Shahjalal University of Science and Technology

Research Proposal for Pathshala.ai

A Computationally Minimal Bangla Language Model for Bangladeshi High School Education

Prepared By:

Abraar Masud Nafiz and Hamim Rahman

4th Year Students, Software Engineering

2019831076, 2019831034

Shahjalal University of Science and Technology (SUST)

Submitted To:

Dr. Ahsan Habib

Associate Professor

Institute of Information and Communication Technology (IICT)

Shahjalal University of Science and Technology, Sylhet-3114, Bangladesh

Phone: +8801915796886

Email: ahabib-iict@sust.edu

Contents

1 Objective

We aim to develop a computationally efficient, minimal Bangla language model as an offline personal tutor for high school students in rural Bangladesh. This AI-driven, accessible educational tool bridges learning gaps for students in classes 6-10 by providing curriculum-aligned assistance.

2 Methodology

2.1 Model Development

- **Architecture Selection:**
  We will begin by selecting a transformer-based architecture that prioritizes computational efficiency, choosing a variant with minimal layers and attention heads to balance performance and resource requirements. For further optimization, we will explore sparse attention mechanisms and quantization techniques to reduce memory usage and computational load.

- **Data Curation:**
  The dataset, sourced from platforms like Kaggle and supplemented with high school-level educational content, will undergo rigorous data-cleaning to remove noise and irrelevant information. This step will ensure that the model is focused solely on curriculum-relevant topics for classes 6–10, maximizing educational relevance.

- **Model Optimization:**
  Once the basic model is functional, we will implement additional optimization techniques, such as knowledge distillation, where a smaller model learns from a larger, pre-trained model. This will further enhance the model's ability to perform well within limited computational constraints. Additionally, lightweight embeddings specific to the Bangla language will be explored to streamline tokenization and processing.

2.2 Data Collection and Training

- **Preprocessing and Segmentation:**
  After initial data curation, preprocessing will include tokenization, part-of-speech tagging, and contextual segmentation tailored to Bangla. A focus on reducing vocabulary size will also help constrain model size. We will test different techniques for embedding segmentation, emphasizing approaches that handle the nuances of Bangla morphology and syntax.

- **Training Process:**
  The model will be trained on a local server with carefully monitored memory usage. Training will involve incremental iterations, with hyperparameter tuning to balance accuracy and computational expense. We will experiment with existing pre-trained Bangla embeddings if they align with our objectives, allowing us to potentially shorten the training time.

2.3 Deployment and Evaluation

- **User Interface Development:**
  A minimalistic, web-based UI will be developed using lightweight frameworks (e.g., Flask or Django) to ensure responsiveness on low-end devices. The interface will include basic tutoring features, with interactive question-answer segments, problem-solving exercises, and feedback for students.

- **Field Testing:**
  A pilot study will involve students from varied rural settings to gather real-world feedback. This will assess usability, effectiveness, and potential accessibility issues, with evaluations on latency, user comprehension, and interaction quality. Feedback will inform iterative UI adjustments and model refinement.

2.4 Comparative Analysis

- **Benchmarking:**
  Pathshala.ai's performance will be compared with larger-scale Bangla language models where possible. Metrics will include response accuracy, memory usage, processing speed, and educational effectiveness. This analysis will highlight the computational savings achieved by Pathshala.ai while gauging its adequacy as an educational tool.

3 Work Schedule

- **Model Development and Data Collection**
  - Develop the core architecture with initial configurations for a small, efficient model.
  - Gather and preprocess educational data, with filtering and segmentation to focus on high school-level Bangla content.
  - Preliminary experimentation with optimization techniques and the inclusion of small-scale embeddings.

- **Data Refinement and Initial Training**
  - Finalize the dataset and conduct data augmentation if necessary.
  - Start initial model training on curated datasets, monitoring for memory efficiency and iterating on model structure as required.

- **Model Optimization and Web UI Development**
  - Further refine the model using distillation and quantization techniques, adjusting for peak efficiency.
  - Begin developing the web-based UI, ensuring it is compatible with low-power devices.
  - Conduct preliminary testing of the UI with a subset of users to gather usability feedback.

- **Field Survey and Comparative Testing**
  - Implement a three-day survey with a diverse sample of students from rural areas to test Pathshala.ai's effectiveness and usability in real-world conditions.
  - Conduct comparative testing with existing models to benchmark Pathshala.ai's educational impact and computational efficiency.
- **Documentation and Reflection**
  - Compile findings, finalize the analysis, and prepare documentation for submission.
  - Reflect on the project's outcomes, noting areas for future improvement and possible extensions to functionality.
- **Submission of Final Report**
- **Thesis Defense Preparation**

4 Future Prospects

If the initial version of Pathshala.ai proves successful, we envision expanding its capabilities to include audio-based interaction for more inclusive intuitive student engagement. With Dr. Habib's expertise in signal processing and experience in Bangla text compression and mobile-based learning, his guidance would be essential as we explore this aspirational feature and further improve accessibility.

5 Conclusion

Pathshala.ai has the potential to impact the educational landscape for rural high school students in Bangladesh. Dr. Habib's guidance will be instrumental in refining our approach and maximizing the project's success. We are eager for your feedback and any additional suggestions you may offer to refine our approach and maximize the impact of our work.