IndicTalk: A Large-Scale Persona-Based Multilingual Conversational Corpus for Indic Languages

Sahil Deepak Gawande
Lingo Research Group
IIT Gandhinagar
sahil.gawande@iitgn.ac.in

Mayank Singh
Lingo Research Group
IIT Gandhinagar
singh.mayank@iitgn.ac.in

Abstract

Large Language Models (LLMs) have transformed conversational AI, yet high-quality multilingual code-mixed dialogue resources remain scarce, particularly for Indic languages where speakers naturally alternate between English and their native language in both native-script and Romanized forms. We present INDICTALK, one of the largest multilingual Indic code-mixed conversational corpora, comprising over 13,28,604 event-grounded multi-turn conversations across 18 language varieties covering 9 Indic languages. The corpus is generated through a fully automated pipeline that combines real-world news grounding, persona-conditioned dialogue generation using multilingual LLMs, and automatic quality validation. Extensive linguistic, automatic, and human evaluations demonstrate that INDICTALK produces fluent, coherent, and naturally code-mixed conversations across both script variants. We will release INDICTALK to support the development and evaluation of multilingual conversational AI for underrepresented Indic languages. Dataset is available at Hugging Face: IndicTalk Dataset

1 Introduction

Large Language Models (LLMs) have transformed conversational AI, enabling increasingly natural and capable dialogue systems. Their success, however, depends critically on the availability of large-scale, high-quality conversational corpora. While English benefits from abundant dialogue datasets, comparable resources remain scarce for multilingual and code-mixed settings. This gap is particularly evident for Indic languages, where bilingual speakers routinely alternate between English and their native language within the same conversation, using both native scripts and Romanized writing in informal digital communication. Developing conversational systems that serve these multilingual users therefore requires dialogue corpora that accurately capture natural code-mixing patterns across multiple turns.

Although code-mixed NLP has received growing attention, existing Indic datasets primarily target discriminative tasks such as sentiment analysis, hate speech detection, part-of-speech tagging, and natural language inference. Conversational resources are comparatively rare and are typically limited to Hinglish, small-scale collections, or manually curated datasets. To the best of our knowledge, no publicly available resource provides large-scale, event-grounded, multi-turn conversations across multiple Indic code-mixed language varieties under a unified generation framework. This lack of conversational data has become a major bottleneck for training, instruction tuning, and evaluating multilingual conversational LLMs.

To address this gap, we introduce INDICTALK, the largest multilingual Indic code-mixed conversational corpora to date. The corpus comprises over 13,28,604 event-grounded multi-turn conversations spanning 18 language varieties across 9 Indic languages, covering both native-script and Romanized code-mixed variants paired with English. Each conversation consists of 6–8 dialogue turns generated through a fully automated pipeline that combines real-world news grounding, persona-conditioned dialogue generation using multilingual LLMs, and automatic quality validation, enabling scalable corpus construction without manual annotation.

We perform extensive automatic and human evaluation to assess both conversational quality and code-mixing characteristics. Our evaluation includes established code-mixing metrics, fluency analysis for native-script variants, LLM-as-a-Judge evaluation using Gemini-2.5-Flash (Comanici et al., 2025) and GPT-OSS-120B (OpenAI et al., 2025), and human evaluation by native or highly proficient speakers. Results consis-

tently demonstrate that INDICTALK contains fluent, coherent, and naturally code-mixed conversations across all supported language varieties.

Our primary contributions are as follows:

1. We introduce INDICTALK, the largest multilingual Indic code-mixed conversational corpora, comprising over 13,28,604 event-grounded multi-turn conversations across 18 language varieties covering 9 Indic languages.
2. We propose a fully automated pipeline for generating multilingual code-mixed conversations that combines real-world news grounding, persona-conditioned dialogue generation, multilingual LLMs, and automatic quality validation.
3. We provide a comprehensive evaluation of the corpus using linguistic code-mixing metrics, automatic fluency measures, LLM-as-a-Judge assessment, and human evaluation, demonstrating the quality and naturalness of the generated conversations.

2 Related Work

Existing work on code-mixed NLP has primarily focused on building datasets for downstream understanding tasks, while comparatively little attention has been devoted to large-scale conversational resources for generative dialogue modeling. We briefly review prior work in three directions: (i) code-mixed conversational datasets, (ii) large-scale Indic conversational resources, and (iii) how our work differs from existing efforts.

2.1 Code-Mixed Conversational Datasets

Most existing code-mixed resources for Indic languages are designed for discriminative NLP tasks such as sentiment analysis, hate speech detection, named entity recognition, and natural language inference (Chakravarthi et al., 2020; Srivastava and Singh, 2020). Although valuable, these datasets consist primarily of isolated sentences or social media posts and are not suitable for training conversational LLMs.

Only a handful of datasets focus on code-mixed dialogue. Banerjee et al. (Banerjee et al., 2018) extended the DSTC2 task-oriented dialogue corpus to four Indic language pairs through human translation. GupShup (Mehnaz et al., 2021) introduced a Hinglish conversational summarization dataset by translating the SAMSum corpus, while KCM (Singh et al., 2022) constructed knowledge-grounded code-mixed dialogues from the Holl-E dataset. More recently, (Singh et al., 2025) demonstrated synthetic Hinglish dialogue generation using LLMs. However, these datasets are restricted to Hinglish or a small number of languages, rely heavily on translation of existing conversations, or target specific domains such as task-oriented or knowledge-grounded dialogue.

