IndicDISCO-MT: A Discourse-Centric Benchmark for Evaluating Discourse Phenomena in Indian Language Machine Translation

Heli Hingrajiya, Vennela Bairi, Vandan Mujadia
Dipti Misra Sharma, Parameswari Krishnamurthy, Vasudeva Varma
Language Technologies Research Center (LTRC)
IIIT Hyderabad, Telangana, India
{heli.hingrajiya, vennela.bairi, vandan.mu}@research.iiit.ac.in
{dipti, param.krishna, vv}@iiit.ac.in

Abstract

Discourse-level translation remains a challenge for machine translation (MT) systems, particularly for translation from Indian languages to English. This challenge is amplified in Indian languages due to rich morphology and discourse complexity. Existing evaluation benchmarks prioritize sentence-level translation quality and fail to capture important discourse phenomena such as pronoun resolution and lexical cohesion. To address this gap, we introduce IndicDISCO-MT, a parallel benchmark dataset covering translations from eight Indian languages to English. On top of this corpus, we introduce DiscoAlign, a human-annotated source-to-target alignment pairs for discourse evaluation. We further propose two evaluation benchmarks, ProAlign and LexiAlign, to evaluate ability of large language models (LLMs) and MT systems to handle personal pronouns and lexical cohesion. Our evaluation of recent LLM and MT systems shows that although models achieve high translation quality, they still struggle to accurately preserve discourse-level phenomena. The dataset is made publicly available. [1]

1 Introduction

In recent years, Neural Machine Translation (NMT) and Large Language Models (LLMs) have significantly improved translation quality at both sentence and document levels. However, their ability to handle discourse-level translation remains limited, especially for Indian languages. Discourse Translation has a significant impact on the NLP tasks such as machine translation, coreference resolution, dialogue systems, text summarization etc. Since discourse translation includes many phenomena, we focus on the personal pronouns and lexical cohesion in Indian languages. Recent LLM and MT systems still struggle to consistently capture discourse-level phenomena such as pronoun resolution and lexical cohesion during translation, which affects the coherence and consistency of the output (Mohammed et al., 2026).

Moreover, the rich morphology and diverse syntactic structures in Indian languages makes it difficult for the models to preserve these discourse phenomena. In addition, there are significant differences between Indian languages and English when considering personal pronominal forms. For example, pronouns in English are generally not inflected, whereas Indian languages are typically inflected with case markers. Furthermore, many Indian languages distinguish between near and distant referents using different pronouns, whereas English uses the same pronouns (e.g., he, she, it) regardless of such distinctions. Furthermore, English languages use 3rd person pronouns to differentiate in genders (he/she/it), while Indian languages usually apply gender neutral forms (Dash et al., 2025). Beyond pronoun-related challenges, maintaining lexical cohesion in terms of name entities and coreference relation across sentences is also crucial in document-level MT (Jiang et al., 2023). Examples of failures in preserving personal pronouns and lexical cohesion during translation from Indian languages to English are shown in Figure 1.

As shown in the Figure 1, The highlighted pronouns reveal common gender and antecedent er-

[1] https://huggingface.co/datasets/HimangY/IndicDISCO-MT

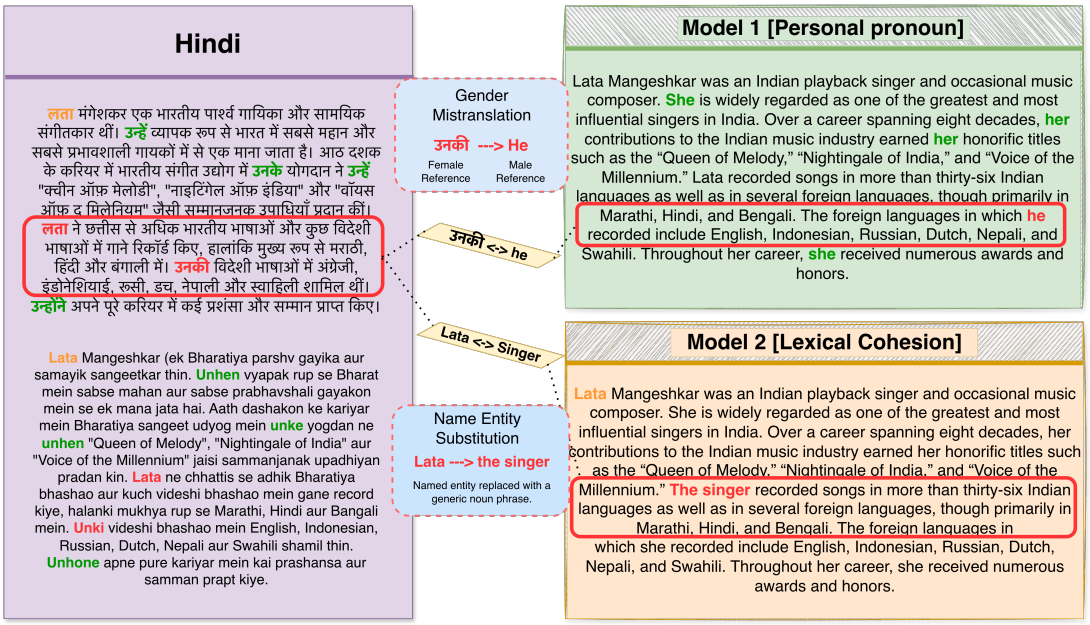

Figure 1: Example of Models failure for preserving personal pronoun and lexical cohesion in translated text.

rors. In the Roman Hindi source, ”unki” refers to the female singer Lata Mangeshkar, but the models incorrectly use masculine pronouns (“he”) that alters the meaning and accurate translation requires feminine pronouns (“she”). Additionally, Model 2 replaces “Lata” with “the singer,” that shows that it fails to preserve the named entity and lexical cohesion.

To systematically study these discourse-level challenges, there is the necessity of evaluating the recent LLM and MT systems before using them for discourse-level translation. Thus in the same direction, the paper makes 5 major contributions:

1. **IndicDISCO-MT:** We introduce the **Indi**an Language Machine Translation benchmarks for **DISCO**urse phenomena (IndicDISCO-MT), a parallel dataset from 8 Indian languages namely: Hindi, Gujarati, Bengali, Marathi, Tamil, Telugu, Kannada and Urdu to English. It includes the paragraphs that are rich in personal pronouns and lexical cohesion. For example, these paragraphs contain multiple entities that are referenced across sentences through pronouns, repeated lexical items, and semantically related expressions which enables us to examine the translation systems in maintaining discourse-level coherence in translation. Although translation is performed sentence by sentence, annotators are instructed to consider the full document context, ensuring cross-sentence coherence, coreference resolution, and lexical consistency.

