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* subtask. 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 the 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 participated in the "Help us break LLMs" subtask at the Workshop on Machine Translation (WMT) 2024 (Kocmi et al., 2024). The primary objective of this subtask is to create a dataset that exposes the difficulties faced by LLM-based MT systems when dealing with domain-specific content. Our approach involves collating a multi-domain dataset that includes sentences from judicial, educational, religious literature, and noisy user-generated content from online product reviews and forums like Reddit.

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 participate in the Break the LLM challenge in WMT24 for English-Hindi language direction, where we submit diverse data 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.

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 341–354
November 15–16, 2024.

2 Related Works

Neural Machine Translation has achieved significant advancements (Vaswani et al., 2017). However, translation of text involving low-resource languages remains a challenge. In low-resource languages, the translations of Indic languages like Hindi is difficult due to the paucity of the high-quality parallel corpus. Existing multilingual models like IndicTrans (Ramesh et al., 2022) and IndicTrans2 (Gala et al., 2023) achieved significant performance gains compared to other models. However, English-Hindi machine translations still have room for improvement.

Moslem et al. (2022) has previously used pretrained Language Models(LM) for domain specific data augmentation for Machine Translation. They simulated the characteristics of a small bilingual dataset or monolingual source text and combined it with back translation to create huge amounts of synthetic in-domain data. Other works involving low-resource languages include translation of chat-based conversation by (Gain et al., 2022) where English Hindi translation was implemented on Chat and question answers in chatbots. In the domain of education, (Behnke et al., 2018) used crowdsourcing English texts to obtain translation into 11 languages for generating NMT data. Similarly, Ramakrishna et al. (2023) introduced the EduMT dataset for improving the English-Hindi translation for educational content.

In a recent study, (Briva-Iglesias et al., 2024) showed that LLMs outperform Google translate when it comes to the Legal domain. (Martínez-Domínguez et al., 2020) implemented machine translation in the legal domain for Italian to Swiss language. For low-resource language, (Poudel et al., 2024) introduced a custom-built dataset for the legal domain for English Nepali language machine translation.

In the Literary domain, (Drobot, 2023) has studied the prospects of neural machine translation. Earlier (Matusov, 2019) has used NMT for translating German literary works to English, and (Kuzman et al., 2019) implemented NMT for the literary domain from English to Slovene. (Yirmibeşoğlu et al., 2023) has implemented NMT in the literary domain for the low-resource language of English-Turkish. (Thai et al., 2022) has also explored document-level literary machine translation for non-English languages. They have also shown that there is a disparity between the automatic evaluation of these machine translations and human evaluation, prompting further improvement of machine translation in this domain.

Noisy or non-standard input text can cause disastrous mistranslations in most modern Machine Translation (MT) systems.Khayrallah and Koehn (2018) has shown in a study the impact of noise on NMT systems. Michel and Neubig (2018) proposed a benchmark dataset for machine translation of noisy texts(MTNT). Herold et al. (2022) has worked on filtering noise from machine translation data for improving the performance of NMT systems.Bolding et al. (2023) has used LLMs to remove noise from the MTNT dataset target sentences and proposed C-MTNT dataset. Machine Translation of noisy text is mainly explored through multimodal translation in English-Hindi (Gain et al., 2021b; Laskar et al., 2021; Gain et al., 2021a; Gupta et al., 2021c; Gain et al., 2023) where images features were used to aid in machine translation from English to Hindi.

Product review is a translation task that is related to the field of e-commerce. (Gupta et al., 2022) explores NMT with sentiment preservation in this domain for the low-resource language of the English-Hindi pair. (Gupta et al., 2021b) and (Gupta et al., 2021a) are some of the other works on online product review translation.

Some other notable works on low-resource languages include (Goyle et al., 2023), (Chowdhury et al., 2022) and (Ranathunga et al., 2023) that have implemented unique NMT techniques to complement the scarcity of data in these languages.

