

Empowering Thai Question Answering: Leveraging Context Generation with Large Language Models

Nipun Angkavichai, Theera-anuchit Chalapinyo

Language and Information Technologies,

Faculty of Arts, Chulalongkorn University

6440126922@student.chula.ac.th, 6440109222@student.chula.ac.th

Abstract

Recently, Large Language Models (LLMs) have found widespread application across various tasks. In the context of Question Answering (QA), previous studies have explored extracting knowledge stored in LLMs using traditional methods such as few-shot and zero-shot prompting. Notably, these investigations have revealed that the performance of context generation methods surpasses that of traditional approaches. In this research, our focus is on enhancing the Thai QA system by examining LLM-stored Thai knowledge. We employ the context generation method and establish baselines comparable to traditional methods for both closed-book question answering and reading comprehension tasks, as no dedicated Thai QA benchmarks currently exist. Our findings indicate nuanced LLM performance as an extractor compared to the traditional finetuning method in reading comprehension tasks. For closed-book question answering tasks, our adapted context generation, inspired by prior research, exhibits outstanding performance across all baseline models in terms of F1-score and Exact Match. The optimal LLM performance achieved is an F1 score of 0.19 and a human evaluation score of 0.24. These results underscore the significant potential for future work in leveraging LLMs for the advancement of Thai QA systems. Additionally, human evaluation metrics reveal that these models were underestimated in previous research, with a performance increase of nearly +30% in reading comprehension tasks.

1 Introduction

In the rapidly evolving nature of Natural Language Processing (NLP), the challenge for enhancing systems in non-English languages remains a field awaiting comprehensive exploration, especially in under-resourced languages like Thai. Although the Thai NLP community has achieved a significant achievement over the past decades, particularly on

NLP upstream tasks such as tokenization, research on downstream tasks such as syntactic parsing and question answering (QA) is still limited. Unlike English having a standard dataset for evaluating, SQuAD, and researching English QA, Thai QA systems lack various explorations as well as researchers and a standard dataset. Mostly because of focusing on using QA systems for application and practical use, Thai QA has not been improving or researching so much. Our study tries to be a pioneer work for Thai QA and to improve Thai QA by using adapted methods that have been done in English before. To initiate this pursuit, we have laid groundwork by establishing baselines using a reading comprehension method. Through comparative analyses with proposed methodologies, we aim to pave the way for enhanced Thai QA systems. Our proposed method based on (Su et al., 2023) finding Context Generation and Answer Prediction (CGAP). Having the assumption that techniques proven effective in English could potentially apply to other languages, including Thai, is applied to this study. By leveraging established methods and adapting them to suit the linguistic nuances of Thai, there's hope for significant progress in Thai QA. For adapting the previous method to suit Thai language, we use GPT, a Thai-understanding Language Model (LLM) and Thai language prompt engineering, to generate contexts and answers. The goal is to demonstrate how established methods, when customized to Thai using GPT, can significantly improve Thai QA systems. Our adapted CGAP extract the relevant knowledge for creating context for each question and use generated context to answer a question. While the outcomes may lack surprise and require further development, this study is viewed as an encouraging foundation for the continued development of Thai QA.

In short, this paper makes the following contributions:

- We propose adapted CGAP that works better

than both zero shot and few shot in closed-book tasks but doesn't achieve a state-of-the-art for Thai QA task, reading comprehension by BERT is still the best.

- We find that, after employing our human evaluation metric, all models are underestimated when utilizing traditional metrics like Exact Match and F1 Score.
- We reveal that LLMs, specifically GPT-3.5, still lack Thai knowledge as we extract in-stored information using our adapted Context Generation and Prompting (CGAP), along with few-shot and zero-shot techniques.

2 Related works

To the best of our knowledge, no prior research has established a benchmark for Question Answering (QA) tasks in the Thai language. However, contemporary Language Model (LLM) approaches, exemplified by ChatGPT, have gained widespread popularity and adoption across diverse industries. Numerous studies have explored the utilization of LLM capabilities, particularly in closed-book QA tasks, leveraging the stored knowledge through in-context learning methods such as few-shot and zero-shot prompting.

Context Generation Improves Open Domain Question Answering (Su et al., 2023) is a closed-book QA framework which employs a coarse-to-fine approach to extract the relevant knowledge and answer a question. Their results on three QA benchmarks show that this method significantly outperforms previous closed-book QA methods either directly finetunes or prompts a pretrained language model (LM) to leverage the stored knowledge. We aim to experiment this method compared to previous methods, few-shot prompting and zero-shot prompting, to exploit the stored knowledge.

