ViMedAQA: A Vietnamese Medical Abstractive Question-Answering Dataset and Findings of Large Language Model

Minh-Nam Tran, Phu-Vinh Nguyen, Long Nguyen*, Dien Dinh
Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{tmnam20,npvinh20}@apcs.fitus.edu.vn, {nhblong,ddien}@fit.hcmus.edu.vn

Abstract

Question answering involves creating answers to questions. With the growth of large language models, the ability of question-answering systems has dramatically improved. However, there is a lack of Vietnamese abstractive question-answering datasets, especially in the medical domain. Therefore, this research aims to mitigate this gap by introducing ViMedAQA [1]. This **Vi**etnamese **Med**ical **A**bstractive **Q**uestion-**A**nswering dataset covers four topics in the Vietnamese medical domain, including body parts, disease, drugs, and medicine. Additionally, the empirical results on the proposed dataset examine the capability of the large language models in the Vietnamese medical domain, including reasoning, memorizing, and awareness of essential information.

1 Introduction

Question-answering (QA) is a Natural Language Processing (NLP) task that aims to generate an appropriate response to a given question. QA systems are categorized based on their answer format. While extractive QA systems return the sub-strings from the provided context as the answer, abstractive QA systems identify keywords within the context and then rewrite this information to answer the question. Several QA datasets are SQuAD (Rajpurkar et al., 2016) and HotpotQA (Yang et al., 2018) for extractive QA, and AQuaMuSe (Kulkarni et al., 2020) and MS MARCO (Nguyen et al., 2016) are notable abstractive QA datasets.

In the field of Vietnamese NLP, various extractive QA datasets exist, including UIT-ViQuAD Nguyen et al. (2020) and VIMQA (Le et al., 2022), both of which serve for general knowledge (open-domain QA). Within the specific context of the Vietnamese medical domain, datasets such as UIT-ViNewsQA (Van Nguyen et al., 2022) and UIT-ViCoQA Luu et al. (2021) are available for extractive QA. However, there is a shortage of a Vietnamese abstractive question-answering corpus, especially in the medical domain.

To address the identified problem, we have developed and introduced ViMedAQA, a Vietnamese medical abstractive QA dataset. The corpus undergoes question-answer generation and human annotation stages to ensure quality while minimizing construction time. This proposed dataset is also leveraged to investigate the reasoning, denoising, and memorizing capabilities of large language models (LLMs) within the Vietnamese medical and healthcare domain. The contributions of this research work are listed as follows:

- Development of a dataset construction pipeline for abstractive QA tasks that utilizes existing LLMs to generate QA pairs from the context, thereby reducing the human effort required for question-answer creation.
- Introduction of ViMedAQA, a dedicated corpus for abstractive QA, encompasses four topics in Vietnamese medical literature: body parts, diseases, drugs, and medicine.
- Analysis of LLMs' reasoning, critical information extracting and memorizing capabilities within the Vietnamese medical domain.

2 Related Work

Extractive and Abstractive QA: Extractive QA systems answer the question by extracting parts of the context (Fajcik et al., 2021). Common extractive QA datasets include SQuAD (Rajpurkar et al., 2016, 2018), Natural Questions by Kwiatkowski et al. (2019), TriviaQA by Joshi et al. (2017) and SearchQA by Dunn et al. (2017). Conversely, the abstractive QA task generates responses using the

* Corresponding author.

