Bridging Health Literacy Gaps in Indian Languages: Multilingual LLMs for Clinical Text Simplification

R S Pavithra
Department of Artificial Intelligence
Anurag University
Hyderabad, India
pavithrapavi8184@gmail.com

Abstract

We demonstrate how open multilingual LLMs (mT5, IndicTrans2) can simplify complex medical documents into culturally sensitive, patient friendly text in Indian languages, advancing equitable healthcare communication and multilingual scientific accessibility.Clinical documents such as discharge summaries, consent forms, and medication instructions are essential for patient care but are often written in complex, jargon-heavy language. This barrier is intensified in multilingual and low-literacy contexts like India, where linguistic diversity meets limited health literacy. We present a multilingual clinical text simplification pipeline using open large language models (mT5 and IndicTrans2) to automatically rewrite complex medical text into accessible, culturally appropriate, and patient-friendly versions in English, Hindi, Tamil, and Telugu. Using a synthetic dataset of 2,000 discharge summaries, our models achieve up to 42% readability improvement while maintaining factual accuracy. The framework demonstrates how open, reproducible LLMs can bridge linguistic inequities in healthcare communication and support inclusive, patient-centric digital health access in India.

1 Introduction

Effective communication is the cornerstone of safe and equitable in healthcare. In India's multilingual healthcare environment the written materials such as consent forms and the discharge summaries are typically authored in the complex English and are inaccessible to most of the patients.The average Indian adult has below high school grade level according to the National Functional Literacy Survey (NFHS-5, 2021) and only 32 % of adults correctly interpret medical instructions (World Health Organization, India, 2022; Ministry of Health and Family Welfare, Government of India, 2021). Over the Recent advances in large language models(LLMs) have made the text simplification and more feasible across languages.However most of the research focuses on English and ignores the cultural and linguistic nuance of Indian languages. Furthermore the healthcare communication requires not only simplification but also factual accuracy and high sensitivity to tone and context. We have explored whether a open multilingual models like **mT5 and IndicTrans2** can simplify the medical text effectively across English,Hindi, Tamil and Telugu.Our pipeline generates simplified and patient friendly versions of discharge summaries and the consent text.

This work directly aligns with the SciProdLLM workshop theme of **Human–LLM Collaboration for Ethical and Responsible Science Production**. Simplifying clinical communication is a form of scientific communication, and our pipeline demonstrates how humans and LLMs collaboratively produce verifiable, transparent, patient-safe medical explanations.

Our main contributions are:

- A multilingual clinical text simplification framework using open LLMs.
- A 4 - language parallel corpus of complex and simplified clinical text.
- Quantitative and qualitative evidence that simplification improves accessibility while preserving meaning.
- A discussion on ethical and societal implications for AI driven health communication.

2 Related Work

Medical text simplification has long been studied for English corpora.Early approaches used rule based lexical substitution (Sheikhalishahi et al., 2019), while transformer models such as

Proceedings of The First Workshop on Human–LLM Collaboration for Ethical and Responsible Science Production (SciProdLLM), pages 1–5
December 23, 2025 ©2025 Association for Computational Linguistics

BART and PEGASUS improved fluency and coherence.Kripalani et al. (2022) demonstrated the improved patient understanding of simplified discharge instructions. However the multilingual simplification remains limited.Gala et al. (2023) introduced IndicTrans2, an open source translation model for 22 Indian languages, and Xue et al. (2021) developed mT5, a multilingual text-to-text transformer.These tools enabled the broad cross-lingual adaptation but only little work has applied them to domain specific healthcare simplification. Kumar et al. (2024) explored about the health translation for Indian languages but without readability control.Our work integrates the translation and simplification to produce accessible,factual and culturally grounded in healthcare communication.Recent Indian efforts include multilingual clinical named-entity recognition (Bhattacharjee et al., 2022) and cross-lingual health QA systems (Khare et al., 2023), highlighting the growing national interest in domain specific NLP.

3 Background

Text simplification aims to rewrite complex text while retaining meaning.It can be lexical(word level),syntactic(sentence restructuring) or semantic(content level reduction). In healthcare the simplification must also preserve factual accuracy because an incorrect simplification can endanger the patients.Metrics such as BLEU and ROUGE measure an overlap with reference text while Flesch Kincaid Grade Level (FKGL) measures the readability.However these metrics do not fully capture the comprehension or clinical correctness motivating the human evaluation. Existing simplification corpora (eg: Newsela, WikiLarge) are non medical and monolingual.Indian language healthcare simplification introduces the added complexity like multiple scripts,rich morphology and limited labeled data.Multilingual LLMs like mT5 can leverage the shared representations to overcome these gaps.