Evaluating Open-Domain Question Answering in the Era of Large Language Models (Kamalloo et al., 2023) is a research on evaluating the performance of open-domain question answering (QA) models, especially LLMs by manually evaluating their answers on a subset of Natural Question-OPEN, a popular benchmark. They find that all models are significantly underestimated. Since the current evaluation method, lexical matching, is not accurate enough for LLMs, as it fails to account for semantically equivalent answers and long-form answers. Our study adopts a similar methodology

to assess the performance of our Thai QA system, aiming to determine if similar challenges persist.

3 Problem statement

We aim to investigate various methods of performance to extract in-stored knowledge within LLMs to examine its ability for question answering tasks without external knowledge. Our main task for using LLM is Prompt engineering.

Prompt engineering is central to our experimental approach in harnessing knowledge within LLMs. Prompt engineering can be defined as the deliberate crafting and refinement of input prompts to elicit desired responses from the language model. It involves formulating prompts in a way that maximizes the model's ability to generate accurate and relevant answers based on its internal knowledge. Essentially, Prompt engineering serves as a key methodology for optimizing the performance of LLMs in specific tasks.

Few-shot prompting involves adapting a language model by presenting it with a small number of task-specific examples, allowing it to generalize and perform related tasks. This technique is particularly beneficial when labeled data for a specific task is limited, enabling the model to adapt its behavior beyond its initial pre-training.

Zero-shot prompting instructs a model to execute a task without providing any task-specific examples. In this scenario, the model relies solely on its pre-existing knowledge acquired during pre-training to infer and generate responses for the given task.

The challenge arises from the token limit constraints imposed by the GPT-3 API ('gpt3.5-turbo'), affecting both input and output token capacities. To adhere to the API constraints, we must restrict the number of input tokens in the few-shot prompting process. Consequently, this limitation necessitates a reduction in the number of samples within the prompt, thereby impacting the performance of few-shot prompting, as observed in prior research that more samples in the prompt increases the performance of few-shot prompting.

4 Methodology

For each question in the test set, three question-context-answer samples are extracted from the training set to form prompts for context generation, answer prediction, and answer extraction. To identify relevant samples for specific test set ques-

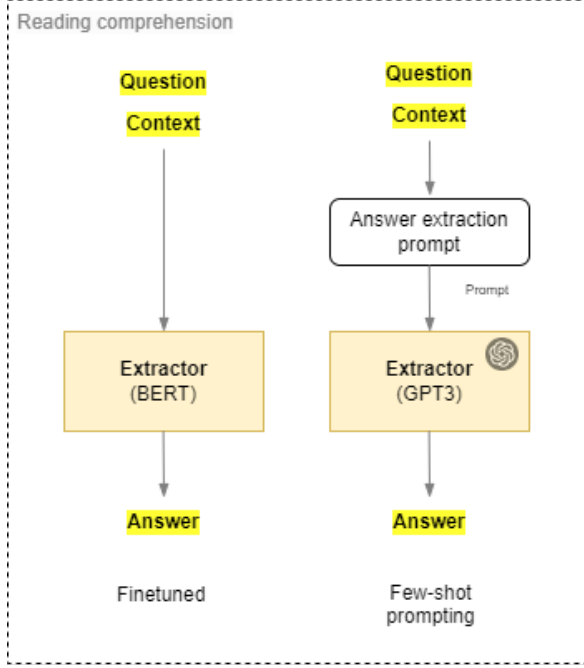


Figure 1: illustration of the architectural overview of our experiment aimed at enhancing the Thai QA system for Reading Comprehension QA tasks. The experiment comprises two methods: the first method employs BERT as an extractor, while the second method leverages a few-shot prompting Large Language Model (LLM), specifically GPT-3, as an extractor.

tions, concatenation of each training set question and context is performed. The sentence transformer model is then employed to embed the concatenated string, along with embedding each query question from the test set. Subsequently, cosine similarity is utilized to compare the embedding similarity score, facilitating the selection of the three most relevant questions, contexts, and answers for each test set question.

$$Score(Q, (q_i; c_i)) = BERT(Q)BERT(q_i; c_i)$$

In our exploration of leveraging Large Language Models (LLM) for the Thai QA system, we conduct experiments in two QA tasks: 1) Reading Comprehension and 2) Closed-book Question Answering.

In Figure 1, dedicated to Reading Comprehension, we employ the Few-shot prompting method for the LLM approach. We construct answer extraction prompts using a single question-context-answer retrieved from relevant samples in the early stage, as illustrated in Table 2.