[1] Source code is available at: https://github.com/trminhnam/vimedaqa and the dataset is published at: https://huggingface.co/datasets/tmnam20/ViMedAQA.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 270–278

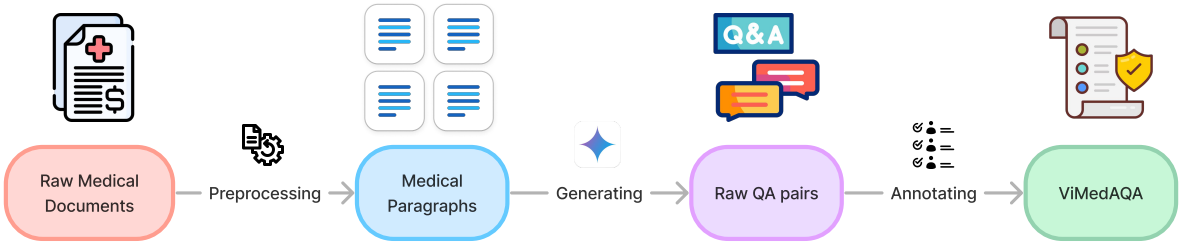


Figure 1: ViMedAQA construction pipeline.

model's knowledge. When provided with context, this task becomes open-book abstractive QA, and in the absence of context, it is closed-book QA (Ciosici et al., 2021). Common datasets for abstractive QA are ELI5 by Fan et al. (2019), AQuaMuSe by Kulkarni et al. (2020), MS MARCO by Nguyen et al. (2016), PolQA by Rybak et al. (2024) and Natural Questions (Kwiatkowski et al., 2019).

Open-Domain and Close-Domain QA: Open-domain QA systems assist with general knowledge. Some common open-domain QA datasets include TriviaQA (Joshi et al., 2017), SearchQA (Dunn et al., 2017), and MS MARCO (Nguyen et al., 2016). Conversely, close-domain QA systems answer questions in specific domains such as healthcare, law, and finance. Close-domain biomedical and healthcare QA datasets include MedQuAD (Ben Abacha and Demner-Fushman, 2019), HealthQA (Zhu et al., 2019), MedMCQA (Pal et al., 2022) and BiQA (Lamurias et al., 2020).

Vietnamese QA Datasets: Multiple QA datasets have been widely published in Vietnamese. UIT-ViQuAD by Nguyen et al. (2020), which follows the SQuAD format, is constructed from Wikipedia text. VIMQA (Le et al., 2022) is a multi-hop extractive QA dataset based on Wikipedia. UIT-ViNewsQA, introduced by Van Nguyen et al. (2022), is built on top of Vietnamese healthcare news articles. UIT-ViCoQA, developed by Luu et al. (2021), is a medical extractive QA dataset for machine reading comprehension evaluation.

3 Dataset

3.1 Dataset Creation Process

The dataset creation process, visualized in Figure 1, contains three steps below.

Data source and preprocessing. Initially, raw documents are sourced from the internet. To ensure quality and credibility, we select only those written by doctors with Master's or PhD degrees in medicine. These documents undergo preprocessing to eliminate HTML tags, links, and non-medical content. Each document is divided into paragraphs according to the article's structure. To respect Vietnam's intellectual property rights, the article URL, the author's name, and the URL are included in each paragraph. Additionally, this dataset is published for educational and research uses only.

Question-answer generation process. Using the parsed paragraph as the context, the Gemini 1.0 language model (Team et al., 2023) generates pairs of question-answer where each answer corresponding to the question must be included in the paragraph. The number of question-answer pairs to request Gemini to generate depends on the number of sentences in the paragraph as `num_pairs=max{3, num_sentences_in_paragraph}`.

Annotation Guideline. The team of annotators consists of five individuals (see Appendix C). Each annotator carefully evaluates the meaning and grammatical correctness of the questions and answers generated for each paragraph. They also verify whether the answer is contained within the context, either implicitly or explicitly. If any question-answer pair is marked with a `Reject` label by an arbitrary labeler, the question-answer pair is removed from the dataset's final version.

Using the outlined pipeline, we constructed and validated the ViMedAQA dataset, represented as $S = \{(p_i, q_i, a_i) \mid 1 \leq i \leq n\}$, where n denotes the total number of samples. For each datapoint, p_i denotes the paragraph, q_i is the corresponding question, and a_i represents the corresponding answer, with key information in the answer a_i sourced directly from the corresponding paragraph p_i.

