BaatCheet: A Multilingual Corpus for Dialogue Translation in Indian Languages

Priyanka Dasari*, **Yuvrajsinh D. Bodana***, **Vandan Mujadia**, **Arafat Ahsan**, **Dipti Misra Sharma**, **Parameswari Krishnamurthy**

Language Technologies Research Centre, IIIT Hyderabad, India

{dasari.priyanka,yuvrajsinh.bodana}@research.iiit.ac.in, vmujadia@gmail.com, {arafat.ahsan,dipti,param.krishna}@iiit.ac.in

Abstract

Existing translation models are typically trained on sentence-level and formal text, limiting their ability to capture everyday conversational dialogue phenomena such as informality, speaker interaction, and discourse coherence. Most existing Indic translation resources and evaluation benchmarks focus on sentence-level or formal text, making it difficult to assess translation quality of the dialogue phenomena. In this work, we introduce *BaatCheet*, a multilingual dialogue corpus named after the Hindi term for *conversation* or *chitchat*, containing approximately 49,000 dialogues for dialogue translation across five translation directions. We fine-tune five open-source LLMs across seven training data configurations and find that fine-tuning yields substantial gains over zero- and few-shot baselines. To comprehensively evaluate dialogue translation quality, we employ multiple evaluation strategies, including automatic metrics, LLM-as-judge, and human assessments using an SQM-guided Direct Assessment (DA) Protocol. The dataset is available here[1].

1 Introduction

Dialogue is a fundamental mode of human communication through which individuals express emotions, intentions, and maintain social relationships (Gumperz, 1982; Halliday, 1987). In the sociolinguistic context of the Indian subcontinent, these interactions are predominantly multi-turn and informal, often involving dynamic code-mixing between regional and global languages. Such dialogues diverge from formal written text, which exhibits fragmented syntax, elliptical structures, and heavy use of colloquial expressions that reflect the spontaneity of natural speech. With the growing need for cross-lingual communication in multilingual societies like India, machine translation (MT) has become pivotal in bridging language barriers. Most existing MT systems are primarily trained on formal, sentence-level parallel corpora derived from news and governance domains (Koehn, 2005; Mujadia and Sharma, 2025). Such systems often struggle to preserve the pragmatic and contextual nuances inherent in natural conversation (Sung et al., 2024) and this gap persists even for LLM-based translation systems (Choudhary et al., 2025), a trend our own baseline evaluation reaffirms for informal Indic dialogue (§3.2).

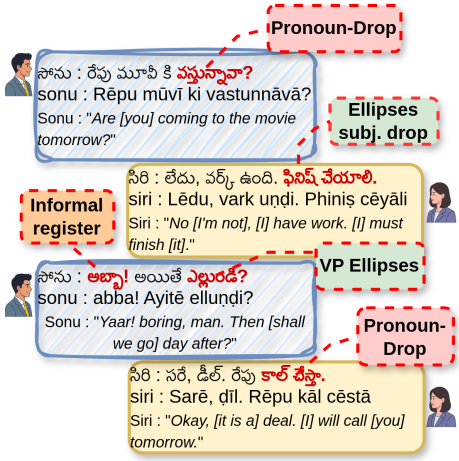

Figure 1: Highlighting Linguistic phenomena in informal Telugu dialogue. Dashed boxes highlight examples of pro-drop, ellipses, and register-specific informal expressions in a sample conversation.

To illustrate these challenges, we present a sample dialogue in Telugu, a major Dravidian language[2], in Figure 1. Unlike the formal datasets commonly used in MT research, this spontaneous conversation is replete with code-mix, pro-drop (where both subjects and objects are null) and contextual ellipsis, where utterances like 'No' (lēdu) and 'Then day after?' (ayitē elluṇḍi?) carry a

*Authors contributed equally.

[1]https://huggingface.co/datasets/HimangY/BaatCheet-Corpus

[2]Throughout this paper we use *Indic* as an umbrella term for the languages of the Indian subcontinent covered in our corpus, spanning both the Indo-Aryan (e.g., Hindi, Gujarati) and Dravidian (e.g., Tamil, Telugu) language families; we use *Dravidian* specifically when referring to that family.

Proceedings of the Eleventh Conference on Machine Translation, pages 116–145
October 28-29, 2026 ©2026 Association for Computational Linguistics

high semantic load that is only recoverable through multi-turn reasoning.

For example, a major challenge in dialogue translation is the encoding of social relationships through honorific pronouns. Unlike English, which uses a single second-person pronoun *you* across all social contexts, Indian languages grammaticalize social distance and respect, directly distinguishing formal and informal pronouns. These distinctions carry pragmatic weight that MT systems routinely collapse (Hingrajiya et al., 2026). This limitation is highly consequential in high-stakes applications (e.g., healthcare, customer support, and travel assistance), where misinterpreting register or speaker intent can degrade communication. To evaluate these phenomena, sentence-level parallel test sets exist for Indic languages (Team et al., 2022), but no dialogue-level test set is available for evaluating informal dialogue translation.

We address this gap by introducing **BaatCheet**, named after the Hindi term for *conversation* or *chitchat*, a multilingual corpus for informal dialogue translation spanning English, Hindi, Tamil, and Telugu. It is built from combining existing dialogue resources and synthetically generated dialogues, as illustrated in Figure 2. The dataset consists of two primary components: (i) **BaatCheet_Gold**, containing gold-standard human translations split into *BaatCheet_Human_Train* and a *BaatCheet_Test*; and (ii) **BaatCheet_Synthetic**, comprising two training subsets: *BaatCheet_Syn_X2IL*- a set whose monolingual dialogue data was collected and created in source languages and translated into target languages and *BaatCheet_Syn_IL2X* whose monolingual dialogue data was collected and created in target languages and those were translated into source languages and column swapped to make it source to target parallel data for fine-tuning purpose.

Using this corpus, we fine-tune five open-source LLMs across seven training data configurations spanning synthetic-only, human-only, and hybrid compositions to study how data provenance, scale, and quality interact with model capacity. To evaluate translation quality beyond automatic metrics, we conduct human and LLM-as-a-judge evaluations using the SQM-guided DA framework (Bojar et al., 2017; Perrella et al., 2024).

The primary contributions of this work are as follows:

1. ***BaatCheet***: A multilingual dialogue corpus for English → {Hindi, Telugu, Tamil} and Hindi → {Telugu, Tamil} featuring human translations and synthetically generated dialogues.
2. ***BaatCheet_Gold***: A curated gold-standard test-set for evaluating informal dialogue machine translation in Indic languages.
3. A structured data creation pipeline employing automatic metric-guided quality selection for high-quality synthetic training corpus curation.
4. A fine-tuning study of five open-source LLMs across multiple data configurations, tailored for informal dialogue translation.
5. A large-scale human and LLM-as-judge evaluation using SQM-guided DA establishing the reliability of automatic and LLM-based assessment for this task.

Specifically, our work addresses the following key research questions:

- **RQ1:** How well do existing open-source LLMs, dedicated MT systems, and proprietary LLMs translate informal dialogues into Indic languages?
- **RQ2:** How do the two synthetic data axes, *BaatCheet_Syn_X2IL* and *BaatCheet_Syn_IL2X*, differ in their effect on fine-tuned dialogue translation quality?
- **RQ3:** How well do automatic metrics, and LLM-based judges agree with human assessment of informal dialogue translation?

2 Related work

Early research on document and dialogue-level machine translation focused on incorporating broader discourse history to improve translation coherence (Maruf et al., 2021). Traditional approaches leveraged source and target-side conversational history to model multi-speaker dynamics (Maruf et al., 2018), or integrated extra-textual speaker identity cues to improve personalized translation (Vincent, 2021). The emergence of multilingual Large Language Models (LLMs) has further advanced this domain by enabling zero-shot context handling (Zhang et al., 2023) and generating high-quality synthetic parallel text to address low-resource gaps (Jiao et al., 2023).

Another line of work focuses on constructing dialogue datasets to support conversational machine translation research. The DailyDialog dataset (Li et al., 2017) provides a widely used multi-turn conversational dataset containing human-written ev-

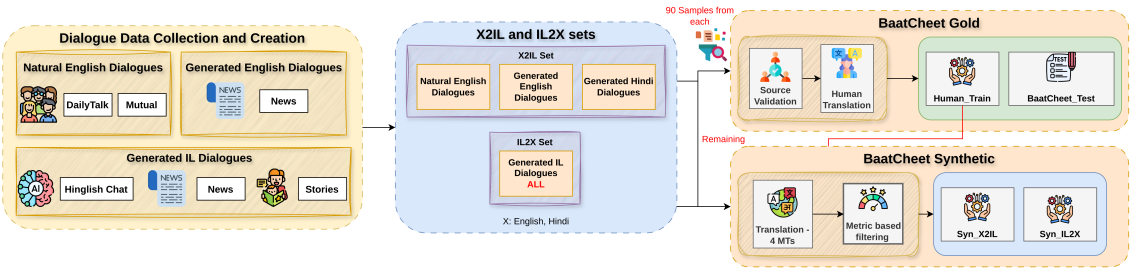

Figure 2: A flowchart of the BaatCheet Corpora preparation process for training and evaluation. The Metric-based methodology for selecting high-quality synthetic is detailed in Section 3.3.

eryday dialogues. Subtitle-based resources such as OpenSubtitles (Lison and Tiedemann, 2016) have also been used extensively for training conversational translation systems. (Wang et al., 2016) proposed methods to automatically construct parallel dialogue corpora by aligning subtitle data with monolingual scripts while preserving speaker and discourse information. More recently, (Majewska et al., 2023) introduced an outline-based approach for generating multilingual dialogue datasets, while (Bawden et al., 2021) developed the *DiaBLa* English–French corpus of informal written dialogues specifically designed, for evaluating context-aware machine translation systems. Additionally, IndicDialogue (Arnob et al., 2024) provides subtitle-based dialogue corpora across multiple Indic languages.

For Indic languages, benchmarks like FLORES-200 (Team et al., 2022) and the conversational subset IN22-Conv (Gala et al., 2023) evaluate translation across standard formal registers.

However, many of these datasets focus primarily on sentence-level translation or subtitle-style dialogues and often lack the informal conversational structures, and code-mixed characteristics that are common in real-world multilingual conversations across Indian languages. Standard reference-based metrics like BLEU and sentence-level COMET (Rei et al., 2022) often fail to capture dialogue-level coherence and register fidelity. Recent advances introduce context-aware metrics like DOC-COMET (Vernikos et al., 2022) alongside LLM-as-a-Judge frameworks for qualitative assessment (Zheng et al., 2023).

3 Building BaatCheet Corpora

BaatCheet Corpora is a multilingual dialogue corpus spanning English, Hindi, Tamil, and Telugu, designed to support training and evaluation of LLMs and MT systems for informal dialogue. In total, the corpus comprises approximately 49,000 dialogues, 383,000 utterances, and 690,500 sentences, organized into two major components: (i) *BaatCheet_Synthetic*, a large-scale machine-translated corpus subdivided into *BaatCheet_Syn_X2IL* and *BaatCheet_Syn_IL2X*; and (ii) *BaatCheet_Gold*, a human-translated set subdivided into *BaatCheet_Human_Train* and *BaatCheet_Test*.

Data Sources. The corpus draws on three categories of dialogue sources spanning all four languages: **(i) Natural English Dialogues:** *DailyTalk* (Lee et al., 2022), and *MuTual* (Cui et al., 2020) are human-authored two-speaker daily conversational corpora. **(ii) Generated English Dialogues:** *News-Eng**[3], informal dialogues generated using GPT-4o-mini from English news articles. **(iii) Generated IL Dialogues:** *Hinglish-Chat* (Khatri, 2024), an existing LLM-generated dialogue dataset subsequently refined via prompting to reduce robotic phrasing and improve conversational fluency; and *News-Hin**, dialogues generated from Hindi news articles, *News-Tam**, *News-Tel** and *Story-Tam**, *Story-Tel** dialogues generated from Tamil and Telugu news articles and short stories, respectively. A full description of each data source, including generation prompts and language coverage, is provided in Table 5 (Appendix C).

Human Validation of Generated Dialogues. Large Language Models often generate synthetically fluent text that may suffer from contextual pronoun mismatches, rigid/overly formal phrasing, or awkward code-mixing. To address this, native speakers validated and edited a small sample of LLM-generated dialogues, which comprise both BaatCheet_Gold source dialogues (Eng-src: News-Eng and Hin-src: Hinglish, News-Hin) and other generated dialogue sources across all lan-

[3]All * datasources are generated dialogues with GPT-4o-mini

guages. Across 344 dialogues containing 4,487 turns, 71.80% of dialogues required at least one human intervention, with 37.24% of overall individual turns modified. Table 6 (Appendix C.1) detailed analysis along with TER and CER.

3.1 BaatCheet_Gold

Dialogue Sampling. For English source dialogues, we stratify a sample 90 dialogues each from *DailyTalk*, *MuTual*, and *News-Eng**. For Hindi, 90 dialogues each are sampled from *Hinglish-Chat* and *News-Hin**. From each source, dialogues are partitioned into *BaatCheet_Human_Train* and *BaatCheet_Test*.

1. ***BaatCheet_Human_Train***: Used both as the reference set for Automatic metric-guided quality selection to identify the highest-quality translations for curating the synthetic training corpus, i.e., *BaatCheet_Synthetic* (detailed in section 3.3), and as gold training data for fine-tuning our selected translation models.
2. ***BaatCheet_Test***: Reserved exclusively as a held-out evaluation set for all translation experiments.

These validated BaatCheet_Gold source dialogues are human-translated by our inhouse translators. Translation guidelines, quality review processes, and detailed source validation metrics are described in Appendix C.4.

3.2 Baselines Evaluation on BaatCheet_Human_Train

With our established human-translated *BaatCheet_Human_Train* reference set, we conduct a systematic evaluation of existing MT systems and open-source and proprietary LLMs on informal Indic dialogue translation across our target directions (Eng → IL and Hin → IL). This evaluation serves a dual purpose: to test current models' capability on informal dialogue, and to identify which systems produce translations of sufficient quality to serve as candidate MT engines for large-scale synthetic data generation.

We evaluate nine systems on *BaatCheet_Human_Train* using reference-based COMET and reference-free COMET-QE across three categories: (i) five open-source LLMs (*Llama-3.1-8B-Instruct*[4] (Meta, 2024), *Llama-3.2-3B-Instruct*[5], *Qwen3-4B-Instruct*[6] (Team, 2025), *Gemma3-4B-IT*[7] (Team et al., 2025), and *Sarvam-Translate* [8].); (ii) three dedicated MT systems (*Google Translate* (GMT)[9], *IndicTrans2* (Gala et al., 2023), and *BhashaVerse*(Mujadia and Sharma, 2025) [10]); and (iii) one proprietary LLM (*GPT-4o-mini* (Jiao et al., 2023)).

Figure 3 reports scores across all five language directions. The three dedicated MT systems alongside *GPT-4o-mini* consistently outperform among all models under both metrics (ref-based avg. 0.803–0.848; ref-free avg. 0.823–0.857). Among open-sourced LLMs, *Gemma3-4B* and *Sarvam-Translate* are the strongest, while *Qwen3-4B* and *Llama-3.2-3B* score lowest (avg. below 0.54 and 0.59 respectively). These results reveal a clear quality gap between established MT systems and open-source LLMs for informal Indic dialogue translation, motivating our approach of using the four top-performing systems to generate large-scale synthetic training data, and subsequently fine-tuning the five open-source LLMs on this curated corpus to evaluate whether the gap can be closed.

3.3 BaatCheet_Synthetic Curation Pipeline

Synthetic Dialogue Sources. To enrich training data diversity and mitigate MT-bias on the Indic target side, we curate synthetic dialogues along two translation axes, as illustrated in Figure 2.