In Figure 2, pertaining to the Closed-book QA task, the Adapted CGAP method utilizes two relevant question-context samples for the context generation prompt, as demonstrated in Table 3. For

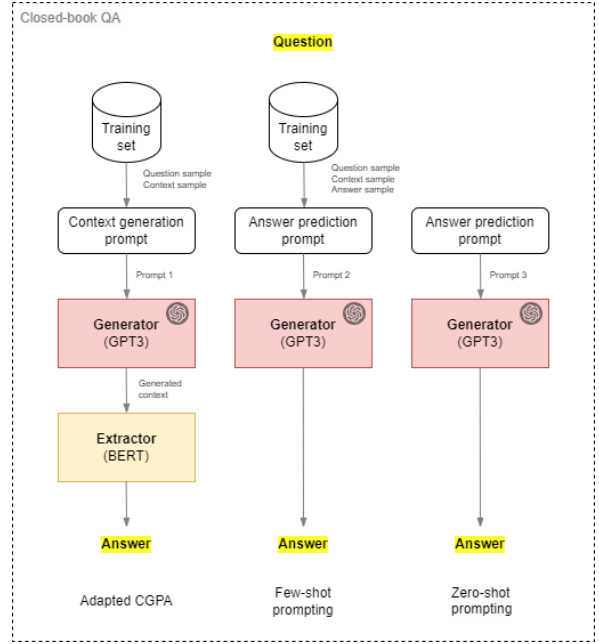


Figure 2: illustration of the architectural overview of our experiment aimed at enhancing the Thai QA system for Closed-book QA tasks. The experiment encompasses three methods: the first method involves adapted CGAP, utilizing LLM (GPT3.5) for context generation and BERT as an extractor for answer extraction; the second method employs few-shot prompting, utilizing question and context pairs from the training set as samples in the prompt to obtain answers from LLM; and the third method employs zero-prompt, incorporating instructions for answering questions.

Few-shot prompting, we incorporate three relevant question-answer samples in the answer prediction prompt, as shown in Table 4. Additionally, for Zero-shot prompting, only an instructional prompt on how to answer a specific question is utilized in the answer prediction prompt, illustrated in Table 5.

5 Experiment

5.1 Dataset

We select questions from the Open Thai Wikipedia QA Dataset created by iApp Technology, which comprises 1,961 documents and 9,170 questions. The dataset includes 739 questions in the test set and 5,761 questions in the training set.

5.2 Preprocessing method

For our experiment’s test set, we systematically assessed the token count of each context generation prompt to accommodate the API’s maximum token limit. Utilizing the tiktoken library with "cl100k-base" encoding to align with GPT-3’s tokenizer, we specifically selected instances with a token count not exceeding 2300, eliminating potential outliers. This process resulted in a curated test set comprising 563 questions. In the construction of our training set for the task of retrieving relevant information pertaining to specific questions within the test set, we employ a few-shot prompting approach for both context generation and answer extraction prompts. To enhance diversity, we identify and remove shared gold contexts among different questions, resulting in a refined training dataset. The dataset now encompasses a reduction from 5761 questions (old) to 1563 questions (new), each associated with distinct contexts.

5.3 Experimental conditions

We hypothesize that Few-shot prompting with Large Language Models (LLM) serves as an effective extractor, leveraging its capability to comprehend human language structure and provide natural responses by incorporating only pertinent information. In our methodology, BERT is chosen as the baseline for comparative analysis against our Few-shot prompting approach.

Reading comprehension: We integrate mDeBERTaV3, a pre-trained multilingual model introduced by Pengcheng He et al. (2023), known for its substantial enhancements over baseline models, particularly in cross-lingual tasks. So we use this

model for reading comprehension tasks and initial baseline.

LLMs: GPT3.5 We leverage the Large Language Model (LLM) GPT3.5 as a foundational baseline, employing it as an extractor similar to a reading comprehension task. The GPT3.5 model serves as a fundamental reference point for our Thai QA experiments due to its extensive pre-training on diverse language data. With the use of LLM GPT3.5, we are able to evaluate the effectiveness and possible areas for development of specific Thai QA models.

For our proposed method in Closed-book QA, we have a hypothesis that the Adapted CGAP framework might be the best method for leveraging stored knowledge in LLM compared to traditional approaches using Few-shot prompting and Zero-shot prompting. The Few-shot prompting technique serves as a robust baseline, exhibiting notable performance, while the Zero-shot prompting method demonstrates a weaker baseline.

