Evaluation of LLM for English to Hindi Legal Domain Machine Translation Systems

Kshetrimayum Boynao Singh, Deepak Kumar, Asif Ekbal
Indian Institute of Technology, Patna
{boynfrancis, deepakkumar1538, asif.ekbal}@gmail.com

Abstract

The study critically examines various Machine Translation systems, particularly focusing on Large Language Models, using the WMT25 Legal Domain Test Suite for translating English into Hindi. It utilizes a dataset of 5,000 sentences designed to capture the complexity of legal texts based on word frequency ranges from 5 to 54. Each frequency range contains 100 sentences, collectively forming a corpus that spans from simple legal terms to intricate legal provisions. Six metrics were used to evaluate the performance of the system: BLEU, METEOR, TER, CHRF++, BERTScore and COMET. The findings reveal diverse capabilities and limitations of LLM architectures in handling complex legal texts. Notably, Gemini-2.5-Pro, Claude-4, and ONLINE-B topped the performance charts in terms of human evaluation, showcasing the potential of LLMs for nuanced translation. Despite these advances, the study identified areas for further research, especially in improving robustness, reliability, and explainability for use in critical legal contexts. The study also supports the WMT25 subtask focused on evaluating the weaknesses of large language models (LLMs). The dataset and related resources are publicly available at https://github.com/helloboyn/WMT25-TS

1 Introduction

Machine Translation (MT) has evolved from basic rule-based and statistical approaches to advanced neural network models, with recent advancements driven by Large Language Models (LLMs) that utilize extensive pretraining datasets and transformer architectures. (Vaswani et al., 2017) The legal domain poses significant challenges for MT, requiring precise handling of context-dependent terminology (Appicharla et al., 2025), complex sentence structures, and the accurate conveyance of cultural and jurisdictional nuances due to varying legal systems. This complexity surpasses that found in general language translation, making high levels of lexical accuracy, logical coherence, and syntactic fidelity essential for effective legal translation.

The text highlights the crucial importance of accuracy and fidelity in legal document translation due to its high-stakes nature. It emphasizes that even minor mistranslations can lead to serious legal and financial consequences such as contractual disputes and judicial errors. Therefore, precise legal translation is essential to support international legal cooperation, manage cross-border litigation, provide equitable access to justice for non-native speakers, and make legal information accessible to a wider audience.(WMT)[1] The series has consistently served as a pivotal platform, instrumental in benchmarking progress and driving innovation within the machine translation research community across a diverse array of language pairs and domains. WMT25 (Kocmi et al., 2024) continues this vital tradition, offering meticulously designed specialized test suites that push the boundaries of current MT technologies and identify areas for future breakthroughs. This paper specifically focuses on the WMT25 Legal Domain Test Suite for English to Hindi [2], embarking on an in-depth investigation into how various LLM-based MT systems perform when compared to more traditional and established hybrid approaches. Our overarching objective is to provide a comprehensive and nuanced analysis of their efficacy in this demanding domain, meticulously identifying the top performing contenders and critically discussing the broader implications of our findings for the future trajectory and practical application of legal machine translation systems, including considerations for deployment and ethical use.

[1] https://www2.statmt.org/wmt25/testsuite-subtask.html
[2] https://github.com/wmt25testsuite/wmt25

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 823–833
November 8–9, 2025.

2 Related Work

The WMT shared tasks have consistently been a primary driving force behind significant advancements in Machine Translation research, fostering innovation and providing a standardized, competitive benchmark for evaluating system performance (Gain et al., 2022). Previous WMT editions, notably those from WMT24 (as evidenced by a series of influential papers such as (Freitag et al., 2024) to (Ármannsson et al., 2024)), have unequivocally showcased the increasing dominance and sophistication of neural MT (NMT) models (Appicharla et al., 2021). These works have meticulously detailed a wide array of architectural innovations, including the widespread adoption of transformer networks, advanced training methodologies such as back-translation and knowledge distillation, and impressive performance gains across diverse language pairs and specialized domains. Key themes emerging from this extensive body of research include the paramount importance of large-scale pre-training on vast textual corpora to learn robust linguistic representations, the efficacy of fine-tuning models on domain-specific data (Bhattacharjee et al., 2024; Moslem et al., 2022) to enhance specialized vocabulary and stylistic nuances (e.g., legal jargon, formal tone), and the continuous refinement of robust evaluation metrics to more accurately reflect human judgment of translation quality. Techniques like data augmentation (e.g., synthetic data generation), transfer learning (Singh et al., 2023a) from high-resource to low-resource languages, and the development of more efficient attention mechanisms have been central to these advancements, enabling NMT models (**?**) to capture intricate linguistic patterns and contextual dependencies with greater precision than their predecessors.

The rapid advancement of Large Language Models (LLMs) like GPT, Gemini, and Claude has significantly transformed Machine Translation (MT) research by challenging existing paradigms. These models, originally crafted for general language tasks, have shown impressive zero-shot and few-shot translation skills due to their training on vast, diverse datasets. They excel in capturing complex semantics and context, making them promising for specialized fields such as legal translation, where precision and adherence to terminology are critical. Research is actively exploring their (Gain et al., 2021) adaptation for specific MT tasks, and it often outperforms traditional Neural Machine Translation (NMT) (Singh et al., 2023b, 2024) models in challenging situations, such as low-resource languages and complex linguistic features (Manakhimova et al., 2024). Nonetheless, adapting LLMs to specific domains poses challenges, such as the risk of losing general linguistic knowledge, generating plausible but incorrect legal outputs, and maintaining strict legal fidelity without creative rephrasing.