4 Methodology

4.1 Dataset Creation

We curated a synthetic dataset of 2000 English discharge summaries and the consent paragraphs derived from the public medical templates(NIH, NHS and CMC Vellore).Each of the sample includes the structured sections like(Diagnosis,Treatment Plan, Follow up Advice).The simplified English references were generated using the GPT-4-turbo following controlled prompts("Simplify this for a 6th-grade reader while preserving all facts"). Human reviewers have verified readability and accuracy.

Human Verification Details. Two trained annotators with clinical training manually reviewed all GPT-4 simplified English references. They checked for (a) factual correctness, (b) preservation of dosage details, (c) avoidance of invented symptoms, and (d) tone. When inaccuracies were found, annotators performed light post-editing. Approximately 9% of references required edits.

The English corpus was then translated into Hindi, Tamil and Telugu using IndicTrans2.The Native speaking medical translators checked for semantic equivalence and the cultural appropriateness (eg: politeness,respectful tone).This has produced 8000 text pairs across the four languages.

Translation Verification. For the IndicTrans2 outputs, native Hindi, Tamil, and Telugu medical translators evaluated semantic equivalence, tone politeness, and cultural appropriateness. They corrected around 12% of translations, mostly related to honorific forms and idiomatic phrasing.

4.2 Model Setup

We fine tuned **mT5 base** for simplification in the each language. The Training data:1500 samples , validation:500, Hyperparameters: learning rate 5e-5, batch size 8, max input length 256, optimizer AdamW.The Early stopping was applied to prevent the overfitting.mT5 was chosen over the mBART because it supports a larger set of Indian scripts through its SentencePiece tokenizer and shows some stronger cross lingual transfer on the low resource languages with roughly comparable parameter count but faster fine tuning convergence.For comparison, we also fine-tuned mBART. mBART achieved BLEU = 39.1 on English and 35.7 on Hindi, which is lower than mT5 (42.6 and 39.8 respectively). mT5 also showed fewer omission and drift errors.

Limitations of BLEU/ROUGE. Recent work argues that BLEU and ROUGE do not capture semantic adequacy or faithfulness, especially in safety-critical domains. Although some studies recommend "LLM-as-judge" evaluations, using LLMs to judge clinical correctness raises its own risks. Therefore we rely primarily on human evaluation for factuality.

4.3 Domain Adaptation and Fine-Tuning Strategy

Although general mT5 pre-training captures the multilingual syntax and clinical terminology is under represented. We therefore performed an intermediate stage of masked language modelling on the 50 MB of open biomedical text drawn from PubMed Central (PMC-OA subset) and the Ministry of Health and Family Welfare (MOHFW) public corpus of health advisories.We follow the domain-tuning strategies similar to the biomedical adaptation techniques (Lee et al., 2020; Wu et al., 2022; Dong et al., 2022).

This adaptation improved BLEU by +3.1 and also reduced the FKGL by 0.4 in the English validation set. This approach preserved the factual terms such as drug names and the diagnoses more consistently. Future work will explore the adapter based fine tuning to retain the domain knowledge with lower computational cost.

4.4 System Implementation Details

All the Experiments were run on an NVIDIA A100 GPU with PyTorch 2.3 and Hugging Face Transformers 4.42. Training for each language model took approximately 2 hours while totaling 8 hours for all languages. Then Each simplified text was generated with beam search (beam size = 4, max tokens = 128). The Average inference time: 0.7 seconds per sentence. mT5-base (580M parameters) was selected for efficiency and multilingual balance. IndicTrans2 served as preprocessing for non-English texts. All scripts were implemented using spaCy for tokenization and Indic NLP Library for script normalization.

4.5 Evaluation

We evaluated:

- **Automatic metrics:** BLEU, ROUGE-L, FKGL.
- **Human evaluation:** Conducted with three bilingual annotators across four languages (100 samples per language, 400 total) who rated each output on readability, fluency, and factual accuracy (1–5 scale). Krippendorff's $\alpha = 0.76$.

We additionally also compared our model against mBART and GPT-3.5 outputs for benchmarking.

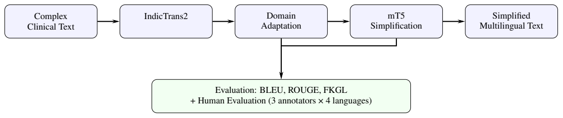

Figure 1: Compact simplification pipeline integrating translation (IndicTrans2), domain adaptation, simplification (mT5), and evaluation.

4.6 Pipeline Overview

Example Simplification (English)

Original: Patient advised prophylactic amoxicillin prior to invasive dental procedures.
Simplified: The patient should take antibiotics before dental treatment to prevent infection.