3 Dataset

Our proposed dataset includes English-Hindi bi-text pairs from six critical domains, chosen for their significance to both the machine translation community and their difficulty of translation. We provide a sample from each domain in Appendix D and some statistics about the datasets in Table 4. It can be noted that the size of each domain is different. We had collected 500 sentences from each domain in the beginning but after filtering out sentences less than 5 words, we arrived at the final size of the dataset.

3.1 Education domain

The education domain plays a crucial role in knowledge dissemination. Enhancing machine translation in education promotes equal access to qual-

Model	Education			General			Judicial		
	BLEU	COMET	HUMAN	BLEU	COMET	HUMAN	BLEU	COMET	HUMAN
Aya23	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	40.90	0.68	2.67	14.68	0.75	2.67	18.45	0.75	2.67
IKUN-C	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	40.47	0.67	2.00	15.41	0.77	4.0	19.12	0.78	3.33
Llama3-70B	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	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

ity learning, supports multilingual environments, and empowers non-native speakers to engage with content. This helps reduce educational disparities and fosters cultural exchange. For this study, 330 English-Hindi language pairs were collected from the EduMT dataset, which focuses on educational content in Indian languages (Appicharla et al., 2021).

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 a diverse range of parallel and monolingual Hindi texts compiled by the Center for Indian Language Technology. It features content from various sources such as news articles, TED Talks, government websites, and Wikipedia. For our study, we randomly selected 500 English-Hindi language pairs from this domain. Improving machine translation in the general domain enhances the accuracy of translations across diverse content, making information more accessible for Hindi-speaking audiences.

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. This can have a significant impact by improving access to legal information, supporting multilingual legal proceedings, and ensuring that individuals who speak Hindi can fully understand and engage with the judicial system.

3.4 Religious Literature domain

The religious literature domain in our dataset consists of 300 pairs: 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 pose unique challenges due to their religious significance and archaic language.

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 retaining the tone and nature of the input sentences. However, they got rid of some noise based on their own discretion. This domain captures the informality of online communication. Improving machine translation in this domain will help models better handle slangs, typos, and non-standard language use, in turn making

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

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

them more robust.

3.6 Online User Review domain

The final domain in our dataset consists of user product reviews from the e-commerce site Flipkart (Gupta et al., 2021b). We included 300 English-Hindi text pairs from this corpus. This domain presents challenges like grammatical errors and code-mixing, where users blend English and Hindi within a sentence. Similar to MTNT, overcoming the challenges in this domain will make the MT systems more robust.

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. We have shared the candidate translations from 3 models, Online-B, Nvidia-Nemo, and INKUN-C in Appendix D. Online B is one of the consistently best performing models across all the domains and metrics among all the submissions. Whereas, Nvidia-Nemo and IKUN-C translations are of lower quality. This table gives us a comparison of the quality of translations by these models.

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 BLEU Scores

The BLEU score measures the quality of machine translations by comparing the output to reference translations based on n-gram similarity. A higher n-gram match leads to a higher score, with a brevity penalty to discourage overly short translations. The score ranges from 0 to 100, with higher values indicating better alignment with the references. We calculate the BLEU score with sacrebleu (Post, 2018) and report corpus_score for the dataset.

