EAI Endorsed Transactions
on Scalable Information Systems
EAI.EU Research Article

VMHQA: A Vietnamese Multi-choice Dataset for Mental Health Domain Question Answering

Tu Anh Hoang Nguyen[1,2], Quang-Dieu Nguyen[1,2], Harius M. Nguyen[1,2], Alfred Hoang Nguyen[3], Loan T.T. Nguyen[1,2,*]

[1]School of Computer Science and Engineering, International University, Ho Chi Minh City 700000, Viet Nam
[2]Vietnam National University, Ho Chi Minh City 700000, Vietnam
[3]Faculty of Information Technology, FPT University, Thu Duc City 71300, Ho Chi Minh City, Vietnam

Abstract

This paper introduces VMHQA, a Vietnamese Multiple-Choice Question Answering (MCQA) dataset designed to address critical mental health resources gaps, particularly in low and middle-income countries like Vietnam. The dataset comprises 10,000 meticulously curated records across 1,166 mental health subjects, including 249 topics in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) and 8,599 contextual paragraphs. Each record adheres to the United States Medical Licensing Examination (USMLE) format, with targeted questions, correct answers, multiple-choice options, and supporting paragraphs from reputable sources such as academic journals and local hospital websites, further inspected by prestigious mental hospitals in Vietnam. VMHQA thus provides a reliable, structured foundation for pre-consultation tools, allowing for early psychological intervention for those concerned about mental health issues. This study also goes beyond data collection to evaluate the effectiveness of VMHQA using cutting-edge machine learning models, such as BERT-based architectures, large language models (LLMs) ranging from 7 to 9 billion parameters, and various generative pre-trained transformer (GPT) frameworks. In addition, we look at how Retrieval-Augmented Generation (RAG) combined with Agentic Chunking can improve the accuracy and interpretability of responses in this specialised domain. The retrieval mechanisms of RAG are examined explicitly for their ability to generate contextually accurate answers sensitive to psychological nuances. Our findings shed light on the effectiveness of these advanced models in handling complex, domain-specific question-answering tasks in mental health, highlighting their potential to make mental health care more accessible and reliable for Vietnamese-speaking communities. VMHQA thus represents a significant step toward making mental health care more accessible, offering hope for improved mental health outcomes.

Received on 28 October 2024; accepted on 22 September 2025; published on 25 September 2025

Keywords: VMHQA, Mental Health Dataset, Vietnamese Multiple-Choice Question Answering (MCQA), BERT-based Models, NLP in Mental Health, Retrieval-Augmented Generation (RAG), Agentic Chunking, Large Language Models (LLMs)

doi:10.4108/eetsis.7678

*Corresponding author. Email:nttloan@hcmiu.edu.vn

1. Introduction

Recent AI-driven diagnostic tools have delivered impressive results across multiple health domains-e.g., deep-learning models with transfer learning for diabetic retinopathy and other ocular disease detection [1], hybrid attention–convolution architectures for acute brain stroke prediction [2], the XCR-Net framework for rapid COVID-19 screening from chest X-rays [3], and GENet networks classifying neurological disorders from EEG data [4]-yet these systems overwhelmingly rely on imaging or bio-signal inputs and remain confined to English-language settings.

Natural language processing (NLP) and machine comprehension have grown due to large-scale question-answering (QA) datasets. Benchmark datasets like SQuAD [5] and RACE [6] are necessary for upgrading AI models' ability to extract information from texts. However, their focus on general knowledge and restricted settings limits these datasets' applicability in particular fields like medicine or psychological wellness. Deep reasoning tasks over complex biomedical texts are shown to AI models in the medical domain by datasets such as HEAD-QA [7], MedQA [8], and MedMCQA [9]. Despite being groundbreaking, these datasets are primarily in English and focus on general medical knowledge, which leaves gaps in more particular areas like mental well-being.

ViMQ [10], VIMQA [11], and UIT-ViQuAD [12] have all contributed to the evolution of the QA dataset landscape in Vietnam. These datasets provide Vietnamese resources in an attempt to close the language gap, but they are still restricted to legal texts and general knowledge. More importantly, they are low in particular areas like mental health. The lack of a thorough Vietnamese mental health QA dataset hinders the creation of AI models that can handle this field. The increased attention paid to mental health globally makes the difference even more noticeable. AI models cannot completely support professionals or patients with the complex, domain-specific logic needed without datasets tailored to specific fields, particularly mental health.

