CoST of breaking the LLMs

Ananya Mukherjee[*], Saumitra Yadav[*], Manish Shrivastava
MT-NLP Lab, LTRC, KCIS, IIIT Hyderabad, India
ananya.mukherjee@research.iiit.ac.in
saumitra.yadav@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

This paper presents an evaluation of 16 machine translation systems submitted to the Shared Task of the 9th Conference of Machine Translation (WMT24) for the English-Hindi (en-hi) language pair using our Complex Structures Test (CoST) suite. Aligning with this year's test suite subtask theme, "Help us break LLMs", we curated a comprehensive test suite encompassing diverse datasets across various categories, including autobiography, poetry, legal, conversation, play, narration, technical, and mixed genres.

Our evaluation reveals that **all the systems struggle significantly with the archaic style of text like legal and technical writings or text with creative twist like conversation and poetry datasets**, highlighting their weaknesses in handling complex linguistic structures and stylistic nuances inherent in these text types. Our evaluation identifies the strengths and limitations of the submitted models, pointing to specific areas where further research and development are needed to enhance their performance. Our test suite is available at https://github.com/AnanyaCoder/CoST-WMT-24-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 diverse 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 the results of using test suites to analyze state-of-the-art machine translation systems across various categories. These evaluations were conducted as part of the theme "Help Us Break LLMs" for the 9th Conference on Machine Translation (WMT24). The test suites were used to evaluate systems submitted for the English-Hindi language pair.

We have curated a unique test suite comprising sentences from 9 categories across 16 sources to evaluate how large language models (LLMs) perform. The diversity of these categories allows us to assess the LLMs' capabilities beyond the typical news or generic domains, which often focus on reporting or narrative writing styles. Details of our 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 16 different machine translation (MT) systems submitted to the General Translation Task at WMT24 (Kocmi et al., 2024a). For referenceless evaluation, we employ COMET-Kiwi (Rei et al., 2022), while (Papineni et al., 2002),

[*] Authors contributed equally

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 299–306
November 15–16, 2024. ©2024 Association for Computational Linguistics

chrF (Popović, 2015, 2017), MEE4 (Mukherjee et al., 2020; Mukherjee and Shrivastava, 2023), BERTScore (Zhang* et al., 2020), and COMET (Rei et al., 2020) are used for reference-based evaluation. Professional English-to-Hindi translators provide the reference translations. Our results indicate that, for the English-to-Hindi language pair, **LLMs show weaker performance on datasets related to poetry, legal, and conversational content**. Details of our evaluation experiments are discussed in Section 3, and our analysis is presented in Section 4.

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 16 submitted MT systems, we performed both automatic and manual evaluations.

3.1 Automatic Evaluation

In automatic evaluation, we leveraged both reference-less 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
	Christmas Carole	923	308
legal	Legal Text	2862	638
	IIT Bombay Jud	167	83
mix	IN22	570	241
conversation	Friends	77	53
play	King Of Dark Chamber	35	22
autobiography	My Reminiscences	109	110
Technical	Technical Papers	185	87
	Total	**6242**	**1947**

Table 1: Data Statistics of **CoST**.

3.1.1 Reference-less Evaluation

For the reference-less automatic evaluation, we utilize COMETKIWI (Rei et al., 2022) 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 [1] (Mukherjee et al., 2020; Mukherjee and Shrivastava, 2023), BERTScore (Zhang* et al., 2020), and COMET (Rei et al., 2020).

3.2 Manual Evaluation

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

4 Results and 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 rank. It is clearly evident that evaluations from all the metrics rank TranssionMT as the best system, followed by ONLINE-B

[1] https://www.kaggle.com/ananyacoder/mee4-metric-run

and Claude-3.5. In contrast, CycleL is ranked the lowest, preceded by IKUN-C and IKUN. We also observe that according to the Preliminary WMT24 Ranking of General MT Systems and LLMs (Kocmi et al., 2024b), Unbabel-Tower70B is listed as the top performer. However, its performance decreases on CoST. For more category-wise informative results, we looked at the performance of systems for each category using lexical-based metric (Figure 2 and 3), embedding- based metric (Figure 4 and 5) and supervised metric (6) and (Figure 1). These results illustrate that **all systems underperform with poetry, legal, and conversation data**. In contrast, the systems consistently exhibit strong performance with autobiography, play, and mixed (IN22) data.