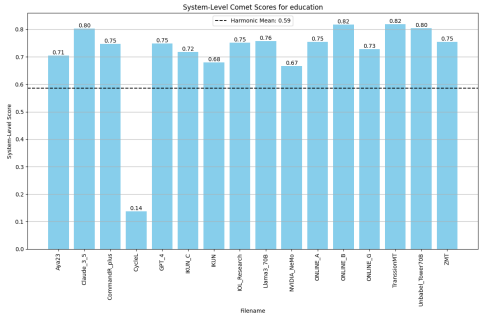

Figure 1: COMET scores in the Education Domain

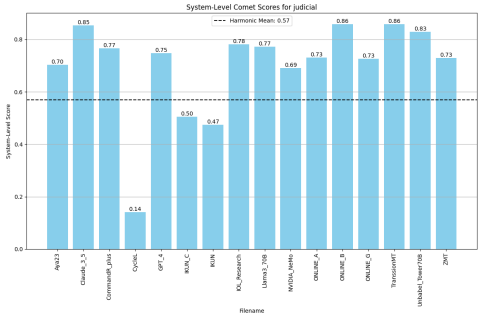

Figure 3: COMET scores in the Judicial Domain

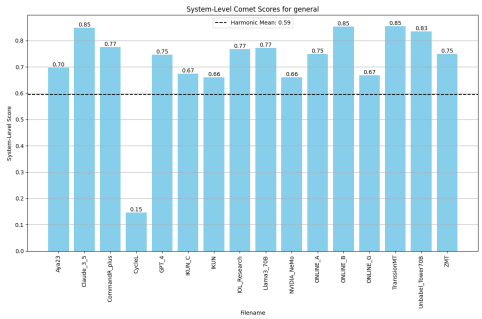

Figure 2: COMET scores in the General Domain

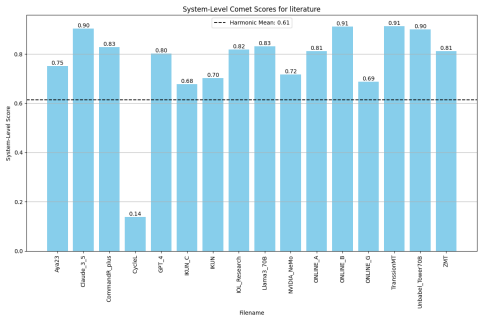

Figure 4: COMET scores in the Literature Domain

4.1.1 Domain wise Overview

The average BLEU scores in the general, judicial, and literature domains are significantly lower, with scores of 15.97, 19.14, and 12.89, respectively. In the literature domain, the frequent use of ornamental language often leads to subjective translations Table 5, causing notable differences between the machine translations and the reference texts. The general domain, encompassing diverse subdomains and characterized by longer sentence lengths and larger data size Figure 17, also contributes to lower BLEU scores, as models struggle with both factors. Similarly, the judicial domain presents challenges due to its specialized terminology and formal tone, which are difficult for models to translate accurately. Additionally, in all three domains, transliteration instead of translation in many cases further impacts the models' performance.

For the education domain, the sentences are relatively straightforward and easier to translate. Interestingly, the models also achieved relatively high BLEU scores for the user-generated data domains, including noisy texts and product review texts.

4.1.2 Model wise Overview

Here we can see the average performance of the models based on all the domains. Models Online-A and ZMT have the best performance, closely followed by Online-B and TransmissionMT, while CycleL has the worst BLEU scores across all the different domains. Note that BLEU is calculated based on N-gram overlaps. Therefore, transliterations of some tokens, even if they are relevant, are not considered. This results in lower BLEU scores in certain models, even if translation quality is acceptable.

4.2 COMET Scores

The COMET score evaluates machine translations using pre-trained language models, focusing on both adequacy (preserving meaning) and fluency (naturalness). It compares machine translations to references and human translations through a regression model trained on human judgments, capturing language nuances that other metrics may miss. The score reflects how closely the machine translation aligns with human preferences. We use xCOMET-XXL to calculate the scores.

4.2.1 Domain wise Overview

The COMET scores of judicial, general, and education domains are the highest. It is easier to retain the adequacy and fluency for these domains compared to the other domains. They have a formal tone to them, and the COMET score does not

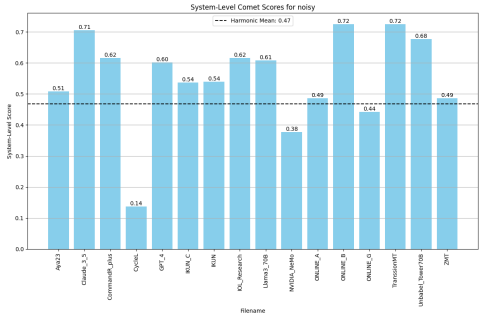

Figure 5: COMET scores in the Noisy Domain

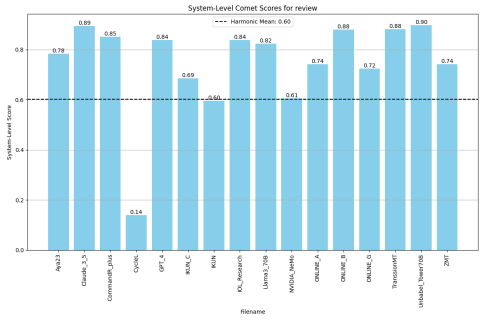

Figure 6: COMET scores in the Product Review Domain

penalize the MT models much for paraphrasing sentences since it is a more robust metric.