1. **Syn_X2IL:** Comprises English source dialogues *DailyTalk*, *MuTual*) and *News-Eng** machine-translated into Hindi, Tamil, and Telugu; and Hindi source dialogues *Hinglish-Chat*, *News-Hin** are machine-translated into Tamil and Telugu.
2. **Syn_IL2X:** Comprises Tamil dialogues (*News-Tam**, *Story-Tam**) and Telugu dialogues (*News-Tel**, *Story-Tel**), machine-translated into English and Hindi; and Hindi dialogues (*Hinglish-Chat*, *News-Hin**) machine-translated into English. All resulting pairs are column-swapped to align with our target language pair direction X→IL (English, Hindi to IL) training direction while fine-tuning our selected open-sourced LLMs, ensuring that the Indic target side in

[4]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[5]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[6]https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507

[7]https://huggingface.co/google/gemma-3-4b-it

[8]https://huggingface.co/sarvamai/sarvam-translate

[9]https://translate.google.co.in/?sl=auto&tl=en&op=translate

[10]https://ssmt.iiit.ac.in/translatev3

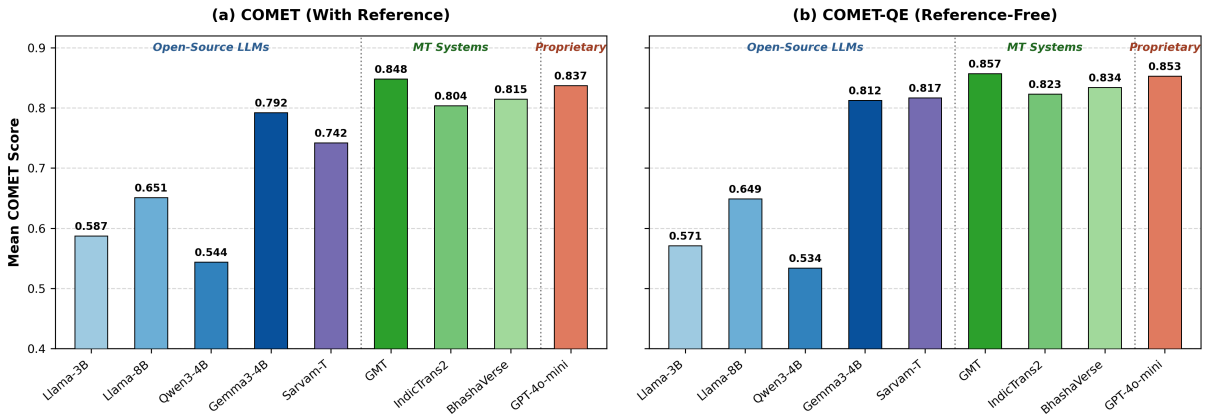

Figure 3: Baseline performance comparison across nine candidate systems on *BaatCheet_Human_Train* using (a)reference-based COMET and (b)reference-free COMET-QE. Dedicated MT systems (*GMT*, *IndicTrans2*, *BhashaVerse*) and *GPT-4o-mini* consistently outperform open-source LLMs across all language directions.

Syn_IL2X comprises dialogues generated natively in Indic style, preserving informal dialogue characteristics rather than just MT-biased output.

Selecting Candidate MT Systems. Based on our baseline evaluation on BaatCheet_Human_Train (Section 3.2, Figure 3), the four top-performing systems, *Google Translate* (GMT), *GPT-4o-mini*, *IndicTrans2*, and *BhashaVerse* are selected as candidate MT outputs.

Each system independently translates all *BaatCheet_Human_Train* dialogues, producing four parallel candidate translations per source utterance across both directions, forming *Human_Train_X2IL* and *Human_Train_IL2X* candidate sets. Language-pair disaggregated COMET scores of all four MT engines on these candidate sets are provided in Appendix D.1.

Automatic Metric Guided Quality Selection. We apply utterance-level reference-based (COMET) and reference-free (COMET-QE) scoring on *Human_Train_X2IL* and *Human_Train_IL2X* candidate sets to select the highest-quality translation per utterance from the four candidates, comparing two selection strategies:

1. ***Utt-Avg* (Utterance-Average):** Selects the single MT system that achieves the highest average utterance score across an entire dialogue.
2. ***Utt-High* (Utterance-Highest):** Selects the highest-scoring candidate translation independently for every individual utterance within a dialogue.

Analysis of MT candidate selection counts (Appendix D.2) reveals that *GMT* and *GPT-4o-mini* are the most frequently selected engines under both strategies across all language directions.

Validation of Selection Strategies and Human Alignment. To determine which filtering strategy yields higher-quality translations on BaatCheet_Human_Train candidates, we compare Utt-Avg and Utt-High using DOC-COMET (Vernikos et al., 2022) and COMTAIL (Ahsan et al., 2025), plus a human binary preference study on 510 dialogues (255 per direction, X2IL and IL2X), where annotators chose between COMET-QE-curated Utt-Avg and Utt-High outputs. Both automatic metrics consistently favor Utt-High (Appendix D.3), while humans preferred Utt-Avg in 58% of dialogues (304/510) versus 42% for Utt-High. This divergence is smaller than it appears: in over 90% of dialogues, a single MT system contributes most utterances, so the two strategies select highly similar data. For example, in a 13-utterance dialogue (Appendix D.4), Utt-High independently picks GMT for 8 turns, and GMT is also the top engine under Utt-Avg's dialogue-level average (0.8697). The two strategies converge on the same dominant engine even though candidate scoring is done turn-by-turn versus dialogue-wide. Disagreement concentrates in the remaining $\sim$10%, where a dialogue-level average can be inflated by a few strong utterances that mask weaker ones (Appendix D.4). Since our downstream goal is SFT data where turn-level accuracy directly shapes model output, we prioritize Utt-High's guaranteed per-utterance quality

BaatCheet Corpus	#Dialogues	#Utterances	#Sentences		Avg Dlg Len	Avg Utt Len		Avg Sent Len	
			Src	Tgt		Src	Tgt	Src	Tgt
BaatCheet_Syn_X2IL	42,148	301,278	533,102	533,494	8.8	14.8	11.8	8.7	6.9
BaatCheet_Syn_IL2X	6,096	73,666	135,894	136,616	12.3	14.6	11.8	8.4	6.9
BaatCheet_Human_Train	510	8,405	16,455	16,077	17.3	17.6	14.1	9.1	7.5
BaatCheet_Test	658	8,904	16,813	16,518	13.7	16.9	14.1	8.9	7.5
TOTAL	**49,412**	**392,253**	**702,264**	**702,705**	**13.0**	**16.0**	**13.0**	**8.8**	**7.2**

Table 1: Overall statistics of the *BaatCheet* corpus across 4 splits.

floor over Utt-Avg's holistic stylistic consistency. We therefore adopt *Utt-High* for selecting the best-scoring translation at each turn via COMET-QE to construct *BaatCheet_Synthetic*.

BaatCheet Corpus Overview. Based on these findings, we curate the final *BaatCheet_Synthetic* dataset using reference-free COMET-QE under the *Utt-High* strategy. Table 1 summarizes *BaatCheet* across its four splits, comprising 49,412 dialogues, 392,253 utterances, and ∼702k sentences. Large-scale synthetic data includes Syn_X2IL (42.1k dialogues) and Syn_IL2X (6.1k dialogues) with native Indic targets to reduce MT bias for our target direction X → IL fine-tuning of models. Human-translated dialogue splits contain 510 dialogues in Human_Train and 658 held-out dialogues in BaatCheet_Test. Full breakdown appears in Table 7, Appendix C.2.

Linguistics Properties. Automatic language-ID tagging (Madhani et al., 2023) measures an average code-mixing rate of ∼20% across language pairs, although this is likely an underestimate because it relies on Roman-script detection and misses the pervasive practice of transliterating English words into native Indic scripts. The corpus also preserves *pro-drop* (null subjects/objects) and syntactic *ellipsis*, especially Tamil and Telugu permit subjects and objects to be omitted when recoverable from the discourse context, as illustrated in Figure 1. The translation challenge is not pro-drop itself, but correctly recovering omitted arguments when the target language requires explicit realization. Statistically evaluating this recovery is, however, not straightforward, as the Auto Pronoun Translation metric (APT (Miculicich Werlen and Popescu-Belis, 2017)) is designed for pro-drop-scarce European languages and is unreliable in conversational Indic settings where omission is grammatical rather than exceptional(Hingrajiya et al., 2026).

4 BaatCheet Experimental Setup

We select a structurally diverse suite of decoder-only instruction-tuned models: *Llama-3.1-8B-Instruct* (Meta, 2024), *Llama-3.2-3B-Instruct*, *Qwen3-4B-Instruct* (Team, 2025), *Gemma3-4B-IT* (Team et al., 2025), and *Sarvam-Translate* [11]. These models are chosen for their pre-training exposure to Indic languages, multilingual baselines, and optimized tokenizers for low-resource scripts.

To establish strong baselines for comparing zero-shot and few-shot translation performance against supervised fine-tuning, we evaluate all models on *BaatCheet_Test* under two settings:

- **Zero-Shot Baseline:** Models translate dialogue prompts directly without in-context examples.
- **Few-Shot Baseline:** Implemented an adaptive retrieval strategy to dynamically select the most contextually relevant in-context examples. Specifically, we generate dense embeddings for all dialogues in the *Human_Train* set using the `all-MiniLM-L6-v2` model, which are indexed using FAISS. For each test dialogue, we retrieve the top-k mathematically closest context examples.

Fine-Tuning Setup. Models are loaded in 4-bit NormalFloat (NF4) quantized format using *bitsandbytes* (Dettmers et al., 2022) and trained using bfloat16 precision for numerical stability. The absolute context length is constrained to 4096 tokens. To prevent truncation of long, multi-turn dialogue contexts, we implement a *length-aware chunking* protocol: dialogues exceeding the maximum token budget are segmented into contiguous sub-dialogues, with each segment prefixed by a custom control token, `<splitted_dialogue>`, to preserve contextual and structural continuity. This chunking mechanism is applied identically during both training and inference. The training prompt template is shown in Appendix section E.2.

[11] sarvamai/sarvam-translate

Model	Baselines		Synthetic-Only		Human-Only	Dual-Hybrid		Multi-Hybrid	
	Zero-Shot	Few-Shot	Syn_X2IL	Syn_IL2X	HT	Syn_X2IL+HT	Syn_IL2X+HT	Syn_Both+HT	Syn_Both+3×HT
(a) COMET (Reference-Based)									
Gemma3-4B	0.5024	0.6410	0.8076	0.7095	**0.8324**	0.8192	0.8223	0.7397	0.6087
Llama-3.2-3B	0.5317	0.5864	**0.7802**	0.7279	0.7139	0.7410	0.7076	0.7647	0.7271
Llama-3.1-8B	0.7113	0.6882	0.8023	0.7418	0.7871	0.7197	0.7210	0.7930	**0.8461**
Qwen3-4B	0.6797	0.6601	**0.8373**	0.7372	0.7959	0.8063	0.7370	0.7695	0.8288
Sarvam-T	0.6168	0.4859	0.8149	0.8057	0.8320	**0.8573**	0.8499	0.8471	0.8411
(b) DOC-COMET (Reference-Based)									
Gemma3-4B	0.4145	0.3755	0.7513	0.6342	**0.7768**	0.7673	0.7704	0.6666	0.5771
Llama-3.2-3B	0.5684	0.5152	0.6467	0.6113	0.6085	0.6418	0.6104	**0.6622**	0.6585
Llama-3.1-8B	0.6485	0.6067	0.6899	0.6418	0.7459	0.6265	0.6324	0.6799	**0.7926**
Qwen3-4B	0.6008	0.5819	**0.7846**	0.6337	0.7408	0.7488	0.6716	0.7097	0.7843
Sarvam-T	0.4954	0.3491	0.7604	0.7608	0.8023	**0.8043**	**0.8043**	0.7916	0.7955
(c) COMTAIL (Reference-Based)									
Gemma3-4B	0.2556	0.4791	0.7017	0.5655	**0.7244**	0.7007	0.6986	0.6011	0.4552
Llama-3.2-3B	0.3615	0.4169	**0.6813**	0.5835	0.5722	0.5999	0.5608	0.6596	0.6037
Llama-3.1-8B	0.5230	0.5489	0.7072	0.5918	0.6481	0.5711	0.5704	0.6945	**0.7563**
Qwen3-4B	0.5648	0.5309	**0.7506**	0.5904	0.6719	0.6810	0.5956	0.6764	0.7372
Sarvam-T	0.4723	0.3126	0.7202	0.6896	0.7328	0.7554	0.7443	**0.7660**	0.7582

Table 2: Performance comparison of five open-source LLMs across fine-tuning configurations on *BaatCheet_Test*, evaluated using (a) COMET DA, (b) DOC-COMET, and (c) COMTAIL DA. `Syn_Both` denotes Syn_[X2IL+IL2X] and `HT` denotes Human_Train. Bold values indicate the best-performing data configuration for each respective model.

Parameter-Efficient Fine-tuning. We employ Parameter-Efficient Fine-Tuning (PEFT) leveraging QLoRA combined with Rank-Stabilized LoRA (rsLoRA) (Kalajdzievski, 2023). Specifically, rsLoRA enhances training stability across linguistically distinct target vocabularies by scaling the adapter updates with a factor proportional to $\alpha/\sqrt{r}$ (where r denotes the LoRA rank and α is the scaling factor), rather than the traditional α/r scaling. LoRA adapter weights, particularly matrix A is Gaussian-initialized to encourage diverse feature representations and matrix B is initialized to zeros. The complete set of training hyperparameters is detailed Table 13 in the Appendix.

4.1 Training Data Configurations

To systematically investigate the effect of dataset composition on informal dialogue translation quality, we define seven configurations spanning five categories:

1. **Baselines:** The zero-shot and few-shot baseline performance of each model.
2. **Synthetic-Only:** Fine-tuning exclusively on each synthetic split (`Synt_X2IL` vs. `Synt_IL2X`) in isolation, to evaluate whether training on target-side machine translations (`Synt_X2IL`) versus column-swapped native Indic target text (`Synt_IL2X`) more effectively boosts fine-tuned translation performance.
3. **Human-Only:** Fine-tuning strictly on gold standard translation set- *Human_Train*.
4. **Dual-Hybrid:** Blending a single synthetic dialogue set with human gold standard - *Syn_X2IL +HT* and *Syn_IL2X +HT* to evaluate the complementary gain from combining machine-translated scale with human-translated dialogues.
5. **Multi-Hybrid:** Blending all synthetic dialogue sets - *Syn_X2IL* and *Syn_IL2X* with human translation sets at two scaling ratios: *Syn_Both +HT* and *Syn_Both + 3×HT* to assess whether increasing the proportion of human-translated data further improves translation quality at scale.

5 Results and Discussion

We evaluate five fine-tuned LLMs across all dataset configurations on *BaatCheet_Test* using three automatic metrics (Table 2). Detailed results across all five LLMs on all configurations are provided in the Appendix Section E.3.

5.1 Metric Behaviour.

COMET (Rei et al., 2022) scores are consistently the highest across all configurations (ranging 0.49–0.86). Being a sentence-level metric, COMET rewards local lexical and semantic adequacy per utterance independently, without modelling cross-turn discourse coherence.

DOC-COMET (Vernikos et al., 2022) scores are systematically lower (0.35–0.79), as it incorpo-

rates up to three preceding utterances as additional context signal. While this enables coarse dialogue-level coherence scoring, its limited context window is insufficient to capture long-range dependencies in multi-turn conversations. Consequently, translations that are locally fluent but contextually incoherent are penalized, resulting in lower overall scores compared to sentence-level COMET.

COMTAIL (Ahsan et al., 2025) produces the lowest scores overall (0.25–0.77). As an Indic- specialized metric, COMTAIL may be more sensitive to Indic language-specific imperfections in lexical choice, morphology, and domain register. However, since COMTAIL is primarily fine-tuned on formal sentence-level parallel data, it may lack calibration for informal dialogue registers, resulting in conservatively lower scores.

5.2 Training Data Configuration Observations

Baselines. Few-shot performance is consistently lower than zero-shot for most models (e.g., *Sarvam-T*: $0.6168 \rightarrow 0.4859$ COMET; *Gemma*: $0.4145 \rightarrow 0.3755$ DOC-COMET). We attribute this to two compounding issues: (i) retrieved in-context examples from *BaatCheet_Human_Train* may not align well in topic or register with test dialogues, introducing distributional noise into the prompt; and (ii) multi-turn dialogue examples significantly expand input length, leading to context window saturation and truncation of earlier dialogue history, forcing the model to translate with incomplete context.

