When Language Models Speak Burmese: Evaluating Hallucination in Low-Resource Domain Question Answering

First Author[1], Second Author[1,2], Third Author[1],
[1]Your University, [2]Another Institution
Correspondence: email@domain

Abstract

Language models have recently gained substantial attention in natural language processing, demonstrating strong performance across a wide range of tasks, including text classification, text generation, language modeling, and question answering (Q&A). Despite these advancements, one of the most pressing challenges in language models is hallucination: the generation of responses that are fluent and plausible-sounding but factually incorrect, irrelevant, or fabricated. This study presents preliminary work investigating the impact of hallucination in Q&A tasks for low-resource languages. Specifically, we evaluate model performance on the MPox-Myanmar dataset, employing both small- and large-scale language models accessible through APIs. Our research contributes by systematically examining observable hallucination across model sizes and prompting strategies, analyzing whether intuition about their behavior holds consistently, and providing explainability behind the observed patterns.

1 Introduction

Large language models (LLMs) have seen a surge in both practical applications and research developments in recent years (Hadi et al., 2023). LLMs are trained on a vast diversity of data and operate on the principle of probabilistic outcomes (Brown et al., 2020). Consequently, hallucination is an inherent phenomenon in language models that cannot be fully eliminated (Xu et al., 2025).

Hallucination occurs across all modalities – text, image, video, and audio (Sahoo et al., 2024). However, it is important to note that hallucination is not always harmful. Hallucinations become problematic when the generated statements are factually inaccurate or conflict with universal human, societal, or specific cultural norms.

The issue becomes critical in mission-specific, high-stakes domains such as finance, medicine, and law, where even a single biased or incorrect decision can pose significant risks and, in some cases, become a matter of life and death (Rawte et al., 2023).

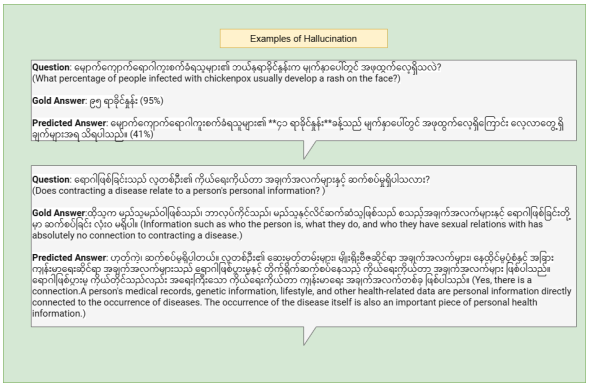

Figure 1: Examples of hallucinations observed in responses by language models from our study on MPox Q&A.

The effects of hallucination also differ between high-resource and low-resource languages (Rohera et al., 2025). High-resource languages such as English, major Western European languages, Chinese, and Japanese dominate the training data of most machine learning systems. Low-resource languages — for example, Burmese, which is the focus of our study – are also included in training but in much smaller proportions compared to high-resource languages.

2 Background

2.1 Hallucination in Early NLP Tasks

Hallucination as a problem was first studied in neural machine translation (NMT) (Lee et al., 2018), as well as in summarization and dialogue systems, before gaining attention in domain-specific Q&A. In these tasks, models often produced plausible-sounding but incorrect outputs with high confidence, creating the illusion for users that the responses were reliable and grounded in truth. In the

case of summarization, hallucinations are typically categorized as intrinsic, where the generated text directly contradicts the source, or extrinsic, where the output introduces facts not supported by the source content. These early findings highlighted that hallucination is not confined to a single NLP task, but rather represents a broader phenomenon inherent to probabilistic sequence generation.

2.2 Hallucination in Domain-Specific Q&A

While hallucination is inherent in language models, prior work has shown that models sometimes demonstrate the ability to correct their own incorrect claims during the process of explanation or justification (Zhang et al., 2023). Hallucinations can be observed in tasks ranging from simple ones – such as counting the number of occurrences of a letter in a word or sentence – to more complex reasoning tasks, such as predicting the number of objects in an image or determining whether a number is prime (Zhang et al., 2023). As noted in the study, the LLM occasionally revises its earlier incorrect response while attempting to explain its reasoning.