2. **DiscoAlign:** For evaluating the personal pronouns and lexical cohesion while translating from Indian languages to English, we introduce source to target alignment benchmark dataset on top of IndicDiSCO-MT dataset, where the source words in different Indian languages and target word in English (DiscoAlign). Using this we can particularly focus on comparing the personal pronouns and lexical entities between gold text and model translated text.

3. **ProAlign:** To evaluate the models capability for preserving personal pronouns in translated text, we extracted only the alignments containing personal pronouns from the DiscoAlign dataset.

4. **LexiAlign:** To evaluate lexical cohesion, named entity consistency, transliteration consistency, and entity substitution errors, we extract lexical alignments from the DiscoAlign dataset for all Indian languages.

5. **Model Comparison:** For evaluating pronoun and lexical cohesion in model translated text,

we have considered - Qwen3-4B-Instruct-2507 (Qwen), GPT-4o (GPT), gemma-3-4b-it (Gemma), sarvam-translate (Sarvam), LlaMa-3.2-1B (LlaMa). We also include Google Translate as a commercial MT system and BhashaVerse as an encoder–decoder MT system. Here we are analysing the strength and limitations of all these models in handling discourse-level challenges while translating from Indian languages to English.

2 Related Works

Discourse-level phenomena such as pronoun resolution and lexical cohesion play a crucial role in ensuring coherence and consistency in machine translation (MT). Prior work has shown that while sentence-level neural machine translation systems achieve strong performance, they often struggle with discourse-level dependencies, particularly in handling inter-sentential context. Early studies (Müller et al., 2018) and (Guillou et al., 2018) demonstrated that state-of-the-art MT systems frequently fail to correctly translate pronouns, especially when cross-sentence context is required. To address this, targeted evaluation benchmarks such as PROTEST (Guillou and Hardmeier, 2016) and multilingual pronoun test suites (Jwalapuram et al., 2019) were introduced to systematically evaluate pronoun translation performance across languages.

Beyond pronouns, other discourse phenomena such as ellipsis and lexical cohesion have also been shown to pose challenges for MT systems. (Khullar, 2021) highlights the importance of ellipsis in translation, demonstrating that discourse-level omissions can significantly affect translation quality. Lexical cohesion, which involves maintaining consistent reference through repetition, synonymy, and semantic relatedness, is similarly difficult to preserve across languages, particularly in multilingual and morphologically rich settings.

Recent advances in large language models (LLMs) have improved translation quality, particularly in document-level and context-aware settings. However, prior work indicates that LLMs still struggle to consistently capture discourse-level phenomena. For example, (Manakhimova et al., 2024) evaluate the linguistic performance of LLMs in machine translation and show limitations in handling discourse-level dependencies. Similarly, (Mohammed et al., 2026) demonstrate that LLMs often fail to adequately model latent discourse structures, particularly for pronoun resolution and cohesion, despite strong overall performance.

In addition to pronoun resolution and lexical cohesion, several studies highlight the importance of broader discourse modeling in translation, including context tracking and cross-sentence dependencies. Context-aware approaches have shown that incorporating document-level information improves translation consistency, but such methods still struggle with long-range dependencies and implicit references (Voita et al., 2018), (Miculicich et al., 2018). Furthermore, recent work on document-level translation using LLMs suggests that while these models benefit from larger context windows, they do not consistently utilize contextual signals for accurate discourse resolution (Wang et al., 2023). These findings reinforce the need for targeted benchmarks that explicitly evaluate discourse-level phenomena rather than relying solely on overall translation quality.

Several automatic evaluation metrics, including BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), COMET (Rei et al., 2020), and BERTScore (Zhang et al., 2019), are widely used for MT evaluation. However, these metrics primarily focus on surface-level similarity and are often insufficient for capturing discourse-level correctness, especially in multilingual and low-resource settings (Yari et al., 2025). This limitation is particularly pronounced for Indian languages, which exhibit rich morphology and complex pronominal systems. Although this linguistic richness increases the difficulty of translation, it also introduces ambiguity at the source level, requiring accurate discourse interpretation when translating into English.

Despite these challenges, there is limited work on evaluating discourse-level phenomena in translation from Indian languages to English. Existing benchmarks largely focus on sentence-level evaluation or high-resource language pairs, leaving a gap in systematic evaluation for multilingual Indian-to-English translation settings. To address this, we introduce IndicDISCO-MT, a unified benchmark for evaluating discourse phenomena such as pronoun resolution and lexical cohesion across multiple Indian languages translated into English.

3 Discourse Benchmark Development

3.1 Dataset Collection for Discourse Phenomena

The end to end pipeline for discourse benchmark dataset creation is given in the Figure 2. For creating the parallel dataset, 85 monolingual English paragraphs were collected from Wikipedia. The selected texts contain rich occurrences of personal pronouns (e.g., I, we, he, she, they, it, you, etc) and instances where pronoun interpretation depends on cross-sentence context, enabling evaluation of referential consistency in translation from Indian languages to English. Moreover, for lexical cohesion we selected the paragraphs that illustrate meaningful links between words through repetition, synonymy, or other semantic relations. Also it includes text where certain lexical items have multiple meanings. For example a word "tree" can refer to a natural object while, "trees" represent the hierarchical data structure in graph theory. Such context-dependent words and rich personal pronouns in text makes it challenging for LLMs and MT systems to preserve coherence and maintain consistent lexical choices throughout the translation.

3.2 Human Translation

Now to create the parallel benchmark dataset from the English monolingual paragraphs, annotators have translated the paragraphs into 8 Indian languages - Hindi, Gujarati, Bengali, Marathi, Tamil, Telugu, Kannada and Urdu. The guidelines are mentioned in Appendix A.1. For each language, three experts were selected based on their proficiency and experience. After the translations were completed according to the given guidelines, each sample was cross-verified by another translator to ensure quality and consistency. The complete manual translation and verification process took approximately one month.

3.3 IndicDISCO-MT Benchmarks

IndicDISCO-MT is the base multilingual parallel corpus containing translations from eight Indian languages to English. On top of this corpus, we build DiscoAlign, a human-annotated alignment layer containing source-to-target alignment pairs. From DiscoAlign, we further derive two phenomenon-specific evaluation subsets: ProAlign for personal pronouns and LexiAlign for lexical cohesion. IndicDISCO-MT is a human translated parallel benchmark corpus covering 8 Indian languages, namely: Bengali, Hindi, Gujarati, Marathi, Tamil, Telugu, Kannada and Urdu, each paired with English. For every language, the corpus contains 85 passages including approximately 1,030 sentences to evaluate the mentioned discourse phenomena such as personal pronoun and lexical cohesion during translation.