Synthetic-Only. Fine-tuning on Syn_X2IL consistently outperforms Syn_IL2X across all models and metrics. In Syn_X2IL, the source side largely comprises naturally occurring dialogues (*DailyTalk*, *MuTual*, *Hinglish-Chat*) with machine-translated Indic targets (Section 3.3). In contrast, Syn_IL2X is entirely synthetic on both sides: its source (English/Hindi) is machine-translated output, while its Indic targets are *GPT-4o-mini*-generated dialogues which, despite being prompted for informal code-mixed style, lack the naturalistic register variation inherent in human-produced dialogues. Combined with its substantially smaller scale (6.1k vs. 42.1k dialogues), this fully synthetic construction limits the quality and diversity of the SFT signal in Syn_IL2X.

Human-Only. Despite comprising only 510 dialogues, *BaatCheet_Human_Train* alone achieves the highest COMET score for *Gemma3-4B* (0.832) and competitive results for *Sarvam* (0.832). Models with strong multilingual pre-training appear capable of efficiently leveraging small but high-quality human-translated corpora.

Dual-Hybrid. Syn_X2IL + Human_Train yields peak performance for three of five models: *Llama-3.2-3B* (0.765 COMET), *Sarvam* (0.857), and *Qwen* (0.806). Syn_IL2X + Human_Train, by contrast, consistently underperforms its counterpart across all models. Beyond the fully synthetic nature of Syn_IL2X, its substantially smaller volume fails to add meaningful linguistic diversity to *Human_Train* during SFT.

Multi-Hybrid. Syn_Both + 3 × Human_Train delivers peak scores for *Llama-3.1-8B* (COMET: 0.846, DOC-COMET: 0.793, COMTAIL: 0.756) and strong gains for *Qwen3-4B* (COMET: 0.829 vs. 0.770 with 1 × Human_Train). However, Syn_Both + 1 × Human_Train noticeably underperforms noticeably *hurts* these models relative to their Syn_X2IL + Human_Train dual-hybrid performance in Gemma, which we attribute to a *gold dilution effect*: the ∼48k combined synthetic volume overwhelms the 510-dialogue *Human_Train* set, exposing the model disproportionately to noisier synthetic signal. Tripling *Human_Train* (3×) restores the quality-to-volume ratio, providing sufficient human-translated exposure for larger-capacity models to calibrate towards authentic informal dialogue quality.

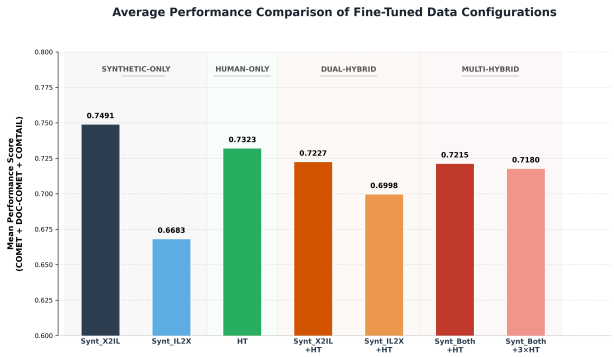

Figure 4: Average translation performance (mean score across all five fine-tuned models and three evaluation metrics: COMET, DOC-COMET, and COMTAIL) for each fine-tuned training configuration.

5.3 Comparative Analysis of Training Data Configurations

To determine which data composition strategy most effectively boosts model performance, Figure 4

		GPT Judge				Gemini Judge				Human Judge			
		Utt-Level	Dia-Level Factors			Utt-Level	Dia-Level Factors			Utt-Level	Dia-Level Factors		
Model	**Configuration**	**Acc**	**Nat**	**Inf**	**Coh**	**Acc**	**Nat**	**Inf**	**Coh**	**Acc**	**Nat**	**Inf**	**Coh**
Gemma	Syn_X2IL+HT	48.58	42.79	48.82	46.66	47.94	45.95	48.86	47.69	49.98	46.45	49.63	47.46
	Human_Train	50.31	48.99	50.52	50.92	50.54	51.62	52.25	52.79	51.52	50.20	50.15	50.67
Llama-3B	Syn_X2IL+HT	43.14	39.55	46.76	37.47	41.31	37.49	41.81	37.13	43.22	43.11	49.23	42.31
	Syn_X2IL	44.41	43.64	41.54	39.82	43.96	41.95	40.78	40.50	43.35	44.88	46.25	43.91
Llama-8B	Syn_Both+3×HT	52.73	55.93	51.35	55.41	53.00	55.45	52.62	55.23	52.48	53.61	50.03	53.67
	Human_Train	47.76	45.09	51.38	48.44	46.52	44.45	51.20	45.58	48.84	46.03	51.67	47.01
Qwen	Syn_Both+3×HT	50.68	52.59	48.68	52.73	51.47	52.06	49.86	52.06	51.42	51.05	48.32	51.25
	Syn_X2IL	51.86	52.61	47.80	52.66	51.73	52.46	49.93	52.60	51.92	51.30	48.50	51.56
Sarvam-T	Syn_X2IL+HT	53.99	58.28	55.64	57.30	54.86	58.18	56.38	57.55	54.76	55.38	52.90	55.32
	Human_Train	**54.76**	**60.55**	**57.62**	**58.66**	**56.18**	**60.59**	**57.32**	**59.20**	**55.53**	**58.04**	**53.92**	**57.18**

Table 3: Qualitative SQM-DA evaluation scores (z-score normalized and scaled) disaggregated by **Utterance-Level Accuracy** (Utt-level) and **Dialogue-Level Factors** (Dia-level): Naturalness (**Nat**), Informality (**Inf**), and Cohesion & Coherence (**Coh**). Evaluated across top-two data configurations per model by GPT-4o-mini, Gemini-Flash-Lite-preview, and Human Judges. Syn_Both denotes Syn_[X2IL+IL2X]. Bold text indicates the higher score per model setup.

compares the mean scores across all fine-tuning configurations (averaged across all models and evaluation metrics). Among all configurations, Synt_X2IL achieves the highest overall average score (**0.7491**), closely followed by the human-only baseline Gold (**0.7323**). Conversely, Synt_IL2X in isolation yields the lowest overall performance (**0.6683**), and incorporating it into hybrid setups consistently degrades overall translation quality compared to Synt_X2IL alone. Crucially, the responsiveness of each model to these data configurations is strongly governed by base architectural factors like model parameter capacity and pre-trained Indic linguistic knowledge.

5.4 Limitations of Automatic Metrics

While automatic metrics confirm substantial and consistent gains from our curated corpus, they do not fully capture pragmatic dimensions critical to informal dialogue translation, including honorific pronoun usage, pro-drop nature, and code-mixed naturalness.

6 LLM-as-Judge and Human Assessment

We conduct a qualitative evaluation using both human judges and LLM judges under a Scalar Quality Metric (SQM)-guided Direct Assessment (DA) protocol.

Experimental Setup and Sampling. For each of the five fine-tuned models, we select its top-two performing training configurations based on DOC-COMET scores, yielding 10 unique model-configuration setups. From each setup, we sample test dialogues across all language directions: 30 dialogues for each English-source pair (Eng → Hin/Tam/Tel) and 20 dialogues for the Hindi-source pair (Hin → Tel). In total, 1,100 dialogues (900 English-source and 200 Hindi-source) are independently evaluated by expert native-speaking linguists and two SOTA LLM judges: *GPT-5-mini* and *Gemini-3.1-flash-lite-preview*.

Scoring Granularity. Evaluations are conducted at two distinct levels: **Dialogue Level:** The judge assesses the full context on three dimensions: *Naturalness*, *Informality*, and *Cohesion & Coherence*. **Utterance Level:** Each sentence pair is independently scored to capture fine-grained translation accuracy. The detailed prompt templates and guidelines are given in section F in the Appendix.

6.1 Qualitative Assessment and Analysis

Cross-Judge Agreement. Table 3 reveals strong directional consistency across all three evaluators, GPT, Gemini, and Human, with judges largely agreeing on both the relative ranking of models and the preferred training configuration per model. This concordance validates the reliability of LLM judges as proxies for human assessment in the informal Indic dialogue translation setting.

Model-Specific Qualitative Trends. Disaggregating scores by model architecture reveals distinct qualitative behaviors:

- For *Gemma3-4B* and *Sarvam-T*, the `Human_Train` configuration consistently yields higher scores than `Syn_X2IL + HT` across all judges and all four evaluation dimensions: utterance accuracy, naturalness, informality, and cohesion. This is consistent with their automatic metric results (Section 5).
- *Llama-3.2-3B* records the lowest qualitative scores across all judges and dimensions, mirroring its modest gains under automatic metrics. This consistent under-performance across both evaluation paradigms confirms that its 3B parameter capacity remains the primary bottleneck, limiting its ability to capture discourse structure regardless of the training data configuration.
- For *Llama-3.1-8B* and *Qwen3-4B*, both evaluated configurations produce closely competitive qualitative scores across all judges, with neither configuration showing a decisive advantage over the other in dialogue-level factors. Notably, for these two models, Gemini and Human judgements exhibit stronger mutual agreement compared to GPT, suggesting that Gemini's evaluation tendencies better align with human preferences for dialogue-level naturalness and coherence assessment in this setting.

Overall SQM-guided DA Analysis. Across all models and judges, *Sarvam-T* under `Human_Train` achieves the highest qualitative scores on all three dialogue-level factors, particularly Naturalness (GPT: 60.55; Gemini: 60.59; Human: 58.04) and Cohesion (GPT: 58.66; Gemini: 59.20; Human: 57.18). This is fully consistent with its consistently strong performance in automatic evaluation (Section 5).

Granular Rubric Distribution for Sarvam-T. At the dialogue level, 85/130 instances were rated as *Grammatical and somewhat natural*, while 64/130 were classified as *Slightly Informal*. Regarding discourse flow, 66/130 dialogues were deemed *Somewhat coherent and cohesive*, with 52/130 achieving *Strong Coherence and Cohesion*. At the utterance level, over 90% of dialogues achieved high accuracy and naturalness (scores $\geq$ 61). A detailed, granular analysis per language pair, along with a sample dialogue error analysis, is provided in Table 17 and Figure 11 in Appendix F.2.

6.2 Inter-Judge Correlation Analysis

Table 4 reports average inter-judge correlations across all five models. Utterance-level accuracy achieves the strongest agreement across all judge pairs, confirming that fine-grained translation accuracy is consistently assessable by both human and LLM evaluators. However, when evaluating holistic dialogue-level factors, correlation drops noticeably across all evaluators, while Informality records the lowest correlations. This reveals a fundamental evaluation gap: while judges reliably agree on local-context utterance-level factors, evaluating on dialogue-level factors introduces subjectivity and the cultural-sensitive nature of Judges. This shows that holistic dialogue quality is far more challenging to standardize than utterance-level accuracy. Full per-model breakdowns are provided in Appendix G.

	GPT vs H		Ge vs Hu		GPT vs Ge	
Factor	r	τ	r	τ	r	τ
Nat	0.552	0.370	0.625	0.428	0.738	0.540
Inf	0.343	0.232	0.332	0.209	0.391	0.266
Coh	0.561	0.364	0.647	0.442	0.743	0.532
Utt	**0.702**	**0.498**	**0.697**	**0.525**	**0.781**	**0.597**

Table 4: Average inter-judge correlation (r = Pearson, τ = Kendall's τ) across all five models. Bold values indicate the highest agreement factor per judge pair(GPT, Gemini-**Ge** and Human-**H**.**Utterance-Level Accuracy** (Utt-level) and **Dialogue-Level Factors** (Dia-level): Naturalness (**Nat**), Informality (**Inf**), and Cohesion & Coherence (**Coh**). **Ge**.

7 Conclusion

We introduce ***BaatCheet***, a multi-turn parallel dialogue corpus for informal Indic translation across five language directions. Our evaluation reveals that while dedicated MT engines and *GPT-4o-mini* lead zero-shot baselines, fine-tuning open-source LLMs—particularly Indic-specialized models like *Sarvam-T* significantly bridges this gap. We demonstrate that synthetic data curated from natural source dialogues (`Synt_X2IL`) provides a far superior training signal compared to synthetic Indic sources (`Synt_IL2X`). Through our evaluations, we show that LLMs can be employed as judges for evaluating dialogue translations as proxies. Future extensions of this study will expand the size of BaatCheet_Human dialogues also to include Bengali, Gujarati, Kannada, and Marathi, maintaining the current source language pairs. Additionally, we aim to develop domain-specific datasets to further refine model performance for specialized downstream applications like healthcare.

Limitations

Synthetic target generation and MT translation may introduce subtle structural artifacts, lacking the spontaneous code-mixing and deep cultural nuances of organic human conversations. Human-translated gold data, while highly effective for fine-tuning, is limited in volume due to annotation costs, making dataset expansion a key priority. Systemic scoring differences between LLM judges (e.g., Gemini vs. GPT) require calibration against human baselines, and budget constraints limited human Direct Assessment to a sampled subset of 1,100 dialogues.

Acknowledgments

This work is supported by HIMANGY (HIndustani Machini ANuvaad TechnoloGY), a consortium-based initiative under BHASHINI, funded by the Ministry of Electronics and Information Technology (MeitY), Government of India. We gratefully acknowledge their support in funding in creating the BaatCheet Corpus.

References

Arafat Ahsan, Vandan Mujadia, Pruthwik Mishra, Yash Bhaskar, and Dipti Misra Sharma. 2025. Crosslingual optimized metric for translation assessment of indian languages. *Preprint*, arXiv:2509.17667.

Noor Mairukh Khan Arnob, A Faiyaz, Md Mubtasim Fuad, Shah Murtaza Rashid Al Masud, Baivab Das, and MF Mridha. 2024. Indicdialogue: A dataset of subtitles in 10 indic languages for indic language modeling. *Data in Brief*, 55:110690.

Rachel Bawden, Eric Bilinski, Thomas Lavergne, and Sophie Rosset. 2021. Diabla: a corpus of bilingual spontaneous written dialogues for machine translation. *Language Resources and Evaluation*, 55:635–660.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ritvik Choudhary, Rem Hida, Masaki Hamada, Hayato Futami, and Toshiyuki Sekiya. 2025. Exploring context strategies in LLMs for discourse-aware machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24382–24391, Suzhou, China. Association for Computational Linguistics.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki 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. *Preprint*, arXiv:2305.16307.

John J Gumperz. 1982. *Language and social identity*. 2. Cambridge University Press.

Michael AK Halliday. 1987. Spoken and written modes of meaning. In *Comprehending oral and written language*, pages 55–82. Brill.

Heli Hingrajiya, Vennela Bairi, Vandan Mujadia, Dipti Sharma, Parameswari Krishnamurthy, and Vasudeva Varma. 2026. IndicDISCO-MT: A discourse-centric benchmark for evaluating discourse phenomena in Indian language machine translation. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 190–204, Tilburg, The Netherlands. European Association for Machine Translation.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *Preprint*, arXiv:2312.03732.

Abhishek Khatri. 2024. Hinglish everyday conversations dataset.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2022. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. *Preprint*, arXiv:2207.01063.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Yash Madhani, Mitesh M. Khapra, and Anoop Kunchukuttan. 2023. Bhasha-abhijnaanam: Native-script and romanized language identification for 22 indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 816–826. Association for Computational Linguistics.

Olga Majewska, Evgeniia Razumovskaia, Edoardo M Ponti, Ivan Vulić, and Anna Korhonen. 2023. Cross-lingual dialogue dataset creation via outline-based generation. *Transactions of the Association for Computational Linguistics*, 11:139–156.

Sameen Maruf, André FT Martins, and Gholamreza Haffari. 2018. Contextual neural model for translating bilingual multi-speaker conversations. *arXiv preprint arXiv:1809.00344*.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Meta. 2024. Llama-3.2-3B-Instruct. `https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct`.

Lesly Miculicich Werlen and Andrei Popescu-Belis. 2017. Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 17–25, Copenhagen, Denmark. Association for Computational Linguistics.

Vandan Mujadia and Dipti Misra Sharma. 2025. Bhashaverse : Translation ecosystem for indian subcontinent languages. *Preprint*, arXiv:2412.04351.

Stefano Perrella, Lorenzo Proietti, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Beyond correlation: Interpretable evaluation of machine translation metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20689–20714, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mingi Sung, Seungmin Lee, Jiwon Kim, and Sejoon Kim. 2024. Context-aware LLM translation system using conversation summarization and dialogue history. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1011–1015, Miami, Florida, USA. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Qwen Team. 2025. Qwen3.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level mt metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128.