While hallucination in simple tasks may have limited consequences, it becomes critical in high-stakes domains such as medicine, finance, and law, or when the language is low-resource. In such settings, the likelihood of generating plausible but incorrector entirely factually inaccurateanswers is significantly higher.

- **Medicine**: Several studies highlight the risks of LLMs in healthcare Q&A. (Siontis et al., 2023) demonstrated that ChatGPT can hallucinate in cardiology-related queries. (Pal et al., 2023) introduced *Med-HALT*, a benchmark specifically designed to evaluate hallucinations in medical reasoning and fact recall. (Zhu et al., 2025) reviewed recent progress on hallucination detection in medical LLMs and LVLMs, providing an overview of available benchmarks. Further, (Jiang et al., 2025) showed that chain-of-thought (CoT) prompting can reduce hallucinations in medical tasks.

- **Finance**: Hallucinations in financial Q&A often involve incorrect numerical values, fabricated company data, or invalid market explanations. (Kang and Liu, 2023) evaluated LLMs on financial tasks and found hallucination rates particularly high when numerical reasoning was required.

- **Law**: In legal Q&A, hallucinations can be especially damaging due to the reliance on case law and statutes. (Mik, 2024) showed that LLMs often invent legal precedents or misinterpret statutes, leading to legal hallucinations.

2.3 Hallucination in Low-Resource Languages

Hallucination detection in low-resource languages for domain-specific Q&A remains underexplored, with most research focusing on high-resource settings. Retrieval-augmented generation (RAG) models (Siriwardhana et al., 2023) have shown promise for domain-specific Q&A by grounding responses in external knowledge sources, with domain adaptation reducing hallucinations and improving factual consistency. However, challenges persist due to limited parallel corpora, absence of language-specific fact-checking resources, and scarcity of annotated hallucination datasets for low-resource domains.

3 Problem Statement

Large language models (LLMs) often hallucinate in Q&A tasks. Hallucinations can occur in both cases: when questions relate to information seen during training, as well as when they concern previously unseen data. Ideally, the model should indicate that it does not know the answer; however, it frequently produces responses that are factually incorrect or misleading. This issue is particularly critical in two scenarios: 1) domain-specific Q&A in sensitive areas such as medicine, finance, and law, where incorrect information can be highly risky, and 2) Q&A in low-resource languages (LRLs), where the training data is significantly smaller than that available for high-resource languages (HRLs).

To address this problem, this paper evaluates LLM performance on a low-resource dataset of critical domain knowledge.

Concisely, we aim to answer the following research questions:

1. What is the baseline hallucination rate of current LLMs when answering questions in Burmese?

2. How do different prompting strategies (zero-shot, one-shot, few-shot, chain-of-thought) affect hallucination rates in low-resource Q&A?