2.2 Indic Conversational Resources

Large-scale conversational corpora have recently become available for monolingual Indic languages. IndicDialogue (Arnob et al., 2024) collects subtitle-based conversations across ten Indic languages, while IndicLLMSuite (Khan et al., 2024) provides large-scale multilingual pretraining and instruction-tuning resources. Although these datasets significantly advance Indic LLM development, they contain predominantly monolingual conversations and do not model natural code-mixing.

2.3 Our Contribution

Our work differs from existing efforts along three important dimensions. First, unlike prior resources that are largely limited to Hinglish or a few language pairs, we introduce one of the largest multilingual Indic code-mixed conversational corpora, spanning 18 language varieties across 9 Indic languages in both native-script and Romanized forms. Second, rather than translating or manually curating conversations, we propose a fully automated pipeline that generates event-grounded, persona-conditioned multi-turn dialogues using multilingual LLMs, enabling scalable corpus creation without manual annotation. Finally, we conduct comprehensive automatic and human evaluation, including linguistic code-mixing analysis, LLM-as-a-Judge assessment, and human evaluation, demonstrating the quality and naturalness of the generated conversations.

3 Methodology

Given a collection of source documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$, our objective is to automatically construct a multilingual code-mixed conversational corpus $\mathcal{C}$ that consists of multi-turn dialogues grounded in real-world events. Here, $|\mathcal{D}| = n$ denotes the total number of source documents. For each document d_i, the pipeline first derives a language-independent semantic representation

and subsequently generates conversations independently for each target language variety. Formally,

$$d_i \xrightarrow{\Phi_s} s_i \xrightarrow{\Phi_g^{(l)}} c_i^{(l)} \xrightarrow{\Phi_v} \hat{c}_i^{(l)},$$

where Φ_s denotes the summarization model that generates a language-independent factual summary s_i from the source document d_i, $l \in \mathcal{L}$ indexes a target language variety, $\Phi_g^{(l)}$ denotes the dialogue generation model for language variety l, $c_i^{(l)}$ is the generated conversation before quality filtering, Φ_v denotes the automatic validation module, and $\hat{c}_i^{(l)}$ is the validated conversation retained in the final corpus.

The final corpus is

$$\mathcal{C} = \left\{ \hat{c}_i^{(l)} \;\middle|\; d_i \in \mathcal{D},\ l \in \mathcal{L} \right\},$$

where $\mathcal{L}$ denotes the set of 18 Indic-English language varieties.

3.1 Source Document Processing

The source corpus consists of a collection of publicly available news articles and blog posts covering diverse domains, including *current affairs* and *politics*, *international affairs*, *technology*, *business*, *sports*, *health*, *entertainment*, *science*, *culture*, *education*, and *lifestyle*. For each document, the primary textual content is extracted after removing boilerplate elements such as navigation menus, advertisements, scripts, and duplicated template content. Documents with insufficient textual content or unsuccessful extraction are discarded to maintain corpus quality. Table 1 presents the composition of the source corpus. Additional implementation details on source document extraction, preprocessing, and data cleaning are provided in Appendix A.

Category	# Articles
Current Affairs & Politics	11,845
International Affairs	10,826
Lifestyle	6,465
Health	6,056
Technology	5,366
Entertainment	5,056
Business	4,690
Education	3,476
Science	3,362
Sports	2,442
Culture	1,562
Other	80,907

Table 1: Composition of the source corpus.

3.2 Language-Independent Semantic Representation

Generating conversations directly from raw documents often introduces unnecessary variability due to differences in document length, writing style, and formatting. Longer documents, in particular, tend to bias LLMs towards producing disproportionately longer or more detailed conversations. We therefore first transform each document into a language-independent semantic representation, which normalizes the input, reduces document-length bias, and enables all language variants to be generated from the same factual content while minimizing stylistic variation.

Given a document d_i, the summarization model produces a concise English summary

$$s_i = \Phi_s(d_i),$$

which serves as the common semantic representation for dialogue generation across all target languages. Using a single English representation avoids generating and maintaining separate summaries for each language, ensuring that every conversation remains grounded in identical factual content while allowing language-specific realization during dialogue generation. We constrain the summary to 6–8 factual sentences to balance contextual completeness with generation stability, providing sufficient information for generating coherent 6–8-turn conversations while avoiding unnecessary details that can lead to topic drift. The summarization prompt preserves named entities, numerical values, temporal expressions, and locations while discouraging unsupported information.

We compare multiple open-source multilingual models through an arena-style blind human evaluation, where annotators independently compare summaries generated by competing models. As shown in Table 2, Sarvam-M (Sarvam AI, 2026) consistently outperforms Llama 7B (Touvron et al., 2023) and Gemma 7B (Team et al., 2024), achieving the highest Elo rating and win rate. We therefore select Sarvam-M (Sarvam AI, 2026) as the summarization model for the remainder of the pipeline. Additional details of the

Model	Elo	W-L-T	Battles
Sarvam-M	1861.5	267-18-36	321
Llama 7B	1582.1	162-104-36	302
Gemma 7B	1056.4	11-318-2	331

Table 2: Arena-style blind human evaluation of three summarization models. Sarvam-M achieves the highest Elo rating and win-loss record.

evaluation protocol are provided in Appendix B.1, while the summarization prompt is included in Appendix B.2.

3.3 Persona Specification

To encourage conversational diversity, each dialogue is conditioned on a pair of personas representing common real-world interactions. We define five persona categories: *Friends*, *Family Members*, *Colleagues*, *Experts*, and *Student–Teacher*. While the *Friends*, *Experts*, and *Student–Teacher* categories use fixed role pairs, the remaining categories randomly sample role pairs and interaction styles from predefined subcategories. The selected personas remain fixed throughout the conversation to ensure consistent context. Table 3 summarizes the persona configurations.

