Domain Dynamics: Evaluating Large Language Models in English-Hindi Translation

Soham Bhattacharjee , Baban Gain
Indian Institute of Technology, Patna

Asif Ekbal
Indian Institute of Technology, Jodhpur
{sohambhattacharjeenghss,gainbaban,asif.ekbal}@gmail.com

Abstract

Large Language Models (LLMs) have demonstrated impressive capabilities in machine translation, leveraging extensive pre-training on vast amounts of data. However, this generalist training often overlooks domain-specific nuances, leading to potential difficulties when translating specialized texts. In this study, we present a multi-domain test suite, collated from previously published datasets, designed to challenge and evaluate the translation abilities of LLMs. The test suite encompasses diverse domains such as judicial, education, literature (specifically religious texts), and noisy user-generated content from online product reviews and forums like Reddit. Each domain consists of approximately 250-300 sentences, carefully curated and randomized in the final compilation. This English-to-Hindi dataset aims to evaluate and expose the limitations of LLM-based translation systems, offering valuable insights into areas requiring further research and development. We have submitted the dataset to WMT24 *Break the LLM* challenge. In this paper, we present our findings. We have made the code and the dataset publicly available at https://github.com/sohamb37/wmt24-test-suite.

1 Introduction

Machine translation (MT) (Bahdanau et al., 2016) has witnessed significant advancements with the advent of Large Language Models (LLMs) (et al., 2024a,b), which leverage extensive pretraining on massive datasets to achieve high performance across various language pairs (Alves et al., 2024; Zhu et al., 2024; Zhang et al., 2023). Despite their remarkable generalization capabilities, LLMs often struggle with domain-specific texts due to a lack of targeted training on such specialized content (Robinson et al., 2023; Jiao et al., 2023; Hendy et al., 2023). Some LLMs (Workshop et al., 2023) generate good translation involving low-resource language when target language is English but not the other way around (Bawden and Yvon, 2023). These challenges are amplified when the domains involved are different from those of training data. This limitation poses a challenge for deploying MT systems in real-world applications where domain-specific accuracy is crucial.

To address this gap, we have collated this dataset that exposes the difficulties faced by LLM-based MT systems when dealing with domain-specific content. We have combined sentences from judicial, educational, religious, literature, and noisy user-generated content domains.

Each domain-specific subset comprises approximately 250-300 sentences, which are then randomized to form the final dataset. This dataset, focusing on the English-to-Hindi translation direction, aims to rigorously test the robustness and adaptability of LLM-based MT systems. By identifying the translation challenges specific to each domain, our study provides valuable insights for improving domain adaptation techniques in machine translation, ultimately contributing to more reliable and accurate MT solutions for specialized applications. Our contributions to the paper are as follows:

- We submit a diverse dataset consisting of six domains.
- We calculate the standard BLEU score as well as the state-of-the-art metric xCOMET-XXL to evaluate the translation quality.
- We perform a tiny scale manual evaluation of the translation outputs.

2 Related Works

Neural Machine Translation (NMT) has made significant progress, especially for high-resource languages, but translating low-resource languages remains a challenge. For example, the translation of Indic languages like Hindi is difficult due to the scarcity of high-quality parallel corpora. Multilingual models like IndicTrans (Ramesh et al., 2022) and IndicTrans2 (Gala et al., 2023) show performance improvements, yet domain-specific performance data is lacking.

For domain-specific augmentation, Moslem et al. (2022) used pre-trained language models to generate synthetic in-domain data through back translation for Arabic-English translation. In the low-resource context, Gain et al. (2022) explored English-Hindi translation in chat-based conversations, while Ramakrishna et al. (2023) introduced the EduMT dataset to enhance English-Hindi translations for educational content. Domain adaptation techniques have also been applied for specialized translations, such as Chemistry related and general English-Hindi texts (Joshi et al., 2020).

In the legal domain, recent studies like Briva-Iglesias et al. (2024) show that LLMs outperform Google Translate for legal texts, and Poudel et al. (2024) developed a custom dataset for English-Nepali legal translation. For the literary domain, NMT has been applied to German-English (Matusov, 2019), English-Slovene (Kuzman et al., 2019), and English-Turkish (Yirmibeşoğlu et al., 2023) translations, with mixed results on automatic versus human evaluation (Thai et al., 2022).