Model	Prompting	Accuracy (%)	Hallucination (%)	Avg. Semantic Similarity
Llama 3.1 8B	Zero-shot	14.10	39.40	0.543
Llama 3.1 8B	One-shot	16.20	32.30	0.562
Llama 3.1 8B	Few-shot	34.30	26.30	0.631
Llama 3.1 8B	CoT-few-shot	13.10	43.40	0.535
Llama 3.3 70B	Zero-shot	19.20	35.40	0.560
Llama 3.3 70B	One-shot	29.20	24.35	0.626
Llama 3.3 70B	Few-shot	29.30	20.20	0.620
Llama 3.3 70B	CoT-few-shot	21.15	29.40	0.530
Gemini-2.5	Zero-shot	28.30	24.20	0.612
Gemini-2.5	One-shot	24.20	26.30	0.597
Gemini-2.5	Few-shot	35.40	30.30	0.614
Gemini-2.5	CoT-few-shot	10.10	39.40	0.521
Gemma 4B	Zero-shot	6.10	44.40	0.513
Gemma 4B	One-shot	16.17	38.37	0.540
Gemma 4B	Few-shot	21.23	29.30	0.598
Gemma 4B	CoT-few-shot	9.13	41.40	0.518
Gemma 12B	Zero-shot	24.20	28.30	0.591
Gemma 12B	One-shot	25.27	29.27	0.573
Gemma 12B	Few-shot	29.30	28.27	0.599
Gemma 12B	CoT-few-shot	14.13	30.30	0.567
SeaLLM 1.5B	Zero-shot	6.10	58.60	0.447
SeaLLM 1.5B	One-shot	11.10	50.50	0.508
SeaLLM 1.5B	Few-shot	7.10	57.60	0.456
SeaLLM 1.5B	CoT-few-shot	9.10	66.70	0.421
SeaLLM 1.5B - Chat	Zero-shot	7.10	51.50	0.476
SeaLLM 1.5B - Chat	One-shot	10.10	53.50	0.497
SeaLLM 1.5B - Chat	Few-shot	6.10	56.60	0.451
SeaLLM 1.5B - Chat	CoT-few-shot	7.10	59.60	0.468

Table 1: Performance comparison of different models and prompting strategies on Burmese MPox Q&A.

3. Can semantic similarity-based evaluation reliably detect hallucination in multilingual contexts?

4 Methodology

4.1 Dataset

We utilized the Mpox-Myanmar dataset (Min-SiThu), a domain-specific question-answering dataset containing 99 question-answer pairs about Mpox (monkeypox) in Burmese language. This dataset represents a critical use case for evaluating LLM performance on:

- Low-resource language: Burmese has limited training data compared to high-resource languages.
- Domain-specific content: Medical/healthcare information requiring factual accuracy.

The dataset contains questions ranging from factual inquiries (symptoms, transmission) to procedural knowledge (prevention measures, treatment guidelines).

4.2 Prompting Strategies

We employed the following prompting strategies in our experiments:

- Zero-shot prompting:
- One-shot prompting:
- Few-shot prompting (3-shot)
- Chain-of-thought prompting combined with few-shot prompting

4.3 Experimental Setup

- **Sentence Language Classification**: A response is classified as *Burmese* if the ratio

Model	Prompting	Correct	Partial	Hallucinated	Refusal	Language Failure
Gemini-2.5	Zero-shot	28	47	24	0	0
Gemini-2.5	One-shot	24	47	26	2	0
Gemini-2.5	Few-shot	35	33	30	1	0
Gemini-2.5	CoT-Few-shot	10	50	39	0	0

Table 2: Answer classification outcomes for Gemini-2.5 under different prompting strategies.

of Burmese characters to total characters exceeds 0.80. If the ratio is between 0.50 and 0.80, it is classified as *Mixed*. Semantic similarity between the generated and expected responses is computed only for Burmese and Mixed sentences; otherwise, the response is marked as a *Language Failure*. The multilingual MiniLM transformer (Reimers and Gurevych, 2019) is used to map sentences into a 384-dimensional dense vector space, enabling the calculation of semantic similarity between the expected and generated response sentences.

- **Sentence Evaluation Classification**: If the similarity score between the generated response and the expected response is greater than 0.75, the response is marked as *Correct*. If the score lies between 0.50 and 0.75, it is marked as *Partially Correct*. In all other cases, the response is classified as a *Hallucination*.

- **API Limit**: The requests-per-minute (RPM) limit varies across models. For the Groq API, which allows 30 RPM, we introduce a 0.5-second delay between queries. For Google Gemini and Gemma models, we set a delay of 4.5 seconds to comply with the limit of 15 RPM.

- **Temperature and Maximum Token Size**: The temperature is fixed at 0.70, and the maximum token size is set to 1024 tokens.

- **Prompting Templates**: Prompting templates for zero-shot, one-shot, few-shot, and chain-of-thought few-shot strategies are shown in Figures 5, 6, 7, and 8.

4.4 Evaluation Pipeline

1. **Dataset Loading**: Load the Mpox dataset from HuggingFace.