A careful review of related works may provide insight into the general problems they are trying to address. For example, Jin et al.'s 2019 introduction of PubMedQA increases AI reasoning over biomedical texts by using PubMed abstracts to structure questions around actual research scenarios. Jin et al.'s 2020 introduction of MedQA, which tests knowledge recall, reasoning, and decision-making skills, adds complexity with questions modelled after medical licensing exams. With more than 194,000 multiple-choice questions spanning 21 medical topics, MedMCQA, first presented by Pal et al. in 2022, broadens the scope of medical QA. These datasets highlight the need for specialised resources in domains like mental health, where AI must deliver significant, context-driven insights rather than just brief answers.

Research Problems and Importance. Despite increasing global attention to mental health, access to reliable, language-specific digital mental health resources remains extremely limited in low- and middle-income countries, including Vietnam. Current multilingual QA datasets predominantly target high-resource languages and general medical domains, with very few tailored to the nuanced challenges of mental health dialogue. This creates critical barriers to developing and deploying NLP-based mental health applications that are both linguistically and culturally contextualized. The lack of a domain-specific Vietnamese mental health dataset prevents AI models from learning the subtle reasoning, symptom overlap, and clinical ambiguities that characterize mental health diagnostics. Addressing this gap is crucial for advancing Vietnamese NLP and empowering scalable, AI-assisted mental health tools, which can facilitate early screening, raise awareness, and extend pre-consultation support to underserved populations.

With over 23,000 question-answer pairs from Wikipedia articles, UIT-ViQuAD in Vietnam still lacks the domain-specific detail required for complex applications like mental health. This gap would be filled using reliable medical sources by creating a Vietnamese mental health QA dataset (VMHQA). This dataset would ensure the precision and practicality of the data, promoting AI models to provide more accurate and understandable responses in this vital field.

To address these issues, we present VMHQA: a Vietnamese Multi-Subject Multi-Choice Question Answering dataset designed for mental health. VMHQA comprises 10,000 high-quality records across 1,166 subjects, including 249 subjects listed in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [13] with 492 related sub-topics. Each record includes the question, answer option, correct answer, and relevant paragraph to provide context and support facts. Source from famous academic books and trusted medical websites, VMHQA ensures reliable content for mental health-related QA.

In addition to introducing VMHQA, we use several cutting-edge models to assess the performance of the dataset. We specifically use generative pre-trained transformer (GPT) models, large language models (LLM) with 7–9 billion parameters, and BERT-based models to address the multiple-choice question task. Given the specialised nature of the field, benchmarking allows us to cast light on its advantages and disadvantages when used for Vietnamese QA focused on mental health.

Furthermore, we explore the use of Retrieval-Augmented Generation (RAG) [14] with Agentic Chunking [15] to enhance the interpretation and accuracy of responses. RAG enables the retrieval of relevant external knowledge, essential in the complex mental health field, where a nuanced understanding of the condition is necessary. Agentic Chunking, which structures the retrieved information into manageable parts, supports the model in concentrating on the most suitable data when generating an answer. By testing RAG on VMHQA, we investigate whether this approach is ideal for handling domain-specific datasets like VMHQA, focusing on improving model performance in both knowledge retrieval. In brief, the contributions of this study are as follows:

- *Comprehensive Dataset*:
  - Extensive Coverage: VMHQA encompasses 10,000 records, offering a robust foundation for Vietnamese mental health question answers.

  - Significant Scale: As the largest Vietnamese QA dataset in the mental health domain, it fills a resource gap.
  - Real-World Relevance: Source from famous books and medical websites, the dataset ensures authenticity and applicability in real-world mental health contexts.

- *Challenging Content*: The questions are designed to evaluate artificial intelligence (AI) models' reasoning abilities and include comprehension, inference, and application across many mental health topics.

- *Quality Annotations*: Each record includes a question, correct answer, multiple distractor options, and relevant paragraphs with supporting facts. This structure enhances the dataset's educational value by promoting nuanced reasoning and evidence-based conclusions.

- *Alignment with Medical Standards*: VMHQA adheres to the United States Medical Licensing Examination (USMLE) format and provides evidence-based questions and answers.