Adapted CGAP: We propose Adapted CGAP framework might be the best method for leveraging stored knowledge in LLM. Because of the limitations, Adapted CGAP is derived from a preceding model by Su et al. (2023). This model operates by retrieving relevant question-context pairs from the dataset, then generating context for each question, and subsequently employing BERT to extract answers from the synthesized context.

Zero shot: Our experimentation with Zero-Shot Prompting involves a scenario where the model is tasked with answering questions without specific training or exposure to related context. This setup might show the model’s innate ability to comprehend and answer questions without any prior tailored training. The Zero-Shot Prompting approach serves as a weaker baseline in Closed-book QA, highlighting the model’s performance limitations in some context.

Few shot: In our experimental setup for Few-Shot Prompting, we employ a strategy where the model receives a limited number of context-question-answer examples for answering questions by retrieving the most relevant examples from the training set. This approach serves as a robust baseline in Closed-book QA, leveraging a restricted but informative set of samples to prompt the model’s understanding and ability to generate accurate responses.

5.4 Implementation

We conducted our implementation on a free Google Colab account, encompassing code development for data processing and API completion using the OpenAI library. For GPT-3 settings, we configured the temperature parameter to zero through the OpenAI library to achieve a focused and deterministic output. The model employed was 'gpt-3.5-turbo.' To retrieve relevant question-context-answer samples from the training set for specific test set questions, we utilized the 'new5558/simcse-model-wangchanberta-base-att-spm-uncased' sentence transformer. This model was applied to embed query questions from the test set and question-context-answer pairs from the training set. The relevance scores were determined using cosine similarity, facilitated by the 'sentence-transformers' libraries, along with the 'sklearn' library for importing the cosine-similarity module.

5.5 Evaluation

F1-Score: Precision and recall are implemented by measuring the overlap between predicted answers and gold standard answers.

$$F1 \text{ Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Exact match: A candidate answer is marked as correct if it can be found in the gold answers. The ratio of correct answers is reported as

$$EM \text{ accuracy} = \frac{\text{All correct answers}}{\text{All answers}}$$

Human evaluation: To assess model performance in closed-book question answering within the open-domain, we utilize a generative model tailored for the Thai language. Conventional metrics like exact match and F1 score might be limited, particularly when candidate answers slightly deviate from the gold answer. To be more close to true performance, we manually evaluate the answers against both the gold answer and provided context. This assessment involves two native Thai speakers, and correctness is determined based on the acceptability of the answer within the context of Thai knowledge and culture, allowing for flexibility beyond strict gold answer matching.

6 Results

In Table 1, the bold texts are the best accuracy scores overall. The fine tuned BERT model performs reasonably well, showcasing an F1 score of

0.566, which is the highest F1 score overall, and an Exact Match of 0.223. Comparatively, the LLM in a few-shot learning demonstrates a slightly lower F1 score of 0.560 but a higher, the highest, Exact Match at 0.325. In human evaluation, the reading comprehension performance remains notably high, with both BERT and LLM achieving scores 0.932 and 0.824 respectively, highlighting their ability to comprehend and respond to the provided context effectively. The underlined text is best accuracy scores from the closed-book method. We see that the proposed method achieves the highest score in both F1 score and human evaluation, but the highest Exact Match is Few-shot, 0.026. When we compared two QA tasks, we found that no Closed-book method achieved a higher score in every way than the reading comprehension task. Even Though the Adapted CGAP gets the highest score in F1 score and human evaluation among Closed-book tasks, it is still far from those in reading comprehension tasks. This outcome does, however, still have some interesting details, which will be discussed in the next section.

7 Discussion

Unlike the CGAP framework, which extensively employs Large Language Models (LLM) to generate multiple contexts and subsequently utilizes few-shot prompting for answer generation, our approach, Adapted CGAP, diverges. Upon discovering that GPT-3.5 has limited proficiency in the Thai language compared to English, we modified the methodology. Instead of generating multiple contexts, we generate only one context in the early stage. Given the comparable performance results between using LLM and BERT as an extractor, we opt for BERT to extract answers from the generated context. This modification is cost-efficient, defining our proposed method as Adapted CGAP.

Bert's superior performance in the reading comprehension section wasn't unexpected given its established proficiency in Thai language tasks. The interesting part, however, is that our suggested model can perform not bad, especially having scores higher than those in zero-shot and few-shot settings, which would be first steps toward Thai closed-book QA development.

