Overview of Changes

The previous version of this manuscript focused primarily on the introduction of the BCoQA dataset. Based on reviewer feedback that the methodology lacked sufficient evaluation and justification, we have fundamentally revised and expanded the paper. The focus has shifted from simply presenting a new dataset to providing a comprehensive, systematic analysis of the pipeline used to create high-quality, machine-translated resources for low-resource languages.

The key revisions are as follows:

1. **Comparative MT System Evaluation (Section 3):** We have moved beyond using a single MT model. The revised paper now includes a systematic comparison of five state-of-the-art English-to-Bangla translation systems (BanglaT5, IndicTrans2, NLLB, M2M100, and Gemma3). We evaluate them across four different test sets using both legacy (BLEU, chrF++) and modern neural metrics (COMET, CometKiwi, BLEURT) to identify the most effective model for this task.

2. **Human Evaluation of QE Methods (Section 4.2):** The original paper used a standard LaBSE cosine similarity filter without validation. We now include a small-scale human evaluation study (500 samples rated by 25 annotators) to directly compare the correlation of LaBSE and the reference-less metric CometKiwi with human judgments of translation quality.

3. **Empirically Grounded Filtering Threshold (Section 4.4):** We replaced the previous ad-hoc LaBSE threshold of 0.7 with a new, empirically-grounded CometKiwi threshold of 0.67. This threshold was determined through a two-stage process: first, analyzing the statistical correlation with our human evaluation data to identify an optimal range, and second, refining the final value through manual inspection of translation quality at different cutoffs.

4. **Recreated Dataset and New Baselines (Section 5 & 6):** The BCoQA dataset has been entirely recreated from scratch using the best-performing MT system (IndicTrans2) and the validated filtering pipeline (CometKiwi > 0.67). We have also updated our benchmarks on the new higher quality dataset and compared them against human performance.

In essence, the paper has evolved from a resource paper into a study on best practices for creating synthetic datasets, using the development of BCoQA as a case study. The result is a more robust, validated, and reproducible contribution to the field.

Addressing Key Concerns from the Previous Submission

Reviewer **Lh7y**

No literature review section.

- We agree this was a significant omission. We have added a much more comprehensive **"Related Works" section.** This section discusses prior work in three key areas: (1) the creation of Bangla datasets using machine translation and filtering, (2) recent advancements in English-to-Bangla MT systems, and (3) the landscape of existing Bangla QA datasets. This allows us to properly situate our work and clearly identify the research gaps we aim to fill, namely the lack of standardized evaluation and validated filtering methods.

Unclear why the dataset is claimed to be "robust".

- We have removed subjective claims like "robust" and instead focused on demonstrating the quality of our dataset through a rigorous, evidence-based methodology. The entire paper has been restructured to present this methodology, which includes:
    - A systematic evaluation of five MT systems to select the best one (**Section 3**).
    - A human evaluation study to validate our choice of QE metric (**Section 4.2, 4.3**).
    - An empirically-grounded approach to setting the filtering threshold (**Section 4.4**). This systematic process provides concrete evidence for the best possible quality of the resulting dataset.

The paper should mention licenses.

- We have added Licensing Information in **Appendix E.**

Mentioning the Original in Figure 1

- We have added the link to the original English passage in the caption of Figure 1

Reviewer **xYGg**

What is the prompt used to convert span-based answers to free-form?

- We have now included an example of the input prompt and generated output in **Figure 2 (page 7)**.

Details about human evaluation participants (volunteers, remuneration, recruitment, instructions).

- We have significantly expanded the details of our human evaluation study to improve reproducibility.
    - We have added a new appendix, **Appendix F**, which includes the complete annotation guidelines and a screenshot of the user interface (**Figure 4a & Figure 4b**) provided to the human evaluators.
    - Details regarding participant recruitment will be included in the final camera-ready version to maintain anonymity during the review process.

Provide translations for examples:

- We have gone through the manuscript and added English translations for all Bangla examples to improve readability for a broader audience.

Reviewer **zCxR**

The paper doesn't explore large multilingual models (>1B).

- The previous version's reliance on a single model was a major limitation. Our revised manuscript now centers around a **comprehensive comparative analysis of five state-of-the-art systems (Section 3.1)**, specifically including large multilingual models as requested.

The process of rewriting extracted answer spans into free-form text is unclear.

- We have added a detailed explanation of this process. This rewriting step is applied *only* to the QuAC dataset, as CoQA already contains free-form answers. **Figure 2 (page 7)** provides a concrete example of the input prompt and the resulting generated answer.

Provide English translations and link figures:

- We have added English translations for all Bangla text in the paper and updated the caption of **Figure 1** to reference its English counterpart in the appendix.

Will thresholding introduce any bias?

- We acknowledge that any filtering process risks introducing bias. Our work aims to make this process more transparent and less arbitrary. The human evaluation in **Section 4.3** demonstrates that a common method like LaBSE penalizes valid translations (e.g., transliterated names), creating a clear bias. Our switch to CometKiwi is motivated by its stronger correlation with human judgments on these exact cases. While this reduces known sources of error, we concede that other, more subtle biases may still exist. We consider the formal analysis of filtering-induced dataset bias an important area for future research.

Meta Reviewer **Area Chair jTWv**

The dataset is available only in one language i.e., Bangla.

- We acknowledge this is true by design. Our primary goal is to address the significant data scarcity in Bangla, a low-resource language. We have now extended this paper to systematically find the best synthetic dataset creation pipeline for Bangla.

Authors do not employ large models for translation.

- We have comprehensively addressed this limitation. For translation, we now evaluate five systems, including **NLLB-200-3.3B**, **IndicTrans2-1B**, and the **Gemma3-12B LLM**, to quantify and find the best English to Bangla translation model (**Section 3.4**).

Limited novelty of the paper.

- This was the most critical feedback, and we have fundamentally restructured the paper to address it. The novelty of the revised manuscript is no longer just the dataset itself, but the **entire standardized pipeline** we propose and validate. Our key novel contributions are:
    - **A Unified MT Benchmark:** We present the first systematic comparison of modern English-to-Bangla MT systems across multiple datasets and metrics, providing a clear picture of the current state-of-the-art (**Section 3**).
    - **An Empirically-Validated QE Method:** We conduct a human evaluation study to show that CometKiwi is a significantly better QE metric than the commonly used LaBSE, and we use this finding to propose a more reliable, empirically-grounded filtering threshold (**Section 4**).
      The paper now presents a generalizable methodology for creating high-quality synthetic datasets, using BCoQA as a concrete case study.

5. Suggested Revision: Some important implementation details are missing.

- **Response:** We have made a concerted effort to add all necessary implementation details for full reproducibility. This includes:
    - The exact prompts used for LLM-based answer generation (**Figure 2**) and translation generation (**Table 8**).
    - The specific model versions and metric signatures used in our experiments (**Appendix A**) for reproducibility.
    - Human DA Scoring with complete guidelines and Human Evaluation Graphical Interface is shown in **Figure 4a** & **4b**
    - Detailed statistical analyses and results in new tables throughout the paper (**e.g., Table 4, Table 5, Table 13, Table 14**).

- To ensure full reproducibility of our comparative analysis, we have released the complete translation outputs from all evaluated models, along with their corresponding metric scores.