Ministral 3 - 14B  | 76.9 (12)     | 82.7 (14)  | 18.6 (16)       | 42.0 (14) | 84.1 (15) | 73.9 (14) | 75.2 (14)  | 75.3 (15) |
| Lumen              | 76.7 (13)     | 83.2 (13)  | 32.2 (5)        | 52.5 (3)  | 87.8 (1)  | 79.6 (7)  | 79.5 (11)  | 81.7 (4)  |
| Gemma 4 - 4B       | 76.4 (14)     | 83.5 (11)  | 31.0 (7)        | 50.0 (7)  | 86.9 (5)  | 79.4 (8)  | 81.5 (6)   | 81.0 (7)  |
| SalamandraTA       | 74.9 (15)     | 79.7 (15)  | 14.1 (18)       | 34.9 (18) | 80.6 (18) | 73.9 (15) | 71.8 (16)  | 73.2 (17) |
| Qwen 3.5 - 9b      | 73.3 (16)     | 77.1 (16)  | 22.0 (13)       | 40.0 (16) | 82.5 (17) | 71.6 (18) | 74.9 (15)  | 73.4 (16) |
| GPT 5.5            | 71.9 (17)     | 74.7 (17)  | 18.4 (17)       | 40.0 (15) | 84.2 (14) | 73.2 (16) | 68.0 (17)  | 76.0 (14) |
| TRIVE              | 69.6 (18)     | 69.5 (19)  | 20.4 (14)       | 38.8 (17) | 83.2 (16) | 71.6 (17) | 65.0 (18)  | 72.8 (18) |
| GPT OSS 20B        | 69.5 (19)     | 69.5 (20)  | 13.3 (19)       | 33.9 (19) | 77.3 (20) | 64.6 (19) | 62.4 (19)  | 63.0 (19) |
| Tiny Aya Global    | 67.2 (20)     | 71.2 (18)  | 4.1 (21)        | 22.5 (22) | 74.8 (21) | 59.3 (20) | 52.5 (20)  | 63.0 (20) |
| GLM 4 - 9B         | 63.2 (21)     | 59.3 (21)  | 11.1 (20)       | 30.7 (20) | 78.6 (19) | 55.7 (21) | 49.8 (21)  | 63.0 (21) |
| Gemini 3.1 Pro     | 58.4 (22)     | 35.6 (22)  | 3.9 (22)        | 26.1 (21) | 65.8 (22) | 53.1 (22) | 37.0 (22)  | 44.8 (22) |

Table 9: System-wise ranking for the *Poetry* category based on reference-free and reference-based metrics.

| Model              | Referenceless |            | Reference-based |           |           |           |            |           |
|--------------------|---------------|------------|-----------------|-----------|-----------|-----------|------------|-----------|
|                    | COMET-Kiwi    | COMTAIL-QE | BLEU            | chrF++    | BERTScore | COMET     | COMTAIL-DA | MEE4      |
| GoogleTranslate    | 84.0 (1)      | 84.1 (1)   | 46.0 (1)        | 68.5 (1)  | 90.8 (1)  | 80.7 (1)  | 79.6 (1)   | 85.5 (1)  |
| Lumen              | 83.7 (2)      | 82.7 (4)   | 36.5 (6)        | 62.2 (5)  | 88.8 (5)  | 79.0 (3)  | 78.3 (4)   | 83.4 (7)  |
| GPT 5.5            | 83.7 (3)      | 81.7 (6)   | 35.9 (8)        | 60.8 (9)  | 88.6 (8)  | 78.7 (6)  | 77.8 (6)   | 83.3 (8)  |
| TRIVE              | 83.6 (4)      | 81.8 (5)   | 40.6 (4)        | 65.1 (4)  | 89.4 (4)  | 79.2 (2)  | 77.5 (9)   | 84.2 (4)  |
| Command A+         | 83.5 (5)      | 83.3 (2)   | 28.3 (12)       | 56.8 (13) | 87.1 (11) | 78.7 (7)  | 78.9 (2)   | 81.2 (14) |
| DeepSeek V4 Pro    | 83.5 (6)      | 82.8 (3)   | 34.0 (9)        | 61.3 (8)  | 88.7 (7)  | 78.5 (9)  | 78.3 (3)   | 83.5 (6)  |
| Cohere CAT+        | 83.3 (7)      | 78.3 (15)  | 25.0 (15)       | 54.0 (15) | 84.7 (16) | 75.7 (16) | 74.7 (15)  | 80.2 (16) |
| GPT OSS 120B       | 83.0 (8)      | 81.2 (9)   | 27.6 (14)       | 56.2 (14) | 86.8 (14) | 78.1 (10) | 77.7 (7)   | 81.5 (12) |
| Dubformer          | 82.9 (9)      | 79.0 (12)  | 36.4 (7)        | 61.7 (7)  | 88.7 (6)  | 78.1 (11) | 76.1 (12)  | 83.7 (5)  |
| Qwen 3.6 - 27b     | 82.4 (10)     | 80.3 (11)  | 29.2 (11)       | 57.1 (12) | 87.0 (13) | 77.2 (12) | 76.6 (11)  | 81.9 (11) |
| Gemma 4 - 4B       | 82.2 (11)     | 81.4 (8)   | 44.0 (2)        | 66.0 (2)  | 90.0 (3)  | 79.0 (4)  | 77.5 (10)  | 84.3 (3)  |
| Mistral Medium 3.5 | 82.2 (12)     | 80.6 (10)  | 33.4 (10)       | 60.2 (10) | 88.3 (9)  | 78.6 (8)  | 77.6 (8)   | 82.9 (9)  |
| Tower 9B           | 82.0 (13)     | 81.6 (7)   | 42.7 (3)        | 65.6 (3)  | 90.2 (2)  | 78.8 (5)  | 78.2 (5)   | 84.8 (2)  |
| GPT OSS 20B        | 81.5 (14)     | 78.8 (13)  | 24.6 (16)       | 53.2 (17) | 85.0 (15) | 76.3 (14) | 74.8 (14)  | 79.4 (17) |
| Gemma 4 - 31B      | 81.3 (15)     | 76.2 (17)  | 39.0 (5)        | 61.9 (6)  | 87.6 (10) | 76.0 (15) | 73.3 (17)  | 82.2 (10) |
| Ministral 3 - 14B  | 80.7 (16)     | 78.4 (14)  | 28.1 (13)       | 57.2 (11) | 87.0 (12) | 76.5 (13) | 76.1 (13)  | 81.3 (13) |
| Qingqiu-MT-9B      | 79.9 (17)     | 70.2 (19)  | 20.8 (18)       | 49.6 (19) | 82.7 (19) | 71.9 (18) | 67.7 (18)  | 80.2 (15) |
| Tiny Aya Global    | 79.6 (18)     | 78.0 (16)  | 21.1 (17)       | 47.6 (20) | 82.8 (18) | 71.6 (19) | 65.5 (19)  | 76.2 (19) |
| Qwen 3.5 - 9b      | 79.1 (19)     | 75.1 (18)  | 20.6 (19)       | 50.4 (18) | 84.0 (17) | 74.5 (17) | 73.8 (16)  | 79.2 (18) |
| Gemini 3.1 Pro     | 77.7 (20)     | 58.9 (21)  | 15.7 (20)       | 53.4 (16) | 80.0 (21) | 69.2 (20) | 62.7 (20)  | 76.0 (20) |
| GLM 4 - 9B         | 70.7 (21)     | 62.4 (20)  | 13.1 (21)       | 40.9 (21) | 80.6 (20) | 65.0 (21) | 60.3 (21)  | 73.2 (22) |
| SalamandraTA       | 65.7 (22)     | 47.0 (22)  | 12.9 (22)       | 31.8 (22) | 78.3 (22) | 62.5 (22) | 51.0 (22)  | 74.1 (21) |