Furthermore, we observed a significant increase in the performance of all models, particularly in reading comprehension tasks, with a nearly +30% improvement when assessed using human evalu-

QA task	Context passage	Extractor	Method	Exact Match	F1 Score	Human Evaluation
Reading comprehension	gold given context	BERT	Fintuned	0.223	0.566	0.932
	gold given context	LLM	Few-shot	0.325	0.560	0.824
Closed-book	generated context	BERT	Adapted CGAP	0.010	<u>0.193</u>	<u>0.243</u>
			Zero-shot	0.000	0.121	0.177
			Few-shot	<u>0.026</u>	0.143	0.154

Table 1: Performance of Each Method in Reading Comprehension and Closed-Book Question Answering Tasks

ation metrics. In our adapted CGAP, the performance increased from 0.193 to 0.243, marking the best performance in closed-book QA tasks. This aligns with prior research on evaluation of Question answering task in English language (Kamalloo et al., 2023), highlighting the limitations of exact match and F1 score in accurately evaluating the true performance of LLM-generated answers.

Our employed Large Language Model (LLM), GPT-3.5, boasts over 175 billion parameters and underwent training on a diverse dataset, incorporating code and text, up to Q4 2021. The training spanned hundreds of billions of words across 95 languages, including Thai, albeit with a relatively smaller dataset compared to major languages such as English, Chinese, and Russian. Consequently, the stored knowledge in Thai language for open-domain question answering is limited, leaving ample room for improvement. Future endeavors could explore fine-tuning LLMs with additional Thai language data to enhance their proficiency in handling open-domain questions.

7.1 Error analysis

We examined 563 questions in our test set and identified 38 questions requiring additional context in the question to answer. We labeled these as incomplete questions and excluded them from our error analysis. In Figure 3, questions are categorized into four types:

- **Foreign knowledge:** External knowledge questions from other countries, such as "What season does South Korea compete in the Olympics?" or "What nationality was Michael Jackson a singer from?"
- **Thai knowledge:** Internal knowledge ques-

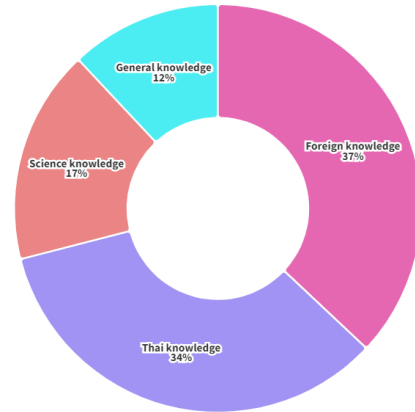


Figure 3: The percentage of required knowledge types to answer each question in our test set.

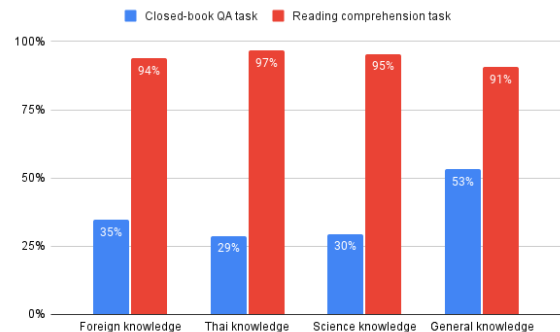


Figure 4: Statistics of corrected answer provided by human evaluation for required knowledge types in both closed-book QA and reading comprehension task.

tions specific to Thailand, requiring experience in Thailand to answer, such as "What does Phra Buddha Mani Rattana Patimakorn look like?" or "Thongchai McIntyre is a singer under what record label?"

- **Science knowledge:** Science-related questions, for example, "Where do anthozoans live?" or "What is the scientific name of the lemongrass plant?"
- **General knowledge:** Generic questions like "What do you browse the website through?" or "What instruments does punk rock consist of?"

In Figure 4, we examine the corrected answers provided by human evaluation for each question type in both closed-book QA and reading comprehension tasks. For reading comprehension tasks, extractive answers are mostly correct across all question types, with a percentage exceeding 90%. In closed-book QA tasks, when considering a model correct if any of the models in our experiment deems it so, we observe that the highest percentage of corrected answers is for questions using general knowledge (53%), followed by foreign knowledge questions (35%), science knowledge questions (30%), and Thai knowledge questions (29%) as the least corrected. This highlights that LLM, GPT-3.5, still lacks knowledge in Thai, particularly in closed-book QA tasks that aim to extract in-stored knowledge from the LLM without relying on external sources. Conversely, LLM performs significantly better for general questions that do not require specific knowledge or experience.

7.2 Limitation

Constrained by the maximum token limit of GPT-3.5, set at 4096 tokens, we are compelled to restrict the examples in both context generation and answer extraction prompts for few-shot prompting. While prior research (Su et al., 2023) demonstrates that increasing examples enhances performance, our budget constraints prevent us from maximizing the CGAP framework.

