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 integrating counterfactual passages and answerability annotations into an existing dataset. In addition, we propose prompting-based pipelines for LLMs to disentangle parametric and contextual knowledge in both factual and counterfactual scenarios. Furthermore, we apply LLM-based evaluation techniques that measure answer quality based on semantic similarity. 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.

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

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 Large language models (LLMs), this emulation has reached new heights for high-resource languages, 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, which occurs because of the imbalance between extensive pre-encoded data and limited contextual input, can lead to the generation of hallucinated answers (Krishna et al., 2021). Some work further shows that contextual questions that contain incorrect assumptions disrupt generation performance (Kim et al., 2021). While 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 probe knowledge prioritization. Moreover, we introduce disentanglement pipelines by leveraging

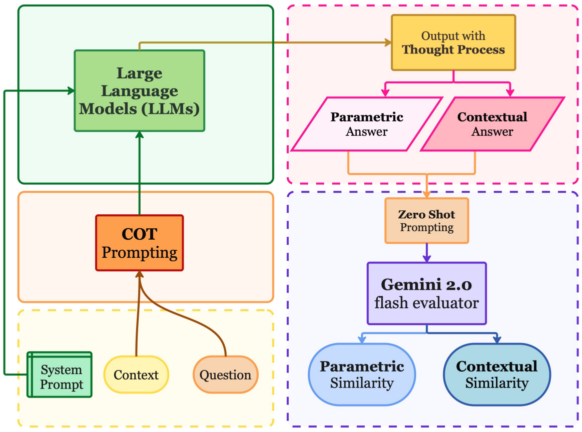

Figure 2: Implementation Pipeline of Large Language Models (LLMs) for Disentangling Parametric and Contextual Knowledge in QA

multiple open-sourced LLMs (LLaMA-3.3B (Touvron et al., 2023), DeepSeek-R1 (DeepSeek-AI et al., 2025), Qwen-2.5-32B (Yang et al., 2024)) 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 for semantic similarity scoring, which outperforms traditional metrics to evaluate the semantic accuracy of Bangla QA responses. Our analysis reveals that integrating counterfactual contexts exhibits strong parametric generation similarity. These findings not only establish a blueprint for low-resource languages and advance QA systems for Bangla, but also emphasizes transparency in knowledge utilization in counterfactual scenarios.

2 BanglaCQA Dataset

The BanglaCQA dataset comprises both factual and counterfactual contexts, along with questions and their corresponding parametric and contextual answers. The factual contexts are sourced from the BanglaRQA (Ekram et al., 2022) dataset (licensed under cc-by-nc-sa-4.0), one of the largest human-annotated Bangla QA factual datasets with contextual answers. However, it lacked distinct parametric and contextual answers to train models in disentangling knowledge sources. Our primary enhancement involves introducing rich, NER-based counterfactual contexts for the low-resource Bangla language, expanding the dataset by **42.28%** to significantly accelerate model performance through alternative scenarios that demand deeper understanding of named entities and their relationships.

Dataset Attribute	Setting
Total QA pairs	21,211
Factual Contexts	14,900
Counterfactual Contexts	6,300
Average Question Word Count	8.26
Average Context Word Count	215.27

Table 1: BanglaCQA dataset summary statistics

2.1 Counterfactual Context Generation

Counterfactual contexts are generated by modifying named entities (e.g., names, dates, quantities) in factual contexts using the NER replacement script (Sarker, 2020), followed by manual validation to ensure semantic coherence. When a named entity appeared in the answer column, it was replaced with alternative entities in both the context and answer columns to ensure consistency. Moreover, the numeric values are altered using regular expressions. To avoid duplication, the data IDs were updated, and each modification was manually reviewed by two of the authors as annotators to ensure contextual relevance and correctness. The whole process ensures precise modification in counterfactual contexts with the aim of challenging LLMs to adapt to contrasting information. Each counterfactual context is paired with its corresponding contradictory parametric and contextual responses, which enables models to learn the disentanglement of knowledge sources.