2. **Inference**: Query the API to obtain a response in Burmese for each question, and compare the models response to the corresponding ground-truth answer in the dataset.

3. **Semantic Evaluation**: Compute the semantic similarity between the generated response and the expected answer.

4. **Classification**: Categorize each generated response into one of the following classes: *Correct*, *Partially Correct*, *Hallucination*, or *Refusal*.

5. **Storage**: Store the outputs in CSV format, containing both the model responses and the expected answers, to enable future comparison and analysis of hallucinated responses.

5 Results and Discussion

Table 1 shows the accuracy, hallucination rate, and average semantic similarity for each model. Gemini-2.5 exhibits the lowest hallucination rate among all models. Table 2 presents the number of correct, partially correct, and hallucinated responses across all prompting strategies, for the best performing model— Gemini-2.5.

The hallucination rate comparison is shown in Figure 2. Across almost all models, one-shot and few-shot prompting tend to result in lower hallucination rates compared to zero-shot and CoT-few-shot prompting. Gemini-2.5 consistently achieves the lowest hallucination rates across all strategies, outperforming Llama, Gemma, and SeaLLM.

The most important observation from Figure 2 is that chain-of-thought prompting leads to the highest hallucination rates across all models. While CoT prompting generally helps in reasoning and logical tasks, in this case it increases hallucination due to the factual nature of the dataset.

Another inference from the plot is that hallucination rates decrease as the parameter size of the model increases, which is intuitive. SeaLLM 1.5B produces the highest hallucination rates, whereas Gemini-2.5 shows the lowest, owing to its large number of training parameters.

Figure 3 validates the inverse relationship between accuracy and hallucination rates across models and different prompting strategies.

Figure 4 compares each model's accuracy under different prompting strategies. It is clearly visible that Gemini-2.5 performs the best compared to all other models. However, it was expected that performance would improve monotonically from zero-shot to CoT-few-shot, with the order CoT-few-shot > few-shot > one-shot > zero-shot. In practice, the results are non-linear, and many models achieve their highest accuracy with one-shot prompting, including Gemini-2.5, Llama 70B, and SeaLLM 1.5B. For Gemma-4B, Gemma-12B, and Llama 8B, few-shot prompting provides the best results.

6 Conclusion

Our research demonstrates the performance of language models on domain-specific, low-resource datasets under different prompting strategies. While it was expected that chain-of-thought prompting would provide the highest accuracy and lowest hallucination rates, it in fact performed the worst. Instead, one-shot and few-shot prompting strategies yielded the most reliable results.

To address the research questions posed in Section 3:

- The baseline accuracy across all models and prompting strategies averages between 35–40%.
- One-shot and few-shot prompting strategies consistently provide the best results. This is likely because they strike a balance: they give models minimal guidance without overwhelming them with examples. In contrast, CoT and higher-example few-shot prompting can introduce unnecessary complexity, increasing the risk of overfitting or generating hallucinations.
- Semantic similarity based evaluation confirms the overall ranking of model performance, with Gemini-2.5 achieving the best results. However, relying solely on semantic similarity to detect hallucinations is insufficient, as it primarily captures lexical or semantic closeness but fails to penalize factual inaccuracies. Complementary evaluation methods, such as human evaluation or factual consistency checks, are necessary for a complete picture.

Overall, this study highlights the need to further develop language models for low-resource languages. Creating models that are both up-to-date and accurate is essential for maximizing their societal benefits.

Limitations and Future Work

This work presents preliminary results on the performance of language models with respect to hallucination in Q&A for the low-resource language Burmese. The key limitations of this study are as follows:

- At present, all models have been evaluated only on Burmese Q&A. We plan to extend the evaluation to additional low-resource languages.
- Our analysis has been restricted to models accessible through free APIs.
- The scope of the current dataset is limited to the MPox domain. Due to rate limits imposed by free APIs, we were unable to evaluate models on other domain-specific datasets, such as those in agriculture and medicine.