3.4 DiscoAlign Benchmarks

In this work, we are presenting the unique method to evaluate the personal pronouns and lexical cohesion when models translate the text from Indian languages to English. To evaluate personal pronouns and lexical cohesion while considering translations from the Indian languages to English, firstly we need each word mapping from source to target languages. Previous studies show the limitation of benchmark dataset for evaluating the capabilities of the recent LLM and MT systems for discourse translation particularly for Indian languages. Thus, we introduce first DiscoAlign, a human-annotated alignment layer built on top of IndicDISCO-MT, where word alignment pairs are created using 1,030 sentences, where the source words are in 8 Indian languages, namely: Bengali, Hindi, Gujarati, Marathi, Kannada, Tamil, Telugu, Urdu and target words are in English. The whole dataset is created by human annotators to ensure accuracy and reliability. To obtain precise mappings between source and target words, we performed the annotation at the sentence level by considering each sentence of a paragraph separately. Thus, DiscoAlign enables detailed analysis of translation behavior and helps identify how models handle discourse-sensitive elements during translation. This approach allows us to obtain the relevant alignment mapping between different languages having diverse linguistic characteristics.

The same annotators, who created IndicDISCO-MT dataset were hired to perform this task. The guidelines given to annotators to perform the source to target alignment pair task from 8 Indian languages to English are given in Appendix A.2. Furthermore, to evaluate the consistency of the alignment annotations, we computed inter-annotator agreement using the F1 score for each language pair. The agreement scores were 86% for Hindi and Bengali, 88% for Gujarati, and 91% for Tamil, while Kannada, Marathi, and Urdu obtained scores of 89%, 87%, and 90%, respec-

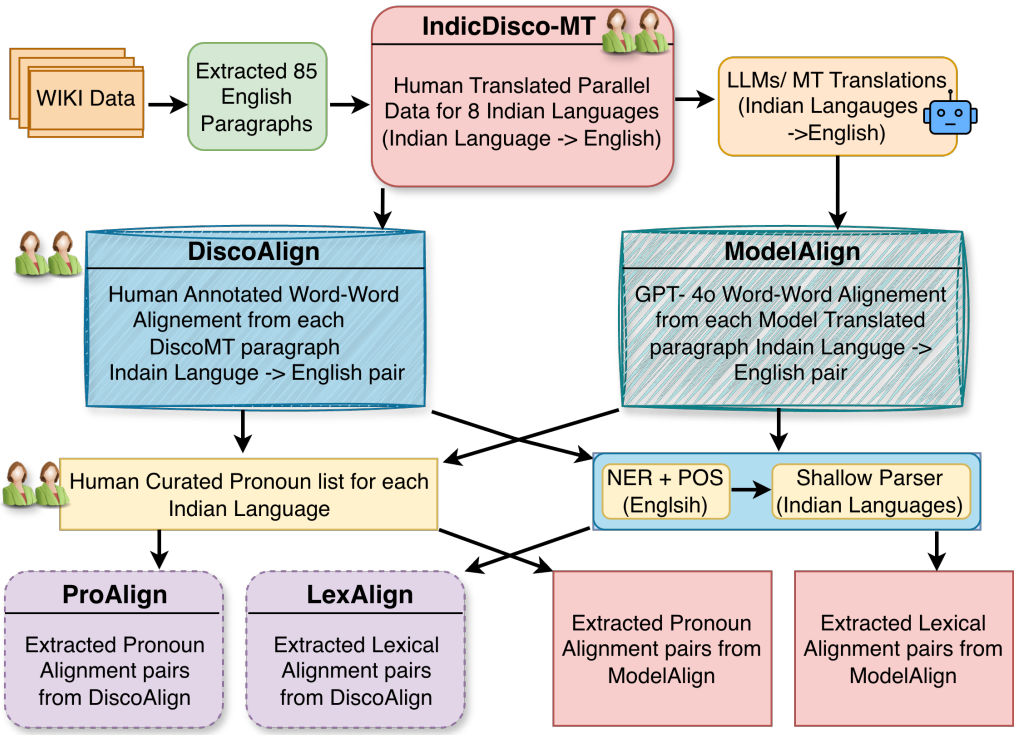

Figure 2: Discourse Benchmark Dataset Creation Pipeline

tively. These results indicate a high level of consistency among annotators in identifying correct alignment pairs. This annotation and verification process took around 3 months to complete.

3.5 Phenomenon-Specific Subsets

3.5.1 ProAlign Benchmark dataset

To evaluate LLMs and MT systems only based on the personal pronouns, firstly annotators have created the list of all personal pronouns for each Indian language by filtering them from a larger list of pronouns. Based on the created list, we fetched only those alignments pairs from DiscoAlign that contained only the identified personal pronouns. Thus, we have created the first ProAlign Benchmark dataset having only personal pronouns alignments pairs from 8 Indian languages to English.

3.5.2 LexiAlign Benchmark dataset

Now here we want benchmark data for evaluating the capabilities of LLMs and MT systems on lexical cohesion preservation while translating from Indian languages to English. For lexical cohesion, we focus on nouns and proper nouns, as they play a crucial role in entity consistency and named entity preservation across sentences. Since nouns typically refer to the same discourse entities throughout a paragraph, their consistent translation helps maintain lexical continuity and discourse coherence in the translated output. Thus, we considered the English text and filtered the words having nouns and proper nouns using POS tags and name entity using Stanza (version-1.11.1) (Qi et al., 2020) and Flair (version-0.15.1) (Akbik et al., 2019) models respectively. On top of this filtering, we have also filtered the words having lexical items particularly nouns and proper nouns in the Indian text using shallow parser (Mishra et al., 2024). Based on this collected lexical items, we extracted the alignment pairs having these lexical items from the DiscoAlign dataset. Thus, we have created the first LexiAlign Benchmark dataset containing only lexical cohesion alignment pairs from 8 Indian languages to English. Statistics for both personal pronouns and lexical items from DiscoAlign are given in Table 1. Also, the sample alignment pairs from each benchmark dataset for a sentence are provided in Figure 6.

3.6 ModelAlign

Now to evaluate the capabilities of LLMs and MT systems for preserving the discourse phenomena, we have taken the source text from the IndicDISCO-MT dataset which is in 8 Indian languages and translated them into English using LLMs and MT systems. For this translation, we

Lang.	No. of Personal Pronouns	No. of Lexical Items
Ban	1449	1682
Hin	1464	1904
Guj	1663	1862
Mar	1424	1836
Kan	1202	1808
Tam	1285	1664
Tel	1520	1680
Urd	1407	1607

Table 1: Statistics of personal pronouns and lexical items per language.

have considered 5 LLMs - GPT, Qwen, Gemma, Sarvam, LlaMa and 2 MT systems - Google Translate (commercial MT system) and BhashaVerse (open source encoder–decoder MT model). The prompt used for translation and detail of all LLMs and MT systems are provided in the Appendix A.4 and Appendix A.5 respectively.