3 Implementation Pipeline

We propose a systematic framework to investigate how LLMs handle parametric and contextual reasoning across factual and counterfactual settings. As shown in Figure 2, each input consists of a context and a question, paired with a system prompt and formatted using either few-shot or CoT (Chain-of-thought) prompting, which instructs models to explicitly articulate intermediate reasoning steps before producing final answers. The full prompt structure and sample outputs are provided in Appendix A for reference. All models were decoded using the same decoding hyperparameters to ensure fair comparison: temperature = 0.1, top-p = 0.1, repetition penalty = 1.02, and max tokens = 1500. Qwen-2.5 and DeepSeek-R1 were used in their non-quantized versions, while LLAMA-3.3 was configured using fp16 quantization due to hardware constraints. All models were hosted on Kaggle with 4xNVIDIA L4 GPUs, each offer-

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

Table 2: Performance of different models under Factual (F) and Counterfactual (CF) settings, evaluated with parametric and contextual similarity. "F" denotes **Factual** contexts and "CF" denotes **Counterfactual** contexts. Bold values indicate the best-performing configurations in each category.

ing 22.5GB of VRAM. Each model generated two distinct outputs: a parametric answer, reflecting pre-trained knowledge, and a contextual answer, reflecting the given input context.Answers generated in non-Bangla languages were reformatted automatically using the Gemini API to ensure cross-lingual consistency in the evaluation phase. Gemini 2.0 Flash was used in a zero-shot evaluation setting to compute semantic similarity between the model's generated answers and gold targets. Two metrics—parametric similarity and contextual similarity—were computed separately to measure alignment between the generated and gold answers across factual and counterfactual settings.[1]

4 Results

We investigate how prompting strategies and model design influence model behavior across factual (F) and counterfactual (CF) settings, by discussing the following research questions:

RQ1: How does the prompting strategy (CoT vs Few-shot) affect the parametric and contextual performance of language models in Bangla across factual and counterfactual settings? Contextual similarity remains consistently high across both prompting strategies **(0.77-0.94)** which indicates that both CoT and Few-shot prompting are similarly effective at utilizing contextual information in both factual (F) and counterfactual (CF) settings. In contrast, parametric similarity improves substantially with CoT prompting, particularly in counterfactual settings. Few-shot prompting results in low parametric similarity scores **(0.24-0.35)**, whereas CoT prompting increases this range to **0.55-0.81**. This indicates that CoT prompting enhances a model's ability to generate correct answers from pre-encoded knowledge while maintaining strong contextual understanding.

RQ2: 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.

RQ3: How well do LLMs adapt to counterfactual contexts in Bangla, and what does this reveal about their sensitivity to narrative contradiction? All models struggle with counterfactual contextual understanding under few-shot prompting for instance: contextual similarity drops below **0.32**, which suggests poor sensitivity to narrative contradiction. CoT prompting alleviates this as it enables models like Qwen-2.5 and DeepSeek-R1 to achieve **0.74 and 0.70** contextual similarity, respectively. This shows that Bangla's syntactic structure requires explicit reasoning to resolve contradictions, and that LLMs, without such support, default to parametric recall even when the context logically invalidates it. This exposes a fundamental limitation in LLMs' default handling of counterfactual semantics.

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

4.1 Error Analysis

While Gemini 2.0 Flash offers a scalable, fast approximation of parametric answer similarity, our evaluation exposes key limitations in counterfactual QA for Bangla. To assess metric reliability, we qualitatively compared its outputs with human judgments—widely regarded as the gold standard in QA (Clark et al., 2021)—revealing three primary sources of discrepancy:

I) Temporal Mismatch (Outdated Targets): We observed that approximately 4% of the randomly selected 200 model-generated answers were more up-to-date than the dataset's reference answers. As shown in Figure 3, Qwen-2.5 produced factually accurate responses, but these were penalized due to the mismatch with stale reference data. This temporal mismatch underscores the limitations of using static reference data. This mismatch may vary across different batches but underscores the need for more dynamic evaluation frameworks.