3 Methodology

3.1 Dataset

The research uses a specialized Legal Domain Test Suite for WMT25 to evaluate translation systems from English to Hindi. This dataset consists of 5000 sentences derived from authentic legal documents on Table 1, reflecting the complexity and diversity of legal texts. It includes sentences varying in length from about 5 words to 54 words.

Word-Count	Sentences	Eng-Token	Hin-Token
5–15	1,600	20000	23046
16–35	1,700	49300	40798
36–54	1,700	78199	69627
Total	5,000	147499	133471

Table 1: Corpus statistics for the English and Hindi legal dataset by word count range.

The variation enables testing of systems' adaptability and robustness across different linguistic complexities, from precise legal terms to complex legal judgments. The dataset tests systems on their ability to handle the unique vocabulary, tone, and structure of legal language, ensuring accurate translation that maintains legal intent and avoids ambiguity.

3.2 Automatic Evaluation Metrics

To deliver a comprehensive and multifaceted (Chen et al., 2023) assessment of the translation quality generated by the various systems, six well-established and complementary automatic evaluation metrics were rigorously utilized. The choice of these metrics was deliberate, with the aim of capturing diverse aspects of translation quality: lexical overlap, semantic equivalence, and character-level accuracy, all of which are essential for the rigorous legal domain.

- **BLEU (Bilingual Evaluation Understudy):** The text examines the BLEU metric used

in machine translation evaluation, noting its strengths in precision, simplicity, and efficiency, which contribute to its widespread adoption. However, it also identifies BLEU's (Papineni et al., 2002) limitations in assessing translation fluency, grammatical correctness, and semantic adequacy, as it emphasizes lexical similarity over meaning. This focus can result in high scores for outputs closely matching references while neglecting valid paraphrases or alternative translations, which is particularly problematic in fields like legal translation, where multiple correct phrasings may exist.

- **METEOR (Metric for Evaluation of Translation With Explicit Ordering):**METEOR improves upon BLEU by using linguistic features such as word stemming, synonymy matching, and chunk-based alignment to better assess translation quality. By focusing on fluency and semantic adequacy, METEOR (Banerjee and Lavie, 2005) aligns more closely with human evaluations. It handles lexical and syntactic variations while penalizing reordering errors, making it particularly effective for domains requiring high semantic precision, such as legal texts.

- **TER (Term Error Rate):** TER, or Translation Edit Rate (Snover et al., 2006), is a metric used to evaluate the quality of machine translation (MT) by measuring the number of edits required to transform an MT output into a perfect, human-quality reference translation. These edits typically include insertions, deletions, substitutions, and shifts of words or phrases. A lower TER score indicates that fewer edits were necessary, meaning the machine translation is closer to the human reference and, thus, of higher quality. Conversely, a higher TER score signifies that many edits were needed, indicating poorer quality of machine translation that deviates significantly from the human standard.

- **CHRF++ (Character n-gram F-score):** The text discusses the CHRF++ (Popović, 2017) metric, which evaluates translation quality by computing the F-score of character n-grams between candidate and reference translations. It is highly regarded for its strong correlation with human judgments and its ability to handle morphological variations and out-of-vocabulary words effectively. CHRF++ is particularly suited for languages with rich morphological systems, such as Hindi, as it captures subtle character-level differences crucial for accurate translations. This makes it especially valuable in legal translation, where precise fidelity and a lack of ambiguity are critical.

- **BERTScore (Bidirectional Encoder Representations from Transformers Score):** BERTScore is a metric that assesses the quality of AI-generated text by measuring the semantic similarity between the generated content and reference texts (Zhang et al., 2020). Unlike traditional metrics that rely on exact word overlap, BERTScore uses the BERT language model to create embeddings of words and sentences, capturing their contextual meaning. A high BERTScore suggests that the generated text successfully conveys similar information and meaning to the reference, indicating good quality, while a low score points to significant differences in meaning, reflecting poor generation quality.

- **COMET (Crosslingual Optimized Metric for Evaluation of Translation):** COMET is an AI-based metric designed to evaluate the quality of machine translations by assessing alignment with high-quality human translations and considering the original source sentence for context. It uses a neural network model trained to align with human judgments, making it more robust and reliable than traditional rule-based metrics. High COMET (Rei et al., 2020) scores indicate superior translations, while low scores suggest poor translation quality.

3.3 Systems Evaluated

The WMT25 Legal Domain Test Suite served as a platform to evaluate a wide range of Machine Translation (MT) systems, reflecting the latest advancements in the field. It featured both proprietary and open-source Large Language Models (LLMs), such as Gemini-2.5-Pro, Claude-4, GPT-4.1, Llama, Mistral, Gemma, and Qwen, showcasing diverse architectures and scales. The evaluation (Manakhimova et al., 2023) also included traditional neural machine translation systems that

have been refined through years of domain adaptation and specialized training, as well as innovative hybrid approaches that incorporate rule-based systems or statistical models with neural components. These evaluations highlight the progress and state-of-the-art techniques in neural language processing, particularly in handling translations in the legal domain.

The paper conducts a comparative analysis of various translation systems—commercial, academic, and open-source large language models specifically within the legal domain. It assesses different model sizes and architectures, exploring the impact of scale, design, and training on translation quality and robustness. The study identifies the strengths and weaknesses of these systems, providing insights for future improvements and applications of machine translation in specialized areas.

4 Results and Observation

The performance of the systems evaluated on the WMT25 Legal Domain Test Suite (English to Hindi) is meticulously summarized below, directly derived from the provided evaluation results:

4.1 Overall Performance and LLM Dominance

The study highlights that LLM-based systems, especially Gemini-2.5-Pro, excel in machine translation within the legal domain, outperforming others across various metrics such as BLEU, METEOR, TER, CHRF++, BERTScore, and COMET. This is due to its extensive pre-training and specialized fine-tuning on legal documents, enhancing its handling of legal terminology and nuances. Other LLMs, such as Claude-4 and Llama-4-Maverick, also demonstrate strong performance, signaling a shift towards general-purpose models that outperform traditional systems in legal translation tasks. This shift offers legal professionals more efficient translation tools but also raises concerns about transparency and potential errors in precision-critical contexts.