The analysis shows a clear trend, i.e., systems struggle with specific genres like poetry, legal, and conversation while excelling in narrative styles such as autobiography and play. This suggests that the training data for these systems may be heavily skewed towards narrative writing, hence strong performance in those areas. The sub-par performance in poetry, conversational and legal texts might reflect challenges in handling diverse linguistic and stylistic features that are less prevalent in the training data.

4.1 Qualitative Analysis

These manual assessments are carried out by professional Hindi speakers who hold graduate-level qualifications and possess good knowledge in the domains covered by our test suite.

4.1.1 Handling Named Entities

Source: *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 . "*

Most models successfully translated "Labanya" correctly, preserving the original name. However, the outputs from Claude-3.5, GPT-4, NVIDIA-NeMo, ONLINE-A, Unbabel-Tower70B, and ZMT show variations or distortions of the name, indicating potential issues with **name recognition or transliteration** in these models.

In another instance, IKUN-C, IKUN, Llama3-70B, NVIDIA-NeMo, ONLINE-A, Unbabel-Tower70B, and ZMT systems have translated 'Phoebe' as Phob, Phobey, Phoyeb, Phoyebe; surprisingly ONLINE-G has generated चाँद (meaning moon, as Phobe is one of the moons of Saturn).

4.1.2 Spelling and Typological Errors

Except for Llama3-70B, IOL_Research, and CommandR-plus, all other models tend to generate हूं instead of हूँ, indicating a recurring spelling error in their outputs.

4.1.3 Omissions

The Hindi translations produced by the IKUN and IKUN-C systems consistently suffer from **incompleteness**, often leaving out key parts of the original sentences, undermining the accuracy and reliability of the translations, making them less effective for conveying the full meaning of the source text.

4.1.4 Incorrect Lexical Word Choices

Choosing the right word in translation is crucial for preserving the essence, tone, and intention of the original sentence. For instance, Unbabel-Tower70B accurately translates "well," whereas all other systems translate it as "alright" or "okay." These alternatives do not fit the context as well, thereby **affecting the tone and overall quality** of the translation.

Source: *I'd be pulling up shoots of grass to use them to check the wind, and looking at maps of ports and piers and roads.*

However, Aya23 and IOL_Research translate it as "removing," while the remaining systems use "pull." These variations of "remove" and "pull" slightly **affect the accuracy and well-formedness** of the Hindi translation.

5 Conclusion

This paper evaluates translations from 16 MT systems submitted to the General Translation Shared Task WMT24 on **Complex Structures Test** suite which was designed to cover various writing styles and domains beyond the typical news and generic data, consisting 1,947 unique sentences selected for their lexical and structural diversity. We conducted automatic reference-based, automatic reference-free, and manual evaluations. Our thorough analysis reveals significant limitations in these LLMs,

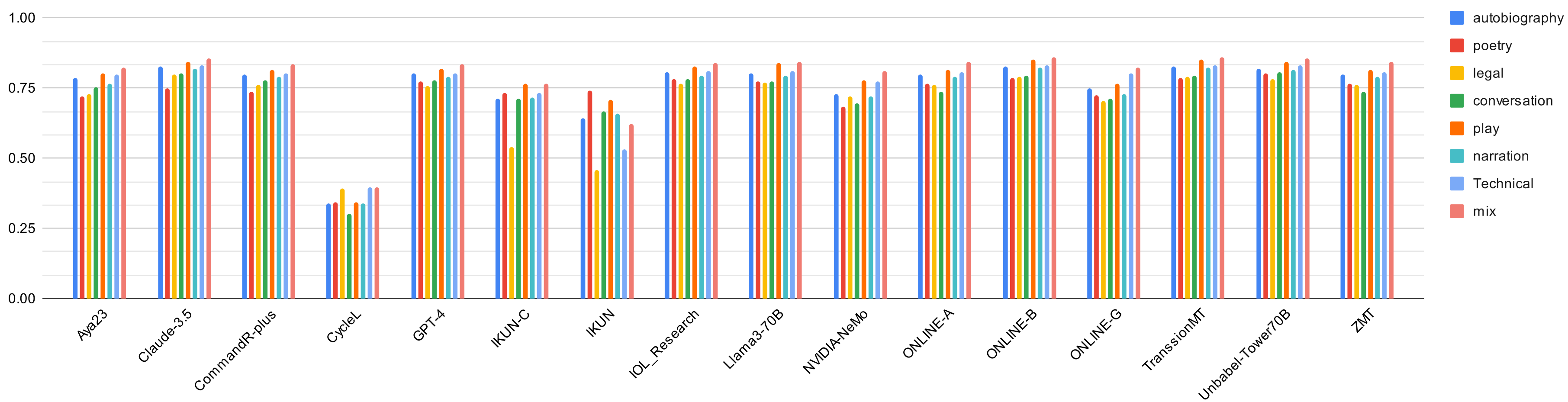

Figure 1: System-wise plots of average COMET-KIWI Scores for each category.