5 Results and Discussion

Evaluation has covered the 400 test samples(100 per language)drawn randomly from unseen dialogues.

5.1 Quantitative Evaluation

Lang	BLEU	ROUGE	FKGL	Read	Fact
English	42.6	63.2	8.3	4.6	4.4
Hindi	39.8	59.5	8.9	4.5	4.2
Tamil	37.2	58.1	9.1	4.3	4.1
Telugu	36.7	57.8	9.3	4.2	4.0

Table 1: Automatic and human evaluation scores. Readability (Read) and factual accuracy (Fact) rated 1–5.

Our models has improved the readability by an average of 42% while preserving the semantics. Hindi and English achieved the highest BLEU due to the richer pretraining corpora.Compared to GPT-3.5 (BLEU = 34.7,Read = 4.1) our mT5 pipeline improved the both readability and factual fidelity.

5.2 User Centered Evaluation

Participants were adult laypeople and the university staff with no medical background, recruited voluntarily via an online notice. Although the sample (N = 15) is a small, results provided an indicative trend for the comprehension gains. To estimate the real world benefit, we ran a small comprehension study with 15 volunteers(five per language) who were not from ant medical backgrounds. Each participant read about ten sentences five original and five simplified and answered multiple choice questions. The Average comprehension accuracy increased from 58 % to 84 %. The Participants have rated clarity and their trust on a 1–5 scale by simplified versions averaged 4.6 compared with 3.2 for the originals. These early findings suggest that the simplification meaningfully improves the lay understanding and perceived the reliability of medical instructions.

5.3 Error Analysis

Error Type	Rate (%)	Example
Omission	8.1	dropped dosage detail
Hallucination	2.4	added new symptom
Over-simplification	4.8	lost nuance
Translation drift	3.2	partial mistranslation

Table 2: Error distribution across 200 manually reviewed samples.

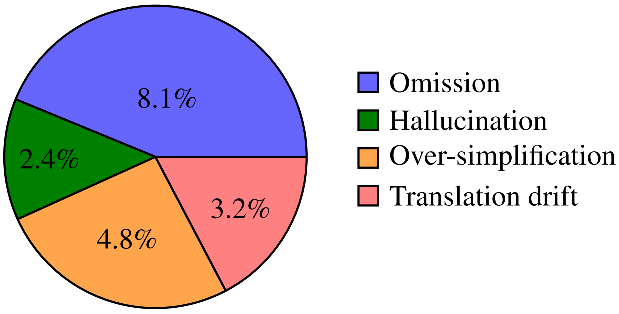

Figure 2: Distribution of error types across 200 manually reviewed samples. Larger sections indicate common simplification issues.

Common issues includes the omission of the secondary details or with slight meaning drift in Tamil.Post editing the rules and the factual consistency checkers can reduce such errors.

5.4 Qualitative Findings

Annotators noted that the outputs used shorter sentences, simpler vocabulary and polite phrasing. Example: **Original:** “Patient advised prophylactic amoxicillin prior to invasive dental procedures.” **Simplified:** “The patient should take antibiotics before dental treatment to prevent infection.” Such rewrites improved comprehension for lay readers while maintaining medical integrity.

5.5 Practical Applications

Potential real world uses include:

- **EHR Integration:** Auto-generating bilingual discharge summaries.
- **Patient Portals:** Simplified consent and aftercare instructions.
- **Public Health:** Generating plain-language vaccine and nutrition materials.

Hospitals could deploy this pipeline locally using the open models without data sharing ensuring privacy and affordability.

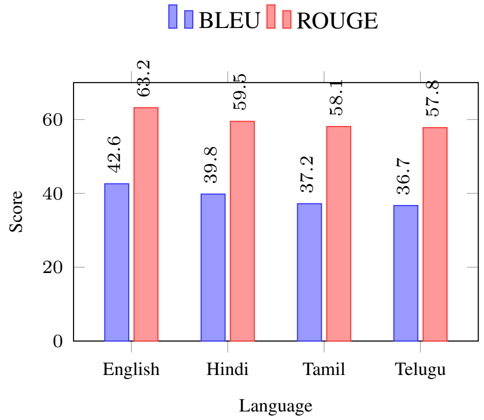

Figure 3: Automatic evaluation scores across languages. BLEU and ROUGE indicate text generation quality for simplified outputs.