4.2 Comparison with Non-LLM Systems

The text highlights that while Large Language Models (LLMs) are dominant in translation tasks, non-LLM or hybrid systems like ONLINE-B and TranssionTranslate also show competitive performance. Traditional Neural Machine Translation (NMT) systems, particularly those optimized for specific domains and language pairs, can achieve state-of-the-art results, offering computational efficiency and control over translation behavior. Hybrid approaches that integrate multiple translation paradigms enhance robustness and accuracy by combining the strengths of various methods, such as rule-based systems and statistical models. These alternatives are particularly viable in resource-limited settings demanding precision, such as legal translation, where accuracy and consistency are crucial.

4.3 Metric-Specific Observations

The evaluation of machine translation metrics highlights the varied strengths and weaknesses of different systems. The BLEU score primarily captures n-gram overlap, but its ability to assess semantic meaning and fluency is limited. Metrics such as BLEU, METEOR, TER, CHRF++, BERTScore, and COMET show different levels of effectiveness, with top models like Gemini-2.5-Pro, ONLINE-B, TranssionTranslate, ONLINE-G, and Claude-4 performing well overall (see Table 2). Gemini-2.5-Pro excels in precision and quality, while METEOR, focusing on semantic and structural accuracy, showcases ONLINE-G's strength. CHRF++ correlates well with translation quality through its character-level focus. A significant performance gap exists between leading and lower-tier models, with weaker systems performing poorly across metrics, indicating insufficient specialization in translation tasks. These metrics emphasize the strengths and constraints of each system in accurately translating legal texts.

4.4 Analysis by Sentence Length

The WMT25 Legal Domain Test Suite evaluates system performance over sentence lengths ranging from 5 to 54 words to assess robustness and adaptability to linguistic complexities. Although the dataset offers an aggregate performance overview, it lacks detailed scores segmented by sentence length. Such a breakdown is important for understanding how language models manage various contextual complexities and for identifying strengths or weaknesses related to sentence length. A more thorough evaluation would categorize sentences and assess performance within each segment.

Small Sentences (5–15 words) The analysis of short legal sentences shows that multiple machine translation (MT) systems excel in this area due to their minimal syntactic complexity. Systems such

Rank	**LLM System**	**BLEU**	**METEOR**	**TER**	**CHRF++**	**BERTScore**	**COMET**
1	Gemini-2.5-Pro	33.35	53.91	55.66	60.95	88.49	72.27
2	ONLINE-B	31.77	52.37	55.69	57.81	87.44	70.96
3	TranssionTranslate	31.65	52.42	55.71	57.83	87.55	71.01
4	ONLINE-G	31.22	57.30	52.07	55.20	86.56	67.37
5	Claude-4	31.09	52.75	57.87	58.46	87.71	70.99
6	Llama-4-Maverick	28.46	54.44	57.15	54.70	86.65	69.86
7	NLLB	27.87	51.55	57.45	53.38	86.11	68.16
8	hybrid	26.97	50.19	62.42	55.47	86.67	71.20
9	DeepSeek-V3	26.65	49.11	62.24	53.94	86.33	69.92
10	GPT-4.1	26.04	48.51	63.18	53.58	86.22	70.23
11	TowerPlus-9B	25.77	48.02	63.58	52.08	85.85	68.79
12	HYT	25.58	48.70	63.58	54	85.93	71.02
13	TMTHY	25.58	48.70	63.58	54	85.93	71.02
14	Shy	25.58	48.70	63.58	54	85.93	71.02
15	CommandA	24.24	47.85	65.11	51.70	85.64	68.85
16	Gemma-3-27B	23.80	46.22	65.91	51.49	85.35	68.96
17	TowerPlus-72B	23.53	46.57	65.42	50.24	85.32	67.76
18	Mistral-Medium	23.32	46.03	66.56	51.09	85.17	68.76
19	Qwen3-235B	22.91	45.75	66.96	50.33	85.14	68.20
20	EuroLLM-22B	22.18	44.72	67.39	48.94	84.76	67.40
21	EuroLLM-9B	21.52	44.65	68.68	48.09	84.35	66.23
22	Gemma-3-12B	21.51	43.81	68.04	49.38	84.47	68.03
23	IR-MultiagentMT	21.42	43.78	67.26	48.26	84.52	68.08
24	CommandA-MT	21.05	44.60	68.78	49.34	84.71	69.94
25	AyaExpanse-32B	20.50	43.75	69.00	47.50	84.28	66.91
26	UvA-MT	19.79	43.46	70.83	47.59	84.29	68
27	IRB-MT	17.26	39.92	84.85	44.15	83.63	67.21
28	AyaExpanse-8B	16.70	39.36	73.32	43.74	83.07	65.30
29	Llama-3.1-8B	15.21	38.54	74.68	42.20	82.03	63.19
30	GemTrans	15.16	38.68	80.62	43.06	82.05	67.76
31	CommandR7B	12.42	34.56	84.17	37.82	81.11	61.44
32	Qwen2.5-7B	8.75	27.88	87.18	33.22	78.52	53.05
33	Mistral-7B	3.03	20.65	177.39	23.19	71.04	41.79
34	Wenyiil	2.68	5.96	107.66	2.20	69.73	41.57
35	Yolu	2.68	5.96	107.66	2.20	69.73	41.57
36	Algharb	2.68	5.96	107.66	2.20	69.73	41.57
37	MMMT	2.68	5.96	107.66	2.20	69.73	41.57

Table 2: The table presents a performance comparison of various machine translation systems, including large language models (LLMs) and traditional neural machine translation (NMT) systems. We evaluate the systems using BLEU (Figure 1), METEOR (Figure 2), TER (Figure 3), CHRF++ (Figure 4), BERTScore (Figure 5), and COMET (Figure 6). The systems are ranked by their BLEU scores, with Gemini-2.5-Pro achieving the highest score, followed by ONLINE-B and TranssionTranslate. The results highlight the varying levels of translation quality across different models.