3.2 Dataset Statistics

The dataset contains $44,313$ $\{p, q, a\}$ triplets divided into train/validation/test sets. It covers four topics in the Vietnamese medical domain, includ-

Model	English Prompt					Vietnamese Prompt				
	BERT	BLEU	MET	ROU	Avg	BERT	BLEU	MET	ROU	Avg
Multilingual LLMs										
Llama3-7B	71.78	30.12	66.83	59.32	**57.01**	71.36	25.33	67.97	55.52	**55.05**
Llama2-7B	49.20	12.03	38.04	35.38	33.66	41.65	6.93	24.36	24.34	24.32
Gemma-2B	63.18	31.89	51.44	52.38	49.72	64.28	32.04	53.48	53.57	50.84
Gemma-7B	64.79	25.73	62.95	53.71	51.80	68.49	31.17	63.52	57.03	**55.05**
Vietnamese LLMs										
PhoGPT-4B	68.60	21.03	59.73	50.52	49.97	68.94	21.06	59.76	50.75	50.13
VinaLlama-7B	73.04	33.69	65.42	59.89	**58.01**	72.47	31.70	64.29	59.08	**56.89**
VinaLlama-2.7B	67.90	23.17	57.36	51.90	50.08	70.09	26.07	59.77	54.96	52.72
ViGPT	58.36	9.98	42.29	33.28	35.98	59.07	10.94	44.39	34.27	37.17

Table 1: Model performance on the test set of ViMedAQA under open-book question-answering task. BERT, MET, ROU, and Avg denote BERTScore, METEOR, ROUGE-L, and Average score, respectively. The best average score across models in each type is shown in **bold**, and the best metric score of each model type is shown in underline.

ing drugs, medicine, body parts, and disease. Further information refers to Table 4 in Appendix D.

The distribution of question types is visualized in Figure 3 in Appendix D. Most questions fall under the "Open-Ended" category, totaling $40,443$, significantly outnumbering other types. The "Yes-No" category has a notable count with $3,740$ questions.

4 Methodology

This study assesses the capabilities of generative language models for learning, memorizing, and understanding medical information in Vietnamese.

Experiments were conducted using the proposed ViMedAQA dataset. Using an open-book QA format, the first experiment examined the model's reasoning ability in the Vietnamese medical and healthcare domain. Each input to the model consisted of a pair, $\{p_i, q_i\}$, where p_i is a paragraph and q_i is a corresponding question. To assess the model's ability to extract essential information to answer q_i, additional unrelated paragraphs, $\{p_j\}$ where $(j \neq i$ and $|\{p_j\}| \in \{0, 1, 2, 4, 8\})$, are included in the prompt. It is hypothesized that adding more noise paragraphs would degrade the model's performance. A second experiment measured the amount of Vietnamese medical knowledge within the language models under a closed-book QA setting. The model is prompted with only the question q_i and answers q_i using its internal knowledge acquired during pretraining and finetuning.

We use greedy search decoding during the evaluation process to prompt the model, resulting in highly reproducible experiments. The BLEU (Papineni et al., 2002; Lin and Och, 2004), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and BERTScore (Zhang* et al., 2020) metrics are utilized to compare model's outputs and labels.

5 Experimental Results and Analysis

This section reports the empirical results for the experiments following setups in Section 4.

All of the experiments use medium to small LLMs, including LLama2-7B and LLama3-7B (Touvron et al., 2023), Gemma-2B and Gemma-7B by Team et al. (2023), PhoGPT-4B by Nguyen et al. (2023a), VinaLlama-7B and VinaLlama-2.7B by Nguyen et al. (2023c), and ViGPT by (Nguyen et al., 2023b). All the models are explored with the chat or instruction tuning versions.

5.1 Reasoning Ability

This experiment examines the models for their reasoning ability through open-book question-answering. The results are reported in Table 1.