8 Conclusion

In this paper, we propose leveraging LLM to improve closed-book question answering for the Thai QA System. We use Adapted Context Generation and Answer Prediction (CGAP), using BERT to extract answers from the generated context by LLM.

Even though our model does not achieve the state-of-the-art overall, it still gets the highest scores in F1 and Human evaluation in closed-book question-answering tasks. In future directions, there is a lot to do to improve the Thai QA system to be equal to English. A key challenge lies in addressing the lack of comprehensive datasets and robust Language Model (LLM) frameworks for Thai. If we can develop in this part, it will have a good consequence for Thai question-answering tasks.

References

- Ehsan Kamaloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. [Evaluating open-domain question answering in the era of large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.
- Dan Su, Mostofa Patwary, Shrimai Prabhumoye, Peng Xu, Ryan Prenger, Mohammad Shoeybi, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2023. [Context generation improves open domain question answering](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 793–808, Dubrovnik, Croatia. Association for Computational Linguistics.

9 Example

<p>Question: ปลาเค้าดำเป็นปลาน้ำจืดที่จัดอยู่วงศ์ของพันธุ์ปลาชนิดใด Gold answer: วงศ์ปลาเนื้ออ่อน (Siluridae)</p>
<p>Question: ปลาแปบควายเป็นปลาที่จัดอยู่ในวงศ์ใด</p> <p>Context: ปลาแปบควาย เป็นชื่อสกุลของปลาน้ำจืดจำพวกปลาแปบ อยู่ในวงศ์ปลาตะเพียน (Cyprinidae) วงศ์ย่อย Cultrinae ใช้ชื่อสกุล Paralaubuca มีรูปร่างโดยรวมคือ มีรูปร่างยาว ลำตัวแบนข้างมาก ปากเล็ก ตาโต ท้องเป็นสันคม ด้านท้องค่อนข้างกว้างออก ครีบออกยาว ครีบหางเว้าลึก เกล็ดเล็กละเอียดสีเงินแวววาว เส้นข้างลำตัวไม่ต่อเนื่องกัน เกล็ดแถวที่อยู่หน้าครีบหลังไม่เลยเข้ามาถึงระหว่างนัยน์ตา เกล็ดตามเส้นข้างลำตัวมีจำนวน 50–85 แถว มีขนาดโตเต็มที่โดยเฉลี่ย 15 เซนติเมตร มักอยู่เป็นฝูงใหญ่ มีการอพยพขึ้นลงตามแม่น้ำเพื่อวางไข่และหากินเป็นฤดูกาล