Now, for evaluating personal pronouns and lexical cohesion, we require an automatic method for generating source-to-target alignment pairs, as large-scale human annotation for this task would be time consuming and expensive. For the automatic source to target alignment pairs, we have used two methods i.e GPT-4o and Awesome aligner. Now to find the best method among both methods for the alignment task, we obtain the source to target alignment pairs for the IndicDISCO-MT dataset, considering Indian languages as source and target as English. The predicted alignment pairs can be compared with DiscoAlign gold annotation to evaluate the effectiveness of both methods. We found that GPT-4o achieved an F1 score of 71.21% for alignment pairs on average for all languages. In contrast, Awesome aligner achieved F1 scores of 52.64%. The accuracy, precision, recall and F1 scores for both GPT-4o and Awesome aligner is given in Table 6. Additionally, studies showed that tools like Awesome-Align may struggle with complex cross-lingual relationships and linguistically diverse language pairs. For example, recent work on Dravidian languages reports (James and Krishnamurthy, 2025) lower alignment accuracy for Awesome-Align, particularly for cross-family pairs such as English–Telugu and English–Tamil. Thus, we choose GPT-4o to generate alignment pairs, where source is Indian languages belonging to the IndicDISCO-MT dataset and target belong to the model generated English text. Additionally, the prompt used for obtaining alignments by GPT-4o is presented in Appendix A.3.

4 Evaluation Metrics and Strategies

4.1 Standard Evaluation Metrics

Before evaluating the ModelAlign data for personal pronoun and lexical cohesion, the translation scores of LLMs and MT systems are calculated to check the translation quality. For this, we have considered standard automatic evaluation metrics, which are commonly used in machine translation research such as BLEU (Papineni et al., 2002), chrF++ (Popović, 2015), and COMET (Rei et al., 2020).

4.2 Task-Specific Evaluation

In this study, we focus on evaluating models for preserving discourse phenomena. We use four metrics to evaluate LLMs and MT systems for personal pronoun and lexical cohesion: accuracy, F1-score, precision, and recall. In this context, precision measures the proportion of model-generated alignment pairs that are correct, capturing overgeneration, while recall measures the proportion of gold alignment pairs successfully identified by the model, capturing undergeneration. Accuracy reflects the overall correctness across alignment instances, penalizing both missed and incorrect predictions. For analysis, we primarily rely on F1-score, as it balances precision and recall and provides a robust measure of alignment quality. However, accuracy is not used as the primary metric, as it does not adequately capture over- and undergeneration effects. Since our evaluation relies on automatically generated alignments, some noise may be introduced by the alignment process. To mitigate this, we compare model outputs against human-annotated gold alignments in DiscoAlign, which exhibit high inter-annotator agreement (86–91%), ensuring reliability. Therefore, our evaluation considers precision, recall, and F1 jointly, providing a balanced assessment that accounts for both alignment noise and coverage.

4.2.1 Personal Pronoun Evaluation

To evaluate personal pronouns in the model prediction, we consider ProAlign dataset as the gold data which is a human-annotated source-to-target word alignment dataset involving personal pronouns across eight Indian languages and English as mentioned in Section 3.5. We then evaluate ModelAlign data for personal pronouns. Thus, we have used the same pronouns list for each language which was used for creating the ProAlign. More-

over, we extracted alignment pairs from ModelAlign that only have identified personal pronouns.

Since gold and model-generated alignments pairs are available, we apply the metrics such as accuracy, F1 score, precision and recall as defined in Section 4.2. As mentioned in the dataset creation, we have used the single sentences for obtaining the relevant alignment pairs. Now for evaluation, we consider the pronoun alignments of all sentences within a paragraph as a single block, aggregating them to compute accuracy, precision, recall, and F1-score. Then to obtain the final scores for each language, we averaged evaluation scores for all the paragraphs.

4.2.2 Lexical Cohesion Evaluation

As we discussed earlier, LexiALign is a human created source to target word alignment dataset which contains only the alignment pairs having lexical items particularly proper nouns, where the source language is 8 Indian languages and the target is English. Now, we evaluate the ModelAlign data for lexical cohesion. Thus, we have used the same lexical items list which was used for creating the LexiAlign for extracting alignment pairs from ModelAlign.

For lexical cohesion evaluation, we apply the same methodology used for pronoun evaluation and use accuracy, F1-score, precision, and recall for comparison. We consider the lexical alignments of all sentences within a paragraph as a single block to compute the evaluation metrics. Prior to evaluation, we remove alignment pairs occurring fewer than three times within a paragraph, as lexical cohesion focuses on recurring entities and lexical items. This filtering also helps reduce noise and capture consistent lexical relationships. Finally, the scores for each language are obtained by averaging the evaluation scores across all paragraphs.

5 Results

5.1 Translation Quality Results

The COMET scores of model translation for all the 8 Indian languages are shown in Figure 3. Moreover, the BLEU, chrF++ scores for all models across all languages are mention in Table 4 and Table 5. From the above figure, we can depict that the translation quality is high for GPT and Google Translate achieving COMET scores of more than 0.86 for all languages. The second highest scores

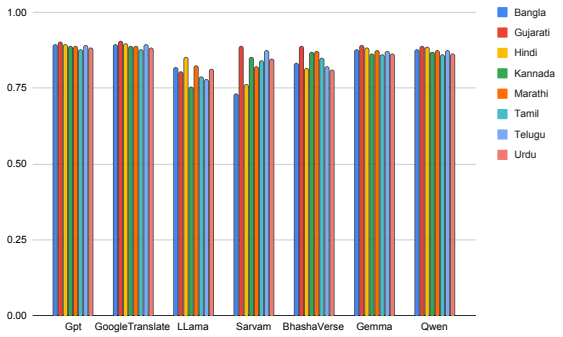

Figure 3: COMET scores for all models

are achieved by models like Gemma, Bhashaverse and Qwen. In contrast, LlaMa and Sarvam obtain comparatively lower scores for languages such as Kannada, Hindi, Bengali, Tamil and Telugu.

5.2 Task-Specific Results

5.2.1 Personal pronoun

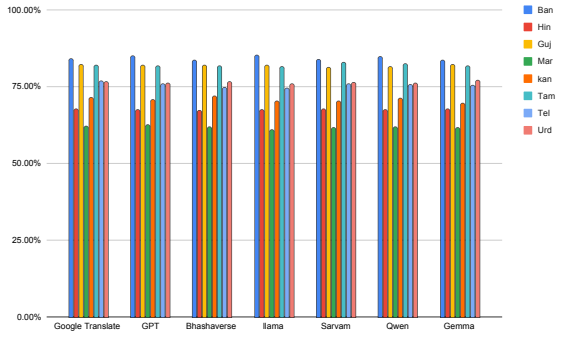

Figure 4: F1 scores for Personal Pronouns across 8 Indian languages and models

The F1 scores and accuracy for personal pronouns evaluation for all the LLMs and MT systems across 8 Indian languages are shown in Figure 4 and Table 2 respectively. Overall the results depict that all models achieve relatively similar performance in handling personal pronouns for each language. Among the evaluated models, Sarvam and Google Translate achieves highest F1 scores of 67.74%, 82.88% and 76.76% respectively among all other models, for the languages such as Hindi, Tamil and Telugu. In contrast, although LlaMa has achieved good COMET scores for languages like Marathi and Urdu, it achieves least F1 scores of 60.93% and 75.87% for these two languages thus it struggles a bit more in capturing the personal pronouns for these languages compared to other models. Also, the accuracy scores of LlaMa are low for the languages like Marathi, Telugu and Urdu.