Acknowledgments

This research received no financial support. The work was carried out independently by the listed authors.

References

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, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, and 1 others. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*.

Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. 2025. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination. *Preprint*, arXiv:2311.15548.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Eliza Mik. 2024. Caveat lector: Large language models in legal practice. *Preprint*, arXiv:2403.09163.

MinSiThu. Mpox-myanmar.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Pritika Rohera, Chaitrali Ginimav, Gayatri Sawant, and Raviraj Joshi. 2025. Better to ask in english? evaluating factual accuracy of multilingual llms in english and low-resource languages. *Preprint*, arXiv:2504.20022.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Konstantinos C Siontis, Zachi I Attia, Samuel J Asirvatham, and Paul A Friedman. 2023. Chatgpt hallucinating: can it get any more humanlike? *European Heart Journal*, 45(5):321–323.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball. *Preprint*, arXiv:2305.13534.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong QingqingLong, Yefeng Zheng, and Xian Wu. 2025. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, Vienna, Austria. Association for Computational Linguistics.

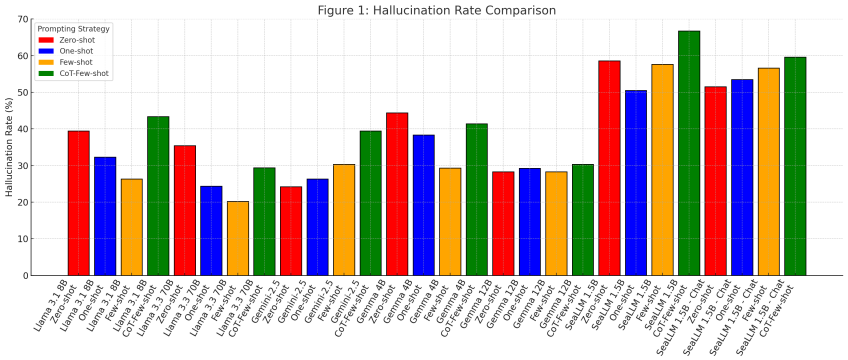

Figure 2: Bar chart for hallucination rate comparison across all models and prompting conditions.

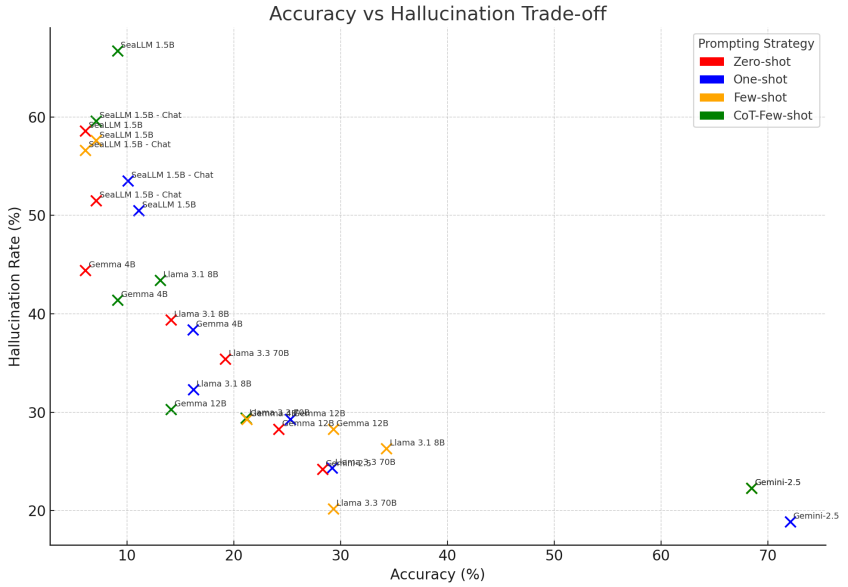

Figure 3: Scatter plot for accuracy vs. hallucination trade-off.