	reference-free	reference-based				
System	**COMET-KIWI**	BLEU	chrF	MEE4	BERTScore	COMET
TransionMT	0.815 (1)	68.399 (1)	81.577 (1)	0.903 (1)	0.942 (1)	0.835 (1)
Claude-3.5	0.815 (1)	43.321 (3)	66.385 (3)	0.85 (3)	0.898 (3)	0.803 (3)
ONLINE-B	0.814 (2)	67.733 (2)	80.768 (2)	0.898 (2)	0.933 (2)	0.83 (2)
Unbabel-Tower70B	0.809 (3)	38.634 (6)	62.811 (6)	0.842 (5)	0.886 (5)	0.799 (4)
Llama3-70B	0.791 (4)	34.164 (9)	58.612 (8)	0.83 (6)	0.874 (7)	0.767 (5)
IOL_Research	0.79 (5)	32.991 (10)	57.244 (10)	0.825 (8)	0.869 (8)	0.765 (6)
ZMT	0.785 (6)	42.277 (5)	65.614 (5)	0.843 (4)	0.893 (4)	0.75 (9)
ONLINE-A	0.785 (6)	42.324 (4)	65.637 (4)	0.843 (4)	0.893 (4)	0.75 (9)
GPT-4	0.785 (6)	31.795 (11)	57.227 (11)	0.826 (7)	0.868 (9)	0.755 (8)
CommandR-plus	0.785 (6)	29.088 (12)	54.918 (12)	0.816 (10)	0.858 (10	0.757 (7)
Aya23	0.761 (7)	27.938 (13)	53.473 (13)	0.81 (11)	0.852 (11)	0.728 (10)
ONLINE-G	0.735 (8)	35.952 (7)	60.861 (7)	0.825 (8)	0.875 (6)	0.669 (12)
NVIDIA-NeMo	0.734 (9)	34.635 (8)	57.977 (9)	0.821 (9)	0.868 (9)	0.689 (11)
IKUN-C	0.658 (10)	10.89 (15)	38.711 (14)	0.693 (12)	0.752 (12)	0.591 (13)
IKUN	0.574 (11)	12.181 (14)	36.159 (15)	0.657 (13)	0.731 (13)	0.546 (14)
CycleL	0.366 (12)	1.77 (16)	16.476 (16)	0.347 (14)	0.665 (14)	0.33 (15)

Table 2: System-wise ranking based on reference-free and reference-based metrics. Top 3 are highlighted in bold. Ranks are mentioned in brackets. The rows are colour coded highlighting the top scores in green and low scores in red.

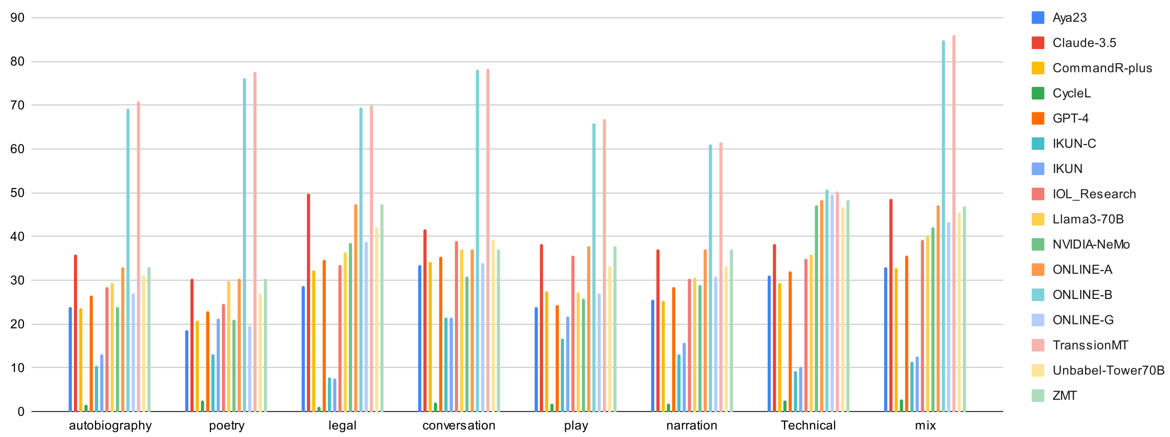

Figure 2: Category-wise plots of average BLEU Scores for all the submitted MT systems.