Sebastian T Vincent. 2021. Towards personalised and document-level machine translation of dialogue. *arXiv preprint arXiv:2102.10979*.

Longyue Wang, Xiaojun Zhang, Zhaopeng Tu, Andy Way, and Qun Liu. 2016. Automatic construction of discourse corpora for dialogue translation. *arXiv preprint arXiv:1605.06770*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Appendix

B Formal vs Informal Translation

Formal language relies on rigid grammatical structures and standardized vocabulary, whereas informal "BaatCheet" prioritizes social flow and spontaneity through pro-drop (omitted subjects) and code-mixing. For instance, in given example Figure 5, the formal translation uses explicit pronouns and literary metaphors like paheliyan sulajhana (solving riddles), which feel robotic in a casual setting. Conversely, the informal translation naturalizes the dialogue by adding social markers like

"Bhai" and utilizing common urban slang like "life set" instead of a literal translation for "figured out." This shift demonstrates that informal translation is a process of pragmatic transcreation, capturing the "vibe" and cultural nuances that formal, sentence-level MT systems often miss.

C BaatCheet Corpora Building

Table 5 presents the overview of dialogue data sources used for constructing BaatCheet Corpora across English, Hindi, Tamil, and Telugu.

C.1 Guidelines for Generated Dialogue Validation

Language experts are provided with the following standardized instructions to validate LLM-generated dialogues:

Validation Guidelines

1. Grammatical correctness and semantic coherence.
2. Contextual pronoun consistency across dialogue turns.
3. Gender agreement and awkward phrasing remediation.
4. Converting grammatically sound yet robotic or overly formal expressions into natural, colloquial style.
5. Refining code-mixed phrasing to ensure conversational naturalness.

Metric Calculation

Dialogue Edit Rate (%): The percentage of dialogues where at least one Utterance was modified by human:

$$\text{Dialogue Edit Rate} = \frac{N_{\text{edited dialogues}}}{N_{\text{total dialogues}}} \times 100$$

Utterance Edit Rate (%): The proportion of individual dialogue Utterances altered during validation:

$$\text{Utterance Edit Rate} = \frac{T_{\text{edited}}}{T_{\text{total}}} \times 100$$

Token Edit Rate (TER %): Corpus-level word-level Levenshtein edit distance (insertions, deletions, substitutions) normalized by total source word count:

$$\text{TER} = \frac{\sum_{i=1}^{T} \text{Levenshtein}_{\text{word}}(\text{Source}_i, \text{Target}i)}{\sum i = 1^{T} \text{WordCount}(\text{Source}_i)} \times 100$$

Character Edit Rate (CER %): Corpus-level character-level Levenshtein edit distance normalized by total source character length:

$$\text{CER} = \frac{\sum_{i=1}^{T} \text{Levenshtein}_{\text{char}}(\text{Source}_i, \text{Target}i)}{\sum i = 1^{T} \text{CharCount}(\text{Source}_i)} \times 100$$

C.2 Detailed BaatCheet Corpus Statistics

The entire *BaatCheet* corpus aggregates to 49,412 dialogues, 392,253 utterances, and over 702,000 parallel sentences as shown in Table 7.

- **Synthetic Volume (Syn_X2IL & Syn_IL2X):** The synthetic component provides the core scale necessary for LLM fine-tuning. BaatCheet_Syn_X2IL accounts for 42,148 dialogues (301,278 utterances), representing standard X $\rightarrow$ IL translations. BaatCheet_Syn_IL2X contributes 6,096 dialogues (73,666 utterances), which are column-swapped to provide natively generated Indic targets and eliminate machine translation bias on the target side.
- **Human Gold Standards (Human_Train & BaatCheet_Test):** The professionally translated human sets consist of 510 dialogues (8,405 utterances) in BaatCheet_Human_Train and 658 dialogues (8,904 utterances) in BaatCheet_Test. Human_Train acts as both a high-fidelity fine-tuning set and the quality-filtering reference for synthetic curation, while BaatCheet_Test is strictly held out for evaluation.

Diversity in Linguistic Properties Comparing Source vs. Target utterance lengths (in words) and sentence lengths reveals fundamental linguistic properties across language families:

Agglutination (Tamil & Telugu). In X $\rightarrow$ IL directions (Eng $\rightarrow$ Tam and Eng $\rightarrow$ Tel), English source utterances average 14.97–17.37 words, whereas Tamil and Telugu target utterances average only 11.18–12.76 words. Similarly, in IL $\rightarrow$ X directions (Tam $\rightarrow$ Eng and Tel $\rightarrow$ Eng), Tamil and Telugu source utterances (9.59–12.12 words) expand significantly when translated into English targets (11.86–16.50 words).

> *Linguistic Cause:* Tamil and Telugu are highly agglutinative Dravidian languages. Case markers, postpositions, tense inflections, and pronominal suffixes are morphologically bound into single word forms, resulting in systemat-

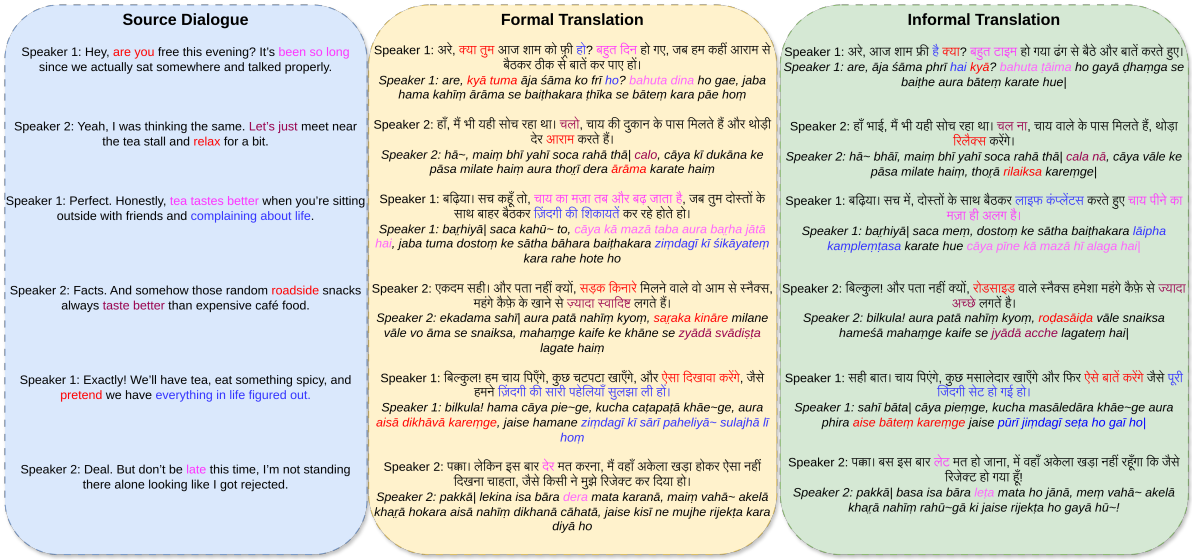

Figure 5: An example of Formal and Informal Translation from English to Hindi along with translitration in IAST format. All the differences are highlited with colors.

ically lower word counts compared to analytic English.

Domain Diversity By combining daily speech (*DailyTalk*), multi-turn reasoning (*MuTual*), code-mixed chat (*Hinglish-Chat*), generated news conversations (*News*), and fictional narratives (*Stories*), *BaatCheet* covers a broad spectrum of registers from task-oriented information exchange to informal chitchat.

C.3 Linguistic Properties: Code-Mixing & Pronoun Ellipsis of BaatCheet Corpus

Code-Mixing. To evaluate the prevalence of code-mixing in our corpus, we processed the datasets using the *Indic ILD* (Madhani et al., 2023) language detection tool. The tool identifies an average code-mixing rate of approximately 20% across our language pairs. However, we note that this baseline is a significant underestimate due to a crucial limitation of current automatic tools: Indic ILD relies primarily on the presence of the Roman script (English characters) to identify code-mixed tokens. In contrast, real-world colloquial Indic speech features extensive code-mixing using English words transliterated directly into native Indic scripts. Consequently, although the corpus exhibits a high density of conversational code-mixing, its true frequency cannot be precisely quantified by script-dependent automatic tools.

Pronoun-Drop and Ellipsis. Prior work in discourse-aware MT has proposed the Auto Pronoun Translation (APT (Miculicich Werlen and Popescu-Belis, 2017), which is specific for European Languages) metric to evaluate pronoun translation quality. We argue that adapting APT as a primary evaluation metric is highly unreliable in conversational Indic settings. Tamil and Telugu, like other pro-drop languages, permit subjects and objects to be omitted when they are recoverable from grammatical or discourse context. In dialogue translation, the challenge is therefore not the presence of pro-drop itself, but correctly recovering omitted information when translating into a language that requires or prefers its explicit realization as shown in the figure 1.

C.4 BaatCheet_Gold Translation Protocal

Prior to human translation, synthetically generated source dialogues of *BaatCheet_Gold* underwent expert editing to ensure naturalness, formality, and coherence, and the edit metrics are shown in Table 6 in section C.1.

Translation Protocol. Validated source dialogues were assigned to professional human translators with native-level proficiency in the target language and at least three years of translation experience. Per language pair, 2–4 translators independently produced target-language translations following strict informal dialogue guidelines.

1. manual translation only (no MT post-editing);
2. use of colloquial vocabulary; avoidance of formal or Sanskritized registers;

Data Source – Src.Lang	Description
Hinglish-Chat – Hin	A synthetically generated Hinglish conversational dataset (Khatri, 2024) featuring everyday life dialogues. The dataset is modified by LLM prompting to *"Remove unnatural nouns and pronouns usage from this dialogue, but it still has to make sense and write it in Hindi (devanagri) script"*
MuTual (Multi-Turn Dialogue Reasoning) – Eng	A Chinese high school English listening comprehension dataset (Cui et al., 2020), adapted, preprocessed, and manually annotated to create a multi-turn dialogue reasoning corpus aimed at evaluating dialogue models' contextual understanding and reasoning capabilities.
DailyTalk – Eng	A conversational speech dataset for TTS systems developed by sampling and modifying dialogues from the open-domain DailyDialog corpus (Li et al., 2017). Dialogues from DailyTalk (Lee et al., 2022) are used here, annotated with speaker identifiers to capture natural conversational flow.
News Articles – Eng, Hin, Tam, Tel	News articles (Mujadia and Sharma, 2025) from diverse domains is used to generate informal dialogues using LLM prompting: *"Generate 10 informal dialogues between two friends on this topic. Keep them short, natural, conversational and direct that concludes meaningfully."*
Short Stories – Tam, Tel	Short stories from online sources such as `tamil_stories`[12] for Tamil and the `Chandamama Kathalu Dataset`[13] for Telugu are used. Each story is prompted as: *"You are given a language story text. Generate one informal dialogue between two language friends discussing the story in detail in a natural, code-mixed (with English) tone in language. The dialogue should be conversational and informal like spoken language. Reach an opinion about the story."*

Table 5: Overview of dialogue data sources used for constructing BaatCheet Corpora across English, Hindi, Tamil, and Telugu.

Datasource	Total Dialogues	Edited Dialogues	Dialogue Edit Rate (%)↓	Total Utts	Edited Utts	Utt Edit Rate(%)↓	TER↓ (%)	CER↓ (%)
News-Eng†	50	4	8.00	516	4	0.78	0.31	0.25
Hinglish†	49	49	100.00	967	638	65.98	21.99	19.60
News-Hin†	49	34	69.39	492	139	28.25	3.96	2.77
News-Tam	48	27	56.25	344	61	17.73	4.81	2.94
News-Tel	48	38	79.17	361	167	46.26	17.87	11.91
Stories-Tam	50	48	96.00	777	330	42.47	9.68	6.21
Stories-Tel	50	47	94.00	1,030	332	32.23	6.88	4.43
TOTAL	**344**	**247**	**71.80**	**4,487**	**1,671**	**37.24**	**10.47**	**7.63**

Table 6: Edit metrics of Human Validation across all data sources. †Indicates BaatCheet_Gold source dialogues; remaining splits are sampled from trainset splits. We report Dialogue Edit Rate, Utterance(Utt) Edit Rate, Token Edit Rate (TER), and Character Edit Rate (CER). Lower is better.

3. contextual adaptation of pronouns and honorifics based on speaker relationships; and
4. preservation of natural turn-taking flow and discourse coherence.

Completed translations underwent an expert review by senior linguists (distinct from the transla-

tors) in the first pass; expert reviewers flagged translations violating any guideline; flagged segments were returned to the translator for revision for targeted revision. This iterative process ensures that *BaatCheet_Gold* constitutes a high-quality gold standard for informal dialogue MT evaluation.

D Baseline Candidate System Evaluation: Disaggregated Results

Table 8 presents the complete language-pair disaggregated evaluation of all nine candidate systems on *BaatCheet_Human_Train*, comparing reference-based COMET and reference-free COMET-QE scores across all five language directions (Eng → Hin, Eng → Tam, Eng → Tel, Hin → Tam, and Hin → Tel).

LLM Translation Prompt Template. For zero-shot baseline evaluation of open-source LLMs (*Llama-3B*, *Llama-8B*, *Qwen3-4B*, *Gemma3-4B*, *Sarvam-Translate*) and *GPT-4o-mini*, we use a simple, direct instruction prompt without in-context examples:

Baseline Evaluation Prompt Template

Translate the following from {source_language} to {target_language}: {source_text}

This simple template ensures that baseline evaluations measure pure zero-shot translation capabilities without prompt-engineering artifacts.

High-Resource Alignment (Eng → Hin). Eng → Hin exhibits the highest scores across all open-source LLMs under both reference-based COMET (0.699–0.804) and reference-free COMET-QE (0.751–0.846). This superior baseline stems from the abundance of English-Hindi parallel training data in open-source LLM pre-training mixtures, enabling stronger zero-shot cross-lingual transfer even for informal dialogue registers.

Dravidian Language Challenge (Eng/Hin → Tam/Tel). For Dravidian target languages (Tamil and Telugu), open-source LLMs experience a sharp performance drop, particularly for Hindi-source directions (Hin → Tam and Hin → Tel). For instance, *Llama-3.2-3B* drops from 0.728 on Eng → Hin to 0.505 on Hin → Tam, while *Qwen3-4B* drops from 0.699 to 0.508. This drop is attributed to morphological complexity (agglutination) and lower representation of Dravidian script in general open-source LLM pre-training.

Robustness of Dedicated MT Systems and GPT-4o-mini. In contrast to open-source LLMs, dedicated MT engines (*Google Translate*, *IndicTrans2*, *BhashaVerse*) and *GPT-4o-mini* demonstrate remarkable stability across all language pairs, maintaining high COMET scores (> 0.80) even for Dravidian targets. *Google Translate* (GMT) achieves peak overall average performance under both reference-based COMET (0.848) and COMET-QE (0.857), followed closely by *GPT-4o-mini* (0.837 and 0.853). This cross-lingual stability justifies using these top four engines as candidate translation generators in our synthetic curation pipeline.

D.1 Candidate MT Evaluation: Disaggregated Results

Table 9 presents the complete language-pair and source-disaggregated evaluation of the four candidate MT engines (*Google Translate*, *GPT-4o-mini*, *IndicTrans2*, and *BhashaVerse*) on *Human_Train_X2IL* and *Human_Train_IL2X* candidate sets.

Engine Performance Trends. *Google Translate* (GMT) and *GPT-4o-mini* consistently emerge as the top-two performing engines across both translation axes. On X2IL directions (X → IL), GMT achieves peak average scores on English-source pairs (Eng → Tam COMET-QE: 0.8704), while *GPT-4o-mini* leads on Hindi-source Indic pairs (Hin → Tel COMET: 0.8659). On IL2X reverse directions (IL → X), *GPT-4o-mini* displays exceptional translation strength into English targets (Tel → Eng COMET: 0.8711).

Source-Level Insights. Conversational corpora (*DailyTalk*, *MuTual*) and news dialogues (*Newsgen*) maintain high translation quality across engines (> 0.80 COMET). Conversely, code-mixed dialogues (*Hinglish-Chat*) present greater cross-lingual difficulty when translating into Dravidian targets (Tam → Hin IndicTrans2 COMET: 0.6911), reflecting the morphological and syntactic challenges of informal code-mixed translation. These disaggregated results validate our selection of GMT and *GPT-4o-mini* as the primary translation generators for synthetic corpus curation.

