A Comparative Analysis of English-to-Bangla Machine Translation Systems and Quality Estimation for Low-Resource Data Creation, Applied to Conversational Question Answering

Anonymous ACL submission

Abstract

Creating datasets for low-resource languages like Bangla often involves machine translation and quality estimation (QE) filtering, but the process currently lacks standardization. Different studies use a variety of translation systems and outdated metrics, making it difficult to compare findings. Likewise, the QE filtering step is often applied using methods and thresholds that have not been systematically tested. To address this, our paper first presents a unified evaluation of English-to-Bangla MT systems using both legacy and modern metrics. We then conduct a small scale human evaluation study to compare automated QE scores with human judgments, which helps us determine the best existing QE system and a more systematically grounded threshold for filtering. Using this improved strategy, we introduce BCoQA, a novel Bangla Conversational Question Answering dataset. We are making the BCoQA dataset and our evaluation scripts publicly available. For complete reproducibility of our study, we also release all model outputs and their corresponding metric scores via this link.

1 Introduction

While recent advances in natural language processing (NLP) have yielded dramatic improvements, these gains have been concentrated in high-resource languages. Low-resource languages like Bangla, however, face significant data scarcity. This is often addressed by creating machine-translated datasets, followed by quality estimation (QE) filtering, typically using embedding cosine similarity. However, this approach suffers from several inconsistencies: existing English-to-Bangla machine translation (MT) systems are evaluated on (1) varying datasets, (2) incomparable metrics, and (3) outdated string overlap-based metrics that inadequately compare diverse MT systems.

This paper addresses these limitations by presenting a unified and systematic analysis of current English-to-Bangla MT systems. We evaluated these systems using both legacy overlap-based metrics (BLEU (Papineni et al., 2002), chrF++ (Popovic, 2017) and modern deep learning-based metrics (COMET (Rei et al., 2022), CometKiwi (Rei et al., 2022), BLEURT (Sellam et al., 2020)), which demonstrate a higher correlation with human judgments.

We also address quality filtering for Bangla machine-translated data. The prevalent method uses Language-Agnostic BERT Sentence Embeddings (LaBSE) (Feng et al., 2020) to calculate cosine similarity between source and translated sentences, discarding pairs that fall below a high, yet arbitrary, similarity threshold. We aim to improve this by: (1) conducting a small-scale human evaluation study to compare LaBSE similarity scores and reference-less CometKiwi (Rei et al., 2023) scores against human quality judgments; and (2) using these findings to determine an optimal filtering threshold that better separates high-quality translations from artifacts.

Using these findings, we introduce BCoQA, the first Bangla Conversational Question Answering dataset. BCoQA is created by translating and filtering the established English conversational question-answering datasets CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018), using our optimized MT and filtering pipeline. Finally, we finetune and evaluate several sequence-to-sequence models on BCoQA, to provide a baseline for future research in this area. Our best model achieves an F1 score of 54.1%, which, when contrasted with the human F1 of 78.7%, highlights the significant room for future improvement.

2 Related Works

Several Bangla datasets have been created using machine translation followed by quality filtering. The BNLI dataset for sequence-pair classification and the SQuAD_bn dataset (Bhattacharjee et al., 2021) were generated using an MT system introduced by Hasan et al. and filtered with a LaBSE (Feng et al., 2020) cosine similarity threshold of 0.7. For the BanglaParaPhrase dataset (Akil et al., 2022), the authors used the same MT model (Hasan et al., 2020) followed by a multi-stage filtering process. After an initial LaBSE filter (0.7) on both translations and back-translations, for the final quality check, they used a BERTScore (Zhang et al., 2020) F1-measure, setting a high threshold of 0.92 that was informed by a small-scale human evaluation. These examples highlight the common, yet often ad-hoc, use of embedding-based similarity for filtering.

There has also been some advancements in machine translation for Bangla. The BanglaNLG benchmark (Bhattacharjee et al., 2022) introduced BanglaNMT, a large-scale machine translation dataset, along with BanglaT5, a model pre-trained on a large Bangla corpus. BanglaT5 achieved state-of-the-art results on various Bangla tasks, including machine translation, demonstrating the effectiveness of monolingual pre-training. Translation quality was evaluated using SacreBLEU (Post, 2018). More recently, IndicTrans2 (Gala et al., 2023), a translation model supporting 22 Indic languages (including Bangla), was introduced alongside the IN22-Gen and IN22-Conv evaluation datasets. IndicTrans2 outperformed the much larger NLLB MoE (54B parameters) (Team et al., 2022) on all 22 languages, including Bangla. This comparison used chrF++ (Popovic, 2017), the same metric employed by Team et al. for NLLB benchmarking.

A recent study Mahfuz et al. (2024) compared NLLB (Team et al., 2022), BanglaT5, and several large language models (LLMs) with multilingual support, using the BLEU metric. NLLB 3.3B outperformed Llama 3.1 70B (Grattafiori et al., 2024) and BanglaT5 in English-to-Bangla translation. The use of disparate metrics across these key studies makes direct comparison of model performance challenging and reinforces the need for a unified evaluation.

While existing Bangla question answering datasets like SQuAD_bn (Bhattacharjee et al., 2021) and BanglaRQA (Ekram et al., 2022) provide valuable resources, they are designed for extractive, single-turn QA. These datasets do not support the multi-turn, conversational interactions that are common in human dialogue. Our work introduces BCoQA, a new dataset designed for conversational question answering. Table 1 provides a comparison of these existing datasets and BCoQA, highlighting the unique characteristics of our contribution.