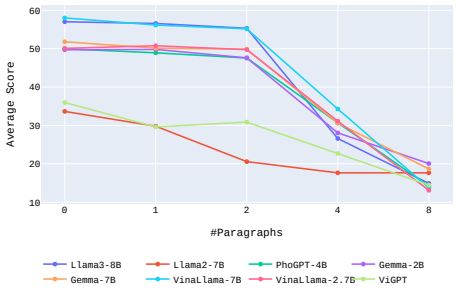
From Table 1, Llama3-7B outperforms other English and Vietnamese prompt template models with average (avg) scores at 57.01 and 55.05, respectively. With Vietnamese prompt, Gemma-7B gets the same mark as Llama3-7B at 55.05. Meanwhile, VinaLlama-7B achieves the highest performance across all Vietnamese LLMs, with a 58.01 average score for the English input template and a 56.89 average score for the Vietnamese input template.

In summary, Llama3-7B and VinaLlama-7B

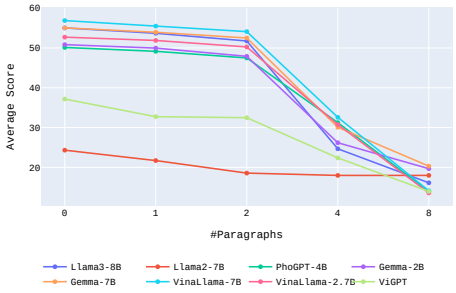
show their strong capability in the Vietnamese medical reasoning task. Additionally, language-specific LLMs slightly outperform multilingual LLMs and using the English template results in a slight performance gain over the Vietnamese prompt template.

5.2 Awareness Ability

Figure 2 illustrates all model's performance when the number of noise paragraphs (denoted as m) provided in the input prompt increases.



(a) English prompt template.



(b) Vietnamese prompt template.

Figure 2: Visualization of the model performance as the number of wrong paragraphs increases for both English and Vietnamese templates. #Paragraphs is the number of noise paragraphs in the model input prompt.

In both languages, the performance of most models demonstrates a shared trend. Performance gradually decreases as m increases from 0 to 2 (first stage), followed by a significant decline as k escalates from 2 to 8 (second stage). Although Llama3-8B exhibits the most robust performance with little noise input ($m \leq 2$), the Gemma model family outperforms other models in scenarios with considerable noise ($m = 8$), suggesting that the Gemma model is superior in extracting crucial information amidst noisy data compared to other models.

5.3 Memorization Capability in Vietnamese Medical and Healthcare Knowledge

Results for this experimental scenario are presented in Table 2. In scenarios where the input prompt only contains a question without context, the model must rely on its internal knowledge for the answer.

Among the models, VinaLlama-7B achieved the highest score of 36.80 with the English prompt template, followed closely by PhoGPT-4B with a score of 36.22. With the Vietnamese instruction template, PhoGPT-4B, scoring 36.88, outperformed all other models. The Gemma models family exhibited balanced performance across both languages, whereas the Llama family underperformed (scoring lower than 30) with the Vietnamese prompt template. These results indicate that, compared to other models, VinaLlama-7B and PhoGPT-4B possess the most extensive Vietnamese medical knowledge.

Model	En	Vi
Llama3-8B	30.95	29.46
Llama2-7B	30.53	17.32
Gemma-2B	34.40	33.28
Gemma-7B	30.71	31.92
PhoGPT-4B	36.22	**36.68**
VinaLlama-7B	**36.80**	35.00
VinaLlama-2.7B	31.99	33.93
ViGPT	31.46	32.49

Table 2: Average scores of the models on the test set when prompting without context.