Noise robustness in NMT is critical, as noisy inputs can degrade translation quality. Studies like Khayrallah and Koehn (2018) explored noise effects, while Michel and Neubig (2018) introduced the MTNT dataset. Recent efforts used LLMs to filter noise and enhance NMT performance (Bolding et al., 2023).

Finally, NMT has also been applied to e-Commerce, particularly to translate product reviews. Gupta et al. (2022) focused on sentiment-preserving translations for English-Hindi, with other works such as Gupta et al. (2021) contributing to the field.

Ranathunga et al. (2023) provides a comprehensive survey of advancements in low-resource NMT, highlighting techniques and offering guidelines for further research. Building on this, Goyle et al. (2023) leveraged transfer learning and back-translation with the mBART model for low-resource languages, while Chowdhury et al. (2022) utilized transfer learning from English-Kannada, English-Gujarati, and English-Marathi models for Lambani, a low-resource tribal language. Additionally, they examined the impact of freezing specific encoder and decoder layers during training.

3 Dataset

Large Language Models (LLMs) excel in general machine translation but struggle with specialized domains. Our dataset includes English-Hindi bitext pairs from six critical domains, aiming to improve LLMs' translation accuracy in these areas, which is vital for advancing their capabilities.

3.1 Education domain

The education domain is crucial for knowledge dissemination, social development, and personal growth. Accurate translation in this field ensures broader access to educational materials, supporting multilingual learning and empowering non-native language communities. This helps reduce educational disparities and promotes inclusivity. Our dataset, sourced from EduMT (Appicharla et al., 2021), includes 330 English-Hindi sentence pairs, enhancing translation performance in education.

3.2 General domain

The general domain in our dataset is sourced from the IIT Bombay English-Hindi Parallel Corpus (Kunchukuttan et al., 2018), which includes diverse content like news, TED Talks, government websites, and Wikipedia. In essence, the general domain is itself composed of diverse mini domains, making translation a challenging task for MT systems. We randomly selected 500 English-Hindi pairs from this corpus.

3.3 Judicial domain

The judicial domain in our dataset is sourced from the IIT Patna Hindi-English Machine Aided Translation (HEMAT) training corpora, which is specifically designed for legal and judicial content. For this domain, we have included 325 sentences in our proposed dataset. Enhancing machine translation performance in the judicial domain is crucial, as it ensures that legal documents, court rulings, and other judicial materials are accurately translated.

3.4 Literature domain

The literature domain in our dataset includes 300 pairs, with 150 Quran verses from the Tanzil Project [1] and 150 Bible verses from the Bible Eudin Project, both sourced from the OPUS collection (Tiedemann, 2012). These texts present unique

[1] https://tanzil.net/docs/tanzil_project

Model	Education			General			Judicial		
	BLEU	COMET	HUMAN	BLEU	COMET	HUMAN	BLEU	COMET	HUMAN
Aya23 (Aryabumi et al., 2024)	36.40	0.71	2.00	14.13	0.70	3.33	17.07	0.70	**4.00**
Claude3.5	46.04	0.80	3.33	19.02	**0.85**	3.67	25.62	0.85	3.67
CommandR-plus	35.33	0.75	3.67	14.39	0.77	3.67	17.64	0.77	3.00
CycleL	0.38	0.72	1.33	1.21	0.15	0.79	1.33	0.14	1.00
GPT-4 (OpenAI, 2023)	40.90	0.68	2.67	14.68	0.75	2.67	18.45	0.75	2.67
IKUN-C (Liao et al., 2024)	28.99	0.75	2.67	11.60	0.67	3.00	8.21	0.50	2.33
IKUN	28.62	0.76	1.33	11.99	0.66	2.33	6.95	0.47	1.00
IOL-Research (Zhang, 2024)	40.47	0.67	2.00	15.41	0.77	4.0	19.12	0.78	3.33
Llama3-70B (Grattafiori et al., 2024)	45.73	0.75	3.00	15.58	0.77	3.0	21.27	0.77	3.00
NVIDIA-NeMo	45.12	**0.82**	3.00	18.12	0.66	3.67	21.21	0.69	1.33
Online-A	**50.27**	0.73	3.00	19.84	0.75	4.0	25.02	0.73	3.33
Online-B	46.19	0.82	4.00	21.36	**0.85**	4.0	25.20	**0.86**	3.67
Online-G	46.19	0.73	2.67	16.49	0.67	3.67	**27.33**	0.73	2.67
TransmissionMT	46.70	**0.82**	3.67	**21.39**	**0.85**	**4.67**	25.25	**0.86**	**4.00**
Unbabel-Tower-70B (Rei et al., 2024)	44.22	0.80	**4.33**	20.50	0.83	**4.67**	22.04	0.83	3.67
ZMT	**50.27**	0.72	3.67	19.83	0.75	4.0	25.01	0.73	3.33

Table 1: Performance of different models across education, general and judicial domains

challenges due to their religious significance and the use of archaic language. We aim to enhance the accurate translation of sacred and classical texts.