- *Advancement of Vietnamese NLP*: By introducing VMHQA, we significantly contribute to Vietnamese NLP and offer a dataset for mental health. This resource enables the AI model to understand complex issues and support mental health professionals.

- *Comparative Analysis of Model Approaches*: Our study evaluates a range of LLMs, comparing open-access models' performance with commercial alternatives like GPT. We assess their capabilities in addressing multiple-choice questions.

- *Agentic Chunking versus. Full Context Chunking*: We also conduct an in-depth comparison of Agentic Chunking with Full Context Chunking using RAG. This comparison assesses the effectiveness of these approaches in handling the retrieval of mental health-related data.

- *Benchmark for Future Research*: Experiments use state-of-the-art models to demonstrate the dataset's effectiveness. The observed performance gap between AI models and human experts highlights improvement opportunities, positioning VMHQA as a challenge.

2. The VMHQA dataset

2.1. Task definition

The VMHQA dataset has undergone a thorough quality and accuracy check by experts at the prestigious National Mental Hospital 2 in Vietnam, ensuring its reliability and validity for mental health research and applications, following the USMLE benchmarks [16], known for their top standards in the medical field. This ensures that the dataset gives high credit and is relevant, particularly for mental health. Considering the USMLE-style framework, VMHQA emphasises multi-layered questions and answers, allowing for an assessment of medical knowledge and reasoning. The dataset is comprehensive, with a Vietnamese vocabulary of 21842 words, and covers 1,166 subjects, including 249 lists in the DSM-5, a globally recognised classification of disorders. Moreover, the dataset incorporates 8,599 paragraphs from over 3,000 famous sources, providing a rich foundation for detailed analysis and model evaluation.

Notably, the dataset has been evaluated and verified by an expert at National Mental Hospital 2, ensuring its accuracy and reliability for mental health research and applications.

A key feature of VMHQA is the Supporting Fact-Explanation, where key sentences are extracted from source materials to show why they chose the chosen answers. This enhances transparency and interpretation, making the dataset effective for training and evaluating AI models. The dataset upgrade analysis of mental health assessments allows models to reach exact conclusions based on factual evidence. The VMHQA task is organised around four main components:

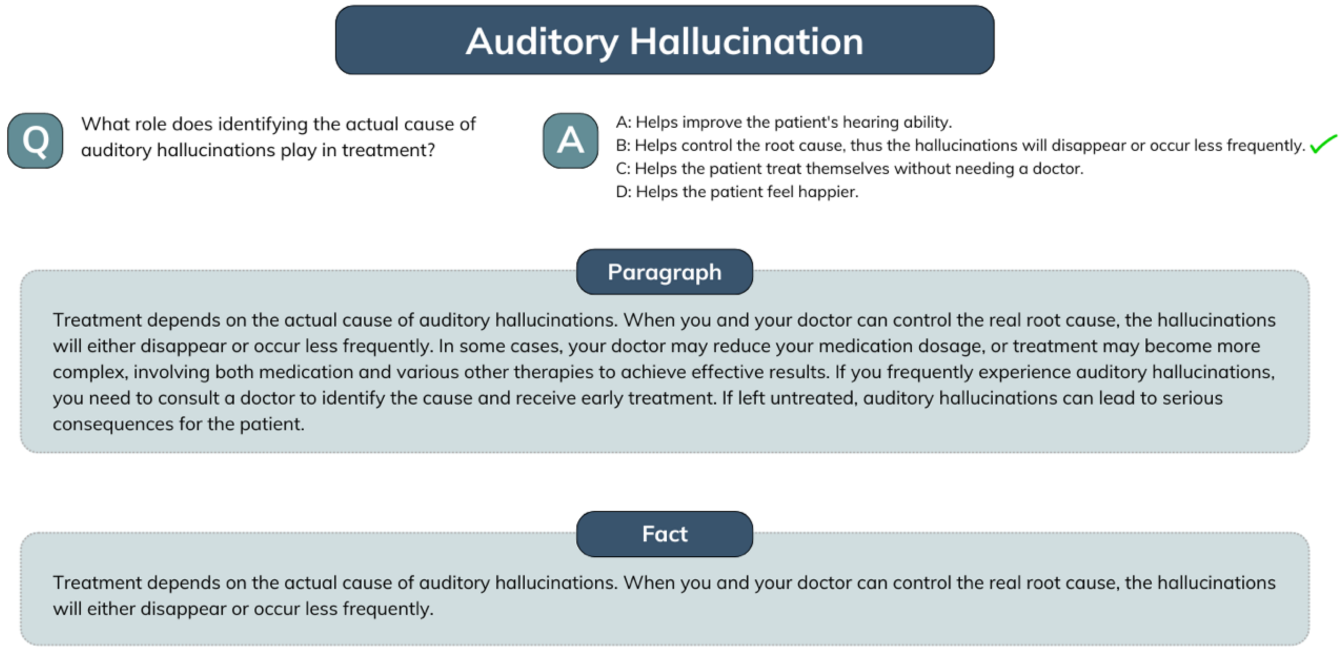
- *Question*: Textual queries range from single sentences to complex case descriptions, reflecting real-world clinical challenges.