Dataset's Target Answer (Factual Context):

Question: টয়োটা মোটর কর্পোরেশনের কারখানাগুলি থেকে বছরে গড়ে কত গাড়ি তৈরি হয়?
Context: পরের বছরেই তারা প্রথমবারের মতো জাপানের বাইরে ব্রাজিলে গাড়ির কারখানা স্থাপন করে। এরপর তারা পর্যায়ক্রমে ছড়িয়ে পরে। এরপরে তারা পর্যায়ক্রমে পরে পৃথিবীতে ছড়িয়ে ....
Target Parametric Answer: **৫৫ লাখ**
Target Contextual Answer: ৯০ লাখ

Qwen-2.5 Generated Answer (Factual Context):

Question: টয়োটা মোটর কর্পোরেশনের কারখানাগুলি থেকে বছরে গড়ে কত গাড়ি তৈরি হয়?
Context: পরের বছরেই তারা প্রথমবারের মতো জাপানের বাইরে ব্রাজিলে গাড়ির কারখানা স্থাপন করে। এরপর তারা পর্যায়ক্রমে ছড়িয়ে পরে। এরপরে তারা পর্যায়ক্রমে পরে পৃথিবীতে ছড়িয়ে ....
Generated Parametric Answer: টয়োটা মোটর কর্পোরেশন প্রায় **১০ মিলিয়ন** বা তার থেকে বেশি গাড়ি উৎপাদন করে
Extracted Contextual Answer: বছরে প্রায় ৯০ লাখ(৯ মিলিয়ন) গাড়ি তৈরি হয়।

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

II) Solution Variation (Multiple Valid Answers): Around 7% of the randomly selected 200 input demonstrate cases where different valid answers are penalized due to lack of lexical overlap. For example, Figure 4 shows a case where the model predicts "23.5°" while the target is "66.5°", both correct, as they represent complementary angles of Earth's axial tilt, but the metric assigns a low score due to the lexical mismatch. This issue may vary across different batches, but it highlights the challenge of accounting for multiple valid solutions in the evaluation process.

III) Length Discrepancy (Verbose but Correct CoT): Around 54% of the randomly selected 200 inputs demonstrate cases where Qwen-2.5 generates longer, more detailed explanations using Chain-of-Thought (CoT) prompting. Even when the final answer is accurate, the inclusion of reasoning reduces the similarity score due to the length discrepancy. As shown in Figure 5, this highlights a mismatch between surface-level similarity and semantic correctness.

Dataset's Target Answer (Factual Context):

Question: পৃথিবী নিজ অক্ষের কত ডিগ্রি কোণে হেলে রয়েছে?
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 ....
Target Parametric Answer: **৬৬.১/২ ডিগ্রি**
Target Contextual Answer: ৬৬.১/২ ডিগ্রি

Qwen-2.5 Generated Answer (Factual Context):

Question: পৃথিবী নিজ অক্ষের কত ডিগ্রি কোণে হেলে রয়েছে?
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 ....
Generated Parametric Answer: পৃথিবী নিজ অক্ষের প্রায় **২৩.৫ ডিগ্রি** কোণে হেলে রয়েছে।
Extracted Contextual Answer: পৃথিবী নিজ অক্ষের প্রায় ৬৬.১/২ ডিগ্রি কোণে হেলে রয়েছে।

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

Dataset's Target Answer (Counterfactual Context) :

Question: রিমোট সেন্সিং বিশেষজ্ঞের কোন কোন শাখার যথেষ্ট জ্ঞান থাকা অত্যাবশ্যক?
Context: দূর অনুধাবন হল কোন বস্তুকে সরাসরি স্পর্শ না করা সেই বস্তু হতে টার গুণাবলি সম্পর্কিত উপাত্ত সংগ্রহ করা এবং পর্যবেক্ষণ করার এক ধরনের কৌশল। এক্ষেত্রে বিভিন্ন সেন্সর.....রিমোট সেন্সিং বিশেষজ্ঞের রসায়ন ও গণিতে যথেষ্ট জ্ঞান থাকা অত্যাবশ্যক । উদাহারন হিসেবে বলা যায়- আকাশে অবস্থিত কৃত্রিম উপগ্রহ বা বিমান থেকে....
Target Parametric Answer: **পদার্থ বিজ্ঞান ও গণিতে**
Target Contextual Answer: রসায়ন ও গণিতে