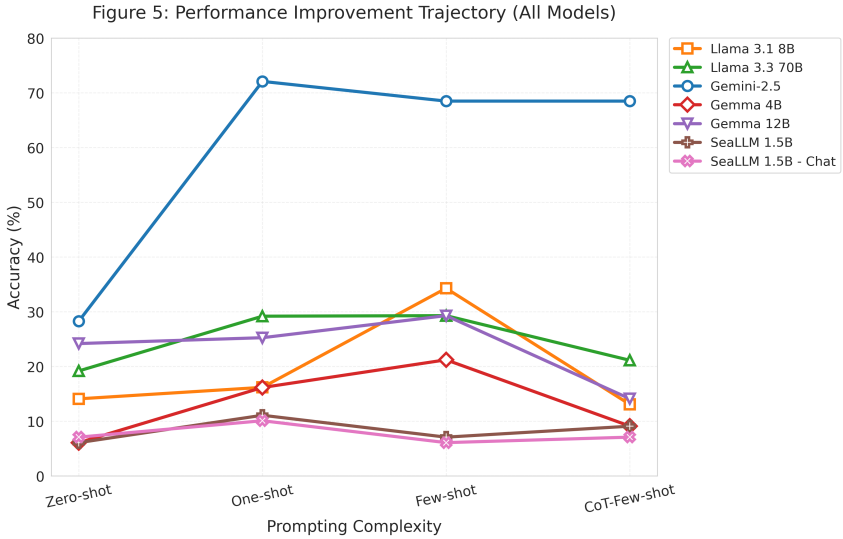

Figure 4: Line chart comparing accuracy of all models across different prompting conditions.

Prompt = f"""Question: {question} Answer the question in Burmese language. Be accurate and concise."""

Figure 5: Zero-shot prompting template.

Prompt = f"""You are being asked questions about Mpox (monkeypox). Answer accurately in Burmese language based on the information you have.

Important guidelines:
- For factual questions with definite answers: Be specific, detailed, and descriptive
- For questions about preventive measures or recommendations: Only provide information you are certain about
- Do not guess or provide uncertain information

Here are some examples:

Question: ကလေးငယ်များ မျောက်ကျောက်ရောဂါ ကူးစက်ခံရနိုင်လား?
Answer: ကလေးငယ်များသည် လူငယ်လူရွယ်များနှင့် လူကြီးများထက် ရောဂါလက္ခဏာ ပိုမိုပြင်းထန်စွာ ခံစားရနိုင်ပါသည်။ ရောဂါပိုးသည် သန္ဓေသား သို့မဟုတ် မွေးကင်းစ ကလေးငယ်မွေးဖွားသည့်အချိန် သို့မဟုတ် မွေးကင်းစကလေးငယ်များအား ကနဦး ကိုင်တွယ်ထိတွေ့ခြင်းစသည်တို့မှ ကူးစက်နိုင်ပါသည်။

Now answer this question:
Question: {question}
Answer:"""

Figure 6: One-shot prompting template.

Prompt = f"""You are being asked questions about Mpox (monkeypox). Answer accurately in Burmese language based on the information you have.

Important guidelines:
- For factual questions with definite answers: Be specific, detailed, and descriptive
- For questions about preventive measures or recommendations: Only provide information you are certain about
- Do not guess or provide uncertain information

Here are some examples:

Question: ကလေးငယ်များ မျောက်ကျောက်ရောဂါ ကူးစက်ခံရနိုင်လား?
Answer: ကလေးငယ်များသည် လူငယ်လူရွယ်များနှင့် လူကြီးများထက် ရောဂါလက္ခဏာ ပိုမိုပြင်းထန်စွာ ခံစားရနိုင်ပါသည်။ ရောဂါပိုးသည် သန္ဓေသား သို့မဟုတ် မွေးကင်းစ ကလေးငယ်မွေးဖွားသည့်အချိန် သို့မဟုတ် မွေးကင်းစကလေးငယ်များအား ကနဦး ကိုင်တွယ်ထိတွေ့ခြင်းစသည်တို့မှ ကူးစက်နိုင်ပါသည်။

Question: မျောက်ကျောက်ရောဂါကို ဘယ်နေရာမှာ စတင်တွေ့ရှိခဲ့သလဲ?
Answer: ဒီနှစ်မတိုင်ခင် ကိုပင်ဟောင်ဖြူရှိ ဓာတ်ခွဲခန်းတွင် စတင်တွေ့ရှိခဲ့သည်။

Now answer this question:
Question: {question}
Answer:"""

Figure 7: Few-shot prompting template.