- *Answer Candidates*: Multiple choice questions challenge model reason that will stimulate decision-making.

- *Document Collection*: A set of trusted reference materials has the knowledge required to identify correct answers and cover mental and DSM-5 categories.

- *Supporting Fact-Explanation*: Key sentence extract to guide the model in narrowing down relevant content and justifying correct answers.

Combining the USMLE-style benchmark with the Supporting Fact-Explanation feature and applying advanced AI technical skills, VMHQA presents a challenging and comprehensive task. It allows the development of AI models that can understand complex medical data and justify their reasoning, making them reliable for applications in the mental health domain.



Auditory Hallucination

Q What role does identifying the actual cause of auditory hallucinations play in treatment?

A A: Helps improve the patient's hearing ability.
B: Helps control the root cause, thus the hallucinations will disappear or occur less frequently. ✓
C: Helps the patient treat themselves without needing a doctor.
D: Helps the patient feel happier.

Paragraph

Treatment depends on the actual cause of auditory hallucinations. When you and your doctor can control the real root cause, the hallucinations will either disappear or occur less frequently. In some cases, your doctor may reduce your medication dosage, or treatment may become more complex, involving both medication and various other therapies to achieve effective results. If you frequently experience auditory hallucinations, you need to consult a doctor to identify the cause and receive early treatment. If left untreated, auditory hallucinations can lead to serious consequences for the patient.

Fact

Treatment depends on the actual cause of auditory hallucinations. When you and your doctor can control the real root cause, the hallucinations will either disappear or occur less frequently.

Figure 1. Example record from the VMHQA Dataset translated from Vietnamese to English with four main components and the topic.

2.2. Data collection

The creation of the VMHQA dataset was a meticulous and thoughtful process. Initially, the VIMQA-Maker tool generated question-answer pairs from Wikipedia based on specific keywords. However, after collecting 250 entries, several key issues became apparent:

- The tool's direct answers were inadequate for nuanced mental health assessments.
- Wikipedia's open-edit nature raised concerns about the credibility of sensitive medical information.
- The keyword-based approach often produced irrelevant paragraphs filled with foreign terms or English keywords, reducing the relevance of Vietnamese materials.

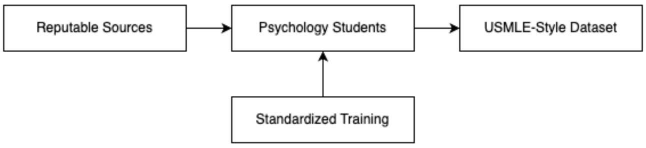


Figure 2. The data collection process of VMHQA.

To overcome these challenges, a revised data collection process was implemented, as shown in Figure 2, which included:

- *Source Compilation*: Reputable psychology and mental health materials were gathered from trusted sources, including academic books, hospital websites, and psychological counselling centres.
- *Expert Annotation Team*: A team of trained psychology students was assembled to construct the dataset. They followed a format that was in alignment with USMLE standards.
- *Standardized Training*: Annotators receive instructions and examples to ensure consistency and accuracy when formulating complex multiple-choice questions and extracting supporting facts.
- *Human-Curated Content*: Each question, along with its answer option, correct answer, and support paragraph, is carefully crafted to fit the required reliability standards.

This human-centred approach ensures that the VMHQA dataset is highly quality, making it a resource for training the AI model. As shown in Table 1, the comparison highlights critical attributes of the VMHQA dataset. What makes VMHQA different from other datasets is the unique question-type-based split method. This approach offers a structured way of organising the data, making it applicable to mental health applications compared to random or exam-based splits in other datasets like MedQA [8] or MedMCQA [9].