Qwen-2.5 Generated Answer (Counterfactual Context):

Question: রিমোট সেন্সিং বিশেষজ্ঞের কোন কোন শাখার যথেষ্ট জ্ঞান থাকা অত্যাবশ্যক?
Context: দূর অনুধাবন হল কোন বস্তুকে সরাসরি স্পর্শ না করা সেই বস্তু হতে টার গুণাবলি সম্পর্কিত উপাত্ত সংগ্রহ করা এবং পর্যবেক্ষণ করার এক ধরনের কৌশল। এক্ষেত্রে বিভিন্ন সেন্সর.....রিমোট সেন্সিং বিশেষজ্ঞের রসায়ন ও গণিতে যথেষ্ট জ্ঞান থাকা অত্যাবশ্যক । উদাহারন হিসেবে বলা যায়- আকাশে অবস্থিত কৃত্রিম উপগ্রহ বা বিমান থেকে....
Generated Parametric Answer: রিমোট সেন্সিং বিশেষজ্ঞের জন্য যথেষ্ট জ্ঞান থাকা প্রয়োজনীয় শাখাগুলো হলোঃ পৃথিবীবিজ্ঞান (জিওসায়েন্স), **পদার্থবিজ্ঞান(ফিজিক্স), গণিত(ম্যাথেমেটিক্স),** পরিসংখ্যান (স্ট্যাটিস্টিক্স), কম্পিউটার বিজ্ঞান (কম্পিউটার সায়েন্স), জিওস্পেসিফিক প্রক্রিয়াকরণ (জিপিসি) এবং রিমোট সেন্সিং এর প্রাথমিক তত্ত্ব ও পদ্ধতি।
Extracted Contextual Answer: রিমোট সেন্সিং বিশেষজ্ঞের রসায়ন ও গণিতে যথেষ্ট জ্ঞান থাকা অত্যাবশ্যক।

Figure 5: Example where the model generates a factually relevant but longer answer than the reference, highlighting penalization due to length mismatch.

5 Conclusion

We propose a study to disentangle knowledge sources in Bangla contextual QA models by introducing a counterfactual extension to the BanglaRQA dataset. This enables differentiation between contextual and parametric answers, which often overlap in factual contexts. We trained encoder-decoder models on the extended dataset and evaluated LLMs (LLAMA-3.3, Qwen-2.5, DeepSeek-R1) using Few-shot and Chain-of-Thought prompting. CoT prompting proved most effective, with Qwen-2.5 excelling in generating parametric answers even under counterfactual settings. Our findings reveal a unique trend in low-resource Bangla, where contextual inputs outweigh parametric knowledge, highlighting Qwen-2.5's robustness for real-world Bangla QA applications.

Limitations

Our work provides valuable insights into Bangla question answering with large language models, but there are naturally some areas to explore further. For example, a few models occasionally gave numeric answers in English even when prompted in Bangla, which reflects the complex multilingual nature of these systems. This is an interesting aspect to investigate more deeply in future work, especially around how models handle different languages during reasoning. In some cases, models produced intermediate reasoning steps without a clear final answer. While this didn't affect the overall evaluation, it suggests there's more to learn about how these models arrive at their conclusions. Our dataset covers a range of question types that mix contextual and knowledge-based information, providing a solid testbed, though future datasets could help sharpen the distinction between these types even more. Finally, because of hardware limits, we ran the larger models using efficient quantization techniques, which worked well for our experiments but leaves room to explore full-precision versions down the line.

Ethics Statement

This research involved manual annotation carried out by the two authors. One author handled the primary annotation, refining the BanglaNER portion to ensure named entity replacements preserved both syntactic and semantic integrity. The other author independently reviewed these edits for contextual and factual consistency across the context and answer fields. For numerical entities, we applied controlled modifications using regular expressions to introduce variation while preserving the original meaning. Additionally, we used AI tools to assist with coding and grammatical corrections throughout the research.

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.

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.

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*.

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.

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.

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.

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 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.

Figure 8 shows the output of Qwen-2.5-32B, which is a representative example; actual outputs and reasoning styles may differ across LLMs.

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.

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.