D.2 Candidate MT Selection Counts Across Filtering Strategies

Table 10 presents the distribution of candidate MT system selections on *BaatCheet_Human_Train* under both filtering strategies: (a) *Utt-Avg* (dialogue-level engine selection across 510 dialogues) and (b) *Utt-High* (turn-by-turn optimal candidate selection across 8,427 utterances).

Engine Selection Trends. Two primary patterns emerge from the selection distribution:

- **Dominance of GMT and GPT-4o-mini:** Across both filtering strategies and directions, *Google Translate* (GMT) and *GPT-4o-mini* collectively account for over 80% of all selected dialogues and utterances, confirming their status as the highest-quality translation engines in our pool.
- **Directional Asymmetry:** On X2IL directions (X $\rightarrow$ IL), GMT is the single most selected engine (accounting for 320/510 dialogues in *Utt-Avg* and 3,506/8,427 turns in *Utt-High* under COMET-QE). Conversely, on IL2X reverse directions (IL $\rightarrow$ X), *GPT-4o-mini* becomes the dominant engine (up to 431/510 dialogues in *Utt-Avg* and 4,725/8,427 turns in *Utt-High*), reflecting GPT-4o-mini's strength in generating fluent English and Hindi target outputs.

D.3 Validation of Filtering Strategies: DOC-COMET and COMTAIL Analysis

Table 11 reports the detailed validation of candidate filtering strategies (*Utt-Avg* vs *Utt-High*) evaluated on *BaatCheet_Human_Train* candidate outputs using context-aware DOC-COMET and Indic-specialized COMTAIL metrics across all language directions.

When candidate selection is guided by reference-free COMET-QE, *Utt-High* consistently outperforms *Utt-Avg* across both evaluation metrics and all language directions:

- **DOC-COMET Alignment:** *Utt-High* achieves higher overall average dialogue scores across both X2IL (0.8854 vs 0.8803) and IL2X (0.8435 vs 0.8392), confirming that selecting the best-scoring candidate translation per turn improves multi-turn discourse consistency.

Conclusion. These results confirm that turn-by-turn COMET-QE selection (*Utt-High*) prevents individual low-quality utterances from degrading dialogue quality, ensuring that no utterance falls below the quality ceiling set by the evaluator. This empirical validation justifies adopting **Utt-High** as the primary filtering strategy for scaling up the construction of *BaatCheet_Synthetic*.

D.4 Human Binary Preference Evaluation of Curation Strategies

To evaluate human perception of dialogue curation strategies, native expert linguists conducted a binary preference study comparing *Utt-Avg* (dialogue-level engine selection) against *Utt-High* (turn-by-turn optimal selection) on 510 sampled dialogues (255 per translation direction). Annotators evaluated pairwise dialogue outputs on overall naturalness, register consistency, and translation accuracy. Table 12 reports the preference counts across both translation axes.

Key Observations.

- **Perceived Naturalness Preference:** Human annotators preferred *Utt-Avg* in 54% of evaluated dialogues (276/510) compared to 37% for *Utt-High* (196/510), primarily due to the single-engine stylistic consistency preserved across dialogue turns under *Utt-Avg*.
- **Translation Axis Consistency:** The preference distribution remains highly consistent across both X2IL (55.9% *Utt-Avg* vs 40.9% *Utt-High*) and IL2X (58.2% *Utt-Avg* vs 36.8% *Utt-High*) directions.
- **Strategic Selection:** As analyzed in Section 3.3, while human evaluators favor *Utt-Avg* for surface naturalness, *Utt-High* guarantees strict turn-level translation accuracy, which is the primary criterion for SFT dataset construction.

Figure 6 provides a concrete illustration of *Utt-High* selection on an English$\rightarrow$Telugu dialogue, showing how different MT engines are selected at different turns based on per-utterance COMET-QE scores.

Translation Direction	Utt-Avg	Utt-High
X2IL Direction (X $\rightarrow$ IL)	123 (55.9%)	90 (40.9%)
IL2X Direction (IL $\rightarrow$ X)	128 (58.2%)	81 (36.8%)
Total	**251 (57.0%)**	**171 (38.9%)**

Table 12: Human binary preference study results comparing *Utt-Avg* and *Utt-High* filtering strategies across 510 sampled dialogues.

Source from X->IL (Eng-> Tel)	Highest Comet scored Target outputs from 4 MTs	COMET-QE	Picked MT System
Amit :: Bro, did you check out the new draft guidelines for retired government employees?	అమిత్ :: బ్రో, రిటైర్డ్ ప్రభుత్వ ఉద్యోగుల కోసం కొత్త డ్రాఫ్ట్ మార్గదర్శకాలను చూశావా?	0.84593	GMT
Sanjay :: Yeah, I saw that. They're planning to regulate the salaries for those re-hired on contract, right?	సంజయ్ :: అవును, చూశాను. కాంట్రాక్ట్‌పై తిరిగి నియమించబడిన వారి జీతాలను నియంత్రించడానికి వారు ప్రణాళికలు రూపొందిస్తున్నారు కదా?	0.876668	GPT
Amit :: Correct. They want to keep the nomination-based appointments minimal.	అమిత్ :: నిజమే. నామినేషన్ ఆధారిత నియామకాలను తక్కువగా ఉంచాలని వారు కోరుకుంటున్నారు.	0.881104	GMT
Sanjay :: That's good. It should be more about merit than just past connections.	సంజయ్ :: అది బాగుంది. ఇది గత సంబంధాల గురించి కాకుండా యోగ్యత గురించి ఎక్కువగా ఉండాలి.	0.867858	GMT
Amit :: Absolutely. They're also fixing the salary by deducting the basic pension from the last salary.	అమిత్ :: ఖచ్చితంగా. వారు చివరి జీతం నుండి ప్రాథమిక పెన్షన్ ను తీసివేయడం ద్వారా జీతాన్ని కూడా నిర్ణయిస్తున్నారు.	0.868421	BV
Sanjay :: So, no increments during the contract?	సంజయ్ :: అయితే, కాంట్రాక్టు సమయంలో ఇంక్రిమెంట్లు లేవా?	0.874282	GMT
Amit :: Nope, the salary remains unchanged for the contract term.	అమిత్ :: లేదు, కాంట్రాక్ట్ కాలానికి జీతం మారదు.	0.881761	GMT
Sanjay :: And what about the contract duration?	సంజయ్ :: మరి కాంట్రాక్ట్ వ్యవధి సంగతి ఏమిటి?	0.878356	GMT
Amit :: Initially up to one year, extendable up to two more years, but not beyond five years post-retirement age.	అమిత్ :: ప్రారంభంలో ఒక సంవత్సరం వరకు, మరో రెండు సంవత్సరాల వరకు పొడిగించవచ్చు, కానీ రిటైర్మెంట్ వయస్సు తర్వాత ఐదు సంవత్సరాలకు మించి కాదు.	0.866749	BV
Sanjay :: Sounds fair. It'll prevent unnecessary extensions.	సంజయ్ :: ఇది న్యాయంగా అనిపిస్తుంది. ఇది అనవసరమైన పొడిగింపులను నివారిస్తుంది.	0.874613	GMT
Amit :: Exactly. It's all about maintaining efficiency and fairness.	అమిత్ః సరిగ్గా. ఇదంతా సమర్థత మరియు న్యాయాన్ని కాపాడుకోవడం గురించి.	0.882973	INT2
Sanjay :: Hope it works out well.	సంజయ్ః ఇది బాగా పనిచేస్తుందని ఆశిస్తున్నాను.	0.879691	INT2
Amit :: Let's see how it unfolds.	అమిత్ :: అది ఎలా జరుగుతుందో చూద్దాం.	0.882759	GMT

Figure 6: An example English→Telugu dialogue from *BaatCheet_Human_Train* illustrating the *Utt-High* COMET-QE selection strategy. Each utterance's Telugu translation is independently selected from four candidate MT engines (GMT = Google Translate, GPT = GPT-4o-mini, BV = BhashaVerse, INT2 = IndicTrans2), with the corresponding COMET-QE score shown per turn. This shows that GMT dominance is higher, and the same is reflected in the utt-avg highest score; GMT is 0.869 in the same dialogue.

E The Fine-Tuning Open-Source LLMs

E.1 Finetuning Parameters

The table 13 shows the hyperparameters used to fine-tune the selected models. BaatCheet_Human_Train configuration we have used 3 epochs to fine-tune the model, and for BaatCheet_[Syn_X2IL+HT] configuration we have fine-tuned the model on BaatCheet_Syn_X2IL first for 1 epoch and then on BaatCheet_Human_Train for 3 epochs.

Hyperparameter	Value
Number of epochs	1
Batch size	2
Gradient accumulation step	4
Learning rate	1e-4
Maximum sequence length	4096
LoRA rank	32
LoRA alpha	64
LoRA dropout	0.05
init_lora_weights	gaussian
Target Modules	q,k,v,o,gate, up,down proj.
Modules to save	lm_head

Table 13: Training Hyperparameters.

E.2 Prompt Used for Finetuning the LLMS

The prompt explicitly instructs the model to produce informal and conversational translations, ensuring that the generated outputs better reflect natural spoken dialogue.

Training Prompt Template

You are an expert in an informal dialogue machine translation assistant. Translate the following dialogue from source_language to target_language into a natural, informal, and colloquial conversational style in target_language.
Source: source_dialogue
Target: target_dialogue

E.3 Detailed Automatic Evaluation Results

The fine-tuned model's performance results for each dataset configuration and detailed language-pairwise scores are provided in Table 14 for COMET, Table 15 for DOC-COMET, and Table 16 for COMTAIL.

F SQM Guided DA Evaluation

We conduct a fine-grained evaluation on these top two configurations of Sarvam-Translate using a Scalar Quality Metric (SQM)-guided Direct Assessment (DA) protocol on a continuous 0 to 100 scale. From each setup, we sample test dialogues across all language directions: 30 dialogues for each English-source pair (Eng $\rightarrow$ Hin/Tam/Tel) and 20 dialogues for the Hindi-source pair (Hin $\rightarrow$ Tel). In total, 1,100 dialogues (900 English-source and 200 Hindi-source) are independently evaluated by expert native-speaking linguists and two SOTA LLM judges:

F.1 LLM Prompting and Rubrics

The LLM judges were provided with a 7-point rating rubric injected into the system prompt to calibrate the 0–100 scale. We utilized few-shot examples to align the scoring distribution. The full prompt used for both dialogue level and utterance-level evaluation is detailed in Figure 7 and Figure 8.

F.2 Human Direct Assessment

Annotator Selection and Screening To ensure high-fidelity annotations, we implemented a strict crowd-annotator screening protocol. Prospective bilingual candidates completed a 12-item source comprehension test with multiple-choice questions evaluating grammar, colloquial register, and translation accuracy. Only candidates scoring greater than or equal to 9/12 were selected for the evaluation task. Human evaluation was conducted on the *posteditme* platform[14]. Annotations were performed over three iterations by professional linguists. Annotation tool samples are shown in Figures 9 and 10

Score Aggregation To eliminate evaluator-specific rating leniency and scale biases, raw scores are standard-normalized (Z-scored) per specific evaluator-language pair group, clipped to $[-3.0, 3.0]$ to suppress extreme outlier variance, and linearly rescaled to an intuitive positive 0–100 range.

G Inter-Judge Correlation Analysis: Per-Model Breakdown

Metrics. We report **Pearson** r (linear magnitude agreement between continuous SQM scores) and **Kendall's** τ (rank-order concordance across all pairwise dialogue comparisons). Together they provide complementary views: r captures score-level agreement while τ captures ordinal consistency under non-normal distributions.

Utterance-Level. The strongest inter-judge agreement is observed at the utterance level (AVG: GPT–Human r=0.702, Gemini–Human r=0.697), confirming that sentence-level translation accuracy provides a more objective and consistently assessable criterion than holistic dialogue-level factors.

Naturalness and Cohesion & Coherence. Both factors achieve moderate-to-good correlations (Naturalness AVG r: 0.55–0.74; Cohesion AVG r: 0.56–0.74 across judge pairs). Gemini–Human correlations consistently exceed GPT–Human across both dimensions, indicating that Gemini's evaluation tendencies better approximate human quality perception for informal Indic dialogue.

Informality. Informality yields the weakest agreement across all models and judge pairs (AVG r<0.40), with Kendall's τ values as low as 0.17. This reflects the inherent subjectivity of register assessment across code-mixed Dravidian languages, where dialectal variation, honorific calibration, and pragmatic code-switching resist standardization within a fixed rubric.

Overall Patterns. Three consistent trends emerge across all factors: (i) GPT–Gemini yields the highest correlations, reflecting shared LLM tendencies in rubric interpretation; (ii) Gemini–Human consistently exceeds GPT–Human, suggesting Gemini is a closer proxy for human judgement in this setting; and (iii) Informality is universally the most disagreed-upon dimension, confirming it as the hardest qualitative factor to standardize across evaluators in multilingual informal dialogue assessment.

[14] https://posteditme.revanai.in/

Dialogue Level DA+SQM-based Scoring Prompt Template for LLM-as-Judge fonttitle

System Prompt

You are an expert evaluator for dialogue translation and have the ability to evaluate the translated dialogues on the basis of naturalness, informality and cohesion and coherence. You will evaluate the provided Target Dialogue (translated from the Source Dialogue). You need to evaluate the entire Target Dialogue on three factors: Naturalness, Informality, and Cohesion & Coherence by understanding the context of the dialogue.

Few-shot Injection

```
You have been provided with some examples of human evaluations below... Do not copy the scores
and patterns from given example.
```

Evaluation Rubrics

Naturalness	Informality	Cohesion & Coherence
00-15: Ungrammatical/Unnatural	**00-15:** Highly Formal	**00-15:** Incoherent/Incohesive
16-30: Ungram. but natural-ish	**16-30:** Very Formal	**16-30:** Coherent but Incohesive
31-45: Minor errors, natural-ish	**31-45:** Moderately Formal	**31-45:** Somew. coherent/cohesive
46-60: Grammatical but unnatural	**46-60:** Neutral	**46-60:** Coherent/partially cohesive
61-75: Grammatical/somew. natural	**61-75:** Slightly Informal	**61-75:** Somew. coherent/cohesive
76-90: Grammatical/natural (minor flaws)	**76-90:** Very Informal	**76-90:** Strong Coherence/Cohesion
91-100: Grammatical and Natural	**91-100:** Extremely Casual	**91-100:** Fully Coherent/Cohesive

Constraint: *DA Score should be on a continuous scale of 0 to 100 representing the overall quality for that factor. Provide precise, granular scores. Do not round to the nearest 5 or 10.*

User Input

Source Dialogue	Target Dialogue (MT)
`{source_text}`	`{mt_text}`

Response Instruction: `Please evaluate the Target Dialogue. Don't provide any other explanations. Give output in the required format.`

Figure 7: The prompt interface used for LLM-as-a-judge evaluation for Dialogue Level Evaluation. Text in brackets denotes dynamic variables.