Table 1. Comparison of VMHQA with several existing MCQA datasets (MedQA [8], HEAD-QA [7], MedMCQA [9], VMHQA) in the medical domain.

Dataset	# Question	# Subject	Publicly Available	Explnation	Split Type	Open Domain
MedQA	270000	-	no	no	random	yes
HEAD-QA	13530	6	yes	no	year-wise	yes
MedMCQA	193155	21	yes	yes	exam-based	yes
VMHQA	**10000**	**1166**	**yes**	**yes**	**question-type-based**	**yes**

2.3. Preprocessing & Quality checks

To ensure the quality and consistency of the VMHQA dataset, which initially comprised 10,120 records manually collected by a team of 10 students, a series of meticulous data-cleaning steps was undertaken:

- *Manual Review and Correction*: Each student carefully reviewed their submissions to ensure linguistic accuracy. The aggregated dataset was then examined again to correct any remaining grammar errors.
- *Handling Missing Values*: Missing values were addressed with great care. Records lacking critical information were excluded, while non-critical fields were filled with placeholders to eliminate ambiguity.
- *Duplicate Removal*: A thorough review was conducted to identify and remove duplicate records. This process resulted in the elimination of 0.16% of the records, ensuring no errors remained.
- *Standardizing Data Formats*: The dataset's structure and columns were standardised for consistency. For example, question types initially indicated by prefixes in the questions were extracted and placed in a separate 'Question Type' column.
- *Normalizing Topics*: Due to the manual collection, inconsistencies in topic labelling occurred (e.g., 'trầm cảm' versus 'Trầm Cảm' for 'depression'). The topics are normalised using the 'GPT-4o' and put into subjects and subtopics.

After these comprehensive cleaning steps, the final dataset was refined to 10,000 high-quality records suitable for subsequent analysis and application. While minor spelling errors may still exist due to the manual collection process, they are minimal and do not significantly impact the dataset's overall quality or reliability.

2.4. Data statistic

The VMHQA dataset consists of 10,000 multiple-choice questions divided into train (8,000), development (1,000), and test (1,000) sets. It contains 21,842 unique tokens, with an average of 19.36 tokens per question, 15.07 tokens per answer, and 80.91 tokens per explanation. The maximum token lengths for questions, answers, and explanations are 1,088, 304, and 3,147.

The underthesea toolkit [17], a Vietnamese NLP package, splits the token. By utilising its train model, the tokenisation of Vietnamese in the dataset was handled, ensuring linguistic breakdowns in a language where NLP resources are limited. As an open-source software promulgated under the GNU General Public License v3.0 license, underthesea is well-suited for handling tasks involving complex token structures in Vietnamese.

A notable feature of the VMHQA dataset is the significant length disparity between questions, answers, and explanations. While questions and answers remain relatively concise, the explanations are much longer and reflect the sensitivity of mental health-related content. This poses a considerable challenge for the natural language model in Vietnamese, where resources are less mature than in other languages.

The variability in token length ranges from questions to extensive explanations, adding a layer of complexity and pushing the limit of the token level. Relying on the underthesea toolkit for token split, the dataset helps build robust question-answer systems capable of balancing deep contextual understanding critical to addressing the sensitive nature of mental health.

Table 2. VMHQA dataset statistics, where Q, A, E, and P represent the Question, Answer, Explanation, and Paragraph, respectively.

	Train	Dev	Test	Total
# Question	8000	1000	1000	10000
Vocab	19899	8061	8123	21842
Paragraph	7221	1314	1354	8599
Max Q tokens	1088	355	378	1088
Max A tokens	304	178	165	304
Max E tokens	3147	978	998	3147
Max P tokens	3047	549	740	3047
Avg Q tokens	19.42	19.11	19.07	19.36
Avg A tokens	15.23	14.32	14.54	15.07
Avg E tokens	81.38	77.55	80.49	80.91
Avg P tokens	96.38	94.76	95.49	96.13

3. Data analysis

3.1. Difficulty and diversity of questions

The VMHQA dataset reflects the complexities of real-world mental health, covering 1,166 subjects, including 249 disorders listed in the DSM-5. This broad scope captures the multifaceted nature of psychological diagnoses, where symptoms often overlap, and diagnoses are rarely straightforward. The dataset's questions are sourced from credible academic references and medical literature, mirroring real-life clinical scenarios where ambiguity and co-occurring symptoms are common.