มักอยู่ในแม่น้ำสายใหญ่ที่ไหลเชี่ยว หากินใกล้ผิวน้ำ พบตั้งแต่แม่น้ำโขงถึงแม่น้ำเจ้าพระยาไปจนถึงมาเลเซีย กินแมลงหรือสัตว์น้ำขนาดเล็กเป็นอาหาร ซึ่งสำหรับในประเทศไทย ปลาแปบควายทั้ง 3 ชนิด ถูกจับขึ้นมาเพื่อทำปลาแห้ง, ปลาร้า และบริโภคโดยปรุงสด อีกทั้งยังเลี้ยงเป็นปลาสวยงามได้อีกด้วย โดยมีชื่อเรียกอื่นอีกว่า "ปลาท้องพลู" และในภาษาอีสานเรียกว่า "ปลาแดบ", "ปลาแดบขาว" หรือ "ปลามะแปบ" เป็นต้น</p> <p>Answer: วงศ์ปลาตะเพียน (Cyprinidae)</p>
<p>จงดตอบคำถาม จาก Question เหมือนตัวอย่างข้างบน โดยตอบแค่คำถามที่ถามเท่านั้น ไม่จำเป็นต้องให้ข้อมูลอื่นประกอบ และไม่ต้องมีครับ/ค่ะลงท้าย</p>
<p>Question: ปลาเค้าดำเป็นปลาน้ำจืดที่จัดอยู่วงศ์ของพันธุ์ปลาชนิดใด</p> <p>Context: ปลาเค้าดำ หรือ ปลาเค้าดำ (ชื่อวิทยาศาสตร์: Wallagonia micropogon) เป็นปลาน้ำจืดชนิดหนึ่ง อยู่ในวงศ์ปลาเนื้ออ่อน (Siluridae) จัดเป็นปลาที่อยู่ในอันดับปลาหนัง (Siluriformes) ปลาเค้าดำมีรูปร่างลำตัวยาวแต่ค่อนข้างป้อม ลำตัวที่อยู่ตอนไปทางหางมีลักษณะแบนข้างมาก พื้นลำตัวสีเทาถึงดำสนิท ส่วนหัวมีขนาดใหญ่และแบน ปากกว้าง ภายในมีฟันซี่เล็ก ๆ อยู่ในขากรรไกรทั้ง 2 ข้าง มีหนวด 2 คู่ โดยคู่ที่อยู่มุมปากมีลักษณะเรียวยาว ส่วนคู่ที่ได้คางจะสั้นและเล็กมาก มีครีบทั้งหมด 7 ครีบ ครีบกันใหญ่และยาวจรดครีบหาง ขอบปลายหางด้านบนจะใหญ่กว่าด้านล่าง ส่วนท้องป่องออก ส่วนหลังยกสูงชันกว่าปลาเค้าขาว (Wallago attu) ซึ่งเป็นปลาที่มีลักษณะใกล้เคียงกัน แต่จัดอยู่ในคนละสกุล มีขนาดลำตัวยาวได้ถึง 1 เมตร น้ำหนักกว่า 50 กิโลกรัม โดยอาจยาวได้ถึงกว่า 2 เมตร น้ำหนัก 80 กิโลกรัม สถิติที่ใหญ่ที่สุดพบที่ปากน้ำโพ จังหวัดนครสวรรค์ พญศิริกรรมตามปกติ มักจะซุกตัวอยู่หนึ่ง ๆ ใต้น้ำ เป็นปลาที่สายตาไม่ดี จึงใช้หนวดในการนำทางและหาอาหาร พบตามแม่น้ำสายใหญ่ ๆ ทั้งภาคกลางและภาคอีสานรวมทั้งภาคใต้ของประเทศไทย เช่น แม่น้ำเจ้าพระยาและสาขา, แม่น้ำโขงและสาขา, แม่น้ำตาปี่รวมทั้งที่ทะเลสาบสงขลาด้วย เป็นต้น ปลาเค้าดำ มีสถานภาพในปัจจุบันใกล้สูญพันธุ์อีกชนิดหนึ่งในธรรมชาติ แต่สามารถเพาะขยายพันธุ์ได้โดยสถานีประมงน้ำจืดจังหวัดสุราษฎร์ธานี ในปี พ.ศ. 2534 โดยได้มีการปล่อยลูกปลาที่เพาะได้กลับคืนถิ่นธรรมชาติ</p> <p>Answer:</p>
<p>Extractive answer: วงศ์ปลาเนื้ออ่อน (Siluridae)</p>