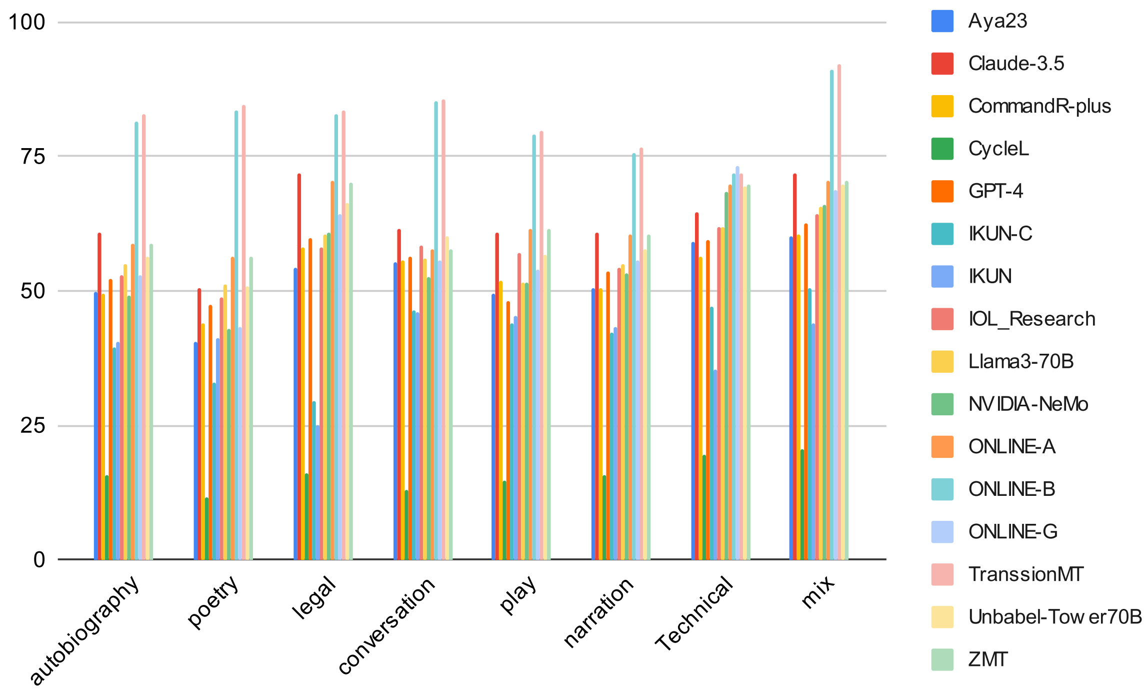

Figure 3: Category-wise plots of average chrF Scores for all the submitted MT systems.

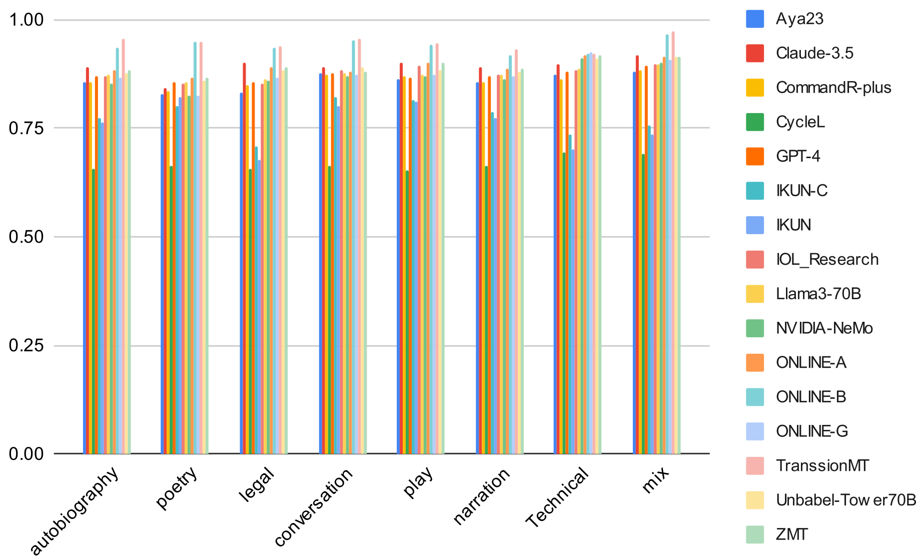

Figure 4: Category-wise plots of average BERTScore Scores for all the submitted MT systems.