Model	Ban	Hin	Guj	Mar	Kan	Tam	Tel	Urd
Google Translate	75.71%	55.48%	72.94%	50.52%	59.23%	73.54%	65.34%	64.37%
GPT	76.76%	55.51%	73.08%	51.26%	58.36%	72.93%	63.90%	64.00%
Bhashaverse	74.88%	55.13%	73.24%	50.47%	59.94%	72.91%	63.21%	64.79%
LlaMa	77.19%	55.31%	72.67%	49.52%	57.92%	72.45%	62.47%	63.56%
Sarvam	75.44%	55.61%	72.03%	50.03%	57.89%	74.24%	64.06%	64.01%
Qwen	76.49%	55.49%	72.28%	50.53%	59.06%	73.56%	63.55%	63.81%
Gemma	75.35%	55.83%	73.08%	50.10%	57.40%	73.14%	63.53%	64.89%

Table 2: Personal Pronoun Accuracy of Models across Indian languages.

Model	Ban	Hin	Guj	Mar	Kan	Tam	Tel	Urd
Google Translate	87.21%	86.12%	88.06%	79.40%	81.13%	85.74%	77.88%	84.84%
GPT	86.46%	86.05%	89.24%	78.68%	82.41%	85.74%	77.67%	84.55%
Bhashaverse	85.95%	85.36%	89.09%	79.14%	81.52%	85.33%	77.59%	85.25%
LlaMa	86.77%	84.98%	88.44%	79.05%	80.20%	84.27%	78.23%	85.90%
Sarvam	87.38%	84.89%	89.00%	78.37%	81.58%	84.94%	77.07%	85.29%
Qwen	86.33%	85.21%	89.73%	79.86%	81.51%	84.88%	78.30%	82.49%
Gemma	86.29%	85.81%	89.26%	78.99%	81.74%	84.98%	78.00%	85.32%

Table 3: Lexical Cohesion Accuracy of Models across Indian languages.

Among all the languages, the F1 scores of Bengali is highest for all the models that suggest that models are capable enough for preserving Bengali personal pronouns compared to other languages. One possible reason for this behaviour is that Bengali has a relatively smaller set of personal pronouns compared to several other Indian languages, which reduces ambiguity and makes pronoun handling easier for translation models. Moreover, F1 scores and accuracy for Gujarati and Tamil languages are also second highest among all models and across all other languages. In contrast, languages such as Marathi and Hindi show comparatively lower F1 scores, typically ranging between 60% and 70% suggesting that all models still struggle with these languages for accurate personal pronoun preservation while translation in certain linguistic contexts. The precision and recall scores for personal pronoun are given in Table 7 and Table 8 respectively.

5.2.2 Lexical Cohesion

The F1 scores and accuracy for lexical cohesion evaluation for all the LLMs and MT systems across 8 Indian languages are shown in Figure 5 and Table 3 respectively. The lexical cohesion results indicates that the F1 scores are more than 85 for all the models and among all languages which suggest that models are more capable to handle lexical cohesion particularly, in terms of entity consistency and repeated content words. Unlike personal pronouns, lexical items are explicitly mentioned in the text and usually retain the same form during translation, making it easier for models to preserve. Among all the LLMs and MT system, Qwen and Google Translate achieves the highest F1 scores for several languages such as Gujarati, Hindi, Marathi, Telugu and Urdu. In contrast, Sarvam and LlaMa together scores least F1 scores for 5 Indian languages, namely: Hindi, Marathi, Kannada, Tamil and Telugu.

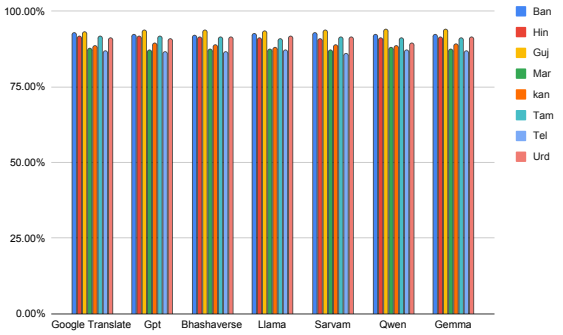

Figure 5: F1 scores for Lexical Cohesion across 8 Indian languages and models.

Across languages, Gujarati and Bengali consistently obtain the highest F1 scores of 92% and above it, suggesting that during machine translation, the lexical items in these languages are preserved more accurately by models. In con-

trast, Telugu, Marathi and Kannada show relatively lower F1 scores and accuracy compared to other languages. Furthermore, the accuracy scores for languages such as Marathi and Telugu are less than 80% among all models. The precision and recall scores for lexical cohesion are given in Table 9 and Table 10 respectively.

Overall, it is observed that although the COMET scores for Google Translate and Sarvam are high for specially for the languages such as Gujarati and Telugu, these models fail for preserving the lexical items compared to other models while translation. The F1 scores for Marathi and Telugu is least for all the models for personal pronouns and lexical cohesion that shows that it is challenging for recent LLM and MT systems to preserve personal pronouns and lexical cohesion in these languages.

6 Conclusion and Future work

Overall, our benchmark dataset reveals that although several models achieve high COMET scores for certain languages such as Telugu and Marathi, their F1 scores for pronoun preservation and lexical cohesion shows a noticeable contradiction. This suggests that it is challenging for models to handles discourse-level phenomena particularly in Indian languages. Furthermore, most models obtain similar F1 and accuracy scores across both discourse tasks, indicating that these systems may rely on mostly similar training data and exhibit comparable limitations in handling discourse-sensitive elements. One possible reason is that the training data may not sufficiently capture diverse pronominal forms and inflections. In future work, models can be trained on datasets that contain more diverse examples of pronouns and lexical items. Additionally, the benchmark can be extended to cover other discourse phenomena and expanded to include more Indian languages.