6 Conclusion

ViMedAQA, a Vietnamese medical abstractive question-answering dataset, is published to mitigate the lack of an abstractive QA corpus for the Vietnamese medical domain. By leveraging the available LLMs, raw question-answer pairs are automatically generated before being verified by expert annotators to ensure the dataset's quality. Additionally, experiments to study the capability of LLMs are examined, including reasoning, memorizing, and awareness of critical information. The empirical results show that VinaLlama-7B is a large language model with powerful reasoning skills in the Vietnamese medical domain, and the Gemma model family is robust in realizing essential information across multiple noise contexts.

There are several potential directions for future work, including (1) expanding the scope of topics covered by ViMedAQA in the Vietnamese medical domain, (2) investigating the performance of LLMs in this domain under different fine-tuning methodologies, and (3) delving into the extraction

of critical data by increasing the number of incorrect paragraphs (m), and exploring solutions to mitigate performance degradation as m increases.

Limitations

Despite the introduction of ViMedAQA to address the absence of a medical abstractive QA dataset in Vietnam, this research has certain limitations.

Firstly, even though the raw documents are sourced from highly reliable resources, the LLMs occasionally fail to generate accurate question-answer pairs. Moreover, LLMs sometimes create similar questions for different paragraphs.

Secondly, there is a lack of experiments involving fine-tuning methods on the proposed dataset for comparison with the prompting method.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20:1–23.

Manuel Ciosici, Joe Cecil, Dong-Ho Lee, Alex Hedges, Marjorie Freedman, and Ralph Weischedel. 2021. Perhaps PTLMs should go to school – a task to assess open book and closed book QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6104–6111, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Martin Fajcik, Josef Jon, and Pavel Smrz. 2021. Rethinking the objectives of extractive question answering. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 14–27, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint arXiv:2010.12694*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Andre Lamurias, Diana Sousa, and Francisco M Couto. 2020. Generating biomedical question answering corpora from q&a forums. *IEEE Access*, 8:161042–161051.

Khang Le, Hien Nguyen, Tung Le Thanh, and Minh Nguyen. 2022. VIMQA: A Vietnamese dataset for advanced reasoning and explainable multi-hop question answering. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6521–6529, Marseille, France. European Language Resources Association.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, Geneva, Switzerland. COLING.

Son T Luu, Mao Nguyen Bui, Loi Duc Nguyen, Khiem Vinh Tran, Kiet Van Nguyen, and Ngan

Luu-Thuy Nguyen. 2021. Conversational machine reading comprehension for vietnamese healthcare texts. In *Advances in Computational Collective Intelligence: 13th International Conference, ICCCI 2021, Kallithea, Rhodes, Greece, September 29–October 1, 2021, Proceedings 13*, pages 546–558. Springer.

Dat Quoc Nguyen, Linh The Nguyen, Chi Tran, Dung Ngoc Nguyen, Nhung Nguyen, Thien Huu Nguyen, Dinh Phung, and Hung Bui. 2023a. Phogpt: Generative pre-training for vietnamese. *arXiv preprint arXiv:2311.02945*.

Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020. A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Minh Thuan Nguyen, Khanh Tung Tran, Nhu Van Nguyen, and Xuan-Son Vu. 2023b. Vigptqa-state-of-the-art llms for vietnamese question answering: System overview, core models training, and evaluations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 754–764.

Quan Nguyen, Huy Pham, and Dung Dao. 2023c. Vinallama: Llama-based vietnamese foundation model. *arXiv preprint arXiv:2312.11011*.

Thien Hai Nguyen, Tuan-Duy H. Nguyen, Duy Phung, Duy Tran-Cong Nguyen, Hieu Minh Tran, Manh Luong, Tin Duy Vo, Hung Hai Bui, Dinh Phung, and Dat Quoc Nguyen. 2022. A Vietnamese-English Neural Machine Translation System. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association: Show and Tell (INTERSPEECH)*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Piotr Rybak, Piotr Przybyła, and Maciej Ogrodniczuk. 2024. PolQA: Polish question answering dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12846–12855, Torino, Italia. ELRA and ICCL.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2022. New vietnamese corpus for machine reading comprehension of health news articles. *Transactions on Asian and Low-Resource Language Information Processing*, 21(5):1–28.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K Reddy. 2019. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference*, pages 2472–2482.