Direction	Data Source	#Dialogues	#Utterances	#Sentences (Src / Tgt)	Avg Dlg Len	Avg Utt Len (Src / Tgt)	Avg Sent Len (Src / Tgt)
(a) BaatCheet_Syn_X2IL (X → IL Direction: English / Hindi ⟶ Indic Language)							
Eng→Hin	DailyTalk	2,446	22,152	32,582 / 32,312	9.06	11.28 / 11.41	7.81 / 7.94
	MuTual	8,346	38,332	68,560 / 68,594	4.59	16.35 / 17.56	9.29 / 9.93
	News	826	7,834	12,690 / 12,663	9.48	11.86 / 13.39	7.34 / 8.30
	Total	**11,618**	**68,318**	**113,832 / 113,569**	**5.88**	**14.97 / 15.97**	**8.84 / 9.39**
Eng→Tam	DailyTalk	2,446	22,152	32,582 / 32,585	9.06	11.28 / 8.52	7.81 / 5.90
	MuTual	8,346	38,332	68,560 / 68,819	4.59	16.35 / 12.31	9.29 / 6.97
	News	826	7,834	12,690 / 12,676	9.48	11.86 / 8.98	7.34 / 5.56
	Total	**11,618**	**68,318**	**113,832 / 114,080**	**5.88**	**14.97 / 11.28**	**8.84 / 6.64**
Eng→Tel	DailyTalk	2,446	22,152	32,582 / 32,668	9.06	11.28 / 8.52	7.81 / 5.87
	MuTual	8,346	38,332	68,560 / 68,698	4.59	16.35 / 12.18	9.29 / 6.89
	News	826	7,834	12,690 / 12,706	9.48	11.86 / 8.95	7.34 / 5.53
	Total	**11,618**	**68,318**	**113,832 / 114,072**	**5.88**	**14.97 / 11.18**	**8.84 / 6.58**
Hin→Tam	HinglishGit	1,558	33,349	75,054 / 75,101	21.41	17.12 / 12.46	7.47 / 5.43
	News	2,089	14,813	20,749 / 20,823	7.09	12.78 / 8.71	9.27 / 6.29
	Total	**3,647**	**48,162**	**95,803 / 95,924**	**13.21**	**14.63 / 10.31**	**8.50 / 5.92**
Hin→Tel	HinglishGit	1,558	33,349	75,054 / 75,057	21.41	17.12 / 12.65	7.47 / 5.51
	News	2,089	14,813	20,749 / 20,792	7.09	12.78 / 8.74	9.27 / 6.32
	Total	**3,647**	**48,162**	**95,803 / 95,849**	**13.21**	**14.63 / 10.41**	**8.50 / 5.97**
(b) BaatCheet_Syn_IL2X (IL → X Direction: Indic Language [Src] ⟶ English / Hindi [Tgt])							
Hin→Eng	HinglishGit	671	14,332	32,177 / 32,121	21.36	17.16 / 15.49	7.50 / 6.78
	News	957	7,055	9,719 / 9,747	7.37	12.68 / 11.36	9.31 / 8.29
	Total	**1,628**	**21,387**	**41,896 / 41,868**	**13.14**	**14.52 / 13.06**	**8.56 / 7.67**
Tam→Eng	News	775	5,568	8,665 / 8,697	7.18	10.18 / 13.64	6.61 / 8.15
	Stories	591	9,025	17,904 / 17,888	15.27	12.09 / 16.50	6.14 / 8.10
	Total	**1,366**	**14,593**	**26,569 / 26,585**	**11.23**	**11.13 / 15.07**	**6.37 / 8.13**
Tel→Eng	News	335	2,705	3,650 / 3,720	8.07	9.59 / 11.86	7.11 / 8.59
	Stories	533	8,842	17,141 / 16,520	16.59	12.12 / 15.35	6.22 / 8.15
	Total	**868**	**11,547**	**20,791 / 20,240**	**13.30**	**11.14 / 14.00**	**6.56 / 8.32**
Tam→Hin	News	775	5,568	8,665 / 8,697	7.18	10.18 / 14.01	6.61 / 9.05
	Stories	591	9,025	17,904 / 17,888	15.27	12.09 / 17.75	6.14 / 9.01
	Total	**1,366**	**14,593**	**26,569 / 26,585**	**11.23**	**11.14 / 15.63**	**6.37 / 9.03**
Tel→Hin	News	335	2,705	3,650 / 3,569	8.07	9.59 / 12.38	7.11 / 9.33
	Stories	533	8,841	17,141 / 17,047	16.59	12.12 / 17.12	6.22 / 8.85
	Total	**868**	**11,546**	**20,791 / 20,616**	**13.30**	**11.14 / 15.29**	**6.56 / 9.03**
(c) BaatCheet_Human_Train							
Eng→Hin	DailyTalk	39	723	1,302 / 1,259	18.54	15.87 / 16.46	8.98 / 9.58
	MuTual	40	447	901 / 895	11.18	21.76 / 22.85	9.56 / 9.98
	News	39	476	888 / 748	12.21	14.38 / 18.07	7.67 / 11.68
	Total	**118**	**1,646**	**3,091 / 2,902**	**13.95**	**17.37 / 19.16**	**8.74 / 10.41**
Eng→Tam	DailyTalk	39	723	1,302 / 1,339	18.54	15.87 / 11.50	8.98 / 6.30
	MuTual	40	447	901 / 907	11.18	21.76 / 15.44	9.56 / 6.84
	News	39	476	888 / 850	12.21	14.38 / 11.27	7.67 / 6.33
	Total	**118**	**1,646**	**3,091 / 3,096**	**13.95**	**17.37 / 12.76**	**8.74 / 6.49**
Eng→Tel	DailyTalk	39	723	1,302 / 1,329	18.54	15.87 / 12.12	8.98 / 6.71
	MuTual	40	447	901 / 920	11.18	21.76 / 14.92	9.56 / 6.53
	News	39	476	888 / 802	12.21	14.38 / 11.17	7.67 / 6.78
	Total	**118**	**1,646**	**3,091 / 3,051**	**13.95**	**17.37 / 12.75**	**8.74 / 6.67**
Hin→Tam	HinglishGit	39	1,369	3,023 / 3,015	35.10	20.82 / 14.90	9.32 / 6.68
	News	39	379	568 / 570	9.72	15.07 / 9.82	9.73 / 6.33
	Total	**78**	**1,748**	**3,591 / 3,585**	**22.41**	**17.94 / 12.36**	**9.53 / 6.50**
Hin→Tel	HinglishGit	39	1,369	3,024 / 2,891	35.10	20.82 / 15.66	9.32 / 7.38
	News	39	379	568 / 552	9.72	15.07 / 10.98	9.73 / 7.46
	Total	**78**	**1,748**	**3,592 / 3,443**	**22.41**	**17.94 / 13.32**	**9.53 / 7.42**
(d) BaatCheet_Test							
Eng→Hin	DailyTalk	51	862	1,430 / 1,458	16.90	15.53 / 16.23	9.58 / 9.85
	MuTual	50	584	1,220 / 1,184	11.68	23.05 / 25.56	10.47 / 11.50
	News	51	526	997 / 973	10.31	14.73 / 17.46	7.77 / 9.42
	Total	**152**	**1,972**	**3,647 / 3,615**	**12.97**	**17.73 / 19.71**	**9.26 / 10.25**
Eng→Tam	DailyTalk	51	862	1,430 / 1,465	16.90	15.53 / 11.42	9.58 / 6.86
	MuTual	50	584	1,220 / 1,232	11.68	22.93 / 17.88	10.40 / 7.96
	News	51	526	997 / 987	10.31	14.73 / 11.58	7.77 / 6.16
	Total	**152**	**1,972**	**3,647 / 3,684**	**12.97**	**17.69 / 13.60**	**9.24 / 6.99**
Eng→Tel	DailyTalk	51	862	1,430 / 1,344	16.90	15.53 / 11.60	9.58 / 7.66
	MuTual	50	584	1,220 / 1,234	11.68	23.05 / 17.55	10.47 / 7.70
	News	51	526	997 / 968	10.31	14.73 / 11.93	7.77 / 6.49
	Total	**152**	**1,972**	**3,647 / 3,546**	**12.97**	**17.73 / 13.67**	**9.26 / 7.28**
Hin→Tam	HinglishGit	50	981	2,106 / 2,101	19.62	16.11 / 11.93	7.33 / 5.43
	News	51	513	830 / 846	10.06	15.28 / 11.83	9.51 / 7.21
	Total	**101**	**1,494**	**2,936 / 2,947**	**14.79**	**15.69 / 11.88**	**8.43 / 6.33**
Hin→Tel	HinglishGit	50	981	2,106 / 1,929	19.62	16.11 / 12.09	7.33 / 6.06
	News	51	513	830 / 797	10.06	15.28 / 11.05	9.51 / 7.22
	Total	**101**	**1,494**	**2,936 / 2,726**	**14.79**	**15.69 / 11.57**	**8.43 / 6.65**

Table 7: Comprehensive statistics for all four *BaatCheet* splits: (a) *BaatCheet_Syn_X2IL*, (b) *BaatCheet_Syn_IL2X*, (c) *BaatCheet_Human_Train*, and (d) *BaatCheet_Test*.

	Open-Source LLMs					Dedicated MT Systems			Proprietary LLM
Lang Pair	**Llama-3B**	**Llama-8B**	**Qwen3-4B**	**Gemma3-4B**	**Sarvam-T**	**GMT**	**IndicTrans2**	**BhashaVerse**	**GPT-4o-mini**
(a) Reference-Based COMET									
Eng→Hin	0.7282	0.7546	0.6989	0.7881	0.8038	0.7957	0.7524	0.7543	0.7857
Eng→Tam	0.5745	0.7141	0.4945	0.8109	0.7725	**0.8623**	0.8079	0.8100	0.8393
Eng→Tel	0.5964	0.6257	0.5023	0.7644	0.8056	**0.8391**	0.8296	0.8210	0.8255
Hin→Tam	0.5054	0.5675	0.5079	0.7998	0.6541	**0.8791**	0.8010	0.8554	0.8686
Hin→Tel	0.5316	0.5939	0.5164	0.7983	0.6748	0.8654	0.8273	0.8327	**0.8659**
TOTAL AVG	0.5872	0.6512	0.5440	0.7923	0.7422	**0.8483**	**0.8036**	**0.8147**	**0.8370**
(b) Reference-Free COMET-QE									
Eng→Hin	0.7745	0.8137	0.7507	0.8383	0.8463	**0.8613**	0.8430	0.8427	0.8498
Eng→Tam	0.5446	0.7118	0.4892	0.8324	0.7741	**0.8704**	0.8462	0.8514	0.8597
Eng→Tel	0.5589	0.6081	0.4914	0.8087	0.8184	**0.8589**	0.8517	0.8462	0.8519
Hin→Tam	0.4637	0.5417	0.4485	0.7771	0.8160	0.8475	0.7710	0.8143	**0.8532**
Hin→Tel	0.5130	0.5687	0.4893	0.8052	0.8284	0.8477	0.8041	0.8147	**0.8500**
TOTAL AVG	0.5710	0.6488	0.5338	0.8124	0.8166	**0.8571**	**0.8232**	**0.8339**	**0.8529**

Table 8: Language-pair disaggregated baseline evaluation across nine candidate systems on *BaatCheet_Human_Train*, evaluated using (a) reference-based COMET and (b) reference-free COMET-QE. Bold text indicates the highest score per language pair.

Lang Pair	Data Source	Reference-Based COMET				Reference-Free COMET-QE			
		GMT	GPT	IndicTrans2	BhashaVerse	GMT	GPT	IndicTrans2	BhashaVerse
(a) Human_Train_X2IL (X → IL Direction)									
Eng→Hin	DailyTalk	**0.7628**	0.7506	0.7350	0.7257	**0.8587**	0.8474	0.8417	0.8297
	MuTual	0.8059	**0.8064**	0.7526	0.7724	**0.8597**	0.8492	0.8329	0.8440
	News-gen	**0.8185**	0.8000	0.7696	0.7649	**0.8655**	0.8527	0.8542	0.8544
	AVG	**0.7957**	0.7857	0.7524	0.7543	**0.8613**	0.8498	0.8430	0.8427
Eng→Tam	DailyTalk	**0.8295**	0.7900	0.8055	0.8010	**0.8693**	0.8502	0.8603	0.8563
	MuTual	**0.8607**	0.8535	0.7890	0.8101	**0.8690**	0.8605	0.8258	0.8473
	News-gen	**0.8969**	0.8743	0.8294	0.8189	**0.8729**	0.8684	0.8525	0.8505
	AVG	**0.8623**	0.8393	0.8079	0.8100	**0.8704**	0.8597	0.8462	0.8514
Eng→Tel	DailyTalk	0.8390	0.7951	**0.8409**	0.8293	**0.8607**	0.8490	0.8544	0.8487
	MuTual	0.8267	**0.8342**	0.8160	0.8065	**0.8549**	0.8489	0.8457	0.8405
	News-gen	**0.8517**	0.8474	0.8319	0.8272	**0.8610**	0.8578	0.8550	0.8493
	AVG	**0.8391**	0.8255	0.8296	0.8210	**0.8589**	0.8519	0.8517	0.8462
Hin→Tam	Hinglish-Chat	**0.8786**	0.8708	0.7844	0.8576	0.8335	**0.8400**	0.7507	0.7998
	News-gen	**0.8796**	0.8665	0.8175	0.8532	0.8614	**0.8664**	0.7914	0.8288
	AVG	**0.8791**	0.8686	0.8010	0.8554	0.8475	**0.8532**	0.7710	0.8143
Hin→Tel	Hinglish-Chat	0.8375	**0.8425**	0.7921	0.8042	0.8339	**0.8387**	0.7789	0.7945
	News-gen	**0.8933**	0.8893	0.8626	0.8612	**0.8615**	0.8614	0.8294	0.8350
	AVG	0.8654	**0.8659**	0.8273	0.8327	0.8477	**0.8500**	0.8041	0.8147
(b) Human_Train_IL2X (IL → X Direction)									
Hin→Eng	DailyTalk	**0.8677**	0.8642	0.8415	0.8136	**0.8537**	0.8513	0.8492	0.8337
	MuTual	0.8595	**0.8735**	0.8510	0.8267	**0.8535**	0.8508	0.8501	0.8354
	News-gen	0.8666	**0.8832**	0.8506	0.8268	**0.8447**	0.8425	0.8325	0.8246
	AVG	0.8646	**0.8736**	0.8477	0.8223	**0.8506**	0.8482	0.8440	0.8312
Tam→Eng	DailyTalk	0.8331	**0.8496**	0.7968	0.7585	0.8286	**0.8335**	0.7912	0.7656
	MuTual	0.8576	**0.8751**	0.8210	0.7899	0.8357	**0.8379**	0.8029	0.7887
	News-gen	0.8517	**0.8766**	0.7887	0.7923	0.8227	**0.8258**	0.7455	0.7797
	AVG	0.8475	**0.8671**	0.8022	0.7802	0.8290	**0.8324**	0.7799	0.7780
Tel→Eng	DailyTalk	0.8571	**0.8707**	0.8341	0.7853	**0.8268**	0.8266	0.8069	0.7492
	MuTual	0.8551	**0.8652**	0.8366	0.7251	**0.8358**	0.8343	0.8154	0.7124
	News-gen	0.8494	**0.8773**	0.8232	0.7872	0.8140	**0.8143**	0.7812	0.7395
	AVG	0.8538	**0.8711**	0.8313	0.7659	**0.8255**	0.8251	0.8012	0.7337
Tam→Hin	Hinglish-Chat	0.8089	**0.8189**	0.6911	0.7553	0.7870	**0.8026**	0.7329	0.7570
	News-gen	0.8170	**0.8351**	0.7367	0.7798	0.7478	**0.7632**	0.6881	0.7047
	AVG	0.8130	**0.8270**	0.7139	0.7676	0.7674	**0.7829**	0.7105	0.7308
Tel→Hin	Hinglish-Chat	0.7837	**0.8096**	0.7091	0.6615	0.8212	**0.8372**	0.7688	0.7044
	News-gen	0.8423	**0.8692**	0.8070	0.7790	0.8363	**0.8462**	0.8153	0.7605
	AVG	0.8130	**0.8394**	0.7581	0.7202	0.8288	**0.8417**	0.7921	0.7324

Table 9: Candidate MT evaluation scores on *BaatCheet_Human_Train* candidate sets disaggregated by language pair and data source for both (a) X → IL and (b) IL → X directions (X are English and Hindi, IL are Hindi, Tamil and Telugu), evaluated using Reference-Based COMET and Reference-Free COMET-QE. Bold values indicate the highest score per row for each evaluation metric block.

MT Engine	(a) Utt-Avg Strategy (Dialogue Counts, $N = 510$)				(b) Utt-High Strategy (Utterance Turn Counts, $N = 8,427$)			
	X2IL Direction (X → IL)		IL2X Direction (IL → X)		X2IL Direction (X → IL)		IL2X Direction (IL → X)	
	COMET-QE	COMET	COMET-QE	COMET	COMET-QE	COMET	COMET-QE	COMET
GMT	**320**	**320**	199	79	**3,506**	**3,449**	3,210	2,811
GPT-4o-mini	154	127	**294**	**431**	2,690	2,541	**3,720**	**4,725**
IndicTrans2	32	37	16	0	1,154	1,186	1,017	573
BhashaVerse	4	26	1	0	1,077	1,251	480	318
Total	**510**	**510**	**510**	**510**	**8,427**	**8,427**	**8,427**	**8,427**

Table 10: Candidate MT system selection counts on *BaatCheet_Human_Train* candidate sets under both (a) *Utt-Avg* (dialogue-level engine selection across 510 dialogues) and (b) *Utt-High* (utterance-level turn selection across 8,427 turns) filtering strategies, evaluated using reference-free COMET-QE and reference-based COMET. Bold text highlights the most selected MT engine per column.