Rank	LLM System	Human Score%
1	Gemini-2.5-Pro	84.67
2	Claude-4	82.00
3	ONLINE-B	81.67
4	TowerPlus-9B	81.33
5	Llama-4-Maverick	81.00
6	GPT-4.1	80.67
7	TranssionTranslate	80.33
8	Qwen3-235B	80.00
9	Mistral-Medium	79.33
10	EuroLLM-22B	79.33
11	NLLB	78.67
12	HYT	78.67
13	ONLINE-G	78.33
14	DeepSeek-V3	78.33
15	TMTHY	78.33
16	CommandA	78.33
17	hybrid	78.00
18	Gemma-3-27B	78.00
19	Shy	77.67
20	TowerPlus-72B	76.74

Table 3: Human evaluation results for the top 20 BLEU-ranked systems on the English→Hindi legal domain dataset. Scores are averaged over two expert annotators.

as ONLINE-G, Llama-4-Maverick, and Claude-4 are identified as top performers, providing accurate and fluent translations. However, these models may face challenges with highly specialized legal jargon or rare terms not extensively covered in their training data, which could impact the precision required for translating legal documents.

Medium Sentences (16–35 words) The text discusses the challenges faced by translation models when dealing with medium-length sentences, which often contain complex structures, such as multiple clauses and conditional statements. These sentences require maintaining logical coherence and resolving anaphora for accurate translation. The models Gemini-2.5-Pro, TranssionTranslate, and ONLINE-B were identified as the most effective in managing these intricacies. Despite the models' suitability for this task, largely due to their transformer-based architectures, their performance showed a slight decline with shorter sentences, indicating that increased complexity still poses a risk of increased errors.

Large Sentences (36–54 words) The primary challenge for the 41 machine translation (MT) systems was translating long sentences characterized by legal jargon and numerous clauses. These sentences posed difficulties in maintaining contextual integrity, causing a significant drop in performance metrics such as BLEU and COMET. EuroLLM-22B, CommandA, and NLLB systems performed slightly better in mitigating this drop. The consistent performance decline underscores the increased risk of "hallucination," context loss, and ambiguity with longer sentences, marking it as a key area for future research and development in MT technology.

Overall System Performance

The evaluation ranks systems in Table 2 based on their robustness across different sentence lengths. Gemini-2.5-Pro is identified as the leading system, showing consistently high performance and the ability to manage various sentence complexities. It is followed by hybrid, with Shy, HYT, and TMTHY tied for third place. The assessment highlights that a system's overall performance is best measured by its consistent quality across all sentence lengths rather than excelling in just one category.

4.5 Human Analysis

In this section, we present the human evaluation conducted on the top 20 systems selected based on their BLEU scores in Table 3 . After identifying these top-performing systems, we carried out a detailed human evaluation specifically within the legal domain. Two linguistic experts proficient in both English and Hindi evaluated the translations of each system. They rated the outputs on a scale from 1 to 100, focusing on both adequacy (how accurately the translation conveyed the source meaning) and fluency (how natural and readable the translation was in Hindi).

We then averaged the scores from both evaluators to produce a final human evaluation score for each system. This human evaluation provides a nuanced measure of translation quality that complements the BLEU-based rankings, helping us identify systems that perform well in real-world, domain-specific scenarios.

5 Conclusion

The WMT25 Legal Domain Test Suite for English to Hindi Machine Translation highlights significant progress made by Large Language Models (LLMs) in specialized domain translation. Notably, Gemini-2.5-Pro excels, outperforming others across multiple evaluation metrics and emphasizing the potential of advanced LLM architectures with domain-specific pre-training or fine-tuning. These

models effectively handle complex legal language and structures, showcasing their sophisticated linguistic capabilities. While LLMs show dominance, traditional and hybrid MT systems also demonstrate competitiveness, indicating their continued relevance. The study underscores the importance of model scale, architecture, and domain adaptation for success in legal MT. It suggests that LLMs will play an increasingly central role in legal translation, advancing accuracy and efficiency. However, ongoing innovations across MT paradigms are needed to balance performance with reliability and ethical considerations, given the high stakes of errors in the legal domain.

Acknowledgement

The authors express their sincere gratitude to the COIL-D Project under Bhashini, funded by MeitY, for their support and resources, which were instrumental in the successful completion of this research.

References

Ramakrishna Appicharla, Asif Ekbal, and Pushpak Bhattacharyya. 2021. EduMT: Developing machine translation system for educational content in Indian languages. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 35–43, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Ramakrishna Appicharla, Baban Gain, Santanu Pal, and Asif Ekbal. 2025. Beyond the sentence: A survey on context-aware machine translation with large language models. *arXiv preprint arXiv:2506.07583*.

Bjarki Ármannsson, Hinrik Hafsteinsson, Atli Jasonarson, and Steinþór Steingrímsson. 2024. Killing two flies with one stone: An attempt to break LLMs using English-Icelandic idioms and proper names. In *Proceedings of the Ninth Conference on Machine Translation*, pages 451–458, Miami, Florida, USA. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Soham Bhattacharjee, Baban Gain, and Asif Ekbal. 2024. Domain dynamics: Evaluating large language models in English-Hindi translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 341–354, Miami, Florida, USA. Association for Computational Linguistics.

Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang, Zongyao Li, Jiaxin Guo, Ning Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. Multifaceted challenge set for evaluating machine translation performance. In *Proceedings of the Eighth Conference on Machine Translation*, pages 217–223, Singapore. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Baban Gain, Ramakrishna Appicharla, Soumya Chennabasavaraj, Nikesh Garera, Asif Ekbal, and Muthusamy Chelliah. 2022. Low resource chat translation: A benchmark for Hindi–English language pair. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 83–96, Orlando, USA. Association for Machine Translation in the Americas.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021. Experiences of adapting multimodal machine translation techniques for Hindi. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 40–44, Online (Virtual Mode). INCOMA Ltd.

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, and 2 others. 2024. Preliminary wmt24 ranking of general mt systems and llms. *Preprint*, arXiv:2407.19884.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 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.

Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of*

the Ninth Conference on Machine Translation, pages 355–371, Miami, Florida, USA. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2022. Domain-specific text generation for machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30, Orlando, USA. Association for Machine Translation in the Americas.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Ningthoujam Justwant Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023a. A comparative study of transformer and transfer learning MT models for English-Manipuri. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 791–796, Goa University, Goa, India. NLP Association of India (NLPAI).

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023b. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.

Ningthoujam Justwant Singh, Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Sanjita Phijam, and Thoudam Doren Singh. 2024. WMT24 system description for the MultiIndic22MT shared task on Manipuri language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 797–803, Miami, Florida, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Evaluation Metrics and Dataset Segmentation

The appendix details benchmarking results for English to Hindi legal domain machine translation systems, evaluated using several metrics, including BLEU, METEOR, TER, chrF++, BERTScore and COMET. The performance of each system is analyzed across three sentence length categories—small, medium, and large—as well as an overall aggregate. Consistently high-performing systems are identified, along with those that rank lower in performance. The results show consistent top-performing systems, such as **Gemini-2.5-Pro**, **Claude-4**, and **TranssionTranslate**, while systems like **MMMT**, **Wenyill**, and **Yolu** consistently rank among the lowest. This segmentation provides deeper insights into system robustness across varying sentence complexities. Furthermore, it highlights the sensitivity of different models to sentence length, revealing cases in which certain systems degrade significantly with longer inputs. These findings underscore the importance of evaluating MT systems with controlled test suites to ensure reliability in specialized domains, such as legal translation.

A.1 Dataset Segmentation

The test data is divided into four buckets based on sentence length:

- **Small (5–15 words):** Marked in green.
- **Medium (16–35 words):** Marked in yellow.
- **Large (36–54 words):** Marked in blue.
- **Overall:** Aggregate results across all lengths, marked in red.

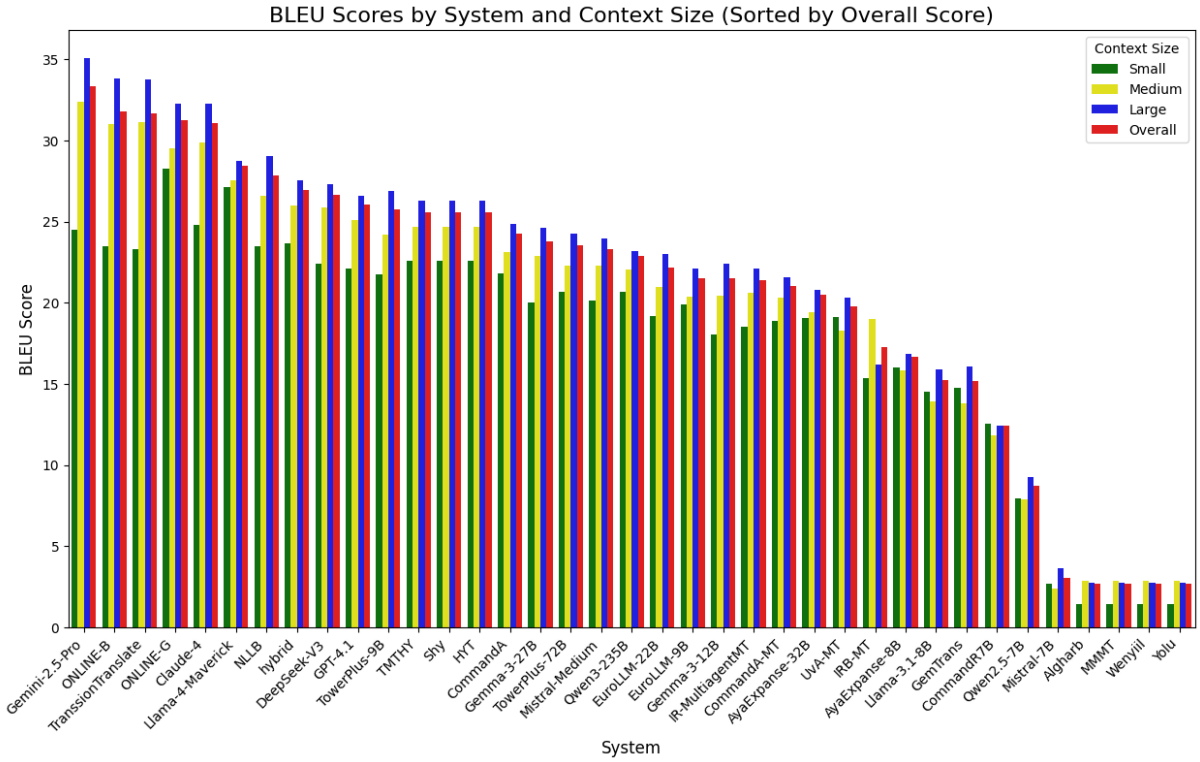

Figure 1: The bar chart displays the **BLEU scores** for various translation systems, broken down by **small, medium, and large context sizes**, as well as an **overall** score. The systems are ranked by their overall score, with **Gemini-2.5-Pro** and **Claude-4** having the highest overall BLEU score and **Wenyiil**, and **Yolu** having the lowest.

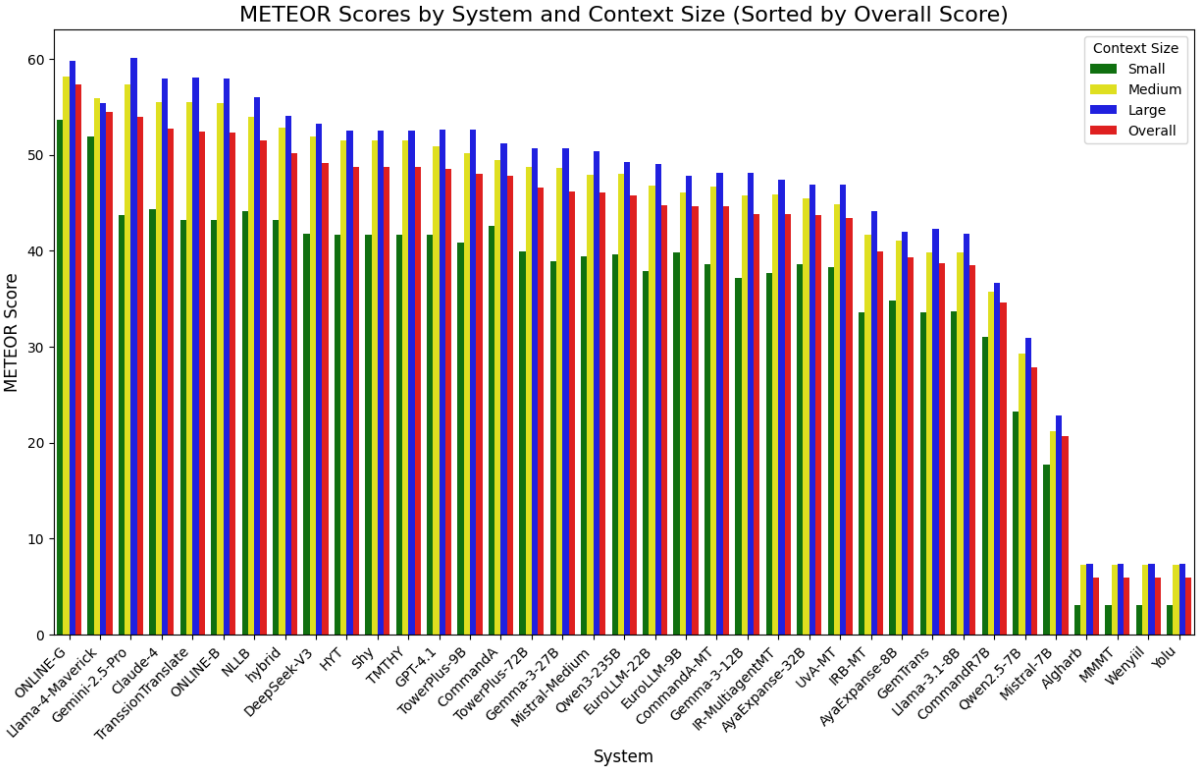

Figure 2: The bar chart shows **METEOR scores** for various translation systems, sorted by their **overall** performance. **ONLINE-G**, **Llama-4-Maverick** and **Gemini-2.5-Pro** have the highest scores, while **MMMT, Wenyiil, and Yolu** have the lowest.