Likewise, the worst COMET scores are obtained for the domains of user-generated data for noisy and product review texts. These texts are more informal in nature and ridden with both spelling and grammatical errors. There could be multiple possible reasons: a) LLMs struggle to translate the noisy texts, resulting in poor quality hypotheses and lower COMET score. b) 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.

4.3 Human Evaluation

The next evaluation method employed is human evaluation. We enlisted the expertise of a linguist in our lab, who randomly selected 3 sentences from each of the 6 domains. For each sentence, the corresponding machine translations from the 16 submitted model outputs were collected, resulting in 288 sentences. These sentences were then rated on a scale from 1 to 5, where 1 indicates the poorest translation, and 5 represents the best possible translation compared to the reference texts. Note that due to such a low number of samples, the results in manual evaluation are very unreliable. However, due to resource constraints, we could not perform a large-scale manual evaluation. Nonetheless, we hope this rating will provide some ideas about the competence of the models when observed along with scores from automated metrics.

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 errors. 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.

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).

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.

Maximiliana Behnke, Antonio Valerio, Antonio Valerio Miceli Barone, Rico Sennrich, Vilelmini Sosoni, Thanasis Naskos, Eirini Takoulidou, Maria Stasimioti, Menno Zaanen, Sheila Castilho, Federico Gaspari, Yota Georgakopoulou, Valia Kordoni, Markus Egg, and Katia Kermanidis. 2018. Improving machine translation of educational content via crowdsourcing.

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.

Sören DREANO, Derek MOLLOY, and Noel MURPHY. 2024. Cyclegn: a cycle consistent approach for neural machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Irina-Ana Drobot. 2023. Translating literature using machine translation: Is it really possible? *Scientific Bulletin of the Politehnica University of Timişoara Transactions on Modern Languages*, 20:57–64.

Denis Elshin, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, Ekaterina Enikeeva, Dmitry Popov, Anton Chekashev, Vladislav Negodin, Vera Frantsuzova, Alexander Chernyshev, and Kirill Denisov. 2024. From general LLM to translation: How we dramatically improve translation quality using human evaluation data for LLM finetuning. In *Proceedings of the Ninth Conference on Machine Translation*, USA. 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.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021a. 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.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021b. IITP at WAT 2021: System description for English-Hindi multimodal translation task. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 161–165, Online. Association for Computational Linguistics.

Baban Gain, Dibyanayan Bandyopadhyay, Samrat Mukherjee, Chandranath Adak, and Asif Ekbal. 2023. Impact of visual context on noisy multimodal nmt: An empirical study for english to indian languages.

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.

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 Gupta, Soumya Chennabasavaraj, Nikesh Garera, and Asif Ekbal. 2021a. Product review translation using phrase replacement and attention guided noise augmentation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 243–255, Virtual. Association for Machine Translation in the Americas.

Kamal Kumar Gupta, Soumya Chennabasavaraj, Nikesh Garera, and Asif Ekbal. 2021b. 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.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021c. ViTA: Visual-linguistic translation by aligning object tags. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 166–173, Online. Association for Computational Linguistics.

Ali Hatami, Shubhanker Banerjee, Mihael Arcan, Bharathi Raja Chakravarthi, Paul Buitelaar, and John Philip McCrae. 2024. English-to-low-resource translation: A multimodal approach for hindi, malayalam, bengali, and hausa. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

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.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. Detecting various types of noise for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2542–2551, Dublin, Ireland. Association for Computational Linguistics.