	(a) DOC-COMET Metric Validation				(b) COMTAIL Metric Validation			
	Selection via COMET		Selection via COMET-QE		Selection via COMET		Selection via COMET-QE	
Lang Pair	**Utt-Avg**	**Utt-High**	**Utt-Avg**	**Utt-High**	**Utt-Avg**	**Utt-High**	**Utt-Avg**	**Utt-High**
(a) Human_Train_X2IL(X → IL)								
Eng→Hin	0.9049	0.9041	0.9057	**0.9082**	0.9177	0.9189	0.9213	**0.9263**
Eng→Tam	0.9012	0.9018	0.9033	**0.9067**	0.8198	0.8209	0.8270	**0.8466**
Eng→Tel	0.8932	0.8958	0.8958	**0.9004**	0.9277	0.9305	0.9320	**0.9371**
Hin→Tam	0.8272	0.8296	0.8412	**0.8489**	0.8925	0.8946	0.8999	**0.9151**
Hin→Tel	0.8538	0.8596	0.8554	**0.8628**	0.9004	0.9011	0.9035	**0.9140**
Overall AVG	0.8761	0.8782	0.8803	**0.8854**	0.8916	0.8932	0.8967	**0.9078**
(b) BaatCheet_Human_Train_IL2X (IL → X)								
Hin→Eng	0.8787	0.8789	0.8808	**0.8840**	0.5508	0.5486	0.5532	**0.5584**
Tam→Eng	0.8543	0.8538	0.8536	**0.8594**	0.5998	0.5996	0.6000	**0.6010**
Tel→Eng	0.8475	0.8469	0.8484	**0.8522**	0.5998	0.5987	0.6015	**0.6055**
Tam→Hin	0.7625	0.7626	0.7686	**0.7733**	0.6748	0.6749	0.6844	**0.7136**
Tel→Hin	0.8448	0.8433	0.8444	**0.8487**	0.7762	0.7809	0.7976	**0.8349**
Overall AVG	0.8376	0.8371	0.8392	**0.8435**	0.6403	0.6405	0.6474	**0.6627**

Table 11: Selection strategy validation on *BaatCheet_Human_Train* candidate outputs comparing *Utt-Avg* vs *Utt-High* filtering strategies under both COMET and COMET-QE candidate selection, evaluated using (a) Context-aware DOC-COMET and (b) Indic-specialized COMTAIL metrics across all language directions. Bold text highlights the winning strategy under COMET-QE selection.

Model	LP	Zero-Shot	Few-Shot	Syn_X2IL	Syn_IL2X	HT	Syn_X2IL+ HT	Syn_IL2X+ HT	Syn_Both +HT	Syn_Both +3*HT
Llama-3B	Eng-Hin	0.496	0.558	**0.751**	0.748	0.748	0.747	0.736	0.712	0.604
	Eng-Tam	0.539	0.572	**0.775**	0.699	0.672	0.733	0.670	0.769	0.727
	Eng-Tel	0.523	0.556	0.784	0.694	0.699	0.711	0.694	**0.786**	0.759
	Hin-Tam	0.567	0.642	**0.790**	0.753	0.723	0.764	0.730	0.778	0.786
	Hin-Tel	0.533	0.605	**0.801**	0.746	0.728	0.751	0.708	0.779	0.760
	AVGs	0.532	0.586	**0.780**	0.728	0.714	0.741	0.708	0.765	0.727
Llama-8B	Eng-Hin	0.633	0.637	0.792	0.770	0.754	0.745	0.752	0.794	**0.802**
	Eng-Tam	0.701	0.680	0.800	0.717	0.790	0.702	0.697	0.792	**0.841**
	Eng-Tel	0.723	0.728	0.812	0.706	0.778	0.700	0.696	0.796	**0.857**
	Hin-Tam	0.741	0.698	0.804	0.757	0.799	0.733	0.742	0.788	**0.857**
	Hin-Tel	0.759	0.698	0.804	0.760	0.815	0.719	0.718	0.795	**0.874**
	AVGs	0.711	0.688	0.802	0.742	0.787	0.720	0.721	0.793	**0.846**
Gemma	Eng-Hin	0.430	0.742	0.765	0.727	**0.800**	0.782	0.792	0.683	0.606
	Eng-Tam	0.480	0.602	0.795	0.630	**0.843**	0.827	0.833	0.694	0.555
	Eng-Tel	0.438	0.557	0.821	0.774	**0.839**	0.819	0.825	0.777	0.547
	Hin-Tam	0.587	0.688	0.808	0.645	**0.834**	0.829	0.833	0.736	0.698
	Hin-Tel	0.578	0.616	**0.850**	0.772	0.845	0.838	0.829	0.808	0.639
	AVGs	0.502	0.641	0.808	0.710	**0.832**	0.819	0.822	0.740	0.609
Sarvam	Eng-Hin	0.590	0.464	0.811	0.779	**0.825**	0.822	0.821	0.805	0.788
	Eng-Tam	0.679	0.527	0.862	0.819	0.866	**0.868**	0.851	0.830	0.854
	Eng-Tel	0.691	0.463	**0.867**	0.771	0.862	0.862	0.848	0.862	0.832
	Hin-Tam	0.537	0.517	**0.871**	0.818	0.747	0.865	0.865	0.863	0.857
	Hin-Tel	0.587	0.457	0.665	0.842	0.860	0.870	0.864	**0.876**	0.875
	AVGs	0.617	0.486	0.815	0.806	0.832	**0.857**	0.850	0.847	0.841
Qwen	Eng-Hin	0.717	0.638	**0.792**	0.750	0.765	0.770	0.727	0.713	0.790
	Eng-Tam	0.630	0.634	0.841	0.718	0.779	0.815	0.719	0.749	**0.849**
	Eng-Tel	0.653	0.638	**0.846**	0.716	0.784	0.794	0.722	0.826	0.838
	Hin-Tam	0.694	0.690	**0.852**	0.746	0.827	0.829	0.766	0.759	0.843
	Hin-Tel	0.705	0.702	**0.855**	0.757	0.824	0.823	0.751	0.801	0.824
	AVGs	0.680	0.660	**0.837**	0.737	0.796	0.806	0.737	0.770	0.829

Table 14: Utterance-level COMET scores on BaatCheet_Test. Shaded cells represent the best performance per language pair. `Syn_Both` denotes Syn_[X2IL+IL2X] and `HT` denotes Human_Train

Model	LP	Zero-Shot	Few-Shot	Syn_X2IL	Syn_IL2X	HT	Syn_X2IL+ HT	Syn_IL2X+ HT	Syn_Both +HT	Synt _Both +3*HT
Llama-3B	Eng-Hin	**0.675**	0.495	0.710	0.700	0.705	0.705	0.688	0.661	0.507
	Eng-Tam	0.556	0.552	0.682	0.577	0.587	**0.675**	0.584	0.688	0.659
	Eng-Tel	0.590	0.525	**0.713**	0.581	0.601	0.650	0.608	0.710	0.699
	Hin-Tam	0.499	0.531	0.555	0.602	0.588	0.590	0.593	0.617	**0.729**
	Hin-Tel	0.523	0.474	0.572	0.597	0.563	0.589	0.578	0.635	**0.700**
	AVGs	0.568	0.515	0.647	0.611	0.609	0.642	0.610	**0.662**	0.658
Llama-8B	Eng-Hin	0.520	0.567	0.747	0.734	0.714	0.702	0.709	0.753	**0.764**
	Eng-Tam	0.640	0.601	0.740	0.611	0.753	0.644	0.639	0.727	**0.785**
	Eng-Tel	0.679	0.654	0.752	0.623	0.739	0.644	0.634	0.734	**0.803**
	Hin-Tam	0.688	0.620	0.597	0.620	0.771	0.558	0.590	0.588	**0.802**
	Hin-Tel	0.715	0.593	0.614	0.621	0.753	0.586	0.591	0.597	**0.810**
	AVGs	0.648	0.607	0.690	0.642	0.746	0.626	0.632	0.680	**0.793**
Gemma	Eng-Hin	0.352	0.259	0.717	0.681	**0.756**	0.737	0.749	0.629	0.620
	Eng-Tam	0.376	0.315	0.737	0.477	**0.793**	0.776	0.781	0.595	0.566
	Eng-Tel	0.331	0.280	0.781	0.704	**0.792**	0.771	0.779	0.710	0.488
	Hin-Tam	0.520	0.524	0.741	0.558	0.775	0.778	**0.783**	0.663	0.626
	Hin-Tel	0.494	0.500	**0.780**	0.751	0.769	0.775	0.761	0.736	0.586
	AVGs	0.415	0.376	0.751	0.634	**0.777**	0.767	0.770	0.667	0.577
Sarvam	Eng-Hin	0.604	0.333	0.766	0.747	0.782	0.776	**0.782**	0.759	0.759
	Eng-Tam	0.475	0.440	0.809	0.746	**0.822**	0.819	0.816	0.772	0.796
	Eng-Tel	0.644	0.384	0.815	0.791	**0.817**	0.815	0.812	0.809	0.807
	Hin-Tam	0.356	0.305	**0.818**	0.743	0.802	0.812	0.817	0.808	0.808
	Hin-Tel	0.398	0.283	0.596	0.776	0.789	0.800	0.794	**0.810**	0.808
	AVGs	0.495	0.349	0.760	0.761	0.802	**0.804**	**0.804**	0.796	0.792
Qwen	Eng-Hin	0.683	0.617	**0.750**	0.710	0.721	0.729	0.698	0.732	0.745
	Eng-Tam	0.522	0.552	0.783	0.600	0.722	0.760	0.661	0.797	**0.797**
	Eng-Tel	0.559	0.552	**0.798**	0.617	0.718	0.734	0.677	0.789	0.789
	Hin-Tam	0.600	0.591	0.798	0.622	0.785	0.778	0.677	**0.799**	0.799
	Hin-Tel	0.640	0.598	**0.795**	0.620	0.759	0.744	0.645	0.793	0.793
	AVGs	0.601	0.582	0.785	0.634	0.741	0.749	0.672	0.784	**0.784**

Table 15: DOC-COMET on BaatCheet_Test. Shaded cells represent the best performance per language pair. `Syn_Both` denotes Syn_[X2IL+IL2X] and `HT` denotes Human_Train

Model	LP	Zero-Shot	Few-Shot	Syn_X2IL	Syn_IL2X	HT	Syn_X2IL+ HT	Syn_IL2X+ HT	Synt _Both +HT	Syn_Both +3*HT
Llama-3B	Eng-Hin	0.394	0.498	**0.755**	0.715	0.723	0.717	0.713	0.698	0.555
	Eng-Tam	0.417	0.434	**0.615**	0.524	0.513	0.558	0.506	0.610	0.553
	Eng-Tel	0.301	0.315	0.700	0.524	0.515	0.536	0.497	**0.701**	0.658
	Hin-Tam	0.390	0.430	**0.621**	0.549	0.529	0.567	0.535	0.609	0.604
	Hin-Tel	0.306	0.408	**0.716**	0.606	0.581	0.621	0.553	0.681	0.648
	AVGs	0.362	0.417	**0.681**	0.584	0.572	0.600	0.561	0.660	0.604
Llama-8B	Eng-Hin	0.480	0.601	**0.806**	0.736	0.735	0.724	0.728	0.804	0.805
	Eng-Tam	0.524	0.534	0.635	0.528	0.586	0.534	0.534	0.624	**0.651**
	Eng-Tel	0.535	0.588	0.745	0.534	0.617	0.507	0.494	0.718	**0.814**
	Hin-Tam	0.512	0.484	0.637	0.548	0.588	0.518	0.531	0.622	**0.694**
	Hin-Tel	0.564	0.539	0.713	0.614	0.715	0.573	0.564	0.705	**0.818**
	AVGs	0.523	0.549	0.707	0.592	0.648	0.571	0.570	0.694	**0.756**
Gemma	Eng-Hin	0.261	0.681	0.755	0.634	**0.789**	0.764	0.779	0.608	0.539
	Eng-Tam	0.217	0.423	0.608	0.406	**0.653**	0.634	0.635	0.513	0.411
	Eng-Tel	0.138	0.397	0.763	0.653	**0.764**	0.718	0.719	0.665	0.341
	Hin-Tam	0.307	0.465	0.603	0.463	**0.649**	0.637	0.635	0.509	0.492
	Hin-Tel	0.355	0.429	**0.780**	0.671	0.768	0.751	0.725	0.711	0.492
	AVGs	0.256	0.479	0.702	0.565	**0.724**	0.701	0.699	0.601	0.455
Sarvam	Eng-Hin	0.532	0.315	**0.826**	0.770	0.825	0.821	0.817	0.820	0.800
	Eng-Tam	0.514	0.389	**0.688**	0.607	0.677	0.671	0.654	0.665	0.676
	Eng-Tel	0.579	0.290	**0.839**	0.707	0.800	0.803	0.781	0.829	0.801
	Hin-Tam	0.364	0.360	**0.707**	0.615	0.581	0.688	0.685	0.696	0.694
	Hin-Tel	0.373	0.209	0.541	0.749	0.780	0.794	0.784	**0.821**	**0.821**
	AVGs	0.472	0.313	0.720	0.690	0.733	0.755	0.744	**0.766**	0.758
Qwen	Eng-Hin	0.698	0.615	**0.807**	0.731	0.752	0.759	0.708	0.720	0.795
	Eng-Tam	0.496	0.494	**0.666**	0.527	0.591	0.615	0.541	0.584	0.663
	Eng-Tel	0.510	0.467	**0.802**	0.542	0.646	0.667	0.550	0.766	0.787
	Hin-Tam	0.534	0.512	**0.687**	0.533	0.641	0.638	0.565	0.599	0.679
	Hin-Tel	0.586	0.566	**0.791**	0.619	0.730	0.727	0.614	0.714	0.762
	AVGs	0.565	0.531	**0.751**	0.590	0.672	0.681	0.596	0.676	0.737

Table 16: COMTAIL Scores on BaatCheet_Test. Shaded cells represent the best performance per language pair. `Syn_Both` denotes Syn_[X2IL+IL2X] and `HT` denotes Human_Train

Utterance-Level DA+SQM-based Scoring Prompt Template for LLM-as-Judge fonttitle

System Prompt

You are an expert evaluator for dialogue translation. You have the ability to evaluate the translated utterances. You will evaluate the provided Target utterances corresponding to Source utterances. You need to evaluate MULTIPLE utterances simultaneously. For each utterance, provide a single DA score (0-100) based on the rubric.

Few-shot Injection (Optional)

```
You have been provided with some examples of human evaluations below... Use this understanding
to calibrate your own scoring for the new utterances.
```

Utterance-Level Scoring Rubric (0-100)

Score Range	Evaluation Criteria
00–15	**Completely unusable:** Nonsensical/total mistranslation; broken grammar; literal word-for-word translation; original meaning cannot be recovered.
16–30	**Very poor translation:** Major errors in meaning/structure; important info added/omitted; extremely awkward word order; general topic only guessable.
31–45	**Meaning partially preserved:** Recognizable source meaning but several errors remain; grammatical/lexical issues affect clarity; intent may be distorted.
46–60	**Meaning correct but unnatural:** Core meaning preserved; grammatically acceptable but sounds literal/stiff; uses formal words in conversational context.
61–75	**Mostly accurate/somewhat natural:** Message correct; minor awkward phrasing or slightly stiff word order; tone and intent generally preserved.
76–90	**Accurate and natural:** Core message/tone clearly preserved; sounds natural and conversational; minor stylistic imperfections that do not affect fluency.
91–100	**Accurate and fully Natural:** Meaning, tone, and intent perfectly preserved; native-like phrasing and word order; correct honorifics and conversational markers.