3.5 Noisy domain

The noisy user-generated data domain in our dataset is sourced from the benchmark dataset for Machine Translation of Noisy Text (MTNT) (Michel and Neubig, 2018). This domain includes 350 English sentences from MTNT, consisting of informal and often error-prone comments made by users on Reddit. Our annotators translated these sentences into Hindi. Capturing the informal and irregular nature of online communication, this domain is critical for improving machine translation models' ability to handle the nuances and challenges of translating user-generated content, which is often rife with slang, typos, and non-standard language usage.

3.6 Online User Review domain

The final domain in our dataset is composed of user product review texts from the e-commerce website Flipkart. This dataset is sourced from the paper "Product Review Translation: Parallel Corpus Creation and Robustness towards User-generated Noisy Text" (Gupta et al., 2021). We have included 300 English-Hindi text pairs from this corpus. The challenges in this domain often stem from grammatical errors and code-mixing, where users blend English and Hindi within the same sentence. Improving machine translation performance in this domain is essential for accurately conveying customer opinions and experiences, which can lead to better user understanding and engagement with

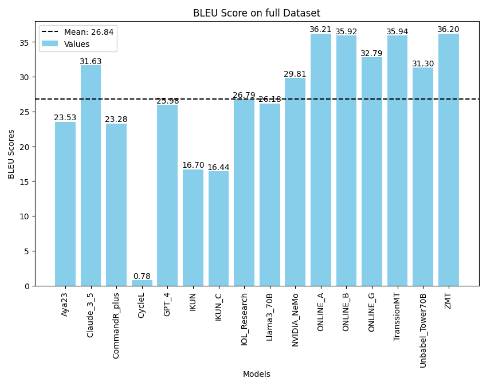

Figure 1: BLEU Score on the Full Dataset

e-commerce platforms, ultimately enhancing the online shopping experience across different languages.

4 Evaluation

In this section, we outline the various evaluation techniques employed to assess the performance of the models based on their outputs. The evaluation metrics considered in this study are the BLEU (Papineni et al., 2002; Post, 2018) score, COMET (Rei et al., 2020; Guerreiro et al., 2023) score, and human evaluation score.

4.1 BLEU Scores

The BLEU score is a metric used to evaluate the quality of machine translations by comparing the generated output to one or more reference translations based on n-gram similarity. We calculate the BLEU score with sacrebleu (Post, 2018) and report corpus_score for the dataset.