Miroslav Hrabal, Josef Jon, Martin Popel, Nam Luu, Danil Semin, and Ondřej Bojar. 2024. CUNI at WMT24 general translation task: Llms, (q)lora, CPO and model merging. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

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.

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

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. 2024. 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.

Minato Kondo, Ryo Fukuda, Xiaotian Wang, Katsuki Chousa, Masato Nishimura, Kosei Buma, Takatomo Kano, and Takehito Utsuro. 2024. NTTSU at WMT2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Keito Kudo, Hiroyuki Deguchi, Makoto Morishita, Ryo Fujii, Takumi Ito, Shintaro Ozaki, Koki Natsumi, Kai Sato, Kazuki Yano, Ryosuke Takahashi, Subaru Kimura, Tomomasa Hara, Yusuke Sakai, and Jun Suzuki. 2024. Document-level translation with LLM reranking: Team-j at WMT 2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

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.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. 2021. Improved English to Hindi multimodal neural machine translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 155–160, Online. Association for Computational Linguistics.

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*, USA. Association for Computational Linguistics.

Rubén Martínez-Domínguez, Matīss Rikters, Artūrs Vasiļevskis, Mārcis Pinnis, and Paula Reichenberg. 2020. Customized neural machine translation systems for the Swiss legal domain. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 217–223, Virtual. Association for Machine Translation in the Americas.

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.

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, 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.

Shaomu Tan, David Stap, Seth Aycock, Christof Monz, and Di Wu. 2024. Uva-MT's participation in the WMT24 general translation shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

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).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Zhanglin Wu, Daimeng Wei, Zongyao Li, Hengchao Shang, Jiaxin GUO, Shaojun Li, Zhiqiang Rao, Yuanchang Luo, Ning Xie, and Hao Yang. 2024. Choose the final translation from NMT and LLM hypotheses using MBR decoding: HW-TSC's submission to the WMT24 general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

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*, USA. Association for Computational Linguistics.

Zebiao Zhou, Xiangxun Zhu, Xiaowei Ji, Li Yang, Fengjie Zhu, and Tuanwei Shi. 2024. Hyper-SNMT at translation task. In *Proceedings of the Ninth Conference on Machine Translation*, 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.

Hao Zong, Chao Bei, Huan Liu, Conghu Yuan, Wentao Chen, and Degen Huang. 2024. DLUT and GTCOM's neural machine translation systems for WMT24. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

A Overall Scores

We report the overall BLEU and COMET scores in Figure 13 and Figure 14. Further, we provide the domain-wise and model-wise average rating by human annotators in Figure 15 and Figure 16.

B Participants

The WMT24 General Translation Task showcased diverse approaches to machine translation. Several teams explored the potential of Large Language Models (LLMs) for translation tasks. IKUN demonstrated the effectiveness of LLMs in multilingual translation, achieving top rankings in multiple language directions (Liao et al., 2024). The IOL Research team leveraged LLMs for continued pretraining and synthetic data generation (Zhang, 2024).

Some teams focused on improving existing neural machine translation (NMT) architectures. HW-TSC combined NMT and LLM-based models using Minimum Bayesian Risk (MBR) decoding (Wu et al., 2024). UvA-MT compared fine-tuned LLMs with traditional encoder-decoder NMT systems (Tan et al., 2024). The DLUT and GTCOM team emphasized back-translation and multilingual models (Zong et al., 2024).

Novel approaches were also presented. CycleGN introduced a cycle-consistent approach for non-parallel datasets (DREANO et al., 2024). Hyper-SNMT proposed embedding sentences in hyperbolic space to better capture language hierarchies (Zhou et al., 2024).

Several teams explored domain-specific adaptations. Team-J incorporated document-level LLM reranking for improved context-aware translations (Kudo et al., 2024). NTTSU focused on speech domain translation for Japanese to Chinese (Kondo et al., 2024).

The Yandex team demonstrated significant improvements using human evaluation data for LLM fine-tuning (Elshin et al., 2024). CUNI explored various techniques including QLoRA, CPO, and model merging (Hrabal et al., 2024).