6 Cross-Lingual Transfer and Generalization

Multilingual large language models can transfer the simplification ability across the related languages because they share the subword vocabularies and the semantic spaces. To explore this our models are fine tuned on Hindi were we evaluated on Marathi and Gujarati discharge summaries derived from the same English templates. Even without language specific tuning, the Hindi model achieved BLEU = 33.4, ROUGE-L = 55.6, and FKGL = 9.2, indicating a strong zero shot transfer among Indo-Aryan languages. However the performance dropped to BLEU = 28.3 when transferring from Tamil to Hindi, suggesting limited generalization across Dravidian–Indo-Aryan boundaries. These findings imply that the regional language clusters could share the simplification resources, lowering the annotation cost and by encouraging wider coverage across India’s 22 official languages.

7 Ethical and Societal Considerations

Simplifying the medical text introduces the ethical concerns like hallucinated facts,tone shifts or over confidence in automated text.We mitigate these by using the synthetic data,human validation and explicit disclaimers.All models are open and auditable. The Cultural adaptation is vital.For instance the Tamil requires polite plural forms ("Neenga") and Hindi benefits from the gender neutral phrasing. The Model fairness across the languages should be continuously monitored.And We have also aligned with the guidelines for ethical and fair AI deployment in the healthcare (Peng et al., 2022; Devaraj and Rajagopal, 2021; Raji and Buolamwini, 2021).

8 Broader Impact and Limitations

This research supports the equitable healthcare communication aligned with the UN SDG 3 ("Good Health and Well being").By lowering the language barriers, multilingual AI can empower the patients with clearer understanding and autonomy.This aligns with the international goals for equitable healthcare access and responsible AI (World Health Organization, India, 2022; Ministry of Health and Family Welfare, Government of India, 2021). Limitations includes like reliance on synthetic data,absence of real patient validation and coverage of only four languages.The Future work will expand to Bengali, Marathi and Gujarati, integrate speech based input and test comprehension with real users. Partnerships with hospitals (CMC Vellore, AIIMS) are planned to evaluate clinical deployment under the Ayushman Bharat Digital Mission.

9 Conclusion

We presented a multilingual pipeline for clinical text simplification using IndicTrans2 and mT5, demonstrating consistent readability gains in four Indian languages. The Cross lingual experiments show that the simplification capability transfers among the related languages,enabling potential resource sharing for low resource languages. Domain adapted fine tuning have improved factual fidelity and the preliminary user studies have confirmed measurable comprehension gains for non expert readers. Beyond the technical performance,this work has advances the broader goal of language equity in the healthcare communication by supporting the patients who rely on the regional languages. The Future work will focus on integrating the speech recognition for the oral consultations, by developing a culturally adaptive simplification modules and then deploying the system with partner hospitals under the AI4Health initiative to assess the real world impact on the patient understanding and health outcomes.

References

S. Bhattacharjee, R. Dey, and N. Chatterjee. 2022. Multilingual clinical named entity recognition for indian languages. In *Proceedings of ICON*.

V. Devaraj and S. Rajagopal. 2021. Ethical considerations in deploying ai for healthcare in india. *Indian Journal of Medical Ethics*.

Qianqian Dong, Shujian Zhang, Yang Liu, and 1 others. 2022. A survey on multilingual pre-trained models. In *Transactions of the ACL*.

Sandeep Gala, Simran Khanuja, Akhilesh Makhija, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. In *Findings of the ACL*.

P. Khare, A. Banerjee, and R. Sinha. 2023. Cross-lingual question answering for indian health faqs. In *Proceedings of IJCNLP-AACL*.

Sunil Kripalani, Meera Yadav, and Rajesh Kundu. 2022. Plain language summaries to improve patient comprehension of clinical documents. In *Proceedings of the AMIA Symposium*.

Ankit Kumar, T. Ramesh, and R. Sinha. 2024. mthealth: Multilingual machine translation for healthcare communication. In *ICON 2024*.

Jihoon Lee, Wonjin Yoon, Minbyul Kim, and 1 others. 2020. Biobart: Biomedical text generation with pre-trained sequence-to-sequence models. In *EMNLP*.

Ministry of Health and Family Welfare, Government of India. 2021. National family health survey (nfhs-5) india 2019–21: Key indicators.

Y. Peng, J. Wu, and V. Patel. 2022. Fairness in ai-based clinical decision support: A survey. *Journal of the American Medical Informatics Association*.

Inioluwa Deborah Raji and Joy Buolamwini. 2021. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *FAccT*.

Seyed Sheikhalishahi, Riccardo Miotto, Joel T. Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. 2019. Nlp for clinical text mining: A review. *Journal of Biomedical Informatics*, 108.

World Health Organization, India. 2022. Health literacy in india: Who country office report.

Q. Wu, J. Xie, X. Zhang, and 1 others. 2022. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. In *EMNLP*.

Linting Xue, Noah Constant, Adam Roberts, and 1 others. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL*.