3.4 Multilingual Dialogue Generation

For each semantic representation s_i, the dialogue generation model $\Phi_g^{(l)}$ independently generates a conversation $c_i^{(l)}$ for every target language variety $l \in \mathcal{L}$ using a multilingual LLM under a self-play framework. The generated conversation before validation is $c_i^{(l)} = \{u_1, u_2, \ldots, u_T\}$, where T denotes the number of dialogue turns. Each utterance is generated autoregressively as $u_t = \Phi_g^{(l)}(s_i, P, u_{<t})$, where P denotes the persona specification and $u_{<t}$ represents the dialogue history up to turn t.

For each Indic language, we generate two complementary code-mixed variants: **Native-Script Code-Mixing**, where the Indic language is written in its native script and English remains in Roman script, and **Romanized Code-Mixing**, where both languages are written entirely in Roman script. Representative examples of both variants are shown in Table 4.

3.5 Automatic Conversation Validation

Despite explicit prompting, multilingual LLMs occasionally generate monolingual or weakly code-mixed conversations. We therefore apply an automatic validation module, Φ_v, to filter generated conversations before constructing the final corpus. A conversation is retained only if it satisfies the following criteria:

1. **Script Validation:** For native-script variants, each utterance must contain at least one token in the target Indic script and one token in the Roman script.
2. **Code-Mixing Threshold:** For native-script variants, every utterance must satisfy a minimum Code-Mixing Index (CMI) (Das and Gambäck, 2014), i.e., $\mathrm{CMI} \geq \tau_{\mathrm{CMI}}$.
3. **Minimum Length:** Each utterance must contain at least τ_{len} tokens.

For Romanized variants, script validation simply verifies the use of Roman characters, while the CMI constraint is omitted because both the Indic language and English share the same script, making token-level language identification unreliable and artificially penalizing natural Romanized code-mixing (Benton et al., 2025). The quality of these conversations is assessed instead through the human evaluation described in Section 4.

3.6 Model Selection for Dialogue Generation

We experimented with several state-of-the-art instruction-tuned language models for multilingual dialogue generation, including GPT-OSS-120B, Qwen2.5-72B, Gemma 4-31B, and their smaller variants. During the iterative development of the generation pipeline, we observed that GPT-OSS-120B consistently produced dialogues that more reliably satisfied our downstream validation criteria, including persona consistency, multilingual fidelity, and structural correctness. In contrast, outputs from the other models more frequently failed one or more validation stages, requiring substantially more regeneration attempts. Based on these empirical observations, GPT-OSS-120B was selected as the primary dialogue generation model for the final dataset. Reference Generation model is present in Appendix E.

4 Dataset Analysis

We analyze INDICTALK from three perspectives. First, we summarize the corpus composition and scale. Second, we characterize its code-mixing properties using established linguistic metrics. Finally, we evaluate conversation quality through automatic metrics, human evaluation, and LLM-as-a-

Category	Representative Role Pairs	Interaction Style
Friends	Friend–Friend	Informal conversations between friends.
Family Members	Mother–Son, Father–Daughter, Siblings, Grandparent–Grandchild, *etc.*	Randomly sampled interaction styles including supportive, curious, advisory, skeptical, optimistic, and debate.
Colleagues	Manager–Employee, Mentor–Mentee, Peer Colleagues, Team Lead–Junior, *etc.*	Professional workplace discussions between hierarchical and peer-level roles.
Experts	Expert–Expert	Technical discussions involving analysis and domain knowledge.
Student–Teacher	Student–Teacher	Educational conversations involve explanation, questioning, and clarification.

Table 3: Persona configurations used for dialogue generation.

Variant	Example Conversation
Native-Script Code-Mixing	**Speaker 1:** গতকাল আমি রাস্তায় গিয়ে একটা সাইকেলিস্টের সামনে হঠাৎই গাড়ি থেমে গেল, আমার মনটা literally froze হয়ে গেল। আমি জানি সেটা AEB system এর thanks, কারণ sensor গুলো ৪৫ mph পর্যন্ত non-moving target detect করতে পারে, আর আমি safety feeling পেয়ে বেশ relaxed লাগলো। যদিও automakers ...
	Speaker 2: ওহ্, তুমি হঠাৎই গাড়ি থামতে দেখলেই আমি totally shocked হলাম! AEB sensor গুলো ৪৫ mph পর্যন্ত non-moving target detect করতে পারে, কিন্তু কি তারা সত্যি ৩৬০ lives yearly বাঁচাতে পারে?regulators ৬২ mph ...
Romanized Code-Mixing	**Speaker 1:** Kal amar choto bhai-er birthday-te sobai baby photo dekhte dekhte, amar mon ektu nostalgic hoye galo, tokhon ei 5-week-old twins er murder suspicion er news shune amar khub shock ...
	Speaker 2: Ami ekta shishu hospital-e volunteer korar somoy ekta choto baby ke niye onek emotional moment dekhlam, tar mayer dukkho dekhte amar mon bhenge gelo. Tokhon ami bujhte parlam je shishuder safety aar ...

Table 4: Representative examples of native-script and romanized code-mixed conversations.

Judge assessment.