Multimodal approaches were also explored, with researchers integrating visual information to enhance translation for low-resource languages (Hatami et al., 2024).

These diverse approaches highlight the ongoing innovation in machine translation, with a notable trend towards leveraging LLMs and exploring novel architectures to improve translation quality across various language pairs and domains.

C Dataset Statistics

Here, we have shared the summary statistics of the lengths of different sentences in each domain. Further we have also shared the harmonic mean of ratio of source to reference text sentence in each domain. From this graph it is evident that general domain has the most disparity in terms of source and reference sentence length. Also, it has the longest sentences compared to the other domains.

D Dataset Example

In Table Table 5, we present examples from the religious domain. This table showcases various outputs relevant to religious texts, highlighting key themes and interpretations.

Table **??** provides examples from the judicial domain. The *Online-B* model has the best quality of translation. The output from the model *Nvidia_Nemo* and *IKUN_C* is inadequate. The original text conveys a universal message about divine provision and the consequences of human actions, while the translation introduces specificity, making it feel more direct and personal.

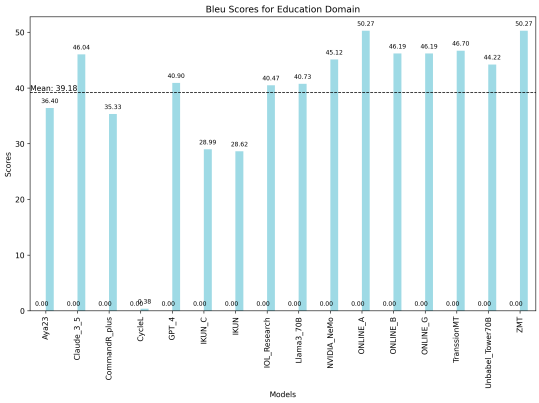

Figure 7: BLEU scores in the Education Domain

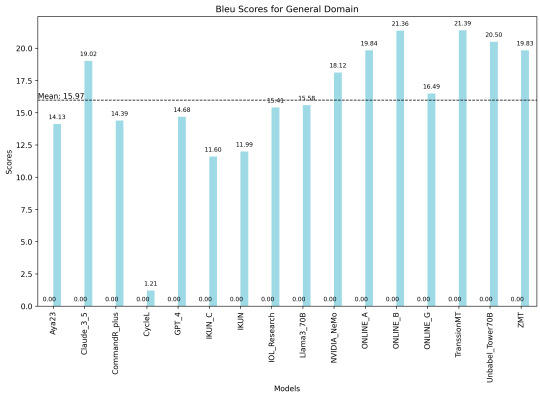

Figure 8: BLEU scores in the General Domain

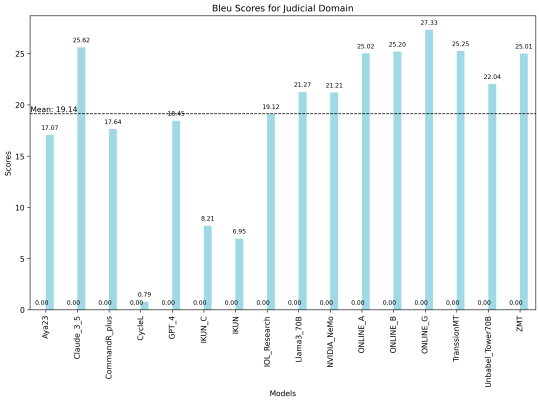

Figure 9: BLEU scores in the Judicial Domain

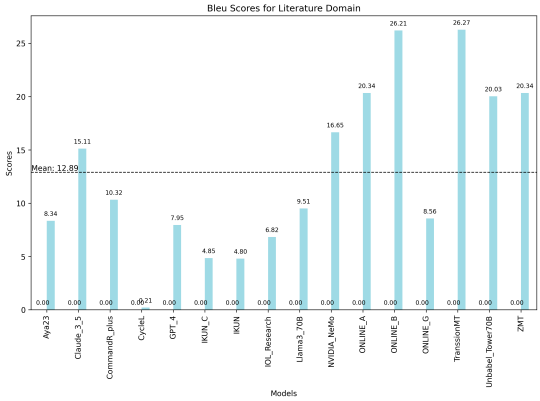

Figure 10: BLEU scores in the Literature Domain

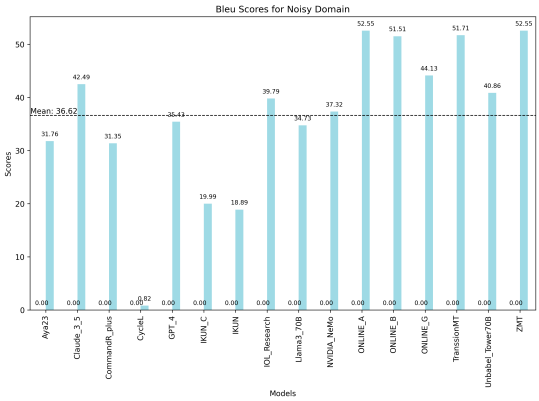

Figure 11: BLEU scores in the Noisy Domain

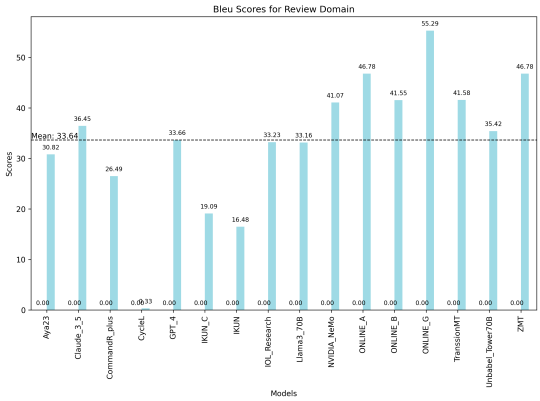

Figure 12: BLEU scores in the Product Review Domain

file	min_nword	max_nword	average_nword
education_source.txt	5	79	25
education_reference.txt	5	80	29
general_source.txt	16	222	29
general_reference.txt	5	195	30
judicial_source.txt	11	39	21
judicial_reference.txt	9	56	24
literature_source.txt	11	38	21
literature_reference.txt	9	63	24
noisy_source.txt	21	49	31
noisy_reference.txt	20	74	38
review_source.txt	11	48	21
review_reference.txt	9	59	25