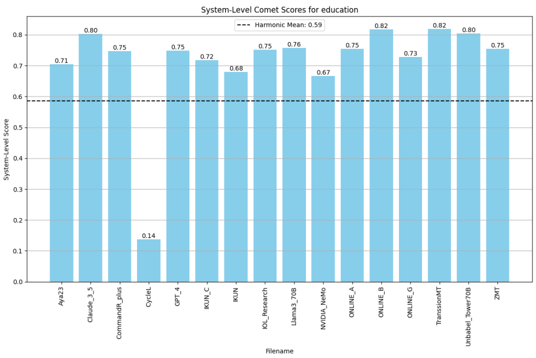

Figure 2: COMET scores in the Education Domain

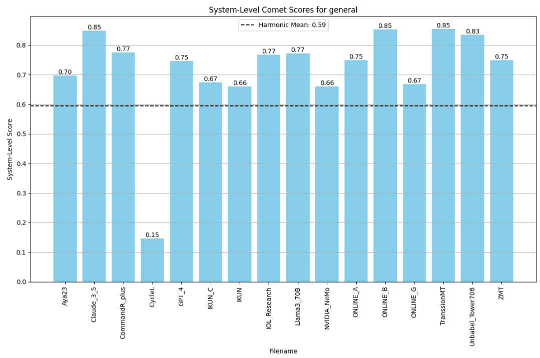

Figure 3: COMET scores in the General Domain

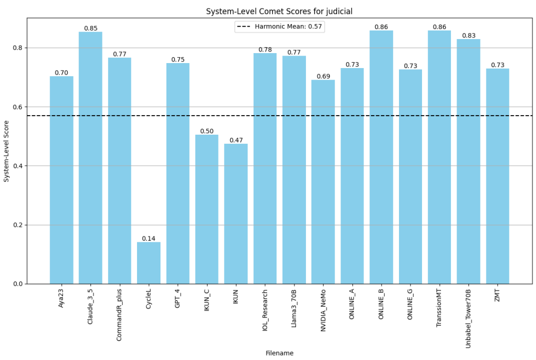

Figure 4: COMET scores in the Judicial Domain

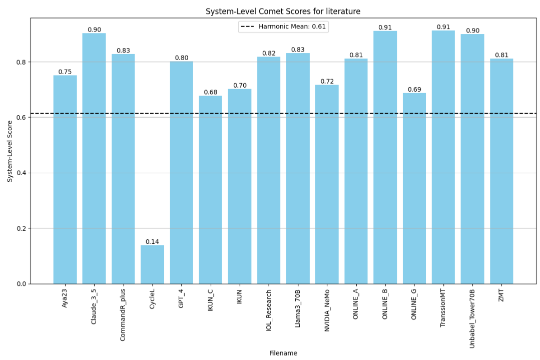

Figure 5: COMET scores in the Literature Domain

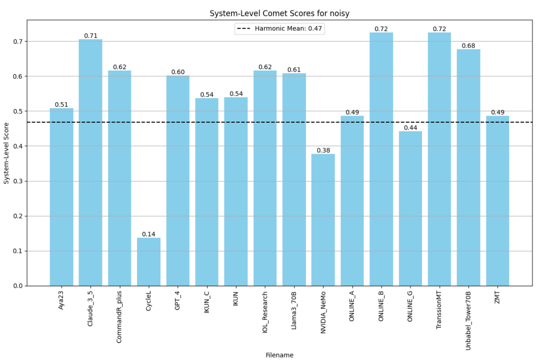

Figure 6: COMET scores in the Noisy Domain

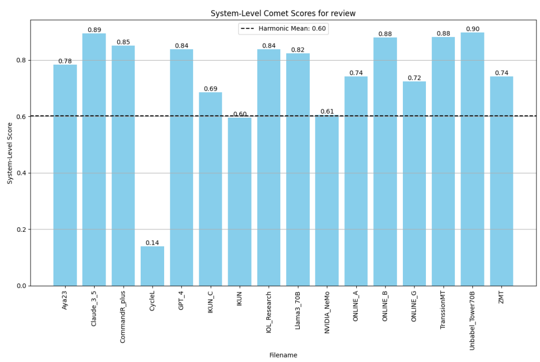

Figure 7: COMET scores in the Product Review Domain

Model	Literature			Noisy			Review		
	BLEU	COMET	HUMAN	BLEU	COMET	HUMAN	BLEU	COMET	HUMAN
Aya23	8.34	0.75	2.67	31.76	0.51	3.00	30.82	0.78	3.00
Claude3.5	15.11	0.90	**3.33**	42.49	0.71	**4.33**	36.45	0.89	3.33
CommandR-plus	10.32	0.83	**3.33**	31.35	0.62	3.67	26.49	0.85	3.33
CycleL	0.21	0.14	1.00	0.82	0.14	1.00	0.33	0.14	1.00
GPT-4	7.95	0.80	2.67	35.43	0.60	3.67	33.66	0.84	2.33
IKUN-C	4.85	0.68	2.0	19.99	0.54	2.33	19.09	0.69	1.33
IKUN	4.80	0.70	1.33	18.89	0.54	2.00	16.48	0.60	1.33
IOL-Research	6.82	0.82	3.00	39.79	0.62	3.33	33.23	0.84	2.67
Llama3-70B	9.51	0.83	2.67	34.73	0.61	3.67	33.16	0.82	2.67
NVIDIA-NeMo	16.65	0.72	1.0	37.32	0.38	2.33	41.07	0.61	2.00
Online-A	20.34	0.81	2.0	**52.55**	0.49	3.00	46.78	0.74	3.00
Online-B	26.21	0.91	**3.33**	51.51	0.72	2.67	41.55	0.88	3.00
Online-G	8.56	0.69	1.67	44.13	0.44	3.33	**55.29**	0.72	**4.00**
TransmissionMT	**26.27**	**0.91**	**3.33**	51.71	**0.72**	3.67	41.58	0.88	3.33
Unbabel-Tower-70B	20.03	0.90	2.67	40.86	0.68	3.00	35.42	**0.90**	**4.00**
ZMT	20.34	0.81	1.67	52.55	0.49	2.67	46.78	0.74	3.00