4.1 Dataset Statistics

Tables 5 and 6 summarize the statistics of IndicTalk. The corpus comprises over 13,28,604 multilingual code-mixed conversations generated from approximately 1,42,053 news articles and blog posts spanning 12 domains. It covers 9 Indic languages, each represented in both native-script and Romanized code-mixed variants, resulting in **18** language varieties. Each conversation contains an average of **7.55** dialogue turns and is grounded in the same source document across language variants, enabling controlled cross-lingual and cross-script comparisons while preserving the underlying factual content.

4.2 Code-Mixing Characteristics

We characterize the code-mixing properties of IndicTalk using four complementary metrics: Code-Mixing Index (CMI) (Das and Gambäck, 2014), Switch Point Fraction (SPF) (Pratapa et al., 2018), Integration Index (I-index) (Guzmán et al., 2017), and Multilingual Index (M-index) (Barnett et al., 2000).

Table 7 summarizes the corpus-level statistics. Across the native-script variants, the corpus achieves an average CMI of 37.81, indicating a high degree of code-mixing according to the formulation of Das and Gambäck (Das and Gam-

Statistic	Value
News Sources / Blogs	1,42,053
Domains Covered	12
Language Variants	9
Script Variants	2
Persona Categories	5
Total Conversations	13,28,604
Average Turns/conv	7.55
Standard Deviation	0.83
Total turns	10,691,164

Table 5: Overall statistics of the proposed multilingual code-mixed conversational corpus.

Table 6: Total Corpus size along with language-wise distribution, CM- Code-Mixing

Languages	Native	Romanized
Bengali	100,168	45,020
Gujarati	105,004	45,012
Hindi	101,303	45,014
Kannada	104,988	45,012
Malayalam	104,273	45,010
Marathi	103,464	45,021
Odia	101819	45,005
Tamil	100,409	45,006
Telugu	102,069	45,007
Total	**9,23,497**	**4,05,107**

bäck, 2014). The average SPF (0.438) and I-index (0.437) indicate that language switches occur frequently throughout the conversations rather than being confined to isolated insertions (Pratapa et al., 2018; Guzmán et al., 2017). Likewise, the average M-index of 0.865 suggests a well-balanced contribution of both languages across the corpus, rather than dominance by a single language (Barnett et al., 2000). Telugu and Kannada exhibit the strongest bilingual mixing, while all native-script variants maintain consistently high values across the four metrics.

For the Romanized variants, the average SPF (0.311), I-index (0.311), and M-index (0.508) remain consistently positive across all languages, confirming that meaningful code-mixing is preserved despite the shared writing system. The lower metric scores compared to native-script variants is expected, as automatic language identification becomes less reliable when both languages are written in Roman script, leading to conservative estimates of language balance. Overall, these statistics demonstrate that INDICTALK captures frequent language alternation and balanced bilingual usage across both writing conventions.

4.3 Fluency Analysis

While the code-mixing metrics quantify bilingual structure, they do not directly measure linguistic fluency. We therefore evaluate the native-script variants using Pseudo Perplexity (PPPL) computed with multilingual BERT (mBERT) following the masked language model scoring approach of Salazar et al. (2020), where lower PPPL indicates greater linguistic fluency.

As shown in Table 7, the native-script variants achieve consistently low PPPL scores (9.14–18.52), indicating fluent and well-formed conversations despite frequent code-switching. PPPL is not reported for Romanized variants because the absence of standardized orthography and high transliteration variability make pseudo-perplexity scores less comparable across languages. We instead assess their fluency through the LLM-as-a-Judge evaluation described in Section 4.4.

4.4 LLM-as-a-Judge Evaluation

Following the LLM-as-a-Judge paradigm (Zheng et al., 2023; Liu et al., 2023), we evaluate a held-out set of 500 conversations per language (9,000 conversations across 18 language varieties) using two independent judge models: Gemini-2.5-Flash(Comanici et al., 2025) and GPT-OSS-120B(OpenAI et al., 2025). Since GPT-OSS-120B is also used for data generation, Gemini-2.5-Flash serves as an independent judge to mitigate potential self-preference bias. Each conversation is rated on five dimensions: *Fluency*, *Coherence*, *Engagement*, *Code-Mixing Naturalness*, and *Overall Quality*, using a 1–5 Likert scale. The complete evaluation prompt used for both judge models is provided in Appendix C for reproducibility.

As shown in Table 8, both judges consistently assign high scores across all 18 language varieties, with the majority achieving an overall quality score above 4.0. Native-script and Romanized variants exhibit comparable quality, indicating that the proposed generation pipeline produces fluent, coherent, and naturally code-mixed conversations under both writing conventions. The overall quality rankings produced by the two judges are highly correlated ($\rho = 0.91$, $p < 0.001$), despite Gemini-2.5-Flash assigning consistently higher absolute

Language	Native Script					Romanized			
	CMI	SPF	I-index	M-index	PPPL[1]	CMI	SPF	I-index	M-index
Bengali	33.05	0.389	0.387	0.786	14.87	20.13	0.322	0.321	0.474
Gujarati	42.56	0.432	0.433	0.943	18.52	13.48	0.235	0.235	0.305
Hindi	32.64	0.394	0.393	0.777	14.36	10.30	0.136	0.139	0.535
Kannada	43.15	0.482	0.480	0.949	10.33	26.81	0.425	0.425	0.642
Malayalam	38.66	0.456	0.455	0.886	9.14	35.17	0.454	0.454	0.831
Marathi	36.74	0.448	0.447	0.859	14.69	13.66	0.236	0.236	0.309
Odia	32.77	0.413	0.411	0.781	14.35	28.42	0.423	0.423	0.683
Tamil	35.17	0.445	0.443	0.827	13.93	10.81	0.198	0.197	0.239
Telugu	45.52	0.485	0.485	0.976	12.00	23.25	0.369	0.369	0.554
Average	**37.81**	**0.438**	**0.437**	**0.865**	**13.58**	**20.23**	**0.311**	**0.311**	**0.508**

Table 7: Automatic analysis of the proposed multilingual code-mixed conversational corpus. CMI measures the degree of code-mixing, SPF measures language switching frequency, I-index captures language integration, M-index quantifies language balance, and PPPL measures conversational fluency computed using mBERT.