3 Comparative Analysis of English-to-Bangla MT Systems

3.1 Systems

We evaluate the following English-to-Bangla machine translation systems:

- **banglat5_nmt_en_bn (bt5):** The BanglaT5 base model (Bhattacharjee et al., 2022), pre-trained on Bangla text, fine-tuned on the BanglaNMT dataset.
- **indictrans2-en-indic-1B (it2):** A model supporting 22 Indic languages, including Bangla, trained on the Bharat Parallel Corpus Collection (BPCC) (Gala et al., 2023).
- **nllb-200-3.3B (nllb):** A 200-language model trained on the NLLBv1 dataset (Team et al., 2022), which includes the largest known Bangla-English parallel text collection (68M pairs).
- **m2m100_1.2B (m2m100):** A multilingual model focused on non-English-centric translation, supporting 9,900 directions across 100 languages, including English-to-Bangla (Fan et al., 2020).
- **gemma3-12b-it (gemma3):** To assess the capabilities of Large Language Models (LLMs) for English-to-Bangla translation, we included the decoder-only gemma3-12b-it model (Team et al., 2025). To manage computational resources and ensure feasibility within our setup, we focused our LLM exploration on models that could be run in 8-bit GGUF format (ggerganov and contributors, 2023) with a context window of at least 1024 tokens. Based on this criterion and an empirical comparison with Llama3.1-8B (Grattafiori et al., 2024) and Qwen3-14B (Yang et al., 2025), gemma3 demonstrated

Dataset	Conversational	Answer Type	Machine Translated
SQuAD_bn (Bhattacharjee et al., 2021)	✗	Spans, Unanswerable	✓
BanglaRQA (Ekram et al., 2022)	✗	Spans, Yes/No, Unanswerable	✗
Tydi QA (Clark et al., 2020)	✗	Spans, Yes/No	✗
QAmeleon (Agrawal et al., 2023)	✗	Free-form Text	✓
BCoQA (this work)	✓	Free-form text, Unanswerable	✓

Table 1: Comparison of BCoQA with existing Bangla reading comprehension datasets.

consistently more coherent Bangla output and was selected for inclusion. During translation, we experimented with three different prompt formats and selected the one that yielded the best performance for our final comparison, the prompts are shown in appendix table 8

3.2 Evaluation Datasets

Name	Samples	Source
bnmt	1000	Bhattacharjee et al.
flores+	997	Team et al.
in22-test-conv	1503	Gala et al.
in22-test-gen	1024	Gala et al.

Table 2: Overview of the evaluation datasets used to assess English-to-Bangla machine translation performance.

Table 2 provides details on the evaluation datasets used in our analysis.

3.3 Evaluation Metrics

3.3.1 Legacy Metrics

We include the widely used BLEU (Papineni et al., 2002) and chrF++ (Popovic, 2017) metrics, calculated using SacreBLEU (Post, 2018), for historical comparison and common practice. However, we acknowledge their limitations, particularly when comparing diverse MT systems, as surface-level metrics like BLEU, chrF++ are known to be less reliable in such cases (Callison-Burch et al., 2006; Kocmi et al., 2024).

3.3.2 Neural Metrics

We also utilize modern neural metrics, which have shown improved correlation with human judgments. Based on a comprehensive comparative analysis (Kocmi et al., 2024), we select CometKiwi (Rei et al., 2023) and BLEURT (Sellam et al., 2020). CometKiwi, a quality estimation metric, is particularly valuable as it does not require reference translations. BLEURT, a reference-based metric with a different architecture, provides a contrasting perspective. We also include the reference-based COMET (Rei et al., 2022) for a more comprehensive evaluation.

3.4 Results

Analyzing Table 3, we can see IndicTrans2 consistently demonstrates strong performance, achieving the highest scores across most datasets and metrics. On the BanglaNMT dataset, BanglaT5 exhibits slightly higher scores in traditional metrics like BLEU and chrF++, as well as COMET and BLUERT, compared to IndicTrans2. However, IndicTrans2 achieves the highest CometKiwi score on this dataset. This pattern on BanglaNMT might be influenced by BanglaT5's potential training on this specific dataset, which could lead to stylistic similarities that are favored by metrics relying on reference translations.

Gemma 3 shows competitive performance, particularly on the flores+ dataset where it achieves the highest CometKiwi score. While it doesn't consistently lead in raw scores, its performance is generally better than M2M100 and NLLB across most metrics and datasets. M2M100 consistently exhibits the lowest performance across all datasets and metrics.

To validate these observations, we performed paired t-tests with bootstrap resampling (Koehn, 2004) on COMET scores ($p < 0.05$). While BanglaT5 shows a numerical advantage on the BanglaNMT dataset, our tests confirm this lead over IndicTrans2 is not statistically significant (Table 4). Across all datasets, the analysis reveals that IndicTrans2's lead is generally robust. Conversely, Gemma 3 achieves statistically significant gains over BanglaT5 on the flores+ and in22-test-conv datasets, but does not significantly outperform IndicTrans2 on any benchmark (see Appendix Tables 9, 10, 11 for full results).