Table 10: System-wise ranking for the *Technical* category based on reference-free and reference-based metrics.

| Error                        | System          | Score | Source                                                                                                                                                                                                                                                                                                           | System Output                                                                                          | Remarks                                                                       |
|------------------------------|-----------------|-------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| Canonical-verse reproduction | GPT-5.5         | 48.44 | <i>Tomorrow's work do today, today's work now; if the moment is lost, the work be done how.</i>                                                                                                                                                                                                                  | <i>Kal kare so āj kar, āj kare so ab; pal meṃ pralay hoegī, bahuri karegā kab.</i>                     | Outputs the canonical Kabir doha rather than translating the given rendering. |
|                              | GPT-5.5         | 48.45 | <i>In anguish everyone prays to Him, in joy does none; To One who prays in happiness, how sorrow can come.</i>                                                                                                                                                                                                   | <i>Dukh meṃ sumiran sab kare, sukh meṃ kare na koy; jo sukh meṃ sumiran kare, to dukh kāhe ko hoy.</i> | Outputs the canonical doha rather than translating the supplied wording.      |
| Wrong-language output        | Tiny Aya Global | 27.86 | <i>In anguish everyone prays to Him, in joy does none; To One who prays in happiness, how sorrow can come.</i>                                                                                                                                                                                                   | <i>Avasādāt sarvajan tyācyākade prārthanā kartāt, ānandāt koṇihī nāhī; ...</i>                         | Generates Marathi instead of Hindi.                                           |
| Meta-response                | GPT OSS 20B     | 13.11 | <i>1125 before embarking on any further discussion of the project of interdicting transfers, as spelt out in section 22, we may read the provision: "where on or after the date of commencement of this Act, but before the notified date, any person has transferred any land held by him by sale, gift..."</i> | <i>Hindī anuvād (Devanāgarī)</i>                                                                       | Returns a translation label rather than a translation.                        |
| Output degeneration          | SalamandraTA    | 4.06  | <i>Ajanta, located in the Aurangabad District of Maharashtra has twenty-nine caitya and vihara caves decorated with sculptures and paintings from the first century B.C.E. to the fifth century C.E.</i>                                                                                                         | <i>Ajanta, jo Maharashtra ke Maharashtra ilaake mein sthit hai, usmein 29-500000000000...</i>          | Degenerates into an unbounded numerical sequence.                             |
| Named-entity distortion      | Gemma 4-31B     | –     | <i>Labanya said to her sister in soothing tones: "Don't be upset about it, dear; I will see what I can do to prevent it."</i>                                                                                                                                                                                    | <i>[distorted rendering of Labanya]</i>                                                                | Proper name is altered during translation.                                    |
|                              | Gemma 4-31B     | –     | <i>Scene: Monica and Rachel's, the next day, Chandler, Monica, and Phoebe are baby-sitting the triplets.</i>                                                                                                                                                                                                     | <i>[distorted rendering of Phoebe]</i>                                                                 | Proper name "Phoebe" is distorted.                                            |
| Lexical selection            | SalamandraTA    | 36.2  | <i>Scene: Monica and Rachel's, the next day, Chandler, Monica, and Phoebe are baby-sitting the triplets.</i>                                                                                                                                                                                                     | <i>triplets</i>                                                                                        | English lexical item is left untranslated.                                    |

Table 11: Qualitative examples of recurring failure modes in LLM translation.