Language	Native		Romanized	
	Gemini	GPT	Gemini	GPT
Bengali	4.42	4.03	4.50	4.03
Gujarati	4.34	4.04	4.28	4.05
Hindi	4.79	4.27	4.79	4.27
Kannada	4.09	4.06	4.10	3.86
Malayalam	4.08	4.01	4.18	4.03
Marathi	4.54	4.22	4.23	3.94
Odia	3.82	3.62	3.55	3.76
Tamil	4.12	4.05	4.18	4.03
Telugu	4.43	4.05	4.23	4.02

Table 8: Mean overall quality scores on a 5-point Likert scale (1 = Very Poor, 5 = Excellent) for native-script and Romanized conversations. Gemini denotes Gemini-2.5-Flash and GPT denotes GPT-OSS-120B. Results are averaged over 500 conversations per language.

scores than GPT-OSS-120B, reflecting differences in judge calibration reported in prior work (Zheng et al., 2023).

4.5 Human Evaluation

To complement the LLM-as-a-Judge evaluation, we conduct a human evaluation on a held-out set of 10 conversations per language variety (180 conversations in total). Each conversation is independently evaluated by at least 2–3 graduate students who are native or highly proficient speakers of the corresponding language. Following the same protocol as the LLM-as-a-Judge evaluation, annotators rate each conversation on *Fluency*, *Coherence*, *Engagement*, *Code-Mixing Naturalness*, and *Overall Quality* using a 5-point Likert scale. The complete annotation guidelines are provided in Appendix D.

Table 9: Mean overall quality scores of Human Evaluation per language variety, evaluated by Human annotators on 10 conversations per language variety

Languages	Native	Romanized
Bengali	4.25	3.4
Gujarati	3.7	3.1
Hindi	4.0	4.4
Kannada	4.3	3.0
Malayalam	3.8	3.3
Marathi	4.4	3.75
Odia	3.6	3.2
Tamil	3.61	3.5
Telugu	3.7	3.0

Table 9 reports the mean overall quality scores across languages and script variants. Consistent with the LLM-as-a-Judge evaluation, native-script variants generally receive slightly higher scores than their Romanized counterparts, although both achieve high overall quality across languages. Marathi and Kannada obtain the highest human ratings among native-script variants, while Hindi receives the highest score for Romanized conversations.

4.6 Overall Evaluation Summary

The three evaluation protocols consistently indicate that INDICTALK contains high-quality multilingual code-mixed conversations across both script variants. The majority of language varieties achieve overall quality scores above 4.0 under both LLM judges and above 3.0 under human evaluation, demonstrating that the proposed generation pipeline produces fluent, coherent, and naturally code-mixed conversations at scale. To quantify agreement, we compute Spearman rank correlations across language varieties. The two LLM judges exhibit strong agreement in their relative quality rankings ($\rho = 0.745$, $p = 0.02$ for native-script and $\rho = 0.778$, $p = 0.01$ for Romanized variants). Human rankings also show moderate positive correlation with both LLM judges($\rho = 0.559$ with GPT-OSS-120B and $\rho = 0.469$ with Gemini-2.5-Flash), indicating broadly consistent quality assessments despite differences in score calibration.

Implementation Details. The complete corpus was generated on 16 NVIDIA H200 GPUs over approximately 10,000 GPU-hours. Models were served through OpenAI-compatible REST APIs and queried via standard HTTP requests with a timeout of 90 seconds per request. Document summarization was performed using Sarvam-M, while dialogue generation employed GPT-OSS-120B. For summarization, we used a low decoding temperature ($T = 0.1$) to produce stable factual summaries. Dialogue generation used a temperature of $T = 0.6$ for the opening utterance and $T = 0.7$ for subsequent turns to encourage conversational diversity while preserving coherence. The maximum generation budget was set to 400 tokens for summarization and 600 tokens per dialogue turn. Thinking mode was disabled for Sarvam-M via the `chat_template_kwargs` parameter.

5 Conclusion

We presented INDICTALK, a large-scale multilingual Indic code-mixed conversational corpus comprising over 13,28,604 event-grounded multi-turn conversations across 18 language varieties covering 9 Indic languages in both native-script and Romanized forms. The corpus is constructed through a fully automated pipeline combining real-world news grounding, persona-conditioned dialogue generation, multilingual LLMs, and automatic quality validation. Extensive automatic and human evaluations demonstrate that INDICTALK contains fluent, coherent, and naturally code-mixed conversations across all supported languages, establishing it as a valuable resource for developing and evaluating multilingual conversational AI for Indic languages.

6 Future Work

Future work includes extending INDICTALK to additional Indic languages, dialects, and script variants, as well as incorporating richer personas and interaction scenarios such as customer support and multilingual debates. Another promising direction is to establish standardized benchmarks for downstream tasks, including conversational summarization, sentiment analysis, and intent detection, enabling systematic evaluation and comparison of multilingual conversational models for Indic code-mixed NLP.