A qualitative analysis of the translation outputs from Gemma 3 and IndicTrans2 reveals that Gemma 3 generally produces translations with

Table 3: Machine Translation Evaluation Results (All scores are scaled to 0-100 for consistency)

Dataset	Model	BLEU	chrF++	COMET	CometKiwi	BLUERT
bnmt	bt5	**25.1**	**58.5**	**92.96**	82.25	**85.39**
	gemma3	21.2	56.0	92.0	82.37	83.22
	it2	23.2	57.7	92.92	**83.66**	85.21
	m2m100	12.6	45.0	87.65	76.38	77.17
	nllb	20.5	54.5	92.20	81.51	84.26
flores+	bt5	15.1	45.4	85.96	72.22	76.40
	gemma3	13.5	44.3	85.96	**77.13**	76.77
	it2	**21.1**	**52.0**	**87.29**	76.49	**77.43**
	m2m100	11.7	40.9	81.92	66.91	69.49
	nllb	16.4	47.1	86.11	74.46	75.96
in22-test-conv	bt5	15.8	43.8	89.41	79.65	79.69
	gemma3	16.4	45.3	86.73	80.98	80.03
	it2	**16.7**	**46.3**	**89.99**	**82.05**	**80.90**
	m2m100	9.4	35.2	84.89	73.35	72.48
	nllb	15.8	43.8	89.40	79.28	79.45
in22-test-gen	bt5	13.7	43.6	85.20	69.96	76.94
	gemma3	13.2	43.3	85.56	73.37	76.57
	it2	**16.4**	**47.6**	**86.75**	**75.15**	**78.50**
	m2m100	7.3	35.1	79.22	64.84	67.93
	nllb	13.1	43.7	85.27	73.19	76.77

Model	bt5	gemma3	it2	m2m100	nllb
bt5	-	True	False	True	True
gemma3	False	-	False	True	False
it2	False	True	-	True	True
m2m100	False	False	False	-	False
nllb	False	True	False	True	-

Table 4: Pairwise t-test results on BanglaNMT (COMET scores). "True" indicates the row model significantly outperforms the column model ($p < 0.05$).

higher fluency and naturalness. However, in scenarios involving complex translations, IndicTrans2 demonstrates a significant advantage in terms of semantic accuracy. Interestingly, despite a substantial difference in model size (IndicTrans2 1B vs Gemma 3 12B parameters), Gemma 3's capabilities as a general-purpose language model, including its ability to engage in coherent conversations in Bangla and comprehend the language's nuances, indicate a considerable potential for applications in Bangla natural language processing.

4 Quality Estimation for Translation Filtering

4.1 Methods

We compare two methods for filtering machine-translated data: LaBSE (Feng et al., 2020), the most common approach, and CometKiwi (Rei et al., 2023), a quality estimation (QE) system shown to align well with human judgments (Mujadia et al., 2023; Kocmi et al., 2024).

4.2 Human Evaluation

To evaluate these methods, we first translated the CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018) datasets using the IndicTrans2 model. We then scored each translation using both LaBSE cosine similarity and CometKiwi. Following Direct Assessment (DA) guidelines (Graham et al., 2013), we recruited 25 human annotators to rate 20 translations each, totaling 500 translated samples on a scale of 0-100. For a representative evaluation, we sampled translations using a stratified approach. We created bins based on the level of agreement between LaBSE and CometKiwi scores, and sampled from each. This ensured our analysis included not only cases where the models agreed, but also a significant number of translations where their quality assessments diverged.

4.3 Correlation Analysis

Table 5 shows the correlation between human judgments (DA scores) and QE scores (CometKiwi and LaBSE). CometKiwi demonstrates significantly higher correlation with human judgments than LaBSE, with high statistical significance across all measures (Pearson, Spearman, and Kendall). Our qualitative analysis reveals key differences in how LaBSE and CometKiwi handle specific translation challenges. LaBSE often assigns high scores to erratic outputs, including

Metric	Pearson (p-value)	Spearman (p-value)	Kendall (p-value)
CometKiwi	0.467 (< 0.001)	0.440 (< 0.001)	0.304 (< 0.001)
LaBSE	0.138 (0.006)	0.128 (0.010)	0.088 (0.010)

Table 5: Correlation between Human Judgments (DA) and QE Scores (CometKiwi and LaBSE)

those with nonsensical repetitions or mixed English and Bangla scripts, while CometKiwi appropriately penalizes such translations. For instance, the nonsensical translation "কি তার হিয়াটাসeneded" (from "What ended her hiatus") received a human DA score of 28, a CometKiwi score of 0.363, but a LaBSE score of 0.893. Furthermore, LaBSE consistently undervalues accurate transliterations of nouns and abbreviations. IndicTrans2 often correctly transliterates terms like "WBC" to "ডাব্লুবিসি" and "Janko Tipsarevic" to "জ্যাঙ্কো টিপসারেভিচ". While human annotators and CometKiwi rated these transliterations highly, LaBSE frequently assigned scores below 0.5, demonstrating a significant misalignment with human judgment.

Beyond our 500 sample human study, a large-scale analysis across the entire dataset reinforces our conclusions (details in Appendix Table 13 and Table 14). LaBSE scores show significantly higher volatility (i.e., a larger standard deviation). Moreover, the score distributions confirm LaBSE's tendency to overestimate quality: it assigns scores in the highest bracket (0.9-1.0) to 46% of the data, compared to just 14% for CometKiwi. This quantitative evidence aligns with our qualitative findings, where LaBSE overvalued flawed translations while undervaluing correct transliterations.

4.4 Threshold calculation

Although 0.7 is a frequently used LaBSE threshold (Bhattacharjee et al., 2021), prior work shows that the optimal threshold is system and language-dependent (Dakwale et al., 2022). To find an optimal threshold, we analyzed the distribution of human DA scores relative to CometKiwi scores. Based on DA guidelines (Graham et al., 2013), which suggest that good quality translations should be rated above 70, we sought a threshold where at least 80% of the translations were rated >70 by our annotators. However, our initial analysis revealed that the optimal threshold was highly sensitive to the percentage target, with different thresholds yielding vastly different results (e.g. 0.58 for 80%, 0.67 for 85%, and 0.81 for 90%). This sensitivity suggests that our approach may lack robustness, and that the optimal threshold may not be easily determined due to the limitations of our annotated dataset. To make our approach more robust, we decided to combine this with manual inspection. Looking closely at the translations that were accepted or rejected at different thresholds, evaluating both the removal rate and the quality of the rejected translations, we chose a threshold of 0.67. This threshold resulted in the removal of 35% of the data, which, although significant, appeared to be a reasonable trade-off for the improved quality of the remaining data, as it struck a balance between removing low-quality translations and preserving acceptable ones. We acknowledge that this threshold may not be optimal, and that a more robust approach to threshold optimization may be necessary to achieve more reliable results.

5 Creation of BCoQA Dataset

5.1 Task Definition

Conversational Question Answering (CoQA) systems facilitate a natural flow of dialogue by understanding and generating responses that align with the context of a conversation. The goal is to answer the current question in conversation, considering the passage and conversation history. If the answer can't be found, the output should be "অজানা" ("Unknown"). Figure 1 shows how the entity of focus[1] changes throughout the conversation.

5.2 Dataset Creation

We chose two conversational QA datasets, CoQA (Reddy et al., 2018) and QuAC (Choi et al., 2018), which share core features like context passages, multi-turn conversations, unanswerable questions, and evidence spans. However, their differing collection methods result in key distinctions. QuAC provides only answer spans, while CoQA includes both spans and free-form answers. QuAC also features more open-ended questions.

Since QuAC lacks free-form answers, we generated them using the Gemma3 12B LLM (Team

[1] a series of pronouns or noun phrases that refer to the same entity or concept in a conversation or text

Split Name	Data points/Conversations	Yes/No	Unknown	Short	Long (>3 words)
Train	12109	42409	11892	59380	48297
Validation	956	3318	1219	4728	3858
Test	50	221	74	466	234
Total	13115				

Table 6: Dataset split analysis with different answer types

Model	Params	EM	F1
Human	-	71.7	78.7
BanglaT5 (Bhattacharjee et al., 2022)	248M	38.3	54.1
mT5-base (Xue et al., 2020)	582M	35.2	42.7
BanglaT5-small (Bhattacharjee et al., 2022)	60.5M	34.0	44.7
mBART-large-50 (Lewis et al., 2019)	611M	32.6	39.6

Table 7: Performance on BCoQA test set (EM: Exact Match, F1: F1 Score).

et al., 2025) (accessed April 18, 2025), providing the context, question, and answer span as input. Figure 2 shows an example of a generated free-form answer.

We then translated both datasets using the IndicTrans2 model (Gala et al., 2023) and filtered out any conversation containing a sentence with a translation score below our chosen threshold of 0.67 as determined earlier through a combination of quantitative and qualitative analysis. Table 6 details the final dataset structure after filtering.

5.3 Benchmarking Existing Models

We framed conversational question answering (CoQA) as a response generation task, fine-tuning sequence-to-sequence (seq2seq) models on BCoQA. We excluded reading comprehension models like BanglaBERT (Bhattacharjee et al., 2022) as they are unsuitable for free-form answer generation. The input was formatted as: P <q> Q_1 <a> A_1 ... <q> Q_{i-1} <a> A_{i-1} <q> Q_i <a> (P: passage, <q>: question, <a>: answer).

We evaluated using macro-averaged F1 score, consistent with CoQA, removing punctuation and stop words. Human performance (10 participants, 5 conversations each) served as a baseline.

Table 7 shows that BanglaT5 achieves the highest scores (EM: 38.3, F1: 54.1), significantly below human performance. Notably, the the smaller version of BanglaT5, with about 1/10th the number of parameters(60.5M), perform comparably to larger multilingual models like mT5-base, mBART-large-50, which have 582M and 611M parameteres, respectively. This highlights the benefit of native Bangla pretraining. Upon closer examination of the detailed results shown in table 12, we observe that the models perform best on yes/no type questions, which is a expected phenomenon for seq2seq models (Feng et al., 2020). This is because yes/no answers often rely on simple factual information or binary decisions, making it easier for the models to predict the correct response. The banglat5 variants also excel in providing accurate long answers (answers longer than 3 words), indicating that Bangla pretraining is essential for generating long coherent responses.

6 Conclusion

This paper presented a comprehensive analysis of English-to-Bangla machine translation, systematically evaluating state-of-the-art systems with a suite of traditional and modern neural metrics, and identifying IndicTrans2 as the most effective. A key contribution was our small-scale human evaluation, which revealed that CometKiwi, a reference-free quality estimation metric, offers significantly better correlation with human quality judgments compared to the prevalent LaBSE-based cosine similarity. Based on this, we proposed an optimized data filtering approach using a CometKiwi threshold of 0.67. Building directly on these advancements, we introduced BCoQA, the first conversational question answering dataset for Bangla, developed through our refined translation and filtering pipeline. We established a baseline F1 score of 54.1% on BCoQA, which, while a solid starting point, falls considerably short of human performance (78.7% F1), underscoring the critical need for further advancements in low-resource Bangla NLP.

ইন্টেল কর্পোরেশন (ইন্টেল নামেও পরিচিত, ইন্টেল হিসাবে শৈলীকৃত) একটি আমেরিকান বহুজাতিক কর্পোরেশন এবং প্রযুক্তি সংস্থা যার সদর দফতর ক্যালিফোর্নিয়ার সান্তা ক্লারায় অবস্থিত (কথোপকথনে "সিলিকন ভ্যালি" হিসাবে উল্লেখ করা হয়) যা গর্ডন মুর (মুরের আইন খ্যাত) এবং রবার্ট নয়েস দ্বারা প্রতিষ্ঠিত হয়েছিল। এটি স্যামসাং দ্বারা ছাড়িয়ে যাওয়ার পরে রাজস্বের উপর ভিত্তি করে বিশ্বের দ্বিতীয় বৃহত্তম এবং দ্বিতীয় সর্বোচ্চ মূল্যবান অর্ধপরিবাহী চিপ প্রস্তুতকারক এবং বেশিরভাগ ব্যক্তিগত কম্পিউটারে (পিসি) পাওয়া প্রসেসর x86 সিরিজের মাইক্রোপ্রসেসরের উদ্ভাবক। ইন্টেল অ্যাপল, লেনোভো, এইচপি এবং ডেলের মতো কম্পিউটার সিস্টেম নির্মাতাদের জন্য প্রসেসর সরবরাহ করে। ইন্টেল মাদারবোর্ড চিপসেট, নেটওয়ার্ক ইন্টারফেস কন্ট্রোলার এবং ইন্টিগ্রেটেড সার্কিট, ফ্ল্যাশ মেমোরি, গ্রাফিক্স চিপ, এমবেডেড প্রসেসর এবং যোগাযোগ ও কম্পিউটিং সম্পর্কিত অন্যান্য ডিভাইসও তৈরি করে। ইন্টেল কর্পোরেশন 18 জুলাই, 1968 সালে সেমিকন্ডাক্টর অগ্রগামী রবার্ট নয়েস এবং গর্ডন মুর দ্বারা প্রতিষ্ঠিত হয়েছিল এবং অ্যান্ড্রু গ্রোভের নির্বাহী নেতৃত্ব ও দৃষ্টিভঙ্গির সাথে ব্যাপকভাবে যুক্ত ছিল। . . .

Q_1: এই প্রবন্ধের বিষয়বস্তু কী?
A_1: ইন্টেল কর্পোরেশন

Q_2: কোম্পানির সদর দপ্তর কোথায়?
A_2: সান্তা ক্লারা, ক্যালিফোর্নিয়া

Q_3: তারা কি একটি বহুজাতিক কর্পোরেশন?
A_3: হ্যাঁ

Q_4: ইন্টেল কি আবিষ্কার করেছে?
A_4: মাইক্রোপ্রসেসরের x86 সিরিজ, বেশিরভাগ ব্যক্তিগত কম্পিউটারে (পিসি) পাওয়া প্রসেসর।

Q_5: এটি কিসে ব্যবহৃত হয়?
A_5: অধিকাংশ ব্যক্তিগত কম্পিউটারে (পিসি)

Q_6: কোম্পানি টি কখন প্রতিষ্ঠিত হয়েছিল?
A_6: জুলাই ১৮, ১৯৬৮

Q_7: একজন প্রতিষ্ঠাতার নাম বলুন।
A_7: রবার্ট নয়েস

Q_8: আর কেউ?
A_8: গর্ডন মুর

Q_9: তিনি আর কি প্রতিষ্ঠা করেন?
A_9: উত্তর নেই।

Figure 1: A conversation from the BCoQA dataset showing entity of focus in colors. For the original English conversation, please refer to Figure 3.

Limitations

This work has several limitations that warrant consideration and future research: Firstly, our comparative analysis focused on existing, pre-trained English-to-Bangla machine translation (MT) systems. We did not train a new MT model by combining all publicly available datasets. Such a comprehensive training approach would likely yield improved translation performance, representing a valuable avenue for future work.

Secondly, the human evaluation study, while crucial for comparing quality estimation (QE)

Input Prompt:

Context: In 1969, still in the Pre-Crisis continuity, writer Dennis O'Neil and artist Neal Adams return Batman to his darker roots. One part of this effort is writing Robin out of the series by sending Dick Grayson to Hudson University and into a separate strip in the back of Detective Comics. The by-now Teen Wonder appears only sporadically in Batman stories of the 1970s as well as a short lived revival of The Teen Titans. In 1980, Grayson once again takes up the role of leader of the Teen Titans, now featured in the monthly series The New Teen Titans, which became one of DC Comics's most beloved series of the era. During his leadership of the Titans, however, he had a falling out with Batman, leading to an estrangement that would last for many years.
Question: What role did he play in Teen Titans?
Answer Span: In 1980, Grayson once again takes up the role of leader of the Teen Titans,
Answer:

Generated free-form answer:

Leader in Teen Titans.

Figure 2: Example of converting answer spans into free-form answers using LLMs.

methods, was limited in scale. A larger-scale study with more annotators and a broader range of translated samples is necessary for a more definitive analysis of optimal filtering thresholds and a more robust validation of QE metrics against human judgments.

Thirdly, our experiments on the BCoQA dataset focused exclusively on fine-tuning sequence-to-sequence (seq2seq) models. While seq2seq models are a natural fit for conversational response generation, we acknowledge that other architectures, including large language models (LLMs) and models designed for extractive question answering, might achieve superior performance. Exploring these alternative architectures on BCoQA is an important direction for future research. The original CoQA paper (Reddy et al., 2018) notes the limitations of seq2seq models, further motivating this exploration.

Acknowledgments

We would like to thank Reddy et al. and Choi et al. for their datasets which are the basis of BCoQA.

References

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2023. QAmeleon: Multilingual QA with only 5 examples. *Transactions of the Association for Computational Linguistics*, 11:1754–1771.

Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. Banglaparaphrase: A high-quality bangla paraphrase dataset. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 261–272. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2022. Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac : Question answering in context.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Preprint*, arXiv:2003.05002.

Praveen Dakwale, Talaat Khalil, and Brandon Denis. 2022. Empirical evaluation of language agnostic filtering of parallel data for low resource languages. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 346–355, Manila, Philippines. Association for Computational Linguistics.

Syed Mohammed Sartaj Ekram, Adham Arik Rahman, Md. Sajid Altaf, Mohammed Saidul Islam, Mehrab Mustafy Rahman, Md Mezbaur Rahman, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2022. BanglaRQA: A benchmark dataset for under-resourced Bangla language reading comprehension-based question answering with diverse question-answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2518–2532, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Preprint*, arXiv:2305.16307.

ggerganov and contributors. 2023. Inference of meta's llama model (and others) in pure c/c++. https://github.com/ggerganov/ggml. Accessed: April 15, 2025.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone,

Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robin-

son, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Tamzeed Mahfuz, Satak Kumar Dey, Ruwad Naswan, Hasnaen Adil, Khondker Salman Sayeed, and Haz Sameen Shahgir. 2024. Too late to train, too early to use? a study on necessity and viability of low-resource bengali llms. *Preprint*, arXiv:2407.00416.

Vandan Mujadia, Pruthwik Mishra, Arafat Ahsan, and Dipti M. Sharma. 2023. Towards large language model driven reference-less translation evaluation for English and Indian language. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 357–369, Goa University, Goa, India. NLP Association of India (NLPAI).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popovic. 2017. chrf ++ : words helping character n-grams.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge.

Ricardo Rei, José G de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F T Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585. Association for Computational Linguistics.

Ricardo Rei, Nuno M Guerreiro, JosÃľ Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G de Souza, and André Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine

Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Pluciska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Reproducibility

To ensure the reproducibility of our experiments, we detail the specific models and configurations used for evaluation.

- **COMET:** We used the Unbabel/wmt22-comet-da model for COMET evaluation.
- **CometKiwi:**
  - Evaluation: wmt23-cometkiwi-da-xl (updated and scaled-up version).
  - BCoQA Quality Estimation: wmt22-cometkiwi-da (due to feasibility constraints with the XL model on the large BCoQA dataset, and as recommended by the authors for sufficient quality).
  - Quality Filtering: wmt22-cometkiwi-da.
- **SacreBLEU:** We used SacreBLEU for BLEU and chrF++ with the following signatures:

  BLEU: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.5.1`

  chrF++: `nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.5.1`

B Prompt and Data Examples

B.1 English Conversation

Figure 3 shows the English version of Figure 1

Intel Corporation (also known as Intel, stylized as intel) is an American multinational corporation and technology company headquartered in Santa Clara, California. It is the world's second largest and second highest valued semiconductor chip makers based on revenue after being overtaken by Samsung, and is the inventor of the x86 series of microprocessors, the processors found in most personal computers (PCs). Intel supplies processors for computer system manufacturers such as Apple, Lenovo, HP, and Dell. Intel also manufactures motherboard chipsets, network interface controllers and integrated circuits, flash memory, graphics chips, embedded processors and other devices related to communications and computing. Intel Corporation was founded on July 18, 1968, by semiconductor pioneers Robert Noyce and Gordon Moore.
. . .

Q_1: What is the subject of the article?
A_1: Intel Corporation

Q_2: Where is the company's headquarters?
A_2: Santa Clara, California

Q_3: Are they a multinational company?
A_3: Yes.

Q_4: What did Intel invent?
A_4: x86 series of microprocessors

Q_5: Where is it used?
A_5: Most personal computers (PCs)

Q_6: When was the company founded?
A_6: July 18, 1968

Q_7: Name One Founder.
A_7: Robert Noyce

Q_8: And the other?
A_8: Gordon Moore

Q_9: What else did he establish?
A_9: Unknown

Figure 3: A conversation from the BCoQA dataset showing coreference chains in colors - Source of figure 1

B.2 Gemma3 Translation Prompts

Table 8 shows the different prompts tested for Gemma 3 translation generation.

Prompt Structure

`Translate the following English sentence to Bangla:`
`English: {English Sentence}`
`Bangla:`

`Translate the following English text into Bangla. Here is an example:`
`English: Hello, how are you?`
`Bangla: হ্যালো, আপনি কেমন আছেন?`
 	
`Now, translate this:`
`English: {English Sentence}`
`Bangla:`

`You are a professional English to Bangla translator. Translate the following sentence accurately and naturally:` `{English Sentence}`

Table 8: Prompt Formats Experimented with for Gemma 3 Translation

C Detailed Test Results

C.1 Pairwise t-test results for flores+, in22-conv, in22-gen

Tables 9, 10, 11 show the pairwise t-test result for flores+, in22-conv and in22-gen dataset consecutively.

Model	bt5	gemma3	it2	m2m100	nllb
bt5	-	False	False	True	False
gemma3	True	-	False	True	True
it2	True	True	-	True	True
m2m100	False	False	False	-	False
nllb	False	False	False	True	-

Table 9: Pairwise t-test results on flores+ dataset

Model	bt5	gemma3	it2	m2m100	nllb
bt5	-	False	False	True	False
gemma3	True	-	False	True	True
it2	True	False	-	True	True
m2m100	False	False	False	-	False
nllb	False	False	False	True	-

Table 10: Pairwise t-test results on in22-conv test dataset

C.2 Answer Specific Performance of BCoQA Finetuned Models

Table 12 shows the answer type specific results of BCoQA fine tuned models.

Model	bt5	gemma3	it2	m2m100	nllb
bt5	-	False	False	True	False
gemma3	False	-	False	True	False
it2	True	True	-	True	True
m2m100	False	False	False	-	False
nllb	False	False	False	True	-

Table 11: Pairwise t-test results on in22-gen test dataset

Model	Yes/No (EM)	Unknown (EM)	Short (F1)	Long (F1)
bt5	78.5	31.2	58.4	39.2
mT5	77.9	14.5	46.3	26.1
bt5-sm	74.2	17.0	48.1	33.5
mBART	76.0	35.9	42.8	20.5

Table 12: Model scores on BCoQA test set by question-answer type.

C.3 Detailed Comparison of CometKiwi and LaBSE QE

Table 13 and Table 14 shows extensive statistical analysis of CometKiwi and LaBSE Quality estimation scores of unfiltered BCoQA dataset.

Table 13: Descriptive statistics and correlation for QE scores on the unfiltered BCoQA dataset.

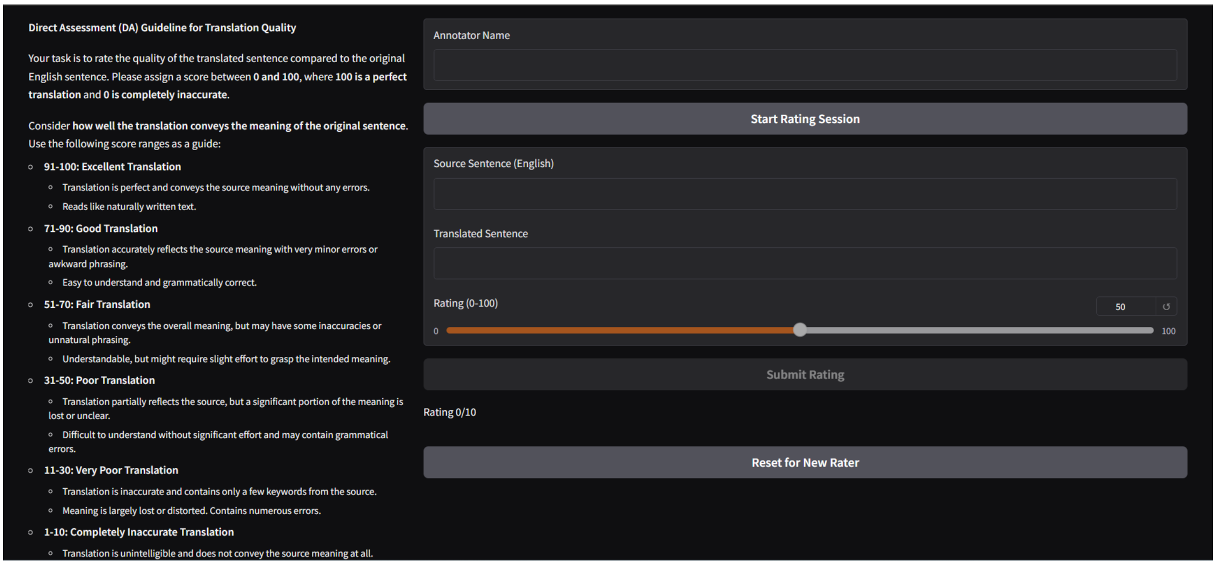
	Train		Validation	
Metric	CometKiwi	LaBSE	CometKiwi	LaBSE
Descriptive Statistics				
Count	705740.0	705740.0	58565.0	58565.0
Mean	0.8628	0.8759	0.8617	0.8721
Std	0.0526	0.1015	0.0538	0.1093
Min	0.1911	-0.3380	0.1479	-0.2987
25%	0.8531	0.8633	0.8516	0.8620
50%	0.8811	0.8962	0.8803	0.8950
75%	0.8945	0.9192	0.8942	0.9180
Max	0.9229	1.0000	0.9226	1.0000
Correlation Analysis				
Pearson	0.4591		0.4839	
Spearman	0.2709		0.2932	
Kendall	0.1859		0.2017	

All correlations are significant (p<0.0001).

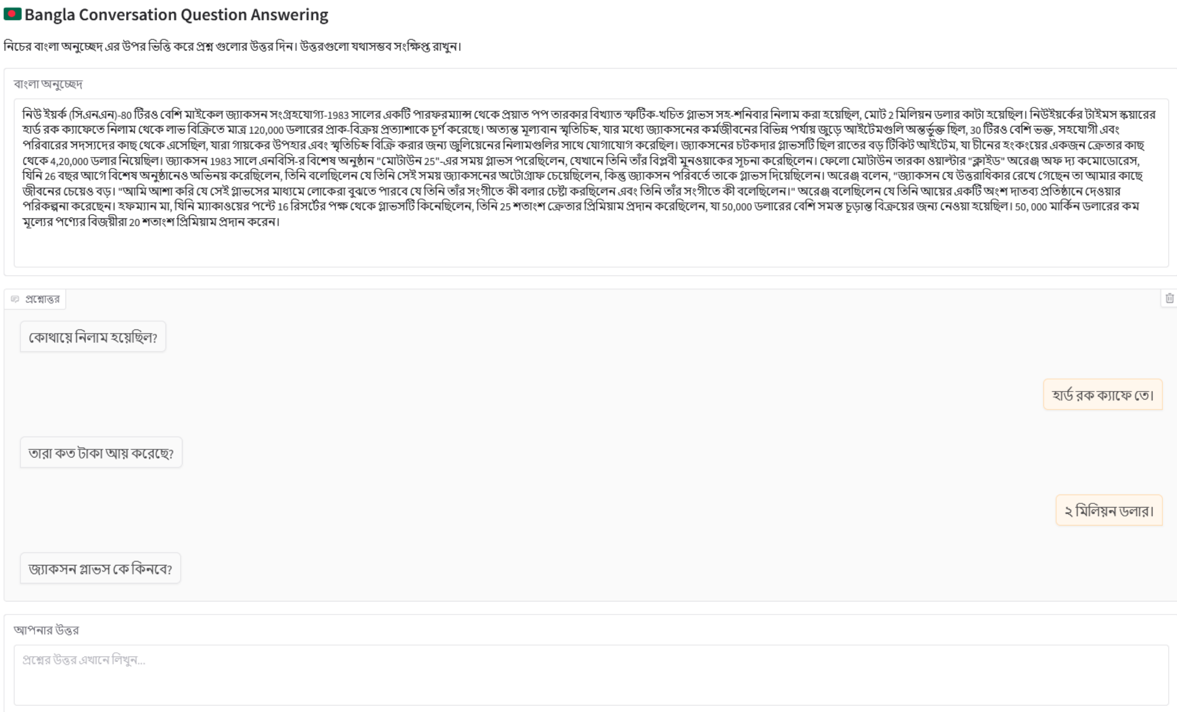
Table 14: Frequency distribution of CometKiwi and LaBSE scores for the training and validation sets.

	Train		Validation	
Score Bin	CometKiwi	LaBSE	CometKiwi	LaBSE
(-0.001, 0.1]	0	18	0	2
(0.1, 0.2]	3	17	2	3
(0.2, 0.3]	38	15196	3	1577
(0.3, 0.4]	105	292	10	20
(0.4, 0.5]	572	514	50	40
(0.5, 0.6]	2403	1325	210	97
(0.6, 0.7]	8895	4455	749	344
(0.7, 0.8]	58776	29476	5142	2311
(0.8, 0.9]	532272	326342	44240	27708
(0.9, 1.0]	102676	326510	8159	26363

D GUI for Human Annotation and Evaluation



(a) UI for Human Direct Assessment Scoring



(b) UI for BCoQA Human Evaluation.

Figure 4: User interfaces developed for human annotation and evaluation tasks.

E Licensing Information

The licenses for the original English data are as follows: QuAC (Choi et al., 2018): The QuAC dataset is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA 4.0). CoQA (Reddy et al., 2018): The CoQA dataset is a compilation of passages from several sources, each with its own license: Literature and Wikipedia passages are licensed under CC BY-SA 4.0. Children's stories from MCTest are licensed under the MSR-LA license. Middle and High school exam passages from RACE are provided under their own specific terms for research use. News passages from the DeepMind CNN/DailyMail dataset are licensed under the Apache License 2.0. Our resulting BCoQA dataset, along with all associated code and evaluation scripts, is made publicly available under the CC BY-NC-SA 4.0 license.

F Human Evaluation Protocol

F.1 Participant Recruitment

A total of 35 participants were recruited on a voluntary basis from undergraduate courses from an university Computer Science and Engineering department. All annotators are native speakers from Bangladesh where Bangla is the primary language, providing the necessary cultural context to judge translation naturalness. The participants were all in the 20-24 age range.

F.2 Task Interface and Procedure

The evaluation was conducted using custom user interface developed with Gradio. The procedure was as follows:

1. Each of the 35 participants was assigned a unique, anonymous ID for tracking purposes.
2. For the translation quality task, 25 participants were presented with a source English sentence and its corresponding machine-translated Bangla output.
3. Following the Direct Assessment (DA) methodology (Graham et al., 2013), they were instructed to rate the quality of the translation on a continuous scale from 0 to 100. Figure 4a shows the UI for this task.
4. For the conversational QA task, 10 participants were tasked with providing baseline human answers. Each was randomly assigned five conversations from the test set. Figure 4b shows the Gradio interface created for this task.

G Computational Infrastructure

All experiments, including model inference, quality estimation, and fine-tuning, were conducted on a single workstation with the following specifications:

- **GPU:** NVIDIA GeForce RTX 4090 with 24 GB of VRAM
- **CPU:** Intel Core i9-9900K
- **RAM:** 128 GB DDR4

H Finetuning Setup

We finetuned our models on the BCoQA dataset using the Seq2SeqTrainer from the Huggingface transformers library. The finetuning setup consisted of:

- 2 epochs of training
- Learning rate of 4e-5
- Maximum sequence length of 1024
- Adafactor optimizer for BanglaT5, BanglaT5 Small, and MT5. AdamW optimizer for MBART.
- Batch size between 4-8 depending on the model size to maximize throughput.