Table 4: Statistics of the domain-wise files

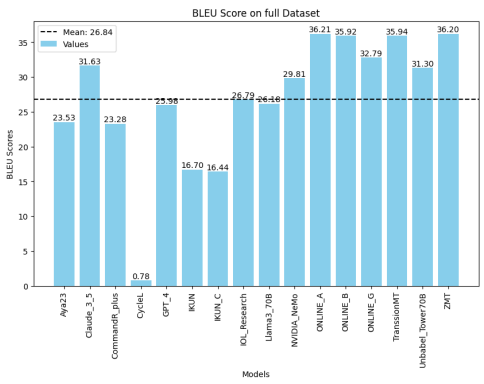

Figure 13: BLEU Score on the Full Dataset

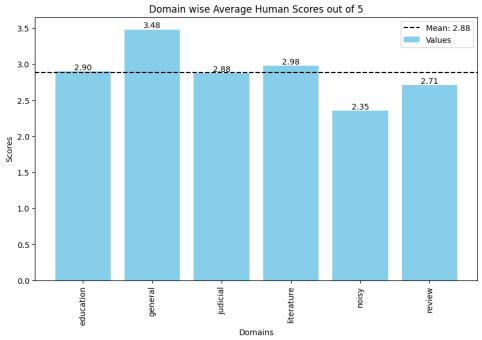

Figure 15: Domain-wise Average Human Score

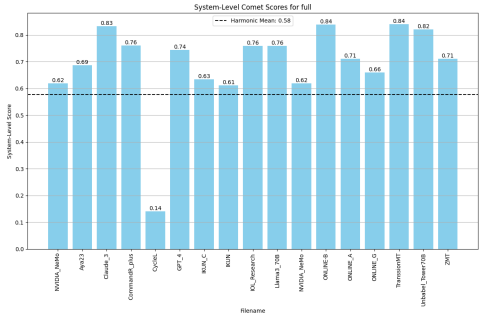

Figure 14: COMET Score on the Full Dataset

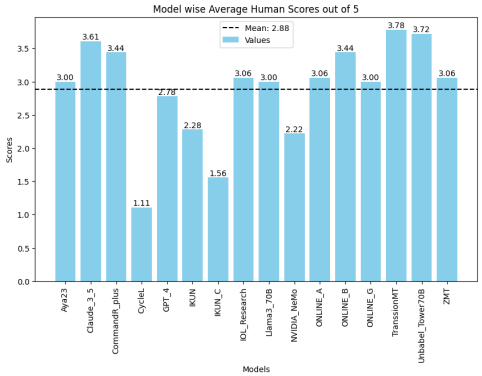

Figure 16: Model-wise Average Human Score

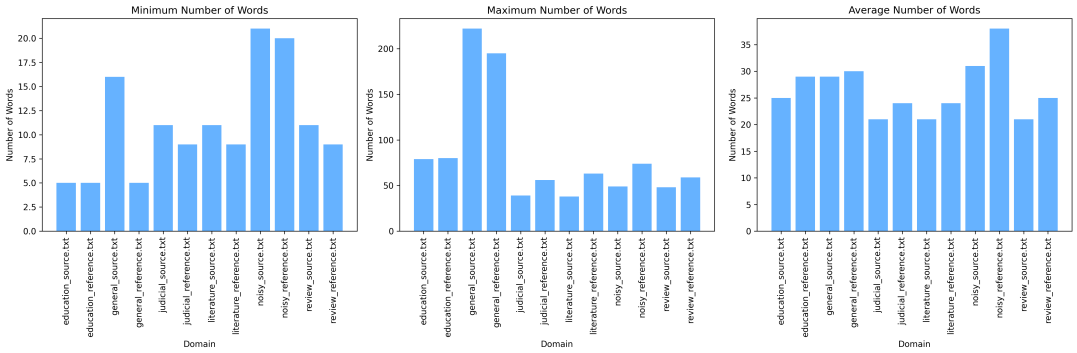

Figure 17: Maximum Minimum and Average Words in File

Model/Text	**Model Output**
Source	And made the cloud spread shade over you, and sent for you manna and quails that you may eat of the good things We have made for you. No harm was done to Us, they only harmed themselves.
Online-B	और तुम्हारे ऊपर बादल छाया कर दिया और तुम्हारे लिए मन्ना और बटेर भेजे, ताकि जो अच्छी-अच्छी चीज़ें हमने तुम्हारे लिए बनाई हैं, उनमें से तुम खाओ। हमें कोई नुक़सान नहीं पहुँचाया गया, बल्कि उन्होंने तो बस अपने आप को नुक़सान पहुँचाया।
Nvidia_Nemo	और हमने तुम्हारे ऊपर बादल की छाया फैला दी और तुम्हारे लिए मन्ना और बटेरें भेजी ताकि तुम उन भलाईयों में से खाओ जो हमने तुम्हारे लिए बनाई हैं और हम पर कोई हानि नहीं हुई, वे अपने आप को ही हानि पहुँचाते थे।
IKUN_C	और बादल को तुम पर छाया डालने के लिए भेजा, और माना और बाज को भेजा कि तुम हमारी ओर से तुम्हारे लिए भेजी हुई अच्छी चीज़ों से खा लो।
Reference	और हमने तुमपर बादलों की छाया की और तुमपर 'मन्न' और 'सलबा' उतारा - खाओ, जो अच्छी पाक चीजें हमने तुम्हें प्रदान की है उन्होंने हमारा तो कुछ भी नहीं बिगाड़ा, बल्कि वे अपने ही ऊपर अत्याचार करते रहे

Table 5: Example of Religious domain output