7 Limitations

Although INDICTALK is generated using a carefully designed pipeline with multiple quality control stages, it remains a synthetic corpus and may not fully capture the pragmatic nuances, dialectal variation, and disfluencies found in naturally occurring bilingual conversations. Since the source documents are primarily drawn from news and blogs, the corpus also exhibits a topical bias toward formal and event-centric discourse. In addition, human evaluation is conducted with 2–3 graduate student annotators per language variety and may not fully reflect the diversity of regional dialects and Romanization conventions. Finally, automatic fluency evaluation is reported only for native-script variants, as pseudo-perplexity is not directly comparable for Romanized Indic text.

8 Ethical Considerations

INDICTALK is generated from publicly available news articles and blogs using a fully automated pipeline. We do not release the source documents and remove personally identifiable information (PII) where detected during preprocessing. Human evaluation was conducted with informed consent from volunteer graduate student annotators. All datasets and models used in this work will be publicly available or used in accordance with their respective licenses. We hope that INDICTALK facilitates more inclusive research on

multilingual code-mixed conversational AI for underrepresented Indic languages.

References

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

Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M. Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3766–3780, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ruthanna Barnett, Eva Codó, Eva Duran Eppler, Montserrat Forcadell, Penelope Gardner-Chloros, Roeland Hout, Melissa Moyer, Maria Torras, Maria Turell, Mark Sebba, Marianne Starren, and Sietse Wensing. 2000. The lides coding manual: A document for preparing and analyzing language interaction data version 1.1–july 1999. *International Journal of Bilingualism*, 4.

Adrian Benton, Alexander Gutkin, Christo Kirov, and Brian Roark. 2025. Improving informally Romanized language identification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2318–2336, Suzhou, China. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India. NLP Association of India.

Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for Modeling Code-Switching Across Corpora. In *Interspeech 2017*, pages 67–71.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun Balan G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. 2024. Indicllmsuite: A blueprint for creating pre-training and fine-tuning datasets for indian languages. *Preprint*, arXiv:2403.06350.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Laiba Mehnaz, Debanjan Mahata, Uma Sushmitha Gunturi, Amardeep Kumar, Rakesh Gosangi, Riya Jain, Gauri Gupta, Isabelle Lee, Anish Acharya, and Rajiv Ratn Shah. 2021. GupShup: Summarizing open-domain code-switched conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6177–6192, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553, Melbourne, Australia. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Sarvam AI. 2026. Open-sourcing sarvam 30b and 105b. Blog post.

Gopendra Vikram Singh, Mauajama Firdaus, Shambhavi, Shruti Mishra, and Asif Ekbal. 2022. Knowing what to say: Towards knowledge grounded code-mixed response generation for open-domain conversations. *Knowledge-Based Systems*, 249:108900.

Sakshi Singh, Abhinav Prakash, Aakriti Shah, Chaitanya Sachdeva, and Sanjana Dumpala. 2025. Sample-efficient language model for hinglish conversational ai. *Preprint*, arXiv:2504.19070.

Vivek Srivastava and Mayank Singh. 2020. PHINC: A parallel Hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 41–49, Online. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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.

Appendix

A Source Data Extraction and Cleaning

For each URL in the source collection, content is fetched using a standard HTTP GET request with a browser-mimicking User-Agent header. HTML pages are parsed using BeautifulSoup, with boilerplate elements including navigation menus, headers, footers, sidebars, scripts, and advertisement blocks removed prior to text extraction. PDF documents are handled using pdfplumber, with text extracted page by page and concatenated.

Documents are discarded if the HTTP request fails or if the extracted text falls below a minimum length threshold of 100 words. The pipeline processes articles in sequential batches over non-overlapping index ranges, allowing multiple instances to run in parallel. Each article is wrapped in fault-tolerant error handling so that individual failures do not interrupt the broader pipeline run.

B Summarization

B.1 Model Selection and Evaluation

We evaluate three multilingual open-source models: Sarvam-M(Sarvam AI, 2026), Llama 7B(Touvron et al., 2023), and Gemma 7B(Team et al., 2024), using an arena-style blind human evaluation to select the summarization model. The evaluation is performed on source documents sampled from four language tracks: English, Gujarati, Hindi, and Tamil, reflecting the linguistic diversity of the corpus.

For each document, annotators compare summaries generated by two models with their identities hidden. The original article and source text are provided to verify factual accuracy, content coverage, and preservation of named entities. Model identities are revealed only after the vote is cast, ensuring unbiased preference judgments.

A total of 509 pairwise preference votes are collected. Sarvam-M achieves the highest Elo rating (1861.5), consistently outperforming Llama 7B and Gemma 7B across all language tracks (Table 2). Based on these results, we use Sarvam-M as the summarization model throughout the pipeline.

B.2 Article Summarization Prompt

The summarization model is prompted to generate a language-independent semantic representation of each source document using the following instruction.

> **Instruction:** Summarize the following article in 6–8 concise, factual, and neutral English sentences. Preserve all important named entities, numerical values, dates, temporal expressions, and locations. Include only information explicitly supported by the article. Do not introduce opinions, speculation, or external knowledge. Write the output as a coherent paragraph without bullet points.
>
> **Article:**
>
> {*Input Article*}
>
> **Summary:**

C LLM-as-a-Judge Prompt

D Human Evaluation Prompt

Prompt Template for Human Evaluation

> You are evaluating a synthetically generated `{script_name}` dialogue between two speakers (Speaker_1 and Speaker_2), simulating a conversation between persons discussing a topic.

For Romanized variants only: The dialogue is written entirely in Roman script, with no native script characters. The speakers mix transliterated native-language words with English, reflecting how many urban Indians text informally in their regional languages.