Constraints: *DA Score should be on a continuous scale of 0 to 100. Provide precise, granular scores. Do not round to the nearest 5 or 10. Provide exactly the same number of scores as the number of utterances provided.*

User Input Prompt

```
Here are {num_utterances} utterances to evaluate:
{paired_utterances_text}
Please evaluate the Target utterances and return a list of scores.
```

Figure 8: The prompt used for utterance-level evaluation. The dynamic variable `{paired_utterances_text}` contains the list of source and target sentence pairs.

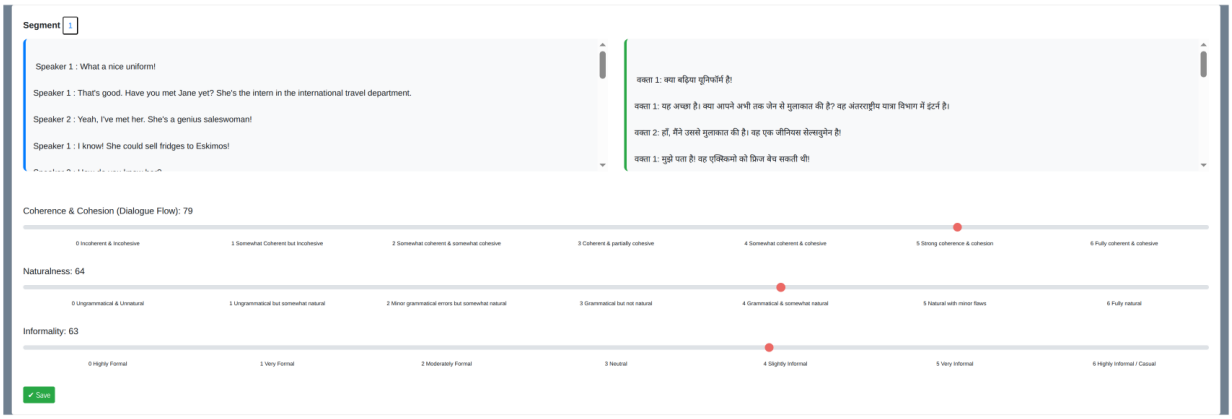

Figure 9: Dialogue-Level Annotation Platform

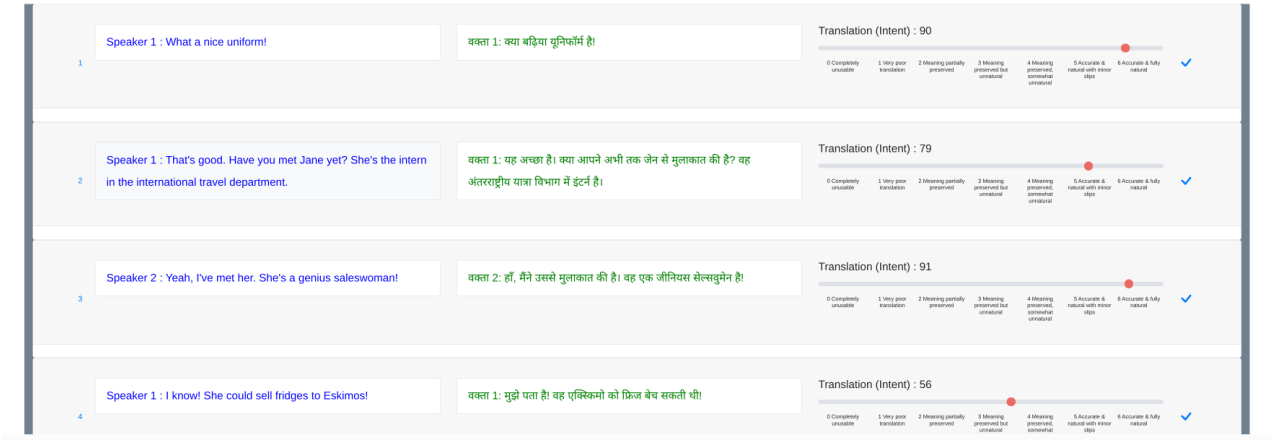

Figure 10: Utterence-Level Annotation Platform

Sample Dialogue from SARVAM_Human_Train Configuration			Utterence_Level_Scores			
SOURCE	REFERENCE	SARVAM_Human_Train	Human	GPT	Gemini	Averages
Speaker 1: Hey Lindaï¼ did you get that letter about the new options for food service next year?	Speaker 1: अरे लिंडा, तुम्हें वो लेटर मिला क्या , जो अगले साल के फूड सर्विस ऑप्शन के बारे में आया था?	स्पीकर 1: हे लिंडा, क्या आपको अगले साल फ़ूड सर्विस के नए ऑप्शन्स वाला लेटर मिला?	97	80	85	87.33
Speaker 2: Not yet, are there a lot of changes?	Speaker 2: नहीं अभी तक तो नहीं मिला, क्या बहुत सारे बदलाव हैं?	स्पीकर 2: अभी नहीं, बहुत सारे बदलाव आ रहे हैं क्या?	95	83	82	86.67
Speaker 1: There sure are, instead of paying one feet to cover all meals for the whole school year. We are now able to choose how many meals a week we want, and can contract just for that amount. We still have to pay the whole year at the beginning, but we can choose to buy 7, 10, 14 or 21 meals per week. They give you a card with a number of meals you get each week marked on it.	Speaker 1: हाँ, काफी हैं। पहले तो पूरे साल के मील्स का एक ही बार में पूरा पेमेंट देना पड़ता था, अब ऐसा नहीं है। अब हम खुद चुन सकते हैं कि हफ्ते में कितने मील चाहिए, और उसी हिसाब से कॉन्ट्रैक्ट कर सकते हैं। हाँ, साल की शुरुआत में पूरा पेमेंट फिर भी देना होगा, लेकिन अब हम 7, 10, 14 या 21 मील हर हफ्ते खरीद सकते हैं। एक कार्ड मिलेगा, जिसमें लिखा रहेगा कि हर हफ्ते तुम्हें कितने मील मिलेंगे।	स्पीकर 1: हाँ, आ रहे हैं। पूरे स्कूल वर्ष के लिए एक ही फीट पे पूरा खाना पेवाना पड़ता था। अब हम हफ़्ते में कितने चाहें उतने वारेन मिल सके, और सिर्फ उतने के लिए कॉन्ट्रैक्ट कर सकते हैं। पूरे साल के लिए शुरुआत में ही पे करना होगा, लेकिन हम हफ़्ते में 7, 10, 14 या 21 वारेन मिल सकते हैं। आपको एक नंबर कार्ड मिलेगा, जिस पर लिखा होगा कि आपको हर हफ़्ते कितने वारेन मिल सकते हैं।	35	48	35	39.33
Speaker 2: That's a big change Tom, and a complicated system.	Speaker 2: ओहो, यह तो काफी बड़ा बदलाव है टॉम, और थोड़ा झंझट वाला भी लग रहा है।	स्पीकर 2: यह तो बड़ा चेंज है, टॉम, और सिस्टम भी थोड़ा जटिल है।	82	80	78	80
Speaker 1: Yeah, but it will be much better for people who don't eat 3 meals a day 7 days a week in the cafeteria, because they don't have to pay for meals they don't eat.	Speaker 1: हाँ, लेकिन उन लोगों के लिए अच्छा है जो कैफेटेरिया में रोज़ दिन में 3 मील्स नहीं खाते। अब उन्हें उन मील्स के पैसे नहीं देने पड़ेंगे, जो वे खाते ही नहीं।	स्पीकर 1: हाँ, लेकिन जो लोग कैफ़ेटेरिया में दिन में 3 बार खाना नहीं खाते उनके लिए यह बहुत अच्छा रहेगा, क्योंकि उन्हें उन वारेन जो खाते हैं, उनके लिए पे नहीं करना पड़ेगा।	52	35	48	45
Speaker 2: So, what's the deal for those who do eat at school all the time?	Speaker 2: अच्छा, तो जो लोग हर रोज़ स्कूल में खाना खाते हैं, उनके लिए क्या डील है?	स्पीकर 2: तो, जो लोग स्कूल में ही खाना खाते हैं उनका क्या?	67	82	80	76.33
Speaker 1: It better for them, too, because the more meals you contract for, the cheaper each one is.	Speaker 1: उनके लिए भी ठीक ही है, जितने ज़्यादा मील्स का कॉन्ट्रैक्ट करोगे, हर मील उतना ही सस्ता पड़ेगा।	स्पीकर 1: यह उनके लिए भी बेहतर होगा, क्योंकि जितने ज़्यादा वारेन आप कॉन्ट्रैक्ट करेंगे उतना ही सस्ता पड़ेगा।	66	62	52	60
Speaker 2: I see, it still sounds rather complicated.	Speaker 2: समझ गई, पर फिर भी थोड़ा कॉम्प्लिकेटेड सा लग रहा है।	स्पीकर 2: मैं समझ गया, यह अभी भी काफी जटिल लग रहा है।	98	76	84	86
Speaker 1: True, it took me several hours to figure it out. I decided to go with the 10 meal plan.	Speaker 1: सही कह रही हो, मुझे भी समझने में कई घंटे लग गए। मैंने तो 10 मील वाला प्लान ले लिया।	स्पीकर 1: यह सच है, इसे समझने में मुझे कई घंटे लग गए। मैंने 10 वारेन वाला प्लान सोचा।	51	60	55	55.33
Speaker 2: Why is that?	Speaker 2: वो क्यों?	स्पीकर 2: ऐसा क्यों?	97	88	88	91
Speaker 1: Well, I never eat breakfast and I often go away on weekends, so the 10 meal plan gives me lunch and dinner every weekday at a fairly low price, and I won't be paying for meals I don't usually eat.	Speaker 1: देखो, मैं तो ब्रेकफास्ट कभी खाता ही नहीं, और वीकेंड्स पर अकसर बाहर चला जाता हूँ। तो 10 मील प्लान से मुझे हर वीक डे पर लंच और डिनर मिल जाता है, वो भी ठीक-ठाक रेट पर। और मैं उन मील्स के लिए पेमेंट नहीं कर रहा, जो मैं खाता ही नहीं।	स्पीकर 1: दरअसल, मैं नाश्ता नहीं करता और अक्सर वीकेंड पर बाहर चला जाता हूँ, इसलिए 10 वारेन वाला प्लान मुझे हर वीकडे में लंच और डिनर काफी कम कीमत पे करा देगा, और मुझे उन वारेन के लिए भी पे नहीं करना पड़ेगा, जो मैं आमतौर पर नहीं खाता हूँ।	49	62	58	56.33

A sample from the best performing model: Sarvam Fine-Tuned on BaatCheet_Human_Train. Figure shows translation from English-->Hindi with MT and reference. Also presents the utterence level and dialogue level DA.

Dialogue_Level_Scores											
Human			GPT			GEMINI			AVERAGES		
Nat	*Inf*	*Coh*	*Nat*	*Inf*	*Coh*	*Nat*	*Inf*	*Coh*	*Nat*	*Inf*	*Coh*
38	66	49	30	52	59	48	68	52	38.67	62	53.33

Figure 11: A sample dialogue in our error analysis suggests that there are still formal-style dialogues. While the baseline semantics are preserved, the translations are only somewhat natural. This finding supports our conclusion that synthetic data alone is insufficient for capturing deep conversational nuances, reinforcing the necessity for larger-scale, human-curated informal dialogue datasets.

7-Scale SQM Rubric	Eng–Hin	Eng–Tam	Eng–Tel	Hin–Tam	Hin–Tel
(a) Utterance-Level Accuracy					
00–15: Completely unusable	0	0	0	0	0
16–30: Very poor translation	0	0	0	0	0
31–45: Meaning partially preserved	0	0	0	0	0
46–60: Meaning correct but unnatural	0	1	0	1	1
61–75: Mostly accurate and somewhat natural	6	**25**	**23**	**13**	**11**
76–90: Accurate and natural (minor slips)	**24**	4	7	6	8
91–100: Accurate and fully Natural	0	0	0	0	0
(b) Dialogue-Level: Naturalness					
00–15: Ungrammatical and Unnatural	0	0	0	0	0
16–30: Ungrammatical but somewhat natural	0	0	0	0	0
31–45: Minor errors but somewhat natural	0	0	2	0	0
46–60: Grammatical but not natural	3	6	8	4	3
61–75: Grammatical and somewhat natural	**18**	**24**	**18**	**12**	**13**
76–90: Grammatical and natural (minor flaws)	9	0	2	4	4
91–100: Grammatical and Natural	0	0	0	0	0
(c) Dialogue-Level: Informality					
00–15: Highly Formal	0	1	0	0	0
16–30: Very Formal	0	8	0	3	0
31–45: Moderately Formal	1	**10**	1	**5**	0
46–60: Neutral	8	7	6	8	0
61–75: Slightly Informal	**20**	4	**22**	4	**14**
76–90: Very Informal	1	0	1	0	6
91–100: Extremely Informal / Casual	0	0	0	0	0
(d) Dialogue-Level: Cohesion & Coherence					
00–15: Incoherent and Incohesive	0	0	0	0	0
16–30: Somewhat Coherent but Incohesive	0	0	0	0	0
31–45: Somewhat coherent and cohesive	0	0	0	0	0
46–60: Coherent and partially cohesive	2	2	3	2	3
61–75: Somewhat coherent and cohesive	8	**17**	**18**	**12**	**11**
76–90: Strong Coherence and Cohesion	**20**	11	9	6	6
91–100: Fully Coherent and Cohesive	0	0	0	0	0
Total Dialogues Evaluated	**30**	**30**	**30**	**20**	**20**

Table 17: SQM score distributions across language directions for *Sarvam-Translate* fine-tuned on *BaatCheet_Human_Train*, disaggregated by (a) Utterance-level Accuracy and Dialogue-level factors: (b) Naturalness, (c) Informality, and (d) Cohesion & Coherence. Bold values indicate peak distribution bands.

Factor	Model	GPT vs Human		Gemini vs Human		GPT vs Gemini	
		r	τ	r	τ	r	τ
Naturalness	Gemma3-4B	0.544	0.361	0.574	0.395	0.687	0.504
	Llama-3.2-3B	0.540	0.382	0.606	0.427	0.758	0.548
	Llama-3.1-8B	0.594	0.422	0.665	0.486	0.788	0.605
	Qwen3-4B	0.543	0.321	0.639	0.397	0.734	0.514
	Sarvam-T	0.539	0.367	0.639	0.435	0.723	0.529
	AVG	**0.552**	**0.370**	**0.625**	**0.428**	**0.738**	**0.540**
Informality	Gemma3-4B	0.283	0.195	0.273	0.167	0.340	0.241
	Llama-3.2-3B	0.385	0.256	0.377	0.264	0.447	0.302
	Llama-3.1-8B	0.356	0.259	0.308	0.195	0.331	0.232
	Qwen3-4B	0.347	0.232	0.320	0.188	0.373	0.236
	Sarvam-T	0.344	0.217	0.383	0.231	0.466	0.321
	AVG	**0.343**	**0.232**	**0.332**	**0.209**	**0.391**	**0.266**
Cohesion & Coherence	Gemma3-4B	0.468	0.323	0.558	0.398	0.624	0.440
	Llama-3.2-3B	0.651	0.462	0.695	0.503	0.869	0.684
	Llama-3.1-8B	0.535	0.377	0.648	0.471	0.738	0.546
	Qwen3-4B	0.605	0.318	0.694	0.431	0.787	0.501
	Sarvam-T	0.547	0.342	0.638	0.408	0.699	0.490
	AVG	**0.561**	**0.364**	**0.647**	**0.442**	**0.743**	**0.532**
Utterance Level	Gemma3-4B	0.714	0.517	0.706	0.541	0.773	0.595
	Llama-3.2-3B	0.818	0.622	0.806	0.624	0.878	0.706
	Llama-3.1-8B	0.682	0.507	0.679	0.539	0.788	0.613
	Qwen3-4B	0.665	0.477	0.662	0.503	0.765	0.584
	Sarvam-T	0.632	0.367	0.630	0.420	0.704	0.485
	AVG	**0.702**	**0.498**	**0.697**	**0.525**	**0.781**	**0.597**

Table 18: Per-model inter-judge correlation (r = Pearson, τ = Kendall's τ) across all four evaluation factors between GPT-4o-mini, Gemini-Flash-Lite, and Human Expert Linguists. Bold rows indicate cross-model averages.