Table 2: Example illustrates answer extraction prompt, in which we use only one sample of question-context-answer for few-shot prompting in Reading comprehension task.

Question: แบคบอยส์ คู่หูขวางนรก 2 เป็นภาพยนตร์แนวอะไร Gold answer: แอคชั่น-คอมดี้
Question: ไอ้แมงปอแมน อีโร่ซูเปอร์ริ้ว เป็นภาพยนตร์แนวอะไร
Context: ไอ้แมงปอแมน อีโร่ซูเปอร์ริ้ว เป็นภาพยนตร์แนวตลกล้อเลียนหนังดัง ๆ หลายเรื่องโดยเนื้อเรื่องเล่าถึง ริกหนุ่มน้อยหน้าละอ่อนจอมซุ่มซ่ามที่บังเอิญโดนแมลงปอในห้องวิทยาศาสตร์ก่อกวนจึงทำให้เขามีพลังเหนือมนุษย์เพื่อคอยปราบเหล่าร้าย แต่เพราะความซุ่มซ่ามของเขาจึงทำให้เกิดเรื่องฮา ๆ ขึ้นมาอีกเป็นขบวน
Question: บุญชู 2 น้องใหม่ เป็นภาพยนตร์ประเทศอะไร
Context: บุญชู 2 น้องใหม่ หรือ บุญชู ภาค 2 เป็นภาพยนตร์ไทย ออกฉายเมื่อปี พ.ศ. 2532 กำกับโดย บัณฑิต ฤทธิ์ถกล โดยค่ายไฟว์สตาร์โปรดักชั่น นำแสดงโดย สันติสุข พรหมศิริ จินตหรา สุขพัฒน์ ส. อาสนจินดา สุเทพ ประยูรพิทักษ์ ญาณี จงวิสุทธิ์ วัชร ปานเอี่ยม เกียรติ กิจเจริญ อรุณ ภาวไล กฤษณ์ สุกระมงคล เกรียงไกร อดายกุล นฤพนธ์ ไชยศ นัย สุขสกุล ธงชัย ประสงค์สันติ พรทิพย์ ประเสริฐยิ่งสุข กัญญาลักษณ์ บำรุงรักษ์ สมเกียรติ คุณานิติพงศ์ ศศิวิมล วิริยานนท์ ส. อาสนจินดา จูรี โอศิริ คร.เสรี วงษ์มณฑา [1] ภาพยนตร์เรื่องนี้ทำรายได้ 16 ล้านบาทในสมัยนั้น
จงสร้าง Context จาก Question เหมือนตัวอย่างข้างบน
Question: แบคบอยส์ คู่หูขวางนรก 2 เป็นภาพยนตร์แนวอะไร
Context:
Generated context: แบคบอยส์ คู่หูขวางนรก 2 เป็นภาพยนตร์แนวตลกล้อเลียนหนังดัง ๆ ที่มีเรื่องราวเกี่ยวกับคู่หูแบคบอยส์ที่ได้รับพลังพิเศษจากการเล่นเบตมินตันอย่างเน้นๆ ในวันหนึ่ง โดยความฮาและสาระที่แปลกปลอมของพล็อตจะทำให้ผู้ชมต้องสงสัยกลับไม่ได้ และเห็นความตลกซึ่งของการใช้พลังพิเศษในสถานการณ์ต่าง ๆ ที่เกิดขึ้น ภาพยนตร์นี้เลียนแบบแนวตลกและฮาของหลายเรื่องดัง ๆ ทำให้เป็นที่น่าสนใจและยังให้ความสนุกสนานในการติดตามเนื้อเรื่องของคู่หูแบคบอยส์น้อย ๆ ที่มีความฮาและความตลกตลอดเรื่องนี้อย่างยิ่ง แม้พล็อตจะเป็นเรื่องฮา ๆ แต่ก็สร้างความสนุกและเป็นเรื่องราวที่น่าจดจำสำหรับผู้ชมทุกคนที่หลงใหลในแบคบอยส์และความตลกขบขันที่ไม่คาดคิด
Extractive answer: เป็นภาพยนตร์แนวตลกล้อเลียนหนังดัง ๆ

Table 3: Example illustrates context generation prompt, in which we use two samples of question-context for our Adapted CGAP framework in Closed-book QA task.

Question: เสลดพังพอนเป็นพรรณไม้จัดอยู่ในวงศ์ของพรรณไม้ใด Gold answer: วงศ์ : Acanthaceae
Question: ต้นชวนชมได้รับสมญานามทนต่อสภาพแห้งแล้งด้วยชื่ออะไร
Answer: "กุหลาบทะเลทราย " (Desert Rose)
Question: ต้นเล็บมือนางมีสรรพคุณทางยาช่วยเหลือสัตว์เลื้อยทางด้านใด อธิบาย
Answer: นำผลตำให้ละเอียด ผสมในอาหารเลี้ยงวัวหรือหมูช่วยฆ่าพยาธิ
Question: ต้นแก้วมีขนาดใบกี่เซนติเมตร
Answer: ใบกว้างประมาณ 2-4 เซนติเมตร ยาวประมาณ3-6 เซนติเมตร
จงตอบคำถาม จาก Question เหมือนตัวอย่างข้างบน โดยตอบแค่คำถามที่ถามเท่านั้น ไม่จำเป็นต้องให้ข้อมูลอื่นประกอบแบบสั้นกระชับ และไม่ต้องมีครับ/ค่ะลงท้าย
Question: เสลดพังพอนเป็นพรรณไม้จัดอยู่ในวงศ์ของพรรณไม้ใด
Answer:
Predicted answer: เสลดพังพอนเป็นพรรณไม้จัดอยู่ในวงศ์ของพรรณไม้วงศ์เสลด (Asteraceae)

Table 4: Example illustrates answer prediction prompt, in which we use three samples of question-answer for few-shot prompting in Closed-book QA task.

Question: การทำให้เป็นอาณานิคมของยุโรปในทวีปอเมริกา เริ่มขึ้นในปีอะไร	Gold answer: ค.ศ. 1492
Question: การทำให้เป็นอาณานิคมของยุโรปในทวีปอเมริกา เริ่มต้นจากใคร	
จงตอบคำถามนี้ อย่างสั้น กระชับ และตอบให้ตรงคำถาม	
Predicted answer: การทำให้เป็นอาณานิคมของยุโรปในทวีปอเมริกาเริ่มขึ้นในปี 1492	

Table 5: Example illustrates answer prediction prompt that contains only instruction for answering question for zero-shot prompting in Closed-book QA task.