Table 2: Performance of different models across literature, noisy, and review domains

Model	BLEU	COMET	HUMAN
Aya23	23.53	0.69	3.00
Claude3.5	31.63	0.83	3.61
CommandR-plus	23.28	0.76	3.44
CycleL	0.78	0.14	1.11
GPT-4	25.98	0.74	2.78
IKUN-C	16.70	0.63	2.28
IKUN	16.44	0.61	1.56
IOL-Research	26.79	0.76	3.06
Llama3-70B	26.18	0.76	3.00
NVIDIA-NeMo	29.81	0.62	2.22
Online-A	**36.21**	**0.84**	3.06
Online-B	35.92	0.71	3.44
Online-G	32.79	0.66	3.00
TransmissionMT	35.94	**0.84**	**3.78**
Unbabel-Tower-70B	31.30	0.82	3.72
ZMT	36.20	0.71	3.06

Table 3: Performance of models on the full dataset

4.1.1 Domain wise Overview

The average BLEU scores for the general, judicial, and literature domains are lower at 15.97, 19.14, and 12.89, respectively. In the literature domain, ornamental language leads to subjective translations, causing discrepancies with reference texts. The general domain, with formal content like news and Wikipedia articles, suffers from the model's difficulty in maintaining a formal tone. The judicial domain poses challenges due to specialized terminology and formality. Transliterations instead of translations also contribute to poor performance in these domains.

In contrast, the models perform better in the education domain, where sentences are simpler, and in user-generated domains like noisy texts and product reviews, where BLEU scores are relatively high.

4.1.2 Model wise Overview

The average performance across all domains shows that Models Online-A and ZMT lead, followed by Online-B and TransmissionMT, while CycleL has the lowest BLEU scores. Since BLEU is based on N-gram overlaps, relevant transliterations are not accounted for, leading to lower scores in some models despite acceptable translation quality.

4.2 COMET Scores

The COMET score is a metric that evaluates machine translation quality using pre-trained language models. Unlike traditional metrics, it assesses both adequacy (meaning preservation) and fluency (naturalness). By comparing machine-generated translations to reference and human translations using a regression model trained on human judgments, COMET captures nuances in language and context. This makes it more context-aware and reliable. We calculate scores using xCOMET-XXL.

4.2.1 Domain wise Overview

The judicial, general, and education domains have the highest COMET scores. Retaining adequacy and fluency is easier in these domains due to their formal tone, and COMET does not penalize models heavily for paraphrasing, as it is a more robust metric.

In contrast, the worst COMET scores are found in user-generated data, such as noisy and product

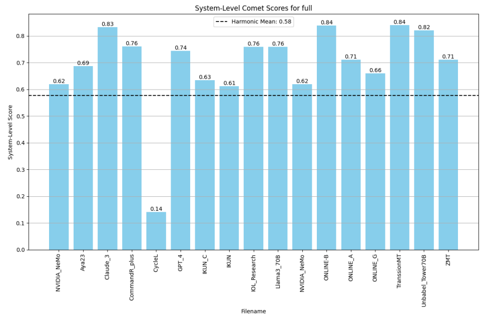

Figure 8: COMET Score on the Full Dataset

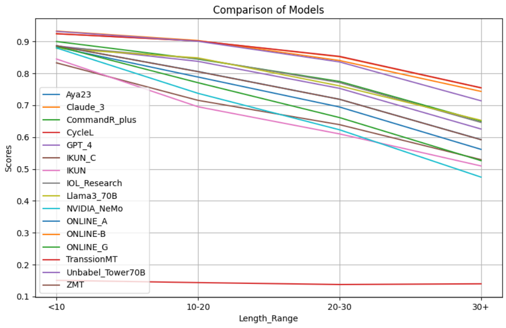

Figure 9: Sentence Length Vs COMET Scores

review texts. These are more informal and often contain spelling and grammatical errors, which present challenges for translation.