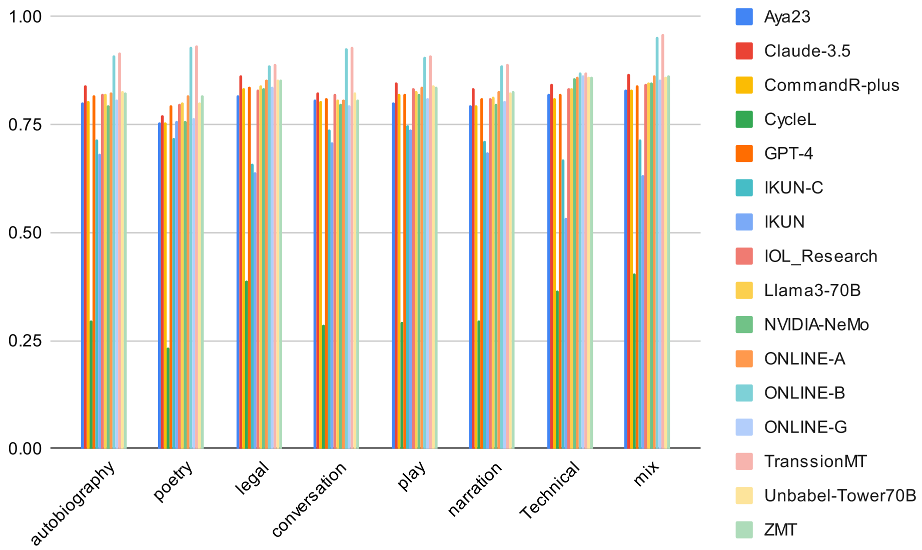

Figure 5: Category-wise plots of average MEE4 Scores for all the submitted MT systems.

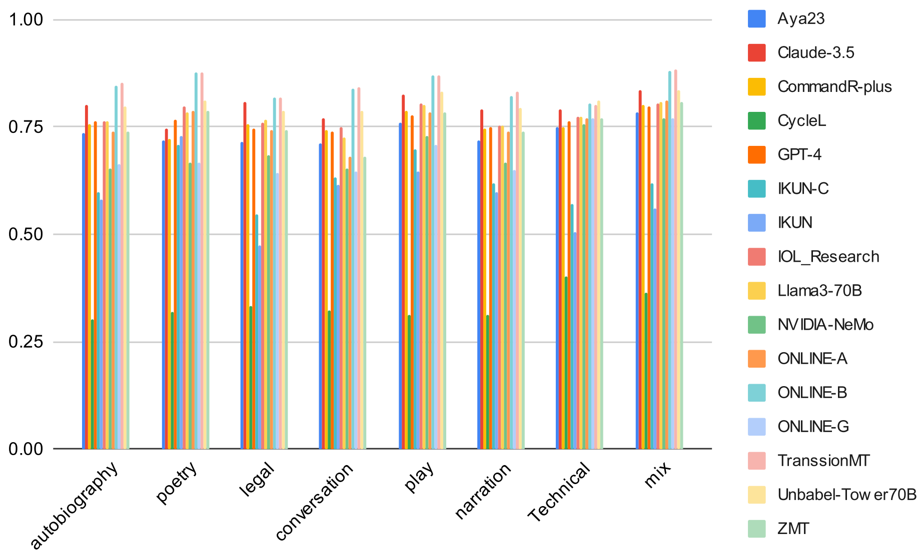

Figure 6: Category-wise plots of average COMET Scores for all the submitted MT systems.

particularly in translating poetry, conversational, and legal texts. Additionally, our manual review uncovered issues such as incorrect word choices, spelling errors, and poor handling of named entities. Despite their advancements, these LLMs show notable weaknesses in handling diverse and complex linguistic contexts. This highlights the need for continued refinement and broader training data to improve their performance across a wider range of text types and domains.

References

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task:
the LLM era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024b. Preliminary wmt24 ranking of general mt systems and llms.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian MÃ¶ller. 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 : Iiit 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.

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*.