7 Limitations

Although this work focuses on discourse phenomena, it considers only two aspects i.e personal pronouns and lexical cohesion while, other discourse elements such as coreference resolution, ellipsis, discourse connectives, and topic continuity are not explored. In addition, the current version of IndicDISCO-MT contains a limited number of paragraphs, which may not fully capture the diversity of discourse structures present in real-world texts. Moreover, we use GPT-4o both for generating word alignments and as one of the evaluated translation systems. Although the alignments are generated independently of the translation outputs, this setup may introduce a potential evaluation bias toward the same model.

References

[Achiam et al.2023] Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[Akbik et al.2019] Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

[Chandra et al.2025] Chandra, Rohitash, Aryan Chaudhari, and Yeshwanth Rayavarapu. 2025. An evaluation of llms and google translate for translation of selected indian languages via sentiment and semantic analyses. *IEEE Access*.

[Dash et al.2025] Dash, Niladri Sekhar, Mendem Bapuji, and Srija Deb. 2025. Exploring the nature of commonalities in personal pronouns in some indian languages.

[Dubey et al.2024] Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

[Guillou and Hardmeier2016] Guillou, Liane and Christian Hardmeier. 2016. Protest: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643.

[Guillou et al.2018] Guillou, Liane, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the english–german mt systems at wmt 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577.

[James and Krishnamurthy2025] James, Antony Alexander and Parameswari Krishnamurthy. 2025. Pos-aware neural approaches for word alignment in dravidian languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 154–159.

[Jiang et al.2023] Jiang, Yuchen Eleanor, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse centric evaluation of machine translation with a densely annotated parallel corpus. *arXiv preprint arXiv:2305.11142*.

[Jwalapuram et al.2019] Jwalapuram, Prathyusha, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2964–2975.

[Khullar2021] Khullar, Payal. 2021. Are ellipses important for machine translation? *Computational Linguistics*, 47(4):927–937.

[Manakhimova et al.2024] Manakhimova, Shushen, 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.

[Miculicich et al.2018] Miculicich, Lesly, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. *arXiv preprint arXiv:1809.01576*.

[Mishra et al.2024] Mishra, Pruthwik, Vandan Mujadia, and Dipti Misra Sharma. 2024. Multi task learning based shallow parsing for indian languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(9):1–18.

[Mohammed et al.2026] Mohammed, Wafaa, Vlad Niculae, and Chrysoula Zerva. 2026. Unlocking latent discourse translation in llms through quality-aware decoding. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4752–4774.

[Mujadia and Sharma2024] Mujadia, Vandan and Dipti Misra Sharma. 2024. Bhashaverse: Translation ecosystem for indian subcontinent languages. *arXiv preprint arXiv:2412.04351*.

[Mujadia et al.2023] Mujadia, Vandan, Ashok Urlana, Yash Bhaskar, Penumalla Aditya Pavani, Kukkapalli Shravya, Parameswari Krishnamurthy, and Dipti Misra Sharma. 2023. Assessing translation capabilities of large language models involving english and indian languages. *arXiv preprint arXiv:2311.09216*.

[Müller et al.2018] Müller, Mathias, Annette Rios Gonzales, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the third conference on machine translation: research papers*, pages 61–72.

[Papineni et al.2002] Papineni, Kishore, 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.

[Popović2017] Popović, Maja. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

[Qi et al.2020] Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pages 101–108.

[Rei et al.2020] Rei, Ricardo, 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.

[Sankalp et al.2025] Sankalp, KJ, Ashutosh Kumar, Laxmaan Balaji, Nikunj Kotecha, Vinija Jain, Aman Chadha, and Sreyoshi Bhaduri. 2025. Indicmmlu-pro: Benchmarking indic large language models on multi-task language understanding. *arXiv preprint arXiv:2501.15747*.

[Sarvam AI and AI4Bharat2025] Sarvam AI and AI4Bharat. 2025. Sarvam-translate. Accessed: 2026-05-11.

[Singh et al.2024] Singh, Harman, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. Indicgenbench: A multilingual benchmark to evaluate generation capabilities of llms on indic languages. *arXiv preprint arXiv:2404.16816*.

[Team et al.2025] Team, Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

[Team2025] Team, Gemma. 2025. Gemma 3.

[Voita et al.2018] Voita, Elena, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274.

[Wang et al.2023] Wang, Longyue, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661.

[Wu et al.2016] Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

[Yang et al.2025] Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

[Yari et al.2025] Yari, Amir Hossein, Kalmit Kulkarni, Ahmad Raza Khan, and Fajri Koto. 2025. Revisiting metric reliability for fine-grained evaluation of machine translation and summarization in indian languages. *arXiv preprint arXiv:2510.07061*.

[Zhang et al.2019] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

A.1 Guidelines for Human Translation for IndicDISCO-MT Benchmark

- Analyze the entire source text before starting.
- Translate sentence by sentence while preserving meaning and clarity.
- Pay attention to pronouns to ensure accurate reference and coherence.
- Maintain lexical consistency for nouns and noun phrases.
- Re-read the translation multiple times to ensure flow and accuracy.
- Compare the final translation with the source to ensure message and context alignment.

A.2 Word-to-Word Alignment Guidelines for Human Annotators for Creating the DiscoAlign Benchmark

To ensure consistency and accuracy in the DiscoAlign dataset, we defined a set of guidelines for validating and correcting word-level alignments between the source sentences and their corresponding English translations.

Objective: The goal of the alignment process is to verify and correct word-level mappings between the source sentence and the corresponding English translation. Each alignment pair should represent the same meaning, grammatical role, and contextual usage.

Alignment Unit: The basic unit of alignment is either a single word or a meaningful phrase. Alignments may occur in the following forms:

- **One-to-one (1:1):** one source word aligned with one target word.
- **One-to-many (1:n) or many-to-one (n:1):** when a word or phrase in one language corresponds to multiple words in the other language.
- **Null alignment:** when a source word does not have an explicit counterpart in the target sentence due to implicit meaning or structural differences.

Semantic Consistency: Annotators prioritize semantic equivalence over literal translation. If a translated word or phrase conveys the intended meaning in context, it is considered acceptable even if it is not a direct literal equivalent.

Handling Multiword Expressions: When a multiword expression in the source language corresponds to a single compound expression in English, the alignment should reflect the combined meaning. For example, a phrase representing age in the source language may align with a compound adjective such as "4-month-old".

Null Alignment Representation: Words whose meanings are implicitly conveyed elsewhere in the translation are marked using a blank (for example: –¿ is). This typically occurs for grammatical particles or morphological markers that do not have a direct English equivalent.

Alignment of Function Words: Function words such as complementizers and relative markers should be aligned when they contribute to the sentence structure or meaning. Repeated function words should also be aligned individually when they appear in the source sentence.

Morphological Equivalence: Inflectional variations should be aligned with their corresponding grammatical forms in English. For example, auxiliary verbs and agreement markers are aligned with the appropriate English verb forms depending on the subject and tense.