- LLMs struggle to translate the noisy texts, resulting in poor quality hypotheses and lower COMET score
- COMET metric is calculated through embeddings. Here, the source side is noisy, which can lead to unreliable embeddings and, therefore, an unreliable COMET score.

4.2.2 Model wise Overview

The best-performing models in terms of COMET scores are Online-B and TransmissionMT, closely followed by Claude-3.5 and Unbabel-Tower-70B. However, the worst-performing model is still CycleL.

From Table 4 and Figure 9, the COMET scores for all LLM translations exhibit a noticeable decline with an increase in source-side sentence length, highlighting that LLMs struggle with translating longer sentences. Among the models, TransmissionMT, Online-B, Claude3.5, and UnbabelTower70B consistently achieve the highest COMET scores across varying sentence lengths.

Source Length	<10	10-20	21-30	30+
Aya23	0.88	0.79	0.69	0.56
Claude3.5	0.93	0.90	0.84	0.74
CommandRplus	0.90	0.84	0.77	0.65
CycleL	0.15	0.14	0.14	0.14
GPT4	0.89	0.84	0.75	0.63
IKUN_C	0.83	0.71	0.64	0.53
IKUN	0.85	0.70	0.61	0.51
IOLResearch	0.88	0.85	0.77	0.65
Llama70B	0.88	0.85	0.77	0.65
NVIDIA_NeMo	0.88	0.74	0.62	0.47
OnlineA	0.89	0.81	0.72	0.60
OnlineB	0.92	0.90	0.85	0.75
OnlineG	0.89	0.77	0.66	0.52
TransmissionMT	0.92	0.90	0.85	0.75
UnbabelTower70B	0.93	0.90	0.84	0.71
ZMT	0.89	0.81	0.71	0.60

Table 4: Change in COMET score on varying source length

Interestingly, while TransmissionMT and Online-B do not achieve the highest COMET scores (0.92) for shorter sentences compared to models like UnbabelTower70B (0.93) and Claude3.5 (0.93), their performance surpasses these models for longer sentences (>30 words), achieving a COMET score of 0.75.

4.3 Human Evaluation

The next evaluation method is human evaluation. We enlisted a linguist to randomly select three sentences from each of the six domains, collecting machine translations from 16 submitted model outputs, resulting in 288 sentences. These were rated on a scale of 1 to 5, with 1 indicating the poorest translation and 5 representing the best compared to the reference texts. Due to the limited sample size, the results are unreliable; however, resource constraints prevented a larger-scale evaluation. We hope these ratings, when considered alongside automated metric scores, will offer insights into the models' competence.

4.3.1 Domain wise Overview

According to the human evaluation, the general domain showed the highest faithfulness to the reference translations. This outcome is expected, as general domain texts are typically easier to translate due to their formal and unambiguous nature, with fewer grammatical, lexical, and spelling er-

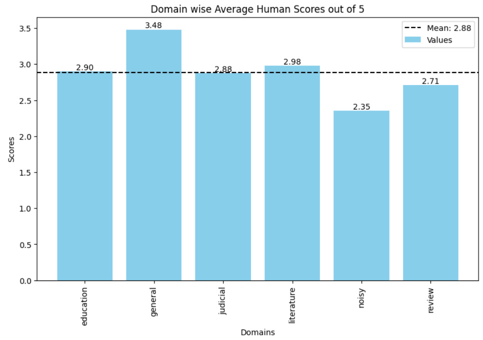

Figure 10: Domain wise Average Human Score

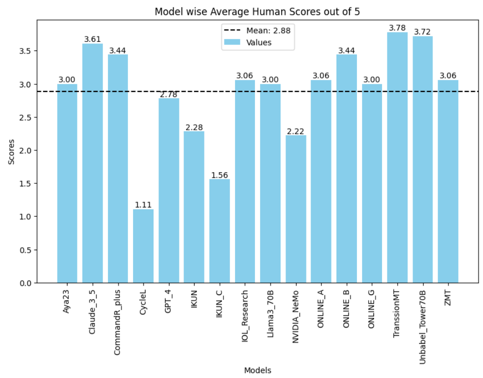

Figure 11: Model wise Average Human Score

rors. Conversely, the noisy domain demonstrated the lowest faithfulness to the reference translations. This is largely attributed to the informal nature of these texts, which often include profanities and internet acronyms like "lol" and "idk" as well as a higher prevalence of errors.