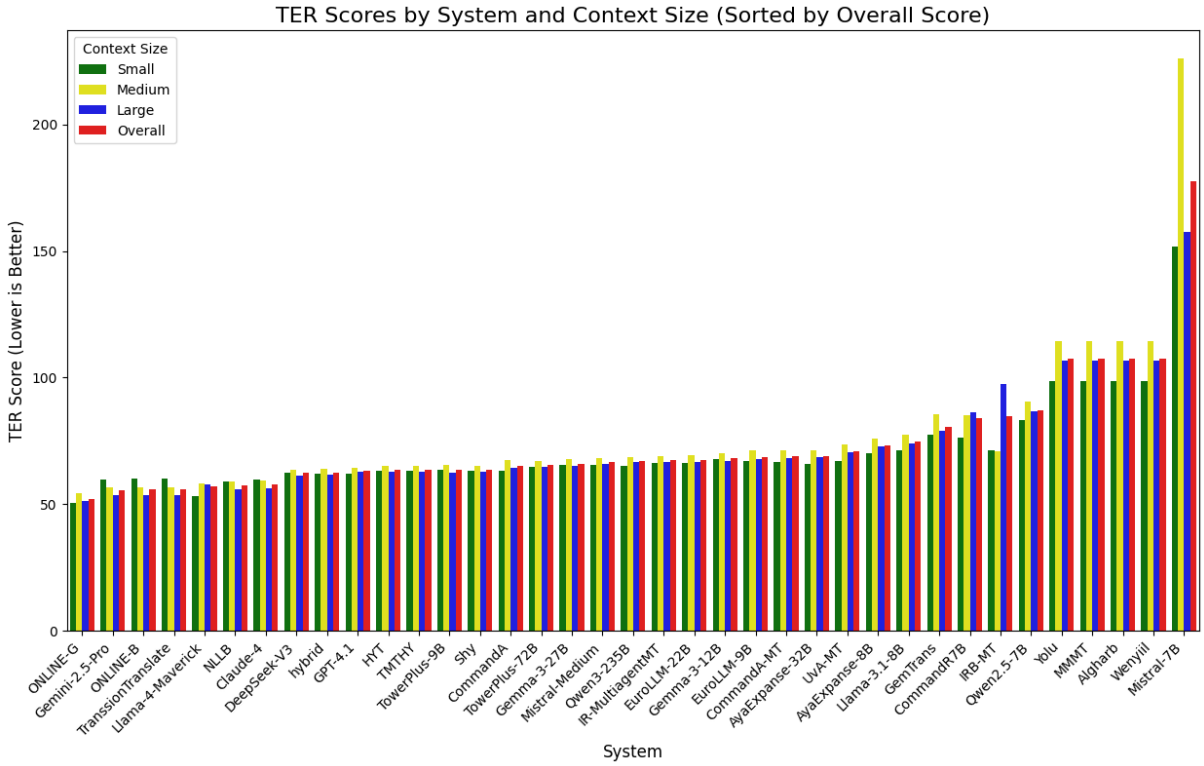

Figure 3: **TER scores** for various translation systems, sorted by their overall performance. **Lower scores are better**. **ONLINE-G, Gemini-2.5-Pro, and ONLINE-B** have the best performance, while **Mistral-7B, Wenyiil, and MMMT** have the worst.

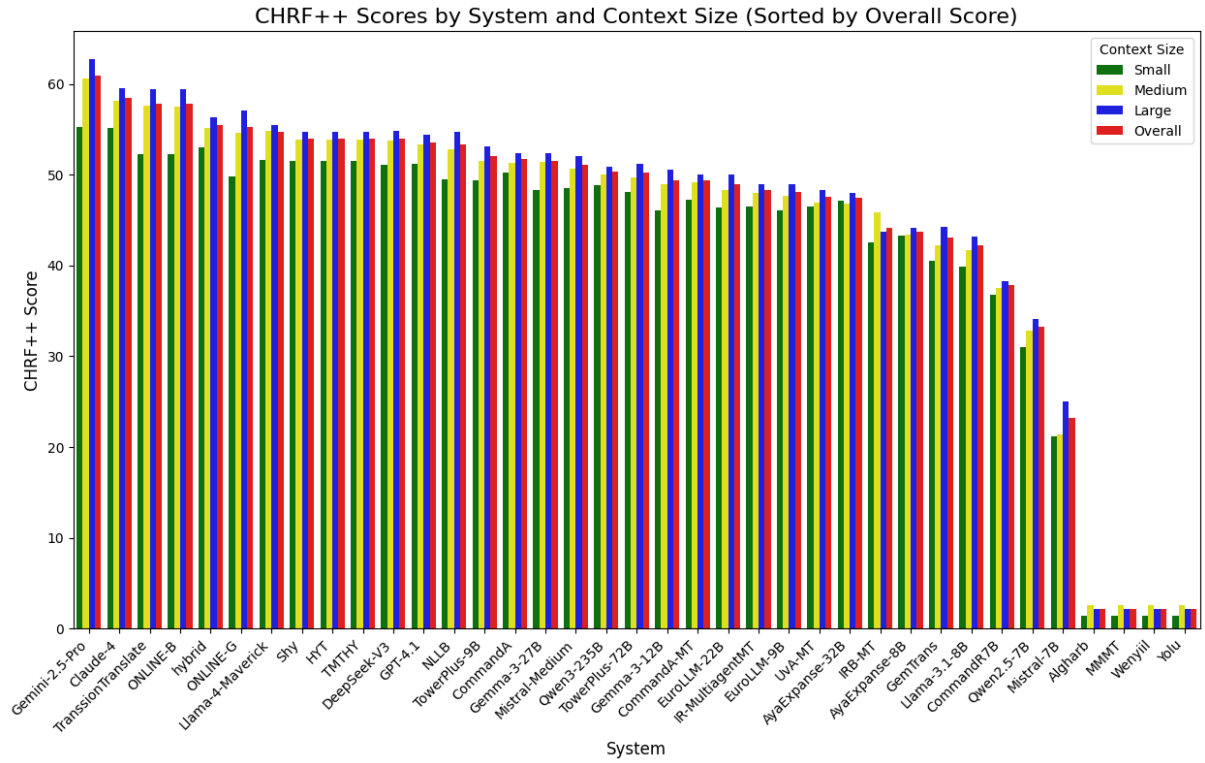

Figure 4: **CHRF++ scores** for various translation systems, sorted by their overall performance. **Higher scores are better**. **Gemini-2.5-Pro, Claude-4, and TranssionTranslate** have the best performance, while **MMMT, Wenyiil, and Yolu** have the worst.

