Understanding QA generation: Extracting Parametric and Contextual Knowledge with CQA for Low Resource Bangla Language

Anonymous ACL submission

Abstract

Question-Answering (QA) models for low-resource languages like Bangla face challenges due to limited annotated data and linguistic complexity. A key issue is determining whether models rely more on pre-encoded (parametric) knowledge or contextual input during answer generation, as existing Bangla QA datasets lack the structure required for such analysis. We introduce BanglaCQA, the first Counterfactual QA dataset in Bangla, by extending a Bangla dataset while integrating counterfactual passages and answerability annotations. In addition, we propose fine-tuned pipelines for encoder-decoder language-specific and multilingual baseline models, and prompting-based pipelines for decoder-only LLMs to disentangle parametric and contextual knowledge in both factual and counterfactual scenarios. Furthermore, we apply LLM-based and human evaluation techniques that measure answer quality based on semantic similarity. We also present a detailed analysis of how models perform across different QA settings in low-resource languages, and show that Chain-of-Thought (CoT) prompting reveals a uniquely effective mechanism for extracting parametric knowledge in counterfactual scenarios, particularly in decoder-only LLMs. Our work not only introduces a novel framework for analyzing knowledge sources in Bangla QA but also uncovers critical findings that open up broader directions for counterfactual reasoning in low-resource language settings.

1 Introduction and Related Work

The domain of Question Answering (QA) is a fundamental area within Natural Language Processing, which aims to train models that emulate human reasoning by mimicking human comprehension and response generation. With the arrival of transformer-based models, this emulation has reached new heights for high-resource languages, specifically for Large Language Models (LLMs),

Factual Context:

Question: এমা গোল্ডম্যান কে ছিলেন? (Who was Emma Goldman?)
Context: এমা গোল্ডম্যান (ইংরেজি:Emma Goldman; জুন ২৭, ১৮৬৯ - মে ১৪, ১৯৪০) **একজন নৈরাজ্যবাদী রাশিয়ান** লেখক যিনি লেখা, বক্তৃতা এবং রাজনৈতিক সক্রিয়তার মাধ্যমে পরিচিত ছিলেন.... (Emma Goldman (English: Emma Goldman; June 27, 1869 – May 14, 1940) was **a Russian anarchist writer** who was known for her writings, speeches, and political activism...)
Parametric Answer: একজন নৈরাজ্যবাদী রাশিয়ান লেখক (a Russian anarchist writer)
Contextual Answer: একজন নৈরাজ্যবাদী রাশিয়ান লেখক (a Russian anarchist writer)

Counterfactual Context:

Question: এমা গোল্ডম্যান কে ছিলেন? (Who was Emma Goldman?)
Context: এমা গোল্ডম্যান (ইংরেজি:Emma Goldman; জুন ২৭, ১৮৬৯ - মে ১৪, ১৯৪০) **একজন নৈরাজ্যবাদী মেক্সিকান** লেখক যিনি লেখা, বক্তৃতা এবং রাজনৈতিক সক্রিয়তার মাধ্যমে পরিচিত ছিলেন.... (Emma Goldman (English: Emma Goldman; June 27, 1869 – May 14, 1940) was **a Mexican anarchist writer** who was known for her writings, speeches, and political activism...)
Parametric Answer: একজন নৈরাজ্যবাদী রাশিয়ান লেখক (a Russian anarchist writer)
Contextual Answer: একজন নৈরাজ্যবাদী মেক্সিকান লেখক (a Mexican anarchist writer)

Figure 1: Parametric vs Contextual Question Answering (QA) in Factual and Counterfactual Settings

as these models demonstrate competitive performance based solely on their pre-encoded knowledge. However, challenges arise in generating accurate responses in contextual QA settings, particularly in counterfactual contexts, due to the interplay of two distinct "knowledge sources": (i) Parametric knowledge, embedded within model parameters through pretraining and (ii) Contextual knowledge, derived from input contexts at execution time (Neeman et al., 2023). Previous work in English QA models has shown that prioritization of parametric knowledge can lead to the generation of hallucinated answers, which occurs because of the imbalance between extensive pre-encoded data and limited contextual input (Krishna et al., 2021). Some work further shows that contextual questions that contain incorrect assumptions disrupt generation performance (Kim et al., 2021). Although some studies show that integrating counterfactual or random contexts into factual datasets improves robustness by disentangling knowledge sources (Hwang et al., 2023), such methods remain largely unexplored for Bangla, a widely spoken yet under-resourced language. Although models evaluated on BanglaRQA (Ekram et al., 2022) and Squad-BN (Bhattacharjee et al., 2022) achieve strong factual QA scores, key challenges remain unsolved: the absence of benchmarks for evaluating parametric and contextual biases as distinct factors, limited insight into counterfactual contexts and unclear methods for tracing knowledge sources.

To address these issues, we present the first Bangla Counterfactual Question-Answering dataset, BanglaCQA, by extending an existing BanglaRQA (Ekram et al., 2022) dataset with answerability, random and counterfactual contexts to analyze the internal or contextual knowledge prioritization. Moreover, we introduce disentanglement pipelines by leveraging multiple encoder-decoder models (BanglaT5-small (Bhattacharjee et al., 2023), BanglaT5-base (Bhattacharjee et al., 2023), mt5 (Xue et al., 2020)) with fine-tuning and decoder-only open-sourced LLMs (LLaMA-3.3-72B (Touvron et al., 2023), DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025), Qwen2.5-32B (Yang et al., 2024), Mistral-3-small (Mistral AI, 2025)) with few-shot (Brown et al., 2020) and Chain-of-Thought (CoT) (Wei et al., 2022) prompting to differentiate parametric and contextual reasoning. To evaluate the results, we use Gemini-2.0-Flash (Hassabis et al., 2024) and GPT-4.1 (OpenAI et al., 2024) for semantic similarity scoring, which outperforms traditional metrics to evaluate the semantic accuracy of Bangla QA responses. Moreover, we applied human evaluation for both the dataset and model's generated answer to maintain accuracy and transparency. Our analysis reveals that integrating counterfactual contexts exhibits strong performance in multiple segments. These findings not only establish a blueprint for low-resource languages and advanced QA systems for Bangla, but also emphasize transparency in knowledge utilization in counterfactual scenarios.

2 BanglaCQA Dataset

We introduce BanglaCQA, the first Bengali QA dataset designed to disentangle parametric and contextual knowledge in language models. For this, we expand the existing BanglaRQA (Ekram et al., 2022) dataset by adding 6.3K counterfactual contexts, an increase of **42.28%** specifically crafted to challenge models on whether they rely on context or fall back on memorized information.[1]

2.1 Counterfactual Context Generation

Counterfactual contexts are derived from factual examples by modifying key named entities using an automated NER pipeline (Sarker, 2020). The script identifies standard named entity types, such as PER (person), LOC (location), ORG (organization), GPE (geo-political entity), DATE or NUM (temporal and numeric expressions) and applies type-consistent substitutions. For example, person names are replaced with other plausible names, locations with alternative locations and organizations with different entities of the same category while ensuring semantic coherence. When named entities appear in both the context and answer fields, replacements are applied consistently. For temporal expressions, if the entity represents a year, only the final digit is altered to preserve plausibility while introducing subtle factual contradictions. In other numerical cases, values are substituted using regular expressions. Each modified row is assigned a unique ID to prevent duplication. These controlled modifications construct hypothetical contradictions while retaining the original sentence structure and allow us to test whether models truly ground their answers in the input context or default to memorized (parametric) knowledge.

Dataset Attribute	**Setting**
Total QA pairs	21,211
Factual Contexts	14,900
Counterfactual Contexts	6,303
Average Question Word Count	8.26
Average Context Word Count	215.27

Table 1: Summary statistics of the BanglaCQA dataset. These statistics highlight the dataset's scale and the relative complexity of its contexts.

2.2 Annotation Quality Assurance

After generating counterfactual passages using the NER script, examples were reviewed by two of the authors of the paper separately. Moreover, to ensure objectivity, two *independent paid annotators*, who were not involved in the construction of counterfactual dataset, further reviewed the dataset for semantic correctness. Disagreements were resolved by consensus and the process yielded a **Cohen's Kappa score of 0.73** which indicates substantial inter-annotator agreement. Additionally, factual rows that were labeled answerable despite lacking valid contextual answers were removed to reduce label noise and enhance overall quality. Further details on annotator roles, requirements are included in the Appendix A.2.

[1] https://anonymous.4open.science/r/banglacqa/

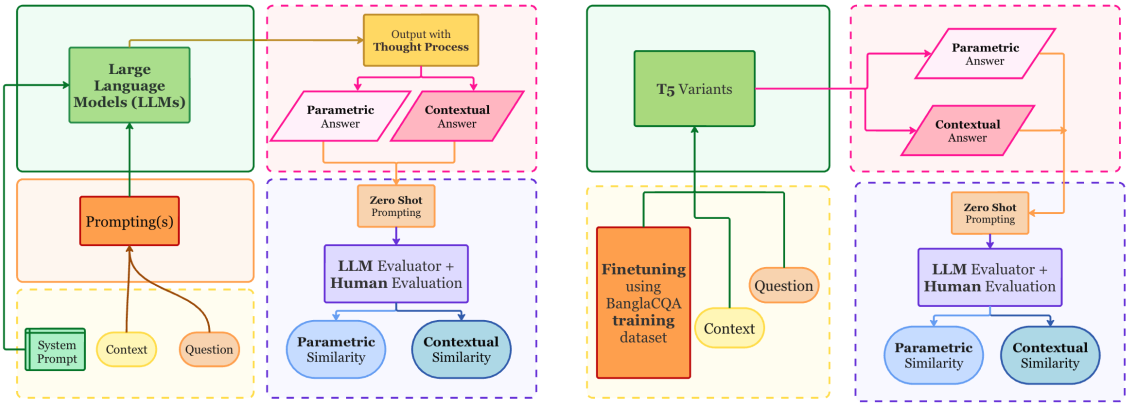


Figure 2: Evaluation pipeline for disentangling parametric and contextual knowledge in QA. **Left:** Prompt-based inference using large language models (LLMs) to generate both parametric and contextual answers. **Right:** Fine-tuning-based evaluation using T5 variants finetuned on BanglaCQA. Both paradigms are evaluated via automated LLM-based and human evaluations to measure answer similarity with respect to both knowledge types.

3 Implementation Pipeline

To identify the most effective model architecture for BanglaCQA, we fine-tuned multiple variants of the T5 (Raffel et al., 2019) framework, namely BanglaT5 (small, base) and mT5, under two configurations: Factual + Answerability (F+A) and Factual + Counterfactual + Answerability (F+CF+A). This dual-configuration strategy enables a focused comparison of how language-specific and multilingual models adapt when exposed to both factual and counterfactual contexts. As shown in Figure 2, each model was trained using a consistent pipeline that emphasizes reproducibility and transparency. Data pre-processing included systematic tokenization and formatting, followed by splitting into training and validation subsets to ensure unbiased evaluation. We adopted a standardized set of hyperparameters: 30 epochs, batch size of 2, learning rate of 5e-5 and 10 warmup steps across all experiments. Early stopping was employed based on validation loss to mitigate overfitting; most models converged by the 15th epoch, optimizing both performance and training efficiency. BanglaT5 models were sourced from the Hugging Face repository of CSEBUET NLP group, while mT5 was obtained from Google's official collection, ensuring credible and community-recognized model baselines.

Model Type	Training Hardware	GPU VRAM
Encoder-decoder	Nvidia RTX 4090 GPU	24GB
Decoder-only	4x Nvidia L4 GPUs	90GB

Table 2: Training hardware and GPU VRAM used for models.

For decoder-only LLMs, we developed a unified inference framework to probe parametric vs. contextual reasoning using few-shot and Chain-of-Thought (CoT) prompting. Each prompt combined a factual or counterfactual context with instruction and a question, structured to stimulate reasoning patterns aligned with the internal knowledge of the model and the external input. To ensure consistency of the evaluation, all models were decoded using the same hyperparameters: temperature = 0.1, top-p = 0.1, repetition penalty = 1.02 and maximum tokens = 1500. We deployed Qwen-2.5 (32B), DeepSeek-R1 (32B), Mistral-3 Small (24B) and LLaMA-3.3 (70B). Due to resource constraints, the LLaMA-3.3 model was quantized using FP16 precision. Crucially, each model produced two separate outputs: one reflecting internal knowledge (parametric) and the other derived from context (contextual). Any non-Bangla output was automatically normalized into Bangla using the Gemini API, enabling cross-lingual evaluation without bias. Semantic alignment was assessed in a zero-shot setting using Gemini 2.0 Flash and GPT-4.1, which we found to be more reliable for Bangla than traditional metrics. We report parametric similarity and contextual similarity separately, offering fine-grained insights into how models interpret and reason across both factual and counterfactual contexts. All encoder-decoder and decoder-only model experiments training hardware and GPU VRAM configurations are shown in Table 2.

Models	Trained on	F Contextual Similarity	F Parametric Similarity	CF Contextual Similarity	CF Parametric Similarity
BanglaT5 Small	F+A	0.77	0.70	0.69	0.11
BanglaT5 Base	F+A	0.82	0.81	0.72	0.13
mT5 Small	F+A	0.84	0.79	0.79	0.09
BanglaT5 Small	F+CF+A	0.83	0.72	0.83	0.19
BanglaT5 Base	F+CF+A	0.86	**0.84**	**0.87**	**0.23**
mT5 Small	F+CF+A	**0.87**	0.81	0.84	0.15

Table 3: Performance of different models under Factual (F) and Counterfactual (CF) settings, evaluated with parametric and contextual similarity using **Gemini-2.0 Flash as an evaluator**. Here, "A" stands for answerability denotes whether the model can generate a grounded response based on the provided context. All reported scores are mean values. "F" denotes **Factual** contexts and "CF" denotes **Counterfactual** contexts.

Models	Trained on	F Contextual Similarity	F Parametric Similarity	CF Contextual Similarity	CF Parametric Similarity
BanglaT5 Small	F+A	0.79	0.74	0.76	0.16
BanglaT5 Base	F+A	0.83	0.80	0.75	0.14
mT5 Small	F+A	0.84	0.79	0.79	0.13
BanglaT5 Small	F+CF+A	0.85	0.79	0.79	0.21
BanglaT5 Base	F+CF+A	0.87	**0.82**	0.84	**0.27**
mT5 Small	F+CF+A	**0.88**	0.80	**0.88**	0.20

Table 4: Performance of different models under Factual (F) and Counterfactual (CF) settings, evaluated with parametric and contextual similarity **using GPT-4.1 as an evaluator**. Here, "A" stands for answerability denotes whether the model can generate a grounded response based on the provided context. All reported scores are mean values."F" denotes **Factual** contexts and "CF" denotes **Counterfactual** contexts.

4 Results

We evaluated the performance of the models in both factual and counterfactual contexts by computing the mean semantic similarity score between generated outputs and target answers. Similarity scores (ranging from 0 to 1) were calculated using Gemini 2.0 Flash and GPT-4.1, which provide more reliable assessments for Bangla text than traditional metrics. For encoder-decoder models, we observed how fine-tuning with counterfactual data influenced performance by comparing the two training configurations. Decoder-only models, evaluated under few-shot and Chain-of-Thought prompting, demonstrated distinct reasoning behaviors reflected in their parametric and contextual outputs. To capture these differences, we separately analyzed parametric responses, which reflect the internal knowledge of the model and contextual responses, which rely on the provided input. This dual evaluation reveals how different architectures and training strategies leverage internal and external information when handling factual and counterfactual queries, offering fine-grained insights into model reasoning and adaptability. We present our findings by discussing the following research questions:

RQ1: What factors contribute to the underperformance of Bangla encoder-decoder models in parametric answer generation in counterfactual contexts, and how can decoder-only LLMs mitigate these challenges? We observe a notable decline in mean parametric similarity scores for counterfactual contexts compared to factual ones across all evaluated encoder-decoder T5 variant models. For instance, using Gemini-2.0-Flash as the evaluator (Table 3), BanglaT5 Small drops from 0.70 (F Parametric) to 0.11 (CF Parametric), while BanglaT5 Base declines from 0.83 to 0.14, both in (F+A) settings clearly illustrating the model's difficulty in generalizing to counterfactual knowledge. The reason is that these models are fine-tuned only on Factual+ Answerability settings, and so their lack of understanding of counterfactual scenarios resulted in such manner. Fine-tuning on both factual and counterfactual data (F+CF+A) improves contextual scores, as seen in BanglaT5 Base rising to 0.86 (F Contextual) and 0.87 (CF Contextual), but this does not sufficiently enhance parametric similarity in CF settings (0.23), reinforcing that fine-tuning aids context understanding more than guides the models to understand the parametric

Models	Prompting	F Contextual Similarity	F Parametric Similarity	CF Contextual Similarity	CF Parametric Similarity
Qwen-2.5	Few-Shot	0.88	0.35	0.79	0.27
DeepSeek-R1	Few-Shot	0.88	0.32	0.81	0.31
LLAMA-3.3	Few-shot	0.84	0.27	0.77	0.24
Mistral-3-small	Few-shot	0.85	0.34	0.79	0.25
Qwen-2.5	COT	0.92	**0.81**	0.86	**0.74**
DeepSeek-R1	COT	**0.94**	0.79	**0.89**	0.70
LLAMA-3.3	COT	0.91	0.69	0.83	0.55
Mistral-3-small	COT	0.90	0.74	0.86	0.64

Table 5: Performance of different decoder-only LLMs under Factual (F) and Counterfactual (CF) settings, evaluated with parametric and contextual similarity **using Gemini-2.0 Flash as an evaluator**. All reported scores are mean values. "F" denotes **Factual** contexts and "CF" denotes **Counterfactual** contexts.

Models	Prompting	F Contextual Similarity	F Parametric Similarity	CF Contextual Similarity	CF Parametric Similarity
Qwen-2.5	Few-Shot	0.89	0.39	0.78	0.31
DeepSeek-R1	Few-Shot	0.83	0.36	0.79	0.30
LLAMA-3.3	Few-shot	0.86	0.29	0.75	0.27
Mistral-3-small	Few-shot	0.87	0.37	0.81	0.26
Qwen-2.5	COT	0.93	**0.84**	**0.88**	**0.78**
DeepSeek-R1	COT	**0.95**	0.81	0.91	0.68
LLAMA-3.3	COT	0.90	0.70	0.84	0.59
Mistral-3-small	COT	0.91	0.73	0.85	0.63

Table 6: Performance of different decoder-only LLMs under Factual (F) and Counterfactual (CF) settings, evaluated with parametric and contextual similarity **using GPT-4.1 as an evaluator**. All reported scores are mean values. "F" denotes **Factual** contexts and "CF" denotes **Counterfactual** contexts.

knowledge. For this reason, when required to produce parametric answers relying on internal knowledge, models tend to hallucinate or conflate contextual cues with facts. In contrast, decoder-only large language models (LLMs), utilize prompting to access a broader and more comprehensive pre-encoded knowledge base. As these models are not fine-tuned, but prompted to complete their tasks, it enables LLMs to better generate accurate parametric answers, particularly in counterfactual contexts. These results highlight a fundamental limitation of Bangla encoder-decoder models: despite fine-tuning improvements in contextual extraction, their constrained internal knowledge restricts generalization to counterfactual reasoning, a gap partially addressed by decoder-only LLMs extensive pre-encoded knowledge.

RQ2: Why does the prompting strategy (CoT vs. Few-shot) affect the parametric and contextual performance of language models in Bangla across factual and counterfactual settings? Our results in Tables 5 and 6 demonstrate that Chain-of-Thought (CoT) prompting leads to statistically significant and practically large improvements in parametric similarity for both factual **(+0.42–>0.44)** and counterfactual **(+0.38–>0.39)** settings. Paired t-tests confirm these gains ($p < 0.01$) with extremely large effect sizes (Cohen's $d > 5$), establishing that the improvements are not due to chance but are practically meaningful (see Table 7). Few-shot prompting inherently lacks an intermediate reasoning phase: models directly predict an answer without explicitly reasoning through the problem. As a result, in counterfactual settings, few-shot models fail to verify the plausibility of the context and default to answers derived from the modified passages, leading to poor parametric similarity. In contrast, CoT prompts explicitly instruct the models to first generate a detailed reasoning chain before producing the final answer (Wei et al., 2022). This structured reasoning step enables the models to differentiate between information derived from the counterfactual context and their encoded parametric knowledge.

These findings align with recent theoretical work showing that transformers without intermediate

Model	Metric	Mean Δ(COT – Few)	t-value	p-value	Cohen's d
Gemini-2.0	F Parametric	+0.44	26.48	0.00012	13.24
Gemini-2.0	CF Parametric	+0.39	11.94	0.00126	5.97
GPT-4.1	F Parametric	+0.42	19.55	0.00029	9.77
GPT-4.1	CF Parametric	+0.38	12.33	0.00115	6.16

Table 7: Parametric similarity evaluation of decoder-only LLMs under **Factual (F)** and **Counterfactual (CF)** contexts, using **Gemini-2.0 and GPT-4.1** as the evaluator. All scores reflect mean differences between Chain-of-Thought (COT) and Few-shot prompting. Positive Δ values indicate improved performance under COT prompting. Statistical significance is shown via t-tests and effect size (d).

reasoning steps are restricted to low-complexity function classes (e.g., AC^0/TC^0 A.1) and fail to solve inherently sequential problems unless their depth or size scales super-polynomially (Peng et al., 2024). By generating intermediate reasoning steps, CoT effectively increases the model's computational depth, allowing it to simulate larger circuits and solve tasks such as arithmetic evaluation and dynamic programming that are otherwise inexpressible for bounded-depth transformers. Recent findings also reveal that CoT benefits arise not only from correct intermediate reasoning but also from structural inductive bias: models achieve up to 90% of CoT gains even with imperfect reasoning if the steps are structurally relevant and correctly ordered (Jin et al., 2024). Furthermore, CoT provides a mechanism for latent state tracking, where each reasoning step encodes an intermediate computation that can be referenced in subsequent steps (Xu et al., 2025). These theoretical insights explain the dramatic gains observed in our results. Bangla question answering requires reasoning over morphologically rich, long contexts (average length = 215 tokens; see Table 1) and counterfactual entity substitutions. Few-shot prompting fails to guide models toward structured inference, resulting in low parametric similarity. CoT enforces a universal reasoning template that bridges the gap caused by the lack of Bangla-specific reasoning supervision during pre-training. Decoder-only models (e.g., Qwen-2.5, DeepSeek-R1) particularly benefit because their training has exposed them to CoT-like reasoning formats. As a result, CoT increases parametric similarity in both factual and counterfactual settings, validating that the gains are statistically significant and theoretically grounded in the expanded expressivity and state-tracking capabilities of CoT-augmented transformers.

RQ3: How do architectural differences among language models affect their ability to integrate contextual and parametric knowledge across factual and counterfactual tasks in Bangla? Qwen-2.5 achieves high similarity scores across both dimensions **(F parametric : 0.81, CF parametric : 0.74; F contextual: 0.92, CF contextual: 0.86)**. This is likely aided by its design for handling long-sequences processing, which aligns well with Bangla's complex and fragmented tokenization. DeepSeek-R1 shows similar improved performance. However, LLAMA-3.3 exhibits a steep decline in CF contextual similarity (0.55) despite a strong factual similarity score (0.91). These findings suggest that architectures optimized for longer contexts are better suited for Bangla's linguistic structure. Details of prompts are shown on Appendix A.3.

4.1 Error Analysis through Human Evaluation

Although Gemini-2.0-Flash and ChatGPT-4.1 provide a scalable and efficient approximation of parametric answer similarity, they exhibit notable limitations in counterfactual QA for Bangla. To assess metric reliability and analyze potential sources of error introduced by the dataset or evaluation metric, we applied human evaluation. Two independent annotators, who were not involved in the dataset creation process, were tasked with evaluating a random subset of 200 model-generated answers. Comparing these human judgments, widely regarded as the gold standard in QA (Clark et al., 2021), with model outputs revealed some discrepancies:

I) Temporal Mismatch (Outdated Targets): Figure 3 presents a counterfactual context, where the numeric value was automatically modified using a Python script and regular expressions as part of the dataset generation pipeline. However, in this instance, the dataset's Target Parametric Answer 5.5 million is factually outdated or incorrect. Despite being given a counterfactual input, the model (Qwen-2.5) successfully generates the correct para-

Dataset's Target Answer (Counterfactual Context):

Question: টয়োটা মোটর কর্পোরেশনের কারখানাগুলি থেকে বছরে গড়ে কত গাড়ি তৈরি হয়? (On average, how many cars are produced annually from Toyota Motor Corporation's factories?)
Context: বর্তমানে জাপানে টয়োটা মোটর কর্পরেশনের নিজস্ব ১২টি কারখানা, ১১টি সাবসিডিয়ারি অ্যাফিলিয়েট কারখানা ছাড়াও বিশ্বের ২৬টি দেশে মোট ৫১টি কারখানা রয়েছে। এগুলোতে গড়ে প্রতি বছর ৯০ লাখ গাড়ি তৈরি হয়....(Currently, Toyota Motor Corporation has 12 of its own factories in Japan, in addition to 11 subsidiary and affiliate factories. Worldwide, it has a total of 51 factories in 26 countries. These factories produce an average of 9 million vehicles per year...)
Target Parametric Answer: **৫৫ লাখ (55 lakh) -> Given on Dataset**
Target Contextual Answer: ৯০ লাখ (9 million) -> Counterfactual Contexts value

Qwen-2.5 Generated Answer (Counterfactual Context):

Question: টয়োটা মোটর কর্পোরেশনের কারখানাগুলি থেকে বছরে গড়ে কত গাড়ি তৈরি হয়?(On average, how many cars are produced annually from Toyota Motor Corporation's factories?)
Context: বর্তমানে জাপানে টয়োটা মোটর কর্পরেশনের নিজস্ব ১২টি কারখানা, ১১টি সাবসিডিয়ারি অ্যাফিলিয়েট কারখানা ছাড়াও বিশ্বের ২৬টি দেশে মোট ৫১টি কারখানা রয়েছে। এগুলোতে গড়ে প্রতি বছর ৯০ লাখ গাড়ি তৈরি হয়....(Currently, Toyota Motor Corporation has 12 of its own factories in Japan, in addition to 11 subsidiary and affiliate factories. Worldwide, it has a total of 51 factories in 26 countries. These factories produce an average of 9 million vehicles per year...)
Generated Parametric Answer: টয়োটা মোটর কর্পোরেশন প্রায় **১০ মিলিয়ন** বা তার থেকে বেশি গাড়ি উৎপাদন করে (Toyota Motor Corporation produces approximately 10 million or more vehicles.) -> **Correct answer generated by LLM**
Extracted Contextual Answer: বছরে প্রায় ৯০ লাখ(৯ মিলিয়ন) গাড়ি তৈরি হয় (Approximately 9 million vehicles are produced per year.) -> Counterfactual Contexts value

Figure 3: Example of temporal mismatch where a model-generated answer is penalized for being more up-to-date than the reference

Dataset's Target Answer (Factual Context):

Question: পৃথিবী নিজ অক্ষের কত ডিগ্রি কোণে হেলে রয়েছে? (At what degree is the Earth tilted on its axis?)
Context: মহাবিশ্বের অন্যান্য বস্তুর সঙ্গে পৃথিবীর সম্পর্ক বিদ্যমান। বিশেষ করে সূর্য ও চাঁদের সঙ্গে এই গ্রহের বিশেষ সম্পর্ক রয়েছে।বর্তমানে নিজ কক্ষপথে মোটামুটি ৩৬৫.২৬ সৌরদিনে বা এক নক্ষত্র বর্ষে সূর্যকে প্রদক্ষিণ করে। পৃথিবী নিজ অক্ষের ৬৬.১/২ ডিগ্রি কোণে হেলে রয়েছে। এর ফলে ....(There is a relationship between Earth and other celestial bodies in the universe. In particular, the Earth has a special relationship with the Sun and the Moon. At present, it revolves around the Sun in approximately 365.26 solar days or one sidereal year. The Earth is tilted at an angle of 66.5 degrees on its axis. As a result...)
Target Parametric Answer: **৬৬.১/২ ডিগ্রি (66.5 degrees)**
Target Contextual Answer: ৬৬.১/২ ডিগ্রি (66.5 degrees)

Qwen-2.5 Generated Answer (Factual Context):

Question: পৃথিবী নিজ অক্ষের কত ডিগ্রি কোণে হেলে রয়েছে? (At what degree is the Earth tilted on its axis?)
Context: মহাবিশ্বের অন্যান্য বস্তুর সঙ্গে পৃথিবীর সম্পর্ক বিদ্যমান। বিশেষ করে সূর্য ও চাঁদের সঙ্গে এই গ্রহের বিশেষ সম্পর্ক রয়েছে।বর্তমানে নিজ কক্ষপথে মোটামুটি ৩৬৫.২৬ সৌরদিনে বা এক নক্ষত্র বর্ষে সূর্যকে প্রদক্ষিণ করে। পৃথিবী নিজ অক্ষের ৬৬.১/২ ডিগ্রি কোণে হেলে রয়েছে। এর ফলে ....(There is a relationship between Earth and other celestial bodies in the universe. In particular, the Earth has a special relationship with the Sun and the Moon. At present, it revolves around the Sun in approximately 365.26 solar days or one sidereal year. The Earth is tilted at an angle of 66.5 degrees on its axis. As a result...)
Generated Parametric Answer: পৃথিবী নিজ অক্ষের প্রায় **২৩.৫ ডিগ্রি** কোণে হেলে রয়েছে।(The Earth is tilted at an angle of approximately 23.5 degrees on its axis.)
Extracted Contextual Answer: পৃথিবী নিজ অক্ষের প্রায় ৬৬.১/২ ডিগ্রি কোণে হেলে রয়েছে। (The Earth is tilted at an angle of approximately 66.5 degrees on its axis.)

Figure 4: Example showing multiple valid answers due to variations in model interpretation and reference grounding.

metric answer: approximately 10 million or more. Due to the dataset's reliance on fixed parametric targets, this correct response is unjustly penalized in automated evaluations. Approximately 4% of the generations were found to be factually superior to the dataset references, particularly in temporally sensitive questions such as population figures or political terms. While this percentage may vary across other subsets, the findings underline a key limitation: static parametric references can fail to reward accurate model behavior, especially when LLMs draw upon up-to-date parametric knowledge.

II) Solution Variation (Multiple Valid Answers): Figure 4 illustrates a case where the model predicts 23.5°, while the dataset target is 66.5°. Both values are scientifically correct as they represent complementary angles of the Earth's axial tilt. However, since the evaluator models compare each answer against a single reference from the dataset's answer, they tend to assign a lower score due to the absence of lexical or numerical overlap. Importantly, in such cases there is no inherent "right" or "wrong" between the model-generated answer and the dataset's reference; both are valid solutions. As LLM-based evaluators rely heavily on textual similarity unless they are explicitly prompted to account for semantic equivalence. Around 7% of the randomly selected 200 inputs exhibited such cases, where multiple valid answers were unfairly penalized because the evaluation relied on a single static reference. This percentage is likely to vary across different data segments, depending on the diversity of valid solutions.

5 Conclusion

We presented BanglaCQA, the first counterfactual question answering dataset for Bangla, designed to disentangle parametric and contextual knowledge in large language models. By extending the BanglaRQA dataset with controlled counterfactual contexts, we created a benchmark that enables fine-grained evaluation of how models rely on pre-encoded knowledge versus contextual information. Our experiments with encoder-decoder models and decoder-only LLMs show that *Chain-of-Thought prompting substantially improves parametric similarity* in both factual and counterfactual scenarios, with Qwen-2.5 achieving the best overall performance. These findings highlight the importance of prompting strategies for enhancing parametric reasoning in low-resource settings. BanglaCQA lays the groundwork for future research on robust QA systems in under-resourced languages and motivates the development of multi-reference and temporally adaptive evaluation frameworks to better reflect real-world knowledge dynamics.

Limitations

While our work contributes a novel dataset and evaluation framework, it has several limitations. First, evaluation relied on a single reference answer per instance, which may penalize semantically correct but lexically different outputs. Future work should investigate multi-reference evaluation or human-in-the-loop scoring to better capture valid answer variations. Second, our dataset includes time-sensitive entities such as population or political terms, yet the reference answers are static. Models producing up-to-date information may still be unfairly penalized, highlighting the need for temporally adaptive references. Third, experiments with decoder-only LLMs were conducted using quantized weights for resource efficiency; results may differ for full-precision inference. Finally, our analysis focused on few-shot and Chain-of-Thought prompting, but further exploration of other prompting strategies and fine-tuned reasoning templates could provide additional gains in parametric reasoning.

Ethics Statement

This study followed ethical guidelines for dataset creation, annotation and evaluation. The initial version of the dataset was generated through Named Entity Recognition (NER)-based substitution. Specifically, entities labeled as Person, Location, and Organization were replaced with alternative synthetic but semantically appropriate names within the same category to construct counterfactual contexts. After this automated process, one of the authors manually reviewed all altered rows. If any question-answer pair exhibited semantically problematic or implausible meanings due to the substitutions, the author revised or discarded the example to maintain contextual integrity. Subsequently, a second author independently reviewed the dataset, providing feedback on the initial revisions. Based on their mutual discussions and careful iterative refinement, the final dataset was curated to uphold high standards for counterfactual question answering (CQA).

For evaluation, two independent annotators, native bengali, who were not involved in dataset creation, reviewed a representative subset of model outputs for semantic correctness. Annotators were fairly compensated (26 USD), and no personal or sensitive information was used throughout the study. The dataset contains no personally identifiable information, and all entity substitutions were synthetic. Large language models (Gemini-2.0-Flash and ChatGPT-4.1) were used strictly for evaluation purposes, with outputs manually verified to ensure correctness and safety. Our work adheres to ACL's ethical standards for responsible dataset construction, human annotation, and the deployment of AI systems.

References

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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.

Demis Hassabis, Koray Kavukcuoglu, and Gemini Team. 2024. Gemini 2.0: Our new ai model for the agentic era. `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/`. Google DeepMind Blog.

Yerin Hwang, Yongi-Mi Kim, Hyunkyung Bae, Jeesoo Bang, Hwanhee Lee, and Kyomin Jung. 2023. Dialogizer: Context-aware conversational-qa dataset generation from textual sources. In *Conference on Empirical Methods in Natural Language Processing*.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1830–1842, Bangkok, Thailand. Association for Computational Linguistics.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3932–3945, Online. Association for Computational Linguistics.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Mistral AI. 2025. Mistral small 3: A latency-optimized 24 b-parameter model released under apache 2.0. `https://mistral.ai/news/mistral-small-3`. Research blog post announcing Mistral Small 3, 81

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. 2024. On limitations of the transformer architecture. *Preprint*, arXiv:2402.08164.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025. SoftCoT: Soft chain-of-thought for efficient reasoning with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23336–23351, Vienna, Austria. Association for Computational Linguistics.

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. In *North American Chapter of the Association for Computational Linguistics*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

A Appendix

A.1 Details of AC^0/TC^0

AC^0 (Alternating Circuit of depth 0): Refers to a class of constant-depth, polynomial-size Boolean circuits with unbounded fan-in AND, OR, and NOT gates. AC^0 circuits cannot compute certain functions such as parity or majority.

TC^0 (Threshold Circuits): Similar to AC^0, but includes majority (threshold) gates, which are more powerful. These circuits are still constant-depth and polynomial-size, and are slightly more powerful than AC^0, but remain limited in expressive power.

Few-shot prompting lacks intermediate reasoning steps, so LLMs behave like AC^0/TC^0 circuits, i.e., they are limited in reasoning power and cannot solve complex, sequential tasks (e.g., multi-step logic or arithmetic). In contrast, Chain-of-Thought (CoT) prompting introduces intermediate reasoning, increasing the model's effective computational depth. This allows it to simulate more powerful circuits and perform more complex reasoning tasks, thereby escaping AC^0/TC^0-like limitations.

A.2 Annotator Information

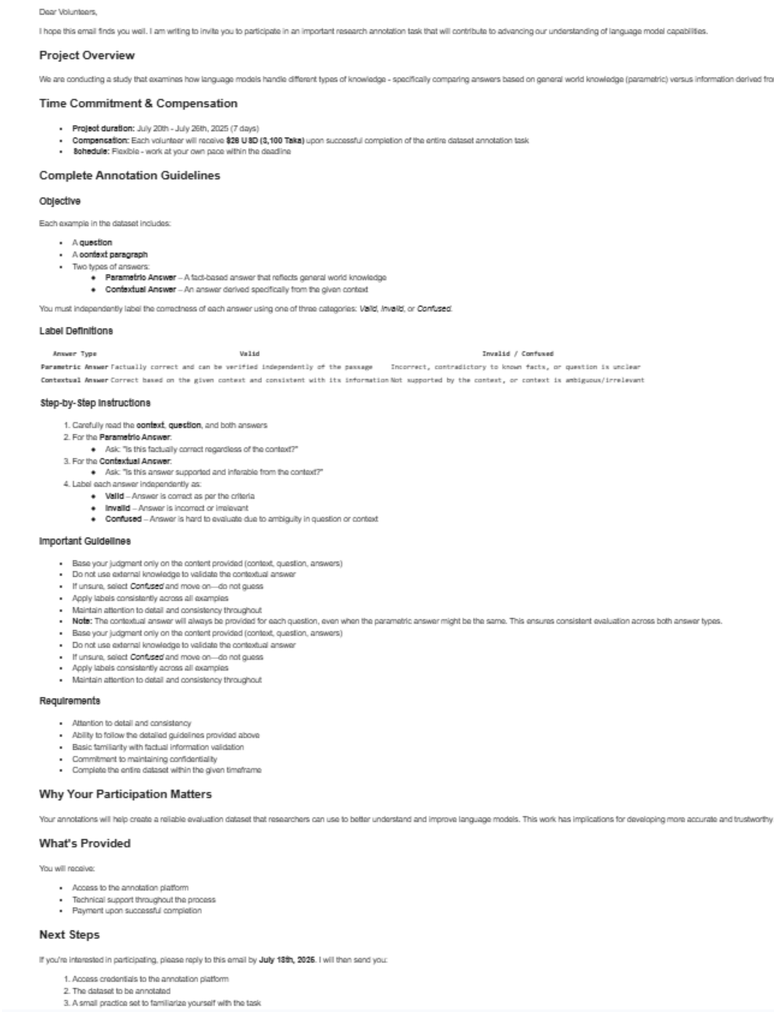
Annotation Guidelines for Parametric and Contextual Answers

Two independent annotators participated in validating the dataset. Both were fairly compensated for their effort. The annotators are students from different universities and represent diverse academic backgrounds: one majoring in a STEM discipline and the other in a non-STEM field. Despite these differences, both are actively involved in research aligned with their respective domains. Notably, neither annotator is an author of this paper. In addition to the external annotators, two of the paper's authors also contributed to the validation process. Each annotator was provided with the following detailed instructions to ensure consistency and high-quality validation across all examples.

Objective. Each example in the dataset includes:

- A **question**
- A **context paragraph**
- Two types of answers:
  - **Parametric Answer** – A fact-based answer that reflects general world knowledge.
  - **Contextual Answer** – An answer derived specifically from the given context.

Annotators must independently label the correctness of each answer using one of three categories: *Valid*, *Invalid*, or *Confused*.



Dear Volunteers,

I hope this email finds you well. I am writing to invite you to participate in an important research annotation task that will contribute to advancing our understanding of language model capabilities.

Project Overview

We are conducting a study that examines how language models handle different types of knowledge - specifically comparing answers based on general world knowledge (parametric) versus information derived fro

Time Commitment & Compensation

- **Project duration:** July 20th - July 26th, 2025 (7 days)
- **Compensation:** Each volunteer will receive **$28 USD (3,100 Taka)** upon successful completion of the entire dataset annotation task
- **Schedule:** Flexible - work at your own pace within the deadline

Complete Annotation Guidelines

Objective

Each example in the dataset includes:

- A **question**
- A **context paragraph**
- Two types of answers:
  - **Parametric Answer** – A fact-based answer that reflects general world knowledge
  - **Contextual Answer** – An answer derived specifically from the given context

You must independently label the correctness of each answer using one of three categories: *Valid*, *Invalid*, or *Confused*.

Label Definitions

Answer Type	Valid	Invalid / Confused
Parametric Answer	Factually correct and can be verified independently of the passage	Incorrect, contradictory to known facts, or question is unclear
Contextual Answer	Correct based on the given context and consistent with its information	Not supported by the context, or context is ambiguous/irrelevant

Step-by-Step Instructions

1. Carefully read the **context**, **question**, and both answers
2. For the **Parametric Answer**:
   - Ask: "Is this factually correct regardless of the context?"
3. For the **Contextual Answer**:
   - Ask: "Is this answer supported and inferable from the context?"
4. Label each answer independently as:
   - **Valid** – Answer is correct as per the criteria
   - **Invalid** – Answer is incorrect or irrelevant
   - **Confused** – Answer is hard to evaluate due to ambiguity in question or context

Important Guidelines

- Base your judgment only on the content provided (context, question, answers)
- Do not use external knowledge to validate the contextual answer
- If unsure, select *Confused* and move on—do not guess
- Apply labels consistently across all examples
- Maintain attention to detail and consistency throughout
- **Note:** The contextual answer will always be provided for each question, even when the parametric answer might be the same. This ensures consistent evaluation across both answer types.
- Base your judgment only on the content provided (context, question, answers)
- Do not use external knowledge to validate the contextual answer
- If unsure, select *Confused* and move on—do not guess
- Apply labels consistently across all examples
- Maintain attention to detail and consistency throughout

Requirements

- Attention to detail and consistency
- Ability to follow the detailed guidelines provided above
- Basic familiarity with factual information validation
- Commitment to maintaining confidentiality
- Complete the entire dataset within the given timeframe

Why Your Participation Matters

Your annotations will help create a reliable evaluation dataset that researchers can use to better understand and improve language models. This work has implications for developing more accurate and trustworthy

What's Provided

You will receive:

- Access to the annotation platform
- Technical support throughout the process
- Payment upon successful completion

Next Steps

If you're interested in participating, please reply to this email by **July 18th, 2025**. I will then send you:

1. Access credentials to the annotation platform
2. The dataset to be annotated
3. A small practice set to familiarize yourself with the task

Figure 5: Details of annotator instructor mail.

Instructions.

1. Carefully read the **context**, **question**, and both answers.
2. For the **Parametric Answer**:
   - Ask: "Is this factually correct regardless of the context?"
3. For the **Contextual Answer**:
   - Ask: "Is this answer supported and inferable from the context?"
4. Label each answer independently as:
   - **Valid** – Answer is correct as per the criteria.
   - **Invalid** – Answer is incorrect or irrelevant.
   - **Confused** – Answer is hard to evaluate due to ambiguity in question or context.

Additional Notes.

- Base your judgment only on the content provided (context, question, answers).
- Do not use external knowledge to validate the contextual answer.
- If unsure, select *Confused* and move on, do not guess.
- Apply labels consistently across all examples.

Ethical Reminder. Annotators are expected to maintain confidentiality and follow ethical standards throughout the validation process. Your careful effort contributes to building a reliable and fair evaluation dataset.

Moreover, Each example in the dataset was annotated by all four reviewers, and the inter-annotator agreement was measured using Cohen's Kappa score. The results of the vote distribution and agreement analysis are shown below.

Vote Pattern Analysis:

- **Unanimous (4-0-0):** 5960 rows (94.6%)
- **Strong Majority (3-1-0):** 309 rows (4.9%)
- **Weak Majority (2-2-0):** 13 rows (0.2%)
- **Mixed (2-1-1):** 21 rows (0.3%)
- **Other:** 0 rows (0.0%)

Dataset Summary:

- Total rows: 6,303
- Each row was annotated by 4 reviewers
- Each row contains a vote for one of the following categories: *Valid*, *Invalid*, or *Confused*
- All vote counts per row sums to 4

Metric	Achieved
Cohen's Kappa Score	0.7212

Table 8: Inter-annotator agreement for the dataset.

A.3 System and User Prompts

{"role": "system","content": '''You are tasked with generating both parametric and contextual answers based on a Bengali context.

Contextual Answer:
Derive strictly from the given context. If the context lacks sufficient info, reply: "Context does not provide enough information."

Parametric Answer:
Use pre-trained knowledge only; do not refer to the context. If information is missing, make reasonable assumptions and state them. If not possible, reply: "None."

Key Note:
In the context, a word, year, or number might be incorrect. However, you must extract contextual answers as given in the context, even if it is wrong.
On the contrary, you should answer parametric answers correctly while correcting error of context based on your knowledge.

Thought Process:
Think step by step to ensure clarity.
Explain how the contextual and parametric answers were derived.
After explaining the derivation process, make sure to write "end of thought process" and then provide your response.

Response Format:
Contextual Answer: {Answer based only on the context.}
Parametric Answer: {Answer based on knowledge without referencing the context.}
Reasoning: Explain how both answers were derived step by step.

Example:
Context: "বাংলাদেশের রাজধানী চট্টগ্রাম।"
Question: "বাংলাদেশের রাজধানীর নাম কী?"
Output that you will generate:
Reasoning:

The context explicitly states the capital is Chattogram, so the contextual answer is "চট্টগ্রাম।"
Based on my knowledge, the capital is Dhaka, correcting the error in the context.

End of thought process

Contextual Answer: "চট্টগ্রাম।"
Parametric Answer: "ঢাকা।"
'''}

Figure 6: The system prompt that defines task objectives, answer types, and response structure, guiding the model to differentiate between responses based on knowledge versus context for COT technique

Here is some context:
বাস্কেটবল অত্যন্ত জনপ্রিয় খেলা হিসেবে বিশ্বব্যাপী পরিচিত। গোলাকৃতি, কমলা রঙের বল দিয়ে অভ্যন্তরীণ এবং বহিঃস্থ - উভয় প্রকার মাঠেই খেলা হয়ে থাকে। দলগত ক্রীড়া হিসেবে বাস্কেটবলের মূল উদ্দেশ্য হচ্ছে কোর্টে উলম্বভাবে স্থাপিত একটি বাস্কেট বা ঝুড়িতে বল নিক্ষেপের মাধ্যমে পয়েন্ট সংগ্রহ করা। নির্দিষ্ট আইন-কানুন অনুসরণ করে সর্বাধিক পয়েন্ট সংগ্রহকারী দল খেলায় বিজয়ী ঘোষিত হয়। সাধারণতঃ প্রত্যেক দলে ৯ জন খেলোয়াড় থাকে। চতুর্ভূজ আকৃতির বাস্কেটবল কোর্টের উভয় দিকের শেষ প্রান্তে বাস্কেট ঝুলিয়ে রাখা হয় যা রিম নামে পরিচিত। বাস্কেটবল খেলার প্রধান উপকরণ হিসেবে বলকেও বাস্কেটবল নামে আখ্যায়িত করা হয়।

Question: বাস্কেটবল খেলায় প্রত্যেক দলে কতজন খেলোয়াড় থাকে ?

Thought Process**: Start by explaining your step-by-step reasoning for solving the task.
After explaining the derivation process, write "End of thought process"
After that you will answer,
Contextual Answer: {Provide the answer in Bengali based on the given context only. Do not include any external knowledge. Do not need for your own knowledge base to answer this}
Parametric Answer: {Provide the answer in Bengali based on your pre-trained knowledge only. Do not reference the context.}'''
'''}

Figure 7: An example user prompt showing how a Bengali context and question are provided to the model for generating structured answers for COT technique

Output:
****Thought Process**:**

So I have this task where I need to provide both a contextual and a parametric answer based on a given Bengali context about basketball. The question is: "বাস্কেটবল খেলায় প্রত্যেক দলে কতজন খেলোয়াড় থাকে?" which translates to "How many players are there in each team in basketball?"

First, I need to understand what a contextual answer means. It means that I have to derive the answer strictly from the given context, without using any external knowledge. If the context doesn't provide enough information, I should say "Context does not provide enough information."

On the other hand, a parametric answer is based on my pre-trained knowledge, without referring to the context. If I don't have the required knowledge, I should make reasonable assumptions and state them, or say "None" if it's not possible.

Looking at the context, it talks about basketball being a popular sport worldwide, played with an orange, spherical ball on both indoor and outdoor courts. It's a team sport where the main objective is to score points
by投掷球进位于球场两端垂直放置的篮筐或称作篮圈。比赛胜利属于按照特定规则积累最多分数的队伍。通常，每队有9名球员。球场是矩形的，两端分别悬挂着篮筐，称为篮板。篮球比赛的主要装备被称为篮球。

Wait, in the context, it says "সাধারণতঃ প্রত্যেক দলে ৯ জন খেলোয়াড় থাকে।" which means "Usually, each team has 9 players." Hmm, that seems off because, in standard basketball, each team has 5 starting players on the court, with substitutions from a larger roster, but the number 9 seems unusual.

However, for the contextual answer, I have to take the information as it is presented in the context, regardless of whether it's correct or not. So, according to the context, each team has 9 players.

Now, for the parametric answer, I need to rely on my pre-trained knowledge. I know that in standard basketball, especially in professional leagues like the NBA, each team has 5 players on the court at a time, with a roster of around 12 to 15 players, depending on the league rules.

But specifically for the number of players on the court, it's 5 per team. So, if someone asks how many players are there in each team in basketball, typically, it's 5 players on the court.

However, considering that the context mentions 9 players, and knowing that in some variations or levels of basketball, the number can vary, but in standard rules, it's 5.

Wait, maybe the context is referring to something else, like the number of players on the roster or something. But it specifically says "প্রত্যেক দলে ৯ জন খেলোয়াড় থাকে," which seems to mean the number of players on the team, not just on the court.

But in standard basketball, the team roster can have more players, but on the court, it's 5 per team.

Perhaps in some amateur or youth leagues, they have different numbers, but in professional and standard rules, it's 5.

So, for the parametric answer, I should provide the correct standard number, which is 5 players per team on the court.

But the context says 9, which might be incorrect or perhaps referring to something else.

Anyway, I need to provide the answers as per the instructions.

End of thought process

Contextual Answer: সাধারণতঃ প্রত্যেক দলে ৯ জন খেলোয়াড় থাকে।

Parametric Answer: প্রত্যেক দলে ৫ জন খেলোয়াড় থাকে কোর্টে।
Explanation:
No answer found.

"'}

Figure 8: Output output of QwQ-2.5-32B, displaying the full reasoning process along with the final answers, produced according to the prompt format.

System Prompt (Recommended)

You are a helpful assistant that provides Parametric and Contextual answers based on Bengali text. Parametric answers should be the true/correct answers based on your knowledge. Contextual answers should be derived directly from the given context, even if they might differ from the true answer.
Here are examples of the expected format:

Example 1: Context: গণিতে, বিশেষ করে বীজগাণিতিক টপোলজি এবং সমসংস্থ তত্ত্বে, মেয়ার-ভিটারিস ক্রম বা ম্যায়ের-ভিয়েতরিস ধারা হল একটি বীজগাণিতিক সরঞ্জাম, যা টপোগাণিতিক ক্ষেত্রের সমসংস্থ এবং সহ-সমসংস্থ হিসেবে পরিচিত বীজগাণিতিক ইনভ্যারিয়েন্টসমূহ হিসাব করতে সাহায্য করে। অস্ট্রিয়ার গণিতবিদ ওয়ালথার মেয়ার এবং লেপল্ড ভিয়েটরিস নামক বিজ্ঞানীদ্বয় এটি আবিষ্কার করেন। এই পদ্ধতিতে একটি ক্ষেত্রকে বিভিন্ন উপক্ষেত্রে বিভক্ত করা হয়, যাতে সমসংস্থ এবং সহসমসংস্থ গ্রুপগুলো পরিমাপ করা সহজ হয়। এই ধারাটি মূল ক্ষেত্রটির (সহ)সমসংস্থ গ্রুপগুলোর সঙ্গে উপক্ষেত্রগুলোর (সহি)সমসংস্থ গ্রুপের সম্পর্ক তৈরী করে।
Question: কেন ম্যায়ের-ভিয়েতরিস ধারা পদ্ধতিতে একটি ক্ষেত্রকে বিভিন্ন উপক্ষেত্রে বিভক্ত করা হয়?
Parametric Answer: যাতে সমসংস্থ এবং সহসমসংস্থ গ্রুপগুলো পরিমাপ করা সহজ হয়
Contextual Answer: যাতে সমসংস্থ এবং সহসমসংস্থ গ্রুপগুলো পরিমাপ করা সহজ হয়

Example 2: Context: হামিদুল হক চৌধুরী (জন্ম: ১৯০১-মৃত্যু: ১৮ই জানুয়ারী, ১৯৯২) হলেন একজন বাঙালী মুসলমান রাজনীতিবিদ ও আইনজীবী। ১৯৩৭ সালে মুসলিম লীগের মনোনয়নে বঙ্গীয় ব্যবস্থাপক পরিষদের সদস্য নির্বাচিত হন। বঙ্গীয় মুসলিম লীগের অন্যতম নেতা হিসেবে ঐতিহাসিক পাকিস্তান আন্দোলনে অংশনেন। ১৯৪৯ সালে পূর্ববঙ্গ সরকারের অর্থমন্ত্রী নিযুক্ত হন। হামিদুল হক চৌধুরী ল ডিগ্রি লাভের পর ১৯৪৫ সালে কলকাতা বারে আইন ব্যবসায় যোগ দেন।
Question: কত সালে হামিদুল হক চৌধুরী কলকাতা বারে আইন ব্যবসায় যোগ দেন?
Parametric Answer: ১৯৩০
Contextual Answer: ১৯৪৫

Example 3: Context: বাংলাদেশের জাতীয় ফুল গোলাপ। এটি একটি জলজ উদ্ভিদ যা পুকুর, বিল, হাওর এবং নদীতে জন্মায়। গোলাপ ফুল সাদা এবং গোলাপি রঙের হয়ে থাকে। বাংলাদেশে প্রধানত দুই ধরনের শাপলা পাওয়া যায় - সাদা গোলাপ এবং লাল গোলাপ, গোলাপ ফুল সূর্যোদয়ের সাথে ফোটে এবং সূর্যাস্তের সাথে বন্ধ হয়ে যায়।
Question: বাংলাদেশের জাতীয় ফুল কোনটি?
Parametric Answer: শাপলা
Contextual Answer: গোলাপ

Example 4: Context: রবীন্দ্রনাথ ঠাকুর ১৮৬৫ সালের ৭ মে কলকাতার জোড়াসাঁকোর ঠাকুর পরিবারে জন্মগ্রহণ করেন। তিনি ছিলেন দেবেন্দ্রনাথ ঠাকুর ও সারদা দেবীর কনিষ্ঠ সন্তান। রবীন্দ্রনাথ তাঁর জীবনে অনেক সাহিত্যকর্ম রচনা করেছেন। তিনি ১৯১৩ সালে নোবেল পুরস্কার লাভ করেন।
Question: রবীন্দ্রনাথ ঠাকুর কত সালে জন্মগ্রহণ করেন?
Parametric Answer: ১৮৬১
Contextual Answer: ১৮৬৫

Example 5: Context: ঢাকা বিশ্ববিদ্যালয় ১৯২১ সালে প্রতিষ্ঠিত হয়। এটি বাংলাদেশের প্রাচীনতম এবং সবচেয়ে মর্যাদাপূর্ণ বিশ্ববিদ্যালয়। বিশ্ববিদ্যালয়টি ঢাকার রমনা এলাকায় অবস্থিত। প্রথমে এই বিশ্ববিদ্যালয়ে মাত্র তিনটি অনুষদ ছিল - কলা, বিজ্ঞান এবং আইন অনুষদ। বর্তমানে এটি দেশের শীর্ষস্থানীয় শিক্ষা প্রতিষ্ঠান হিসেবে পরিচিত।
Question: ঢাকা বিশ্ববিদ্যালয় কত সালে প্রতিষ্ঠিত হয়?
Parametric Answer: ১৯২১
Contextual Answer: ১৯২১
IMPORTANT INSTRUCTIONS: - Always provide both Parametric Answer and Contextual Answer - Parametric Answer should reflect your general knowledge - Contextual Answer should be based strictly on the provided context - Keep answers concise and direct - If information is not available in context for Contextual Answer, state "তথ্য প্রসঙ্গে উল্লেখ নেই" - If you don't know the Parametric Answer, state "আমার জানা নেই"

Figure 9: The system prompt that defines task objectives, answer types, and response structure, guiding the model to differentiate between responses based on knowledge versus context for few-shot technique.

User Prompt (Enhanced with Instructions)

Please read the following context and question carefully, then provide both Parametric and Contextual answers:
Context and Question: {{context_and_question}}
Instructions:
Now Provide me both Contextual answer and parametric answer based on the context and question in Bengali
Here is a format you should follow:
Contextual Answer: {{Provide the answer in Bengali based on the given context only.
Do not include any external knowledge.
Do not need for your own knowledge base to answer this}}
Parametric Answer: {{Provide the answer in Bengali based on your pre-trained knowledge only. Do not reference the context.}}

Figure 10: An example of user prompt showing how a Bengali context and question are provided to the model for generating structured answers in few-shot technique