4.3.2 Model wise Overview

Almost consistent with the COMET metrics, we can see that the TransmissionMT, Unbabel-Tower-70B, and Claude-3.5 have the best human-evaluated scores, whereas CycleL again scored the least favorably.

5 Conclusion

This paper presents a comparison of various model submissions for the WMT Shared Task 2024. We proposed a dataset with domain-wise segregation and conducted a domain-specific analysis of the submitted models. Our comprehensive evaluation using BLEU, COMET, and human assessments of the machine-translated hypotheses identified Claude 3.5, TransmissionMT, Unbabel Tower 70B, Online-A, and Online-B as some of the top-performing models for machine translation using LLMs. The analysis revealed that the formal domains of general and education are the easiest for models to handle, whereas the noisy and review domains proved to be the most challenging. This study highlights that while LLMs show proficiency in machine translation, there is still significant room for improvement.

Acknowledgment

The authors gratefully acknowledge project “VIDYAAPATI: Bidirectional Machine Translation Involving Bengali Konkani Maithili and Hindi" under Bhashini funded by MeitY to carry out the research.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

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).

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom.

Quinten Bolding, Baohao Liao, Brandon James Denis, Jun Luo, and Christof Monz. 2023. Ask language model to clean your noisy translation data.

Vicent Briva-Iglesias, Joao Lucas Cavalheiro Camargo, and Gokhan Dogru. 2024. Large language models "ad referendum": How good are they at machine translation in the legal domain?

Amartya Chowdhury, Deepak K. T., Samudra Vijaya K, and S. R. Mahadeva Prasanna. 2022. Machine translation for a very low-resource language - layer freezing approach on transfer learning. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 48–55, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Abhimanyu Dubey et al. 2024a. The llama 3 herd of models.

OpenAI et al. 2024b. Gpt-4 technical report.

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.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Vakul Goyle, Parvathy Krishnaswamy, Kannan Girija Ravikumar, Utsa Chattopadhyay, and Kartikay Goyle. 2023. Neural machine translation for low resource languages.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

Kamal Kumar Gupta, Soumya Chennabasavaraj, Nikesh Garera, and Asif Ekbal. 2021. Product review translation: Parallel corpus creation and robustness towards user-generated noisy text. In *Proceedings of the 4th Workshop on e-Commerce and NLP*, pages 174–183, Online. Association for Computational Linguistics.

Kamal Kumar Gupta, Divya Kumari, Soumya Chennabasavraj, Nikesh Garera, and Asif Ekbal. 2022. Reviewmt: Sentiment preserved e-commerce review translation system. In *Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, CODS-COMAD '22, page 275–279, New York, NY, USA. Association for Computing Machinery.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Ramchandra Joshi, Rushabh Karnavat, Kaustubh Jirapure, and Raviraj Joshi. 2020. Domain adaptation of nmt models for english-hindi machine translation task at adapmt icon 2020.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *NMT@ACL*.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Taja Kuzman, Špela Vintar, and Mihael Arčan. 2019. Neural machine translation of literary texts from English to Slovene. In *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9, Dublin, Ireland. European Association for Machine Translation.

Baohao Liao, Christian Herold, Shahram Khadivi, and Christof Monz. 2024. IKUN for WMT24 general MT task: LLMs are here for multilingual machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 263–269, Miami, Florida, USA. Association for Computational Linguistics.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. 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.

OpenAI. 2023. Gpt-4 technical report.

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.

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

Shabdapurush Poudel, Bal Krishna Bal, and Praveen Acharya. 2024. Bidirectional English-Nepali machine translation(MT) system for legal domain. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 53–58, Torino, Italia. ELRA and ICCL.

Anil Ramakrishna, Rahul Gupta, Jens Lehmann, and Morteza Ziyadi. 2023. INVITE: a testbed of automatically generated invalid questions to evaluate large language models for hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5422–5429, Singapore. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Ricardo Rei, Jose Maria Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. de Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2023 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. 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.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

BigScience Workshop, :, and Teven Le Scao et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Zeynep Yirmibeşoğlu, Olgun Dursun, Harun Dallı, Mehmet Şahin, Ena Hodzik, Sabri Gürses, and Tunga Güngör. 2023. Incorporating human translator style into english-turkish literary machine translation.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Wenbo Zhang. 2024. IOL research machine translation systems for WMT24 general machine translation shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 147–154, Miami, Florida, USA. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis.