Sample Efficient Language Model for a Conversational Hinglish Chatbot
Project Status Report

Sakshi Singh sakshisi@usc.edu
Abhinav Prakash ap42546@usc.edu
Aakriti Shah shahaakr@usc.edu

Chaitanya Sachdeva csachdev@usc.edu
Sanjana Dumpala dumpala@usc.edu

1 Tasks That Have Been Performed

We have made some progress on the tasks outlined in our proposal. First we sat down and tried to refine our goal to make it more clear and streamline our tasks with more specific details (i.e., exact models and how we will get them, exact datasets, etc). We are developing a Hinglish conversational chatbot and are currently identifying a suitable model to fine-tune using both open-source Hinglish data and synthetically generated Hinglish data.

During the data collection and augmentation phase, we reviewed existing datasets from our proposal – Hinglish-TOP [6], FLORES [2], and PHINC [1] – for integration. However, Hinglish-TOP and PHINC did not offer the conversational Hinglish data we required, and FLORES is geared towards benchmarking multilingual translations. Consequently, we shifted our focus to HinGE [8], LinCE [7], and IITH [3], which provide more relevant conversational Hinglish content. Our initial exploration confirmed gaps in Hinglish representation, and while we considered web scraping, the existing data proved sufficient. We also generated a synthetic Hinglish corpus to effectively meet all of our project's essential conversational needs.

Fine-tuning approaches are currently being explored to improve contextual accuracy and code-switching naturalness. Moving on to the model selection and fine-tuning phase, we evaluated multiple pretrained cross-lingual language models – including Gemma3-1B, Deepseek-1.5B, DistilBERT [4], Gemma 4B, GPT-2, and mt5 [5] – to assess their capabilities in handling Hinglish and code-switching scenarios.

Through this, we found that leveraging Gemma-1B or mt5's existing cross-lingual knowledge, combinded with our fine-tuning strategy, would best adapt the model to our Hinglish dataset. We are still considering other models, but Gemma-1B and mt5 are still the most effective.

Here are the models and their corresponding parameter counts:

Model	Number of Parameters
Gemma 4B	4 billion
Deepseek-1.5B	1.5 billion
Gemma3-1B	1 billion
mT5	300 million - 13 billion
GPT-2	120 million
DistilBERT	66 million

Table 1: Models with their Parameter Counts

2 Risks and Challenges

1. Lack of Hinglish Conversational Datasets: Existing datasets primarily consist of isolated, one-off utterances rather than full conversations.

2. Noisy and Romanized Spelling Variants: Hinglish text in Roman script often has inconsistent spellings (e.g., “bahut”, “bhot”, “bahout”), introducing noise that hinders the model's ability to learn consistent lexical patterns.

3. Training Limitations of Small Parameter Models: Small models, though efficient, often struggle to capture the complexity of noisy code-mixed data, leading to lower-quality and less fluent responses.

4. Inadequacy of Standard Evaluation Metrics: Conventional evaluation metrics such as BLEU and ROUGE, developed for monolingual text, often fail to reflect the quality of code-mixed outputs. They are sensitive to spelling variations, informal structure, and lexical diversity, making them suboptimal for evaluating Hinglish dialog systems.

3 Mitigation Plan for Risks

1. **Addressing the Lack of Conversational Data:**

   To address the lack of multi-turn Hinglish datasets, we:

   - Generated synthetic dialogues from existing datasets using GPT-style prompts
   - Crowdsourced responses via platforms like Google Forms
   - Plan to generate additional data through web scraping and public sources

2. **Normalizing Noisy Hinglish Input:**

   To reduce linguistic noise, we implemented data normalization strategies including:

   - Standardizing common romanized Hindi word variants
   - Removing attached punctuation and emoji clutter
   - Filtering excessively noisy or ambiguous examples

3. **Enhancing Small Model Performance:**

   We plan to address small model limitations through:

   - LoRA/QLoRA for efficient fine-tuning with reduced memory usage
   - Domain-specific pretraining on unlabeled Hinglish text to improve language understanding

4. **Enhancing Evaluation with Code-Mixed Metrics:**

   Recognizing the limitations of traditional metrics, we plan to complement BLEU and ROUGE with:

   - Code-Mixing Index (CMI) to quantify the degree of language mixing
   - Embedding-based metrics (e.g., BERTScore, cosine similarity) to capture semantic similarity
   - Human evaluation protocols to ensure alignment with conversational fluency and cultural appropriateness

4 Individual Contributions

All five members contributed equally to data preprocessing, model evaluation, and documentation. Sakshi designed experiments for fine-tuning Gemma-1B and mt5 on the Hinglish dataset, optimized hyperparameters, and explored techniques to improve contextual fluency and code-switching quality. Abhinav evaluated multiple pretrained language models (e.g., Gemma3-1B, Deepseek-1.5B, GPT-2) for their suitability in Hinglish scenarios and conducted technical benchmarks to finalize Gemma-1B and mt5. Aakriti generated synthetic Hinglish dialogues using GPT-style prompting, implemented data normalization and augmentation strategies, and contributed to reference compilation. Chaitanya supported synthetic data generation and helped build a unified preprocessing pipeline to clean and integrate multiple datasets including HinGE, LinCE, and IITH. Sanjana searched and evaluated available Hinglish datasets (e.g., Hinglish-TOP, PHINC) for conversational coverage and co-developed the preprocessing pipeline. Sakshi and Aakriti wrote Section 1 and compiled references, while Abhinav, Chaitanya, and Sanjana worked on Sections 3 and 4. All members collaborated on overall planning, troubleshooting, and refining the project direction throughout.

References

[1] Mrutyunjay Biswal. 2020. Phinc parallel hinglish corpus - machine translation. Kaggle. Accessed: 2025-03-03.

[2] Open Language Data. 2024. Flores+ dataset. https://huggingface.co/datasets/openlanguagedata/flores_plus. Accessed: 2025-03-05.

[3] Drimpossible. 2021. Iiith codemixed dataset. Accessed: 2025-04-02.

[4] Hugging Face. 2020. Distilbert. https://huggingface.co/docs/transformers/en/model_doc/distilbert. Accessed: 2025-03-05.

[5] Google Research. 2021. mt5-small. https://huggingface.co/google/mt5-small. Accessed: 2025-03-05.

[6] Google Research. 2024. Hinglish-top dataset. https://github.com/google-research-datasets/Hinglish-TOP-Dataset. Accessed: 2025-03-05.

[7] TheDevastator. 2021. Unlock universal language with the lince dataset. Accessed: 2025-04-02.

[8] Mayank Singh Vivek Srivastava. 2021. Evaluating models for code-switching and natural language processing in hinglish. Accessed: 2025-04-02.