By incorporating a question that involves multiple conditions and overlapping symptoms, the dataset challenges models to develop a nuanced understanding of mental health. Conditions like depression, anxiety, and personality disorders present similar symptoms, making an accurate diagnosis hard. VMHQA helps address these challenges and models to handle ambiguity.

This complexity is not a limitation but a key strength, as it exposes models to realistic scenarios and prepares them for use in clinical settings. The VMHQA dataset is thus a tool for advanced model evaluation to navigate the nature of mental health diagnosis in real-world practice.

3.2. Question types

The analysis centres on question types, classified using the 'List of Vietnamese Central Question Words' in Table 3 as adapted from the VIMQA research paper. VMHQA follows and modifies these classifications, adding complex questions that combine multiple clauses or question types. This structured schema ensures comprehensive analysis and understanding of simple questions in the VMHQA dataset.

Table 3. List of Vietnamese central question words (CQW).

Group	English CQW	Vietnamese CQW
Yes/No	Copulas (is, are) Aux (does, did)	Phải không, Đúng không
What	What	Là gì
Which	Which	Là cái nào, điều nào, điều gì, cái gì
Who	Who	Là ai
	Whom	Bởi ai
How	How many	Bao nhiêu
	How often	Bao lâu một lần
	How long	Bao lâu
	How	Như thế nào
When	When	Khi nào
Where	Where	Ở đâu
Why	Why	Tại sao, Vì sao

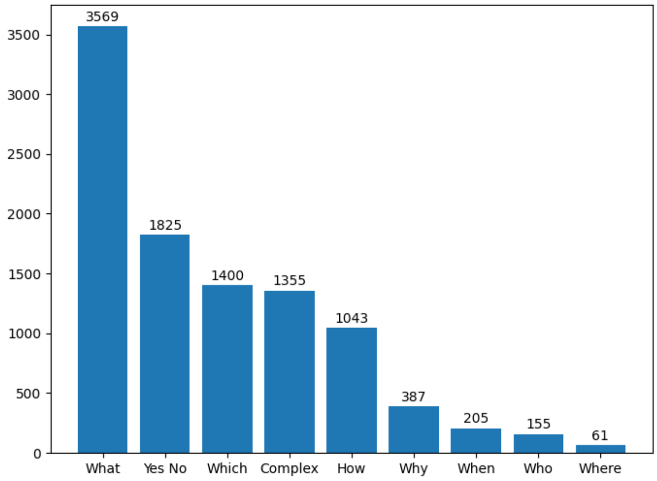


Figure 3. The distribution of question types in VMHQA.

Based on Figure 3, the distribution of question forms is presented in the figure, showing the breakdown of question types in the VMHQA dataset. 'What' questions dominate with 3,569 occurrences, emphasising a focus on factual information. 'Yes/No' questions follow with 1,825 instances, and 'Which' questions appear 1,400 times. Complex questions, which combine multiple types, occur 1,355 times. 'How' questions are frequent, with 1,043 occurrences, while 'Why' questions appear 387 times. Less common are 'When' (205), 'Who' (155), and 'Where' (61) questions. This distribution highlights the dataset's emphasis on factual, procedural, and explanatory queries, focusing on detailed and nuanced information.

3.3. Answer types

As shown in Figure 4, the distribution of answer types reveals a slight imbalance despite the collectors' efforts to maintain label balance. Answer option A constitutes the most significant share at 33.6%, while options B, C, and D account for 23.5%, 22.9%, and 20.0%, respectively. Although this imbalance is not severe, it could potentially introduce bias in the model's predictions, favouring answers associated with option A. This imbalance will be evaluated in the experimental section to determine whether it significantly impacts model performance and if adjustments are necessary.

Regarding answer length, the variability in the maximum token counts (304 for the training and complete sets, 178 for development, and 165 for the test) suggests potential differences in answer complexity across the splits. However, the average token counts across sets, ranging from 14.32 to 15.23, indicate consistency in answer length overall, which may reduce the risk of length-based bias. These factors will be considered when evaluating model behaviour and prediction reliability.