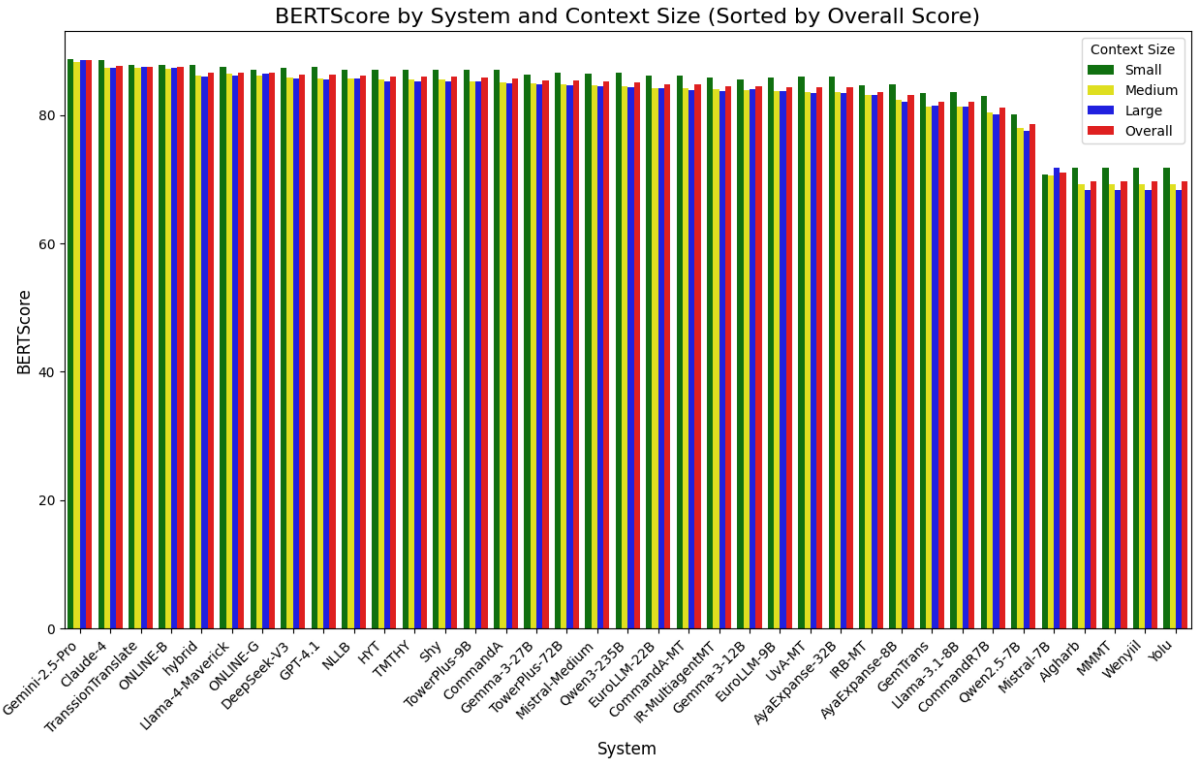

Figure 5: **BERTScore** for various translation systems, sorted by their overall performance. **Higher scores are better**. **Gemini-2.5-Pro, Claude-4, and TranssionTranslate** have the best performance, while **MMMT, Wenyiil, and Yolu** have the lowest.

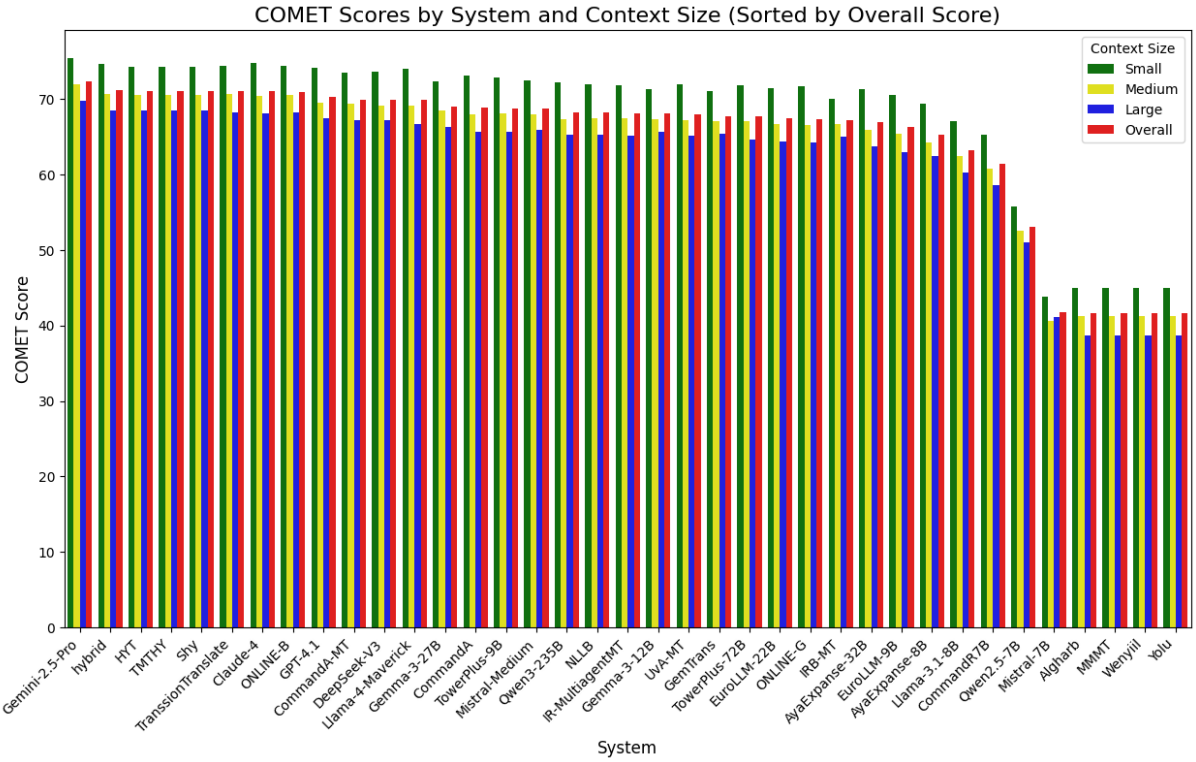

Figure 6: **COMET scores** for various translation systems, sorted by their overall performance. **Higher scores are better**. **Gemini-2.5-Pro, Hybrid, and HYT** have the best overall performance, while **MMMT, Wenyiil, and Yolu** have the lowest.