A Potential Risk

Given the rapid advancements in the medical field, the knowledge contained in ViMedAQA may become outdated. Therefore, this dataset should primarily be used to research the capabilities of QA systems in the Vietnamese medical domain rather than to develop a particular application. Any misuse of the dataset for illegal purposes is strictly prohibited. The dataset should be appropriately used to contribute to the advancements in NLP.

B License

The raw documents are crawled from YouMed.vn[2]. The term of use is available on the YouMed Term Of Use webpage [3], which states that "The information included on this website is strictly for informational and educational purposes." Hence, this research does not violate YouMed's terms of use.

Following the YouMed term of use, the dataset is published under the Creative Commons NonCommercial 4.0 license, which requires users to use it for non-commercial purposes only.

C Annotator List

The academic qualifications of data annotators are:

- Annotator 1 - Associate Professor in Computer Science and Comparative Linguistics.
- Annotator 2 - PhD in Computer Science and Natural Language Processing (NLP).
- Annotator 3 - PhD candidate in Comparative Linguistics.
- Annotator 4 - Undergraduate Student majoring in Computer Science and NLP.
- Annotator 5 - Undergraduate Student majoring in Computer Science and NLP

D Dataset Statistics

Figure 3 visualizes the number of samples for each question type and subtype. Additionally, Table 3 categorizes the questions into Yes-No and Open-Ended types, with further subtypes under Open-Ended, such as Why, What, When, How, How Much/Many, Which, Who, and Where. An additional category labelled Other is included.

[2]https://youmed.vn/

[3]https://youmed.vn/tin-tuc/term-of-use/

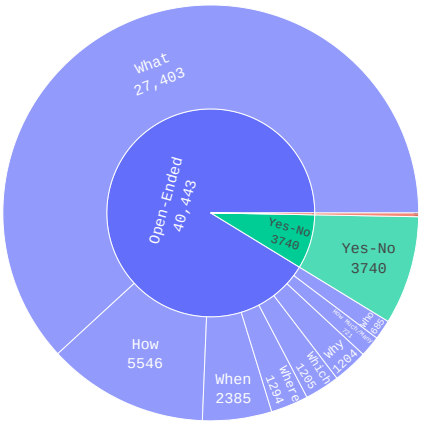


Figure 3: Distribution of question types. Questions are translated to English using the method proposed by Nguyen et al. (2022) before being categorized.

Table 4 shows the distribution of the dataset across training, validation, and test sets, categorized by topic. The total number of samples for each category is 44,313, with Body part, Disease, Drug, and Medicine having 4,970, 15,690, 9,780, and 13,873 examples, respectively.

Main Type	Subtype	#Questions
Yes-No	Yes-No	3,740
Open-Ended	Why	1,204
Open-Ended	What	27,403
Open-Ended	When	2,385
Open-Ended	How	5,546
Open-Ended	How Much/Many	721
Open-Ended	Which	1,205
Open-Ended	Who	685
Open-Ended	Where	1,294
Other	Other	130
	Total	44,313

Table 3: Distribution of questions by Type and Subtype in the proposed ViMedAQA dataset.

Each sample in the proposed ViMedAQA dataset has the following fields:

- **question_idx**: The index of the sample.
- **question**: The question to be answered.
- **answer**: The answer to the question.
- **context**: The context or paragraph that contains the information to answer the question.
- **title**: The title of the corresponding article from which the context was taken.

Topic	Train	Val	Test	Total
Body part	4,473	248	249	4,970
Disease	14,121	784	785	15,690
Drug	8,802	489	489	9,780
Medicine	12,485	694	694	13,873
Total	39,881	2,215	2,217	44,313

Table 4: Number of samples across train, validation (Val), and test subsets by medical topic.