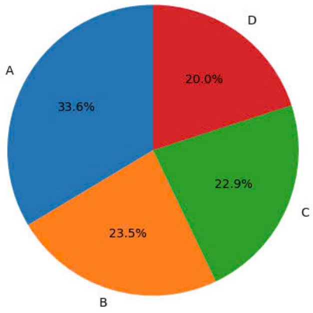


Figure 4. The distribution of answer types in VMHQA.

3.4. References

A team of professionals meticulously curated the VMHQA dataset to ensure high-quality content and reliable data. Unlike many Vietnamese QA datasets that rely heavily on Wikipedia, VMHQA is distinguished by drawing from diverse sources. Specifically, the dataset is built on 35 carefully selected academic books on psychology, mental health, and neuroscience. These sources provide a foundation for addressing mental health problems.

In contrast to datasets with an over-reliance on Wikipedia, VMHQA limits its use of Wikipedia to only 50 posts, ensuring that most of the content is derived from more specialised resources. A portion of the data, 2,961 web pages, was collected from the websites of reputable hospitals, mental health institutions, and therapy consultant centres. These sources ensure that the information within VMHQA is academically relevant and aligned with real-world clinical practices and the latest advancements in mental health care.

By prioritising data from well-regarded academic and clinical sources, VMHQA guarantees high-quality content that is reliable for training AI models and real-world mental health applications. This approach enhances the value, making it a tool for developing AI systems capable of understanding complex mental health scenarios and clinical decision-making with credibility.

4. Methodology

4.1. Data splitting

The VMHQA dataset was partitioned into training, development, and test sets following an 8:1:1 ratio. Given that the dataset comprises two main types of questions, a standard (e.g., 'what,' 'how,' and 'where') and a complex stratified splitting method based on question type was employed. This thorough stratification ensures that each subset maintains a proportional representation of standard and complex questions, preserving consistency across the split. An efficient distributed question type effectively mitigates the risk of over-fitting. It ensures the model is trained across the full spectrum of question complexities.

4.2. Retrieval-augmented generation

The VMHQA dataset contains 8,599 paragraphs, and the RAG system utilises two chunking methods: Paragraph Chunking and Agentic Chunking. The system uses DataStax's Astra Database with vector search capabilities to retrieve the data efficiently. The 'text-embedding-3-small' embedding model from GPT [18], which produces embedding with 1,536 dimensions, is used. Because the search method is based on cosine similarity, the system can find the most critical passage that answers the questions about mental health. This method guarantees effective dataset information retrieval.

Paragraph chunking. Paragraph Chunking was one of the two primary techniques used, in which every finished paragraph was embedded without being split. This method keeps the complete context of each paragraph while allowing the model to extract data from a sizable dataset efficiently. The model preserved the information's integrity by embedding entire paragraphs and storing them in the Astra Database, allowing for more contextually aware answers to complex queries. This approach made better comprehension and knowledge retrieval possible, which decreased the possibility of losing important context and enhanced the model's capacity to handle complex, domain-specific multiple-choice questions in the Vietnamese mental health dataset.

Agentic chunking. Using GPT-4o mini, the Agentic Chunking pipeline first divides paragraphs into propositions. This proposition, which stands for distinct facts or concepts taken from the original paragraphs, is then grouped according to similarities. Similar positions are organised in this clustering step, which is made possible by methods like Agglomerative Clustering, Uniform Manifold Approximation and Projection, and the SimCSE pre-trained model [19]. Following sorting, the propositions go through an evaluation process called Agentic Chunking. The proposition is added if a relevant chunk is found, and GPT-4o mini creates a summary for the modified chunk to ensure the updated data is accurately reflected. If no relevant chunk is found, GPT-4o mini generates an appropriate title.

This dynamic process, driven by the model's ability to analyse and categorise information, ensures that the result chunks are comprehensive and logically organised. The output includes a Chunk ID, Title, Summary, and a collection of propositions. The goal is