Structural Transformations: When the syntactic structure differs between the source and target sentences, the alignment should still capture

the semantic correspondence. In such cases, individual components of a phrase may be aligned with the relevant parts of the translated expression.

A.3 GPT-4o Prompt for Source-to-Target Alignment Task

The following prompt was used with GPT-4o to generate word-level alignments between Indian language sentences and their corresponding English translations.

```
You are a word alignment model for
Indian languages.
TASK:
Align source words from the Indian
language with their
closest corresponding English
target words.
OUTPUT FORMAT (MANDATORY):
source word -> target word
ALIGNMENT RULES:
1. No index numbers, tables, or
explanations.
2. Align word-by-word as much as
possible.
3. If a source word does NOT have
an exact standalone
  English equivalent:
   -> Merge 2-3 source words
together and align them
    to ONE English word.
4. Do NOT use "--" or indicate
missing alignments.
5. Do NOT perform named-entity
recognition.
6. Treat names, dates, places, and
numbers as normal words,
  align them as-is.
7. If multiple English words
match, choose the closest single
word.
8. Output EXACTLY one alignment
per line.
9. Maintain script accuracy, do
not transliterate unless
  the source itself uses Roman
script.
SOURCE ([language name]): [source]
TARGET (English): [target]
```

Figure 6 presents example source-to-target word alignments from the DiscoAlign benchmark, along with examples from its derived subsets, ProAlign and LexiAlign. In this example, the source language is Hindi and the target language is English. The figure demonstrates how personal pronouns and lexical entities in the source text are aligned with their corresponding English translations. While ProAlign focuses on evaluating pronoun preservation, LexiAlign targets lexical cohesion and entity consistency, enabling a more fine-grained assessment of discourse phenomena in machine translation.

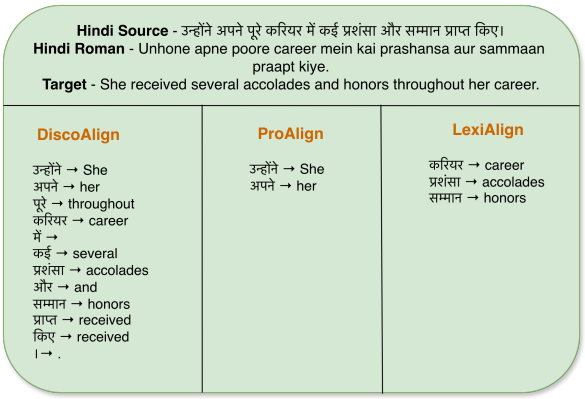

Figure 6: Example of alignment pairs from DiscoAlign, ProAlign and LexiAlign benchmark dataset

A.4 Translation Prompt for All Models

The following prompt was used to generate translations from all evaluated models in our experiments.

```
You are a professional multilingual
translator specializing
in translating from [language-name]
to English.
# Task
Translate the following
[language-name] text into English,
maintaining:
1. The same meaning and tone.
2. Correct pronoun and tense
agreement.
3. Natural fluency without
paraphrasing unnecessarily.
Return your answer strictly as a
Python dictionary:
{
  "translated": "<only the English
translation>"
}
```

A.5 Considered LLMs/MT models

In this paper, we considered both decoder-only large language models (LLMs), designed for a wide range of language understanding and generation tasks, and encoder–decoder machine translation (MT) systems, which are specialized for translation. Among the considered models, ChatGPT-4o, Google Translate, and Bhashaverse are closed-source, while Qwen3-4B-Instruct-2507, LlaMa-3.2, Gemma-3-4B-IT and Sarvam are open-source. These models were selected based on re-

cent evaluations demonstrating strong performance on Indian-language tasks (Mujadia et al., 2023), (Sankalp et al., 2025), (Singh et al., 2024) and (Chandra et al., 2025). Furthermore, we included newer and less-explored models such as Sarvam and Gemma to ensure broader coverage.

Sarvam[2] (Sarvam AI and AI4Bharat, 2025) is a document-level translation model developed by Sarvam AI in collaboration with AI4Bharat, designed to support 22 officially recognized Indian languages. The model is built on Gemma-3-4B-IT (Team, 2025) and fine-tuned through a two-stage process: first on a large multilingual dataset to improve translation capability, and then using LoRA fine-tuning on a curated dataset to enhance format preservation and style consistency.

Gemma (Team et al., 2025) is a multimodal model capable of handling text and image inputs. It is available in multiple sizes (1B, 4B, 12B, and 27B parameters) and trained on web documents covering over 140 languages. In our experiments, we used the gemma-3-4b-it[3] variant, trained on 4 trillion tokens with a 128K token context window.

LlaMa (Dubey et al., 2024) is a multilingual auto-regressive transformer with Grouped-Query Attention (GQA), pretrained on up to 9 trillion tokens and further refined using supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF). We used the LlaMa-3.2-3B-Instruct[4] variant, optimized for diverse multilingual tasks.

Qwen (Yang et al., 2025) is trained with a causal language modeling (CLM) objective. We used the Qwen3-4B-Instruct-2507[5] variant, which contains 4B parameters across 36 transformer layers, uses GQA for efficient inference, and supports a native context length of 262,144 tokens.

ChatGPT-4o (Achiam et al., 2023) is a large-scale generative language model developed by OpenAI, based on the GPT-4 architecture. It is optimized for instruction-following and dialogue tasks, with extensive pretraining on diverse multilingual text corpora.

Google Translate (Wu et al., 2016) is a widely used commercial machine translation (MT) system supporting over 130 languages. Its underlying architecture is not publicly disclosed and may involve a combination of neural machine translation and large language model components.

BhashaVerse (Mujadia and Sharma, 2024) is a 2B-parameter multilingual sequence-to-sequence model supporting translation across 36×36 language pairs, including English and 35 Indian languages. It follows a transformer-based encoder-decoder architecture and is trained on 10 billion parallel sentences. It is included due to its strong Indic language coverage despite its smaller size.

A.6 Translation Quality Scores - BLEU and ChrF++

To provide a broader comparison of translation quality across models and languages, we report BLEU and chrF++ scores for evaluated systems. BLEU measures n-gram overlap between model outputs and reference translations, while chrF++ captures character- and word-level similarity, making it suitable for morphologically rich languages. The BLEU and chrF++ scores are presented in Table 4 and Table 5 respectively. Overall, Google Translate and GPT-4o achieve strong performance, while Qwen and Gemma also demonstrate competitive results. In addition, Table 6 presents the alignment quality comparison between GPT-4o and Awesome-Align against the human-annotated DiscoAlign benchmark. GPT-4o generally achieves higher precision and F1-scores across language pairs. Furthermore, Table 7 and Table 8 report precision and recall scores for personal pronoun evaluation, while Table 9 and Table 10 provide scores for lexical cohesion evaluation across languages and models. The results indicate that preserving lexical cohesion and pronoun consistency across Indian languages remains challenging.