You are being asked questions about Mpox (monkeypox). Answer accurately in Burmese language based on the information you have.

Important guidelines:
- Show your reasoning process step-by-step before giving the final answer
- For factual questions with definite answers: Be specific, detailed, and descriptive
- For questions about preventive measures or recommendations: Only provide information you are certain about
- Do not guess or provide uncertain information

Here are examples showing the reasoning process:

Question: ကလေးငယ်များ မျောက်ကျောက်ရောဂါ ကူးစက်ခံရနိုင်လား?

Reasoning: ဤမေးခွန်းကို ဖြေရန် အောက်ပါအချက်များကို စဉ်းစားရမည်-
1. ကလေးငယ်များသည် ရောဂါကူးစက်မှုအတွက် အန္တရာယ်ရှိသည်။
2. ကလေးငယ်များတွင် ကိုယ်ခံအားစနစ် အားနည်းနေသောကြောင့် ရောဂါလက္ခဏာ ပိုမိုပြင်းထန်နိုင်သည်။
3. ကူးစက်မှုလမ်းကြောင်းများမှာ မိခင်မှ သန္ဓေသားသို့၊ မွေးဖွားစဉ်၊ နှင့် ထိတွေ့မှုတစ်ဆင့် ဖြစ်နိုင်သည်။

Answer: ကလေးငယ်များသည် လူငယ်လူရွယ်များနှင့် လူကြီးများထက် ရောဂါလက္ခဏာ ပိုမိုပြင်းထန်စွာ ခံစားရနိုင်ပါသည်။ ရောဂါပိုးသည် သန္ဓေသား သို့မဟုတ် မွေးကင်းစ ကလေးငယ်မွေးဖွားသည့်အချိန် သို့မဟုတ် မွေးကင်းစကလေးငယ်များအား ကနဦး ကိုင်တွယ်ထိတွေ့ခြင်းစသည်တို့မှ ကူးစက်နိုင်ပါသည်။

Question: မျောက်ကျောက်ရောဂါကို ကာကွယ်နိုင်တဲ့ ကာကွယ်ဆေးရှိသလား?

Reasoning: ကာကွယ်ဆေးများအကြောင်း စဉ်းစားရာတွင် - (၁) ကျောက်ကြီးရောဂါ ကာကွယ်ဆေးများ အကျိုးပြုနိုင်သည်၊ (၂) အထူးပြု ကာကွယ်ဆေးအသစ်များ ရရှိနိုင်သည်၊ (၃) ကာကွယ်ဆေးများကို WHO မှ အတည်ပြုထားသည်။

Answer: ကျောက်ကြီးရောဂါ (smallpox) ကာကွယ်ဆေးတော်တော်များများသည် မျောက်ကျောက်ရောဂါ ကူးစက်မှုမှ တစ်စိတ်တစ်ဒေသကာကွယ်မှု ပေးနိုင်ပါသည်။ ကျောက်ကြီးရောဂါကို ကာကွယ်ရန်အတွက် ထုတ်လုပ်ထားသည့် ကာကွယ်ဆေးအသစ်တစ်မျိုး (MVA-BN, Imvamune, Imvanex or Jynneos) သည် မျောက်ကျောက်ရော ဘ်ပါ ကူးစက်ခြင်းမှကာကွယ်ရန် ၂၁ခု ခုနှစ်တွင် ခွင့်ပြုချက်ရရှိထားပါသည်။

Now answer this question by showing your reasoning first, then providing the final answer:
Question: {question}

Figure 8: Chain-of-thought few-shot prompting template.