Please read the conversation carefully and rate it on the following five dimensions. Score each dimension on a scale of 1–5 (1 = very poor, 5 = excellent). Be strict in your assessment—penalize grammatical errors, spelling mistakes, unnatural expressions, and implausible code-switching patterns as you would encounter them in real informal digital communication.

- **Fluency**: Is the dialogue grammatically correct and natural? Penalize spelling mistakes, morphologically incorrect word forms, and ungrammatical constructions.
- **Coherence**: Does each turn logically follow from the previous ones? Does the conversation maintain a consistent topic throughout?
- **Code-Mixing Naturalness**: Does the switching between the native language and English feel natural and contextually appropriate? Does it resemble how people actually write informally, or does it feel forced and artificial?
- **Engagement**: Is the conversation interesting and dynamic? Does it feel like a realistic exchange between two people, or does it feel mechanical and repetitive?
- **Overall Quality**: Taking all four dimensions into account, how would you rate the overall quality of this conversation?

Conversation

`{conversation}`

For each dimension, provide an integer score from 1 to 5 and a brief justification in one or two sentences explaining your rating.

E Qwen2.5-72B Vs Gemma 4-31B Vs GPT-OSS-120B

Example Conversations. To qualitatively compare the conversational capabilities of different LLMs, we present representative outputs generated by Qwen2.5-72B refer Table 10, Gemma 4 31B refer Table 11, and GPT-OSS-120B refer to Table 12 . Among the three, GPT-OSS-120B consistently produces the most coherent, contextually grounded, and naturally code-mixed conversations, while the other models exhibit relatively weaker grounding or conversational consistency.

Prompt Template for LLM-as-Judge Evaluation

You are an expert linguistic annotator evaluating a synthetically generated `{script_name}` dialogue between two speakers (Speaker_1 and Speaker_2), simulating a conversation between different persons discussing on a topic.

For Romanized variants only: This is a fully Romanized (Latin script) code-mixed dialogue. The speakers use transliterated native-language words mixed with English entirely in Roman script, with no native script characters. This style reflects how many urban Indians text in their regional languages.

Evaluate the conversation below on the following five metrics. Score each metric on a scale of 1–5 (1 = very poor, 5 = excellent).

- **Fluency**: Whether the dialogue is grammatically correct, natural, and easy to read.
- **Coherence**: Whether each utterance logically follows the previous turns and maintains a consistent topic.
- **Code-Mixing Naturalness**: Whether switching between English and the native language is natural, contextually appropriate, and resembles real bilingual communication.
- **Engagement**: Whether the conversation is interesting, interactive, and resembles a realistic exchange rather than a mechanical question-answer pattern.
- **Overall Quality**: A holistic assessment considering fluency, coherence, code-mixing naturalness, and engagement.

Conversation

`{conversation}`

Return a valid JSON object containing, for each metric, an integer score and a one-sentence justification.

Figure 1: Prompt template used for LLM-as-Judge evaluation. The same prompt was used for Gemini-2.5-Flash and GPT-OSS-120B across all language varieties. For Romanized variants, the italicized instruction was additionally included. Both judges used greedy decoding ($T = 0.0$), and invalid outputs were retried up to three times.

Table 10: Example Hindi-English code-mixed conversation generated by the proposed pipeline, model-Qwen 2.5-72B.

Speaker	Utterance
Speaker_1	यह news थोड़ा suspicious लग रही है, यानी Israel की defense minister ने responsibility ली है, but कुछ details missing हैं। Larijani's death ने तो इरान की leadership में एक बड़ा वाक्यांश छोड़ दिया है, और उनकी retaliation की threat serious है।
Speaker_2	Absolutely, Larijani की death ने इरान की internal dynamics पर बहुत impact डाला है। मेरे एक friend की family इरान में रहती है, उन्होंने बताया था कि tension काफी high है, और people scared हैं कि what next होगा।
Speaker_1	Larijani की death के बाद, मेरा cousin जो Dubai में lives, उसने इरान से अपने family members को बुलाने की plan बना रखी है, क्योंकि वह सोचता है कि situation worse हो सकती है। वह चिंतित है कि इरान की retaliation क्या होगी, और उससे उनकी safety पर क्या होगा।
Speaker_2	Wow, यह तो really serious है! Your cousin's concern understandable है, situation इतनी unpredictable हो गई है कि safety first होनी चाहिए।
Speaker_1	I get where your cousin is coming from, but इरान की retaliation ज़रूरी है, यह उनका आत्म–सम्मान और सुरक्षा का मुद्दा है। फिर भी, हर किसी को अपनी safety के बारे में ध्यान रखना चाहिए, लेकिन panic मत करना चाहिए।
Speaker_2	I see your point, but panic नहीं करना चाहिए, इससे स्थिति control से बाहर निकल सकती है। Safety measures लेना ज़रूरी है, लेकिन calm रहना और सोच–समझकर action लेना बेहतर होगा।

Table 11: Example Hindi-English code-mixed conversation generated by the proposed pipeline,model-Gemma 4-31B.