- **keyword**: The related disease/drug/body part in the question. Such as "heart attack."
- **topic**: The topic of the question/context. It can be one of the following: Body part, Disease, Drug and Medicine.
- **article_url**: The URL of the original article.

E Experiment Setup

In the scope of this research, the models are not trained, and their weights are not modified by gradient descent. The experiments are conducted by prompting the model directly with greedy decoding, which is more reproducible and requires less computational resources than sampling methods.

E.1 Computational Resources

The experiments are conducted on a single machine with Intel i5-14500 CPU, 32 GB RAM, and duo NVIDIA GeForce RTX 4060 Ti 16 GB cards.

E.2 Softwares

The `transformers` library by Wolf et al. (2020) and the `datasets` library by Lhoest et al. (2021) are used to load the model and datasets from HuggingFace [4], respectively. The `evaluate` [5] library, the `bert-score` framework and the `rouge_score` library are used to evaluate the model's outputs.

E.3 Model Configurations

The number of parameters of the LLMs used in the experiments is reported in Table 5.

E.4 Prompt templates

The prompt templates are provided in Listing 1 for the English template. When the model does not use a system prompt (like Gemma), it is concatenated with the user prompt before feeding to the model.

[4] https://huggingface.co/
[5] https://github.com/huggingface/evaluate

Model	#Params
Llama3-8B	8.0B
Llama2-7B	7.0B
PhoGPT-4B	4.0B
Gemma-2B	2.0B
Gemma-7B	7.0B
VinaLlama-7B	7.0B
VinaLlama-2.7B	2.7B
ViGPT	6.2B

Table 5: Number of parameters (#Params) per model used in the experiment stage. "B" denotes billion.

E.5 Experiment Running Time

The running times of the models for the three experiment scenarios are provided in Table 6, Table 7 and Table 8. The time format is "HH:MM:DD".

Model	En	Vi
Llama3-8B	01:03:08	01:15:24
Llama2-7B	01:43:09	01:44:18
PhoGPT-4B	06:46:20	06:37:16
Gemma-2B	00:17:23	00:18:32
Gemma-7B	01:15:17	01:04:28
VinaLlama-7B	01:15:37	01:18:28
VinaLlama-2.7B	00:36:39	00:36:33
ViGPT	01:59:54	01:43:24

Table 6: Running time of all models in Section 5.1.

Model	En	Vi
Llama3-8B	01:09:17	01:19:23
Llama2-7B	01:43:59	01:48:21
PhoGPT-4B	05:40:11	04:32:31
Gemma-2B	00:27:35	00:34:55
Gemma-7B	01:29:20	01:21:31
VinaLlama-7B	01:15:41	01:21:52
VinaLlama-2.7B	00:50:51	00:51:08
ViGPT	02:01:32	01:54:27

Table 7: Running time of all models in Section 5.2.

```
{
    "with_context": {
        "system_prompt": "Based on the following context and your knowledge, answer the following
↪   question in Vietnamese.",
        "user_prompt": "### Context:\n{example['context']}\n\n### Question:\n{example['question']}"
    },
    "without_context": {
        "system_prompt": "Based on your knowledge, answer the following question in Vietnamese.",
        "user_prompt": "### Question:\n{example['question']}"
    }
}
```

Listing 1: English prompt template.

Model	En	Vi
Llama3-8B	01:48:19	02:00:34
Llama2-7B	01:42:39	01:45:48
PhoGPT-4B	01:39:48	01:08:55
Gemma-2B	00:21:59	00:24:30
Gemma-7B	01:53:20	01:46:24
VinaLlama-7B	01:39:33	01:45:28
VinaLlama-2.7B	00:24:45	00:24:58
ViGPT	01:20:02	01:05:03

Table 8: Running time of all models in Section 5.3.