[2]https://huggingface.co/sarvamai/sarvam-translate
[3]https://huggingface.co/google/gemma-3-4b-it
[4]https://huggingface.co/meta-LlaMa/LlaMa-3.2-1B
[5]https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

BLUE	Ban	Guj	Hin	Kan	Mar	Tam	Tel	Urd
GPT-4o	58.90	60.69	57.98	54.42	55.52	47.33	55.29	58.67
Google Translate	60.61	65.59	60.07	58.77	57.57	51.01	59.79	60.60
LlaMa	24.08	18.91	31.73	14.50	22.27	16.06	16.15	26.58
Sarvam	30.55	58.31	35.87	50.24	39.05	39.84	53.38	48.56
Bhashaverse	31.42	55.78	29.71	49.39	48.51	41.91	43.15	32.15
Gemma	21.60	28.14	32.12	27.37	28.87	25.11	27.46	25.99
Qwen	35.83	37.58	39.64	30.03	34.04	26.49	32.82	36.66

Table 4: BLEU scores of all models across languages

chrF++	Ban	Guj	Hin	Kan	Mar	Tam	Tel	Urd
GPT	78.20	79.01	77.46	75.59	76.26	70.48	76.05	77.66
GoogleTranslate	79.27	81.95	78.64	78.01	77.11	72.39	78.40	78.71
LlaMa	46.94	41.99	55.79	36.23	47.01	38.87	38.88	50.27
Sarvam	54.88	77.51	60.61	71.73	65.41	64.55	73.99	71.23
BhashaVerse	52.51	76.78	50.81	72.67	72.03	67.39	65.68	52.43
Gemma	45.24	51.79	55.12	52.63	53.69	50.51	52.37	49.71
Qwen	63.86	64.85	66.38	59.67	62.56	56.87	61.50	64.04

Table 5: chrF++ scores of all models across languages

Model	Metric	Ban	Guj	Hin	Mar	Kan	Tam	Tel	Urd
GPT-4o	Accuracy	72.01	68.10	63.77	60.61	52.06	55.08	52.68	73.00
	Precision	90.94	88.48	76.21	85.65	85.45	79.99	80.56	81.88
	Recall	72.44	68.10	66.64	60.72	52.06	56.95	54.09	74.85
	F1	80.35	76.92	69.97	70.98	64.63	65.33	63.74	77.74
Awesome Aligner	Accuracy	54.40	49.41	46.10	50.24	47.63	36.92	49.15	46.82
	Precision	64.25	61.16	61.34	62.34	61.62	42.95	59.76	57.43
	Recall	54.77	49.41	46.58	50.34	47.63	38.46	50.91	47.23
	F1	58.90	54.63	52.67	55.65	53.70	39.81	54.11	51.67

Table 6: Accuracy, Precision, Recall and F1 scores of GPT-4o and Awesome Aligner across all languages

Models	Ban	Guj	Hin	Kan	Mar	Tam	Tel	Urd
Bhashaverse	84.65%	82.56%	64.43%	71.18%	59.60%	81.96%	77.71%	75.23%
Google Translate	85.09%	82.47%	64.75%	70.50%	59.89%	82.35%	79.23%	75.07%
GPT-4o	86.02%	82.15%	64.66%	70.15%	60.24%	81.85%	78.87%	74.76%
LlaMa	86.29%	82.60%	64.64%	69.93%	58.87%	81.84%	77.21%	74.52%
Sarvam	84.94%	81.40%	64.85%	69.94%	59.18%	83.17%	78.63%	74.85%
Qwen	85.79%	81.77%	64.66%	70.75%	59.64%	82.47%	78.57%	74.62%
Gemma	84.75%	82.68%	64.76%	69.26%	59.62%	82.01%	78.19%	75.53%

Table 7: Precision scores of all models for personal pronouns across all models and languages

Models	Ban	Guj	Hin	Kan	Mar	Tam	Tel	Urd
Bhashaverse	82.66%	81.46%	70.58%	73.95%	64.81%	81.42%	82.66%	81.46%
Google Translate	83.16%	81.89%	70.95%	73.27%	65.34%	81.65%	83.16%	81.89%
GPT-4o	84.37%	81.93%	70.77%	72.66%	65.54%	81.67%	84.37%	81.93%
LlaMa	84.40%	81.60%	70.74%	71.83%	63.89%	81.16%	84.40%	81.60%
Sarvam	82.97%	81.04%	71.20%	71.95%	64.83%	82.87%	82.97%	81.04%
Qwen	84.21%	81.29%	70.71%	73.16%	64.82%	82.41%	84.21%	81.29%
Gemma	82.90%	81.93%	70.94%	71.25%	64.32%	81.67%	73.00%	78.41%

Table 8: Recall scores of all models for personal pronouns across all models and languages

Models	Ban	Guj	Hin	Kan	Mar	Tam	Tel	Urd
Bhashaverse	90.43%	93.54%	90.22%	89.82%	88.42%	89.83%	85.60%	88.42%
Google Translate	90.53%	92.98%	90.30%	89.36%	88.76%	89.87%	84.99%	88.41%
GPT-4o	90.30%	93.68%	90.09%	89.92%	88.54%	89.73%	85.29%	88.66%
LlaMa	90.96%	93.26%	90.05%	89.40%	88.43%	89.23%	85.83%	91.17%
Sarvam	90.60%	93.84%	89.57%	89.60%	88.38%	89.69%	84.96%	89.69%
Qwen	90.28%	94.00%	89.41%	89.19%	88.86%	89.94%	85.80%	88.83%
Gemma	90.52%	93.58%	90.13%	89.99%	88.82%	89.80%	85.44%	89.69%

Table 9: Precision scores of all models for lexical cohesion across all models and languages

Models	Ban	Guj	Hin	Kan	Mar	Tam	Tel	Urd
Bhashaverse	94.27%	94.27%	94.25%	88.39%	87.04%	94.54%	88.94%	95.38%
Google Translate	95.58%	93.37%	94.93%	88.24%	87.02%	94.81%	89.51%	94.77%
GPT-4o	94.89%	94.29%	95.09%	89.08%	86.30%	94.61%	89.15%	93.51%
LlaMa	94.68%	93.87%	93.88%	87.30%	86.84%	93.66%	89.48%	92.76%
Sarvam	95.64%	93.91%	94.12%	88.56%	86.32%	94.05%	88.63%	93.55%
Qwen	94.97%	94.53%	94.80%	88.41%	87.59%	93.78%	89.47%	91.46%
Gemma	94.62%	94.49%	94.64%	88.42%	86.69%	94.04%	89.42%	93.74%

Table 10: Recall scores of all models for lexical cohesion across all models and languages