Speaker	Utterance
Speaker_1	मुझे नहीं लगता कि यह airstrike इतनी simple थी, definitely कुछ internal intelligence leak हुआ है। Israel की intelligence capabilities impressive हैं, पर hideout का exact location पता होना suspicious लगता है। यह whole situation अब बहुत unpredictable हो गई है।
Speaker_2	Exactly, मुझे भी यही लगता है because precise targeting बिना inside information के impossible है। My cousin works in a security agency and he always says that human intelligence is the biggest weak point in any high-profile operation. यह incident prove करता है कि Israel का network Iran के inside बहुत deep है। Do you think Iran will now launch a massive counter-attack?
Speaker_1	मेरे एक friend ने recently एक security conference attend की थी, जहाँ उन्होंने discuss किया था कि कैसे high-ranking officials की digital footprints उन्हें vulnerable बनाती हैं। He mentioned कि modern warfare में physical walls से ज़्यादा cybersecurity important हो गई है। शायद Larijani की location भी किसी encrypted communication leak की वजह से expose हुई होगी।
Speaker_2	यह point एकदम valid है because digital leaks आजकल common हो गए हैं। My old college roommate used to work in a tech firm and he once showed me how easily public Wi-Fi can expose a person's real-time location. यह देखकर मैं shock हो गया था कि basic mistakes कितनी dangerous हो सकती हैं।क्या आपको लगता है कि Iran अपनी security protocols को completely change करेगा?
Speaker_1	मुझे नहीं लगता कि वे security protocols को completely change करेंगे, because it's almost impossible to overhaul an entire system overnight. वे शायद सिर्फ high-level officials के communication channels को narrow कर देंगे ताकि leaks कम हों। system change करने के बजाय वे revenge लेने पर ज़्यादा focus करेंगे।
Speaker_2	I disagree with this perspective because only revenge लेने से problem solve नहीं होगी। Iran को realize करना होगा कि their current security infrastructure is fundamentally broken and needs a complete overhaul. अगर वे sirf communication narrow करेंगे, तो Israels intelligence network still find a way to penetrate their system Israel's intelligence network क्या आपको नहीं लगता कि without a systemic change, वे और भी ज़्यादा vulnerable रहेंगे?

Table 12: Example Hindi-English code-mixed conversation generated by the proposed pipeline- model GPT-OSS-120B.

Speaker	Utterance
Speaker_1	यह news सुनकर मुझे बहुत shock लगा कि Israel ने सीधे Iran के सबसे senior political figure, Ali Larijani, को airstrike से मार दिया, और साथ में Basij commander Gholamreza Soleimani भी eliminated बताया गया । इस तरह का high-profile targeted killing पहले कभी नहीं हुआ, और ये Iran की national security council ने भी confirm किया है, जो war के third week में पहले ही 2,000+ casualties के बीच है। अब IRGC की political influence में बड़ा void create हो गया है, और regional dynamics पूरी तरह shift हो सकते हैं ।
Speaker_2	मैं समझता हूँ आपका shock, लेकिन यह कहना सही नहीं कि ऐसा पहले कभी नहीं हुआ; पिछले साल में भी कई senior Iranian officials को targeted killings का सामना करना पड़ा था। साथ ही, IRGC की political influence में पूरी तरह void नहीं बनेगा, क्योंकि उनका institutional network बहुत deep और resilient है। फिर भी, इस नए void से regional dynamics में ज़रूर कुछ बदलाव आने की संभावना है।
Speaker_1	ये बात सुनते ही दिमाग में ये सवाल उठ रहा है कि क्या Israel अब पूरे Middle-East में ऐसे high-value targets को एक–एक करके eliminate कर रहा है? अगर ऐसा है तो IRGC की internal hierarchy में कौन–से नए power-centers उभर सकते हैं, इसका अंदाज़ा लगाना मुश्किल है। साथ ही, Hormuz Strait की बंदिशें कितनी देर तक टिकेंगी, इस पर भी बहुत curiosity है, क्योंकि global oil flow में इसका असर बहुत बड़ा होगा।
Speaker_2	मैं नहीं मानता कि Israel अब पूरे Middle-East में हर high-value target को systematic तरीके से eliminate कर रहा है, क्योंकि ऐसी strategy उनका operational bandwidth और diplomatic cost बहुत बढ़ा देगी। वास्तव में, वे अक्सर specific intelligence-driven missions पर focus करते हैं, जहाँ political payoff सबसे ज़्यादा हो, न कि व्यापक regional sweep । Hormuz Strait की बंदिशें भी शायद कुछ हफ़्तों तक ही रहेंगी, क्योंकि global oil market का pressure दोनों पक्षों को जल्दी ही negotiation की दहलीज़ पर ले आएगा ।
Speaker_1	क्या Israel अगले हफ़्ते तक किसी और high-value target को भी pinpoint कर लेगा, या ये सिर्फ एक strategic message है?अगर IRGC के अंदर नए power-centers उभरते हैं, तो उनका कौन–सा faction सबसे ज़्यादा influence हासिल करेगा, इस पर भी बहुत curiosity है। इस बीच Hormuz की बंदिशें कब तक चलेंगी, और oil markets में इसका real-time impact कैसे दिखेगा, देखना काफ़ी interesting रहेगा।
Speaker_2	मैं मानता हूँ कि यह सिर्फ एक संदेश नहीं, बल्कि एक calculated move है, लेकिन मैं इस बात से असहमत हूँ कि Israel अगले हफ़्ते तक नया high-value target ले लेगा; उनके operational tempo और diplomatic backlash अक्सर उन्हें ऐसा जल्दी करने से रोकते हैं। IRGC में कौन–सा faction हावी होगा, इसका अनुमान अभी premature है क्योंकि कई parallel networks अभी भी सक्रिय हैं । Hormuz की बंदिशें संभवतः कुछ हफ़्तों में ही ढीली होंगी, क्योंकि oil market की volatility दोनों पक्षों को जल्दी pressure में लाती है।