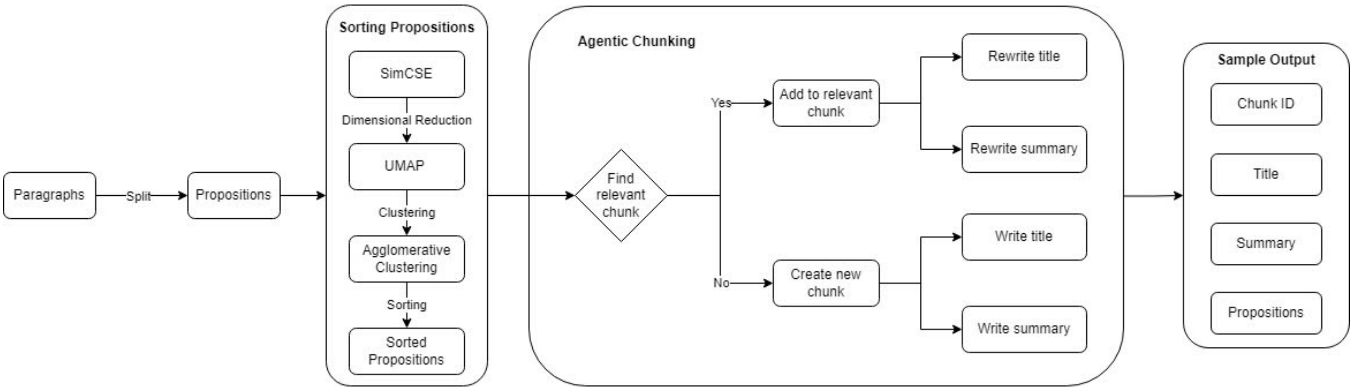


Figure 5. Agentic chunking pipeline.

to manage the information efficiently while preserving each proposition's contextual relevance.

4.3. BERT-based models

The baseline experiments aim to estimate the performance of BERT-based models in addressing mental health-related QA tasks in the context of the Vietnamese language. This evaluation will provide insights into methodologies and guide future advances in the field. The models considered in this study vary in their pre-training approaches, from general-purpose models to highly domain-specific ones.

Each model is based on the Bidirectional Encoder Representations from Transformers (BERT) architecture, fine-tuned for specific tasks relevant to mental health, focusing on question answering. The diversity in pre-training from broad, general-domain corpora to more specialised, in-domain data highlights the limits of each model in tackling mental health-related tasks.

PhoBERT [20] is a pre-trained language model arranged explicitly for Vietnamese, showing considerable enhancements in tasks such as Named Entity Recognition (NER) and Natural Language Inference (NLI), outperforming multilingual models like XLM-R. BioBERT [21], on the other hand, is tailored for biomedical text mining, excelling in tasks such as biomedical NER, Relation Extraction (RE), and Question Answering, though its reliance on static data limits its usefulness. ClinicalBERT [22] focuses on clinical notes and effectively predicts hospital readmission. Similarly, BlueBERT [23], trained on PubMed abstracts from the MIMIC-III dataset, improves NER and RE tasks but is constrained by outdated training data. MentalBERT [24] is highly specialised for mental health applications, excelling in detecting mental disorders and suicidal ideation, making it a promising tool for addressing Vietnamese mental health questions. BioMed-BERT [25], trained on the BREATHE dataset, achieves state-of-the-art results in biomedical QA and Information Retrieval (IR). At the same time, SciBERT [26], pre-trained on scientific publications, enhances performance in various NLP tasks such as sequence tagging and dependency parsing, particularly within scientific domains.

4.4. Large language models

The selection of models ranging from 7 to 9 billion parameters was driven by the need to evaluate their performance on the VMHQA dataset. This parameter range balances computational efficiency and the ability to handle complex, nuanced language tasks. Testing these models provided a clear understanding of their standalone capabilities in processing mental health-related questions. Additionally, the evaluation sought to explore whether these models could enhance performance by combining external knowledge with the model's pre-trained data when integrated with an RAG system. This dual approach aimed to determine whether a single model could handle the task effectively.

Llama 3.1 8B [27] was chosen due to its remarkable ability to handle multilingual tasks, with a focus on adaptability. With 8 billion parameters and a large amount of pre-training, the model dealt with the VMHQA dataset's complex, domain-specific mental health questions. It was a strong option for handling the complicated issues posed by mental health-related multiple-choice questions in Vietnamese because of its ability to understand complex language and focus on specific topics.

Gemma 2-9B [28] was chosen for its impressive balance of size and performance. After being trained on 8 trillion tokens from various sources, such as code, scientific articles, and web documents, Gemma 2 was appropriate for tasks that needed to be completed quickly and precisely. It is very effective at answering questions about mental health because of its architecture, which combines local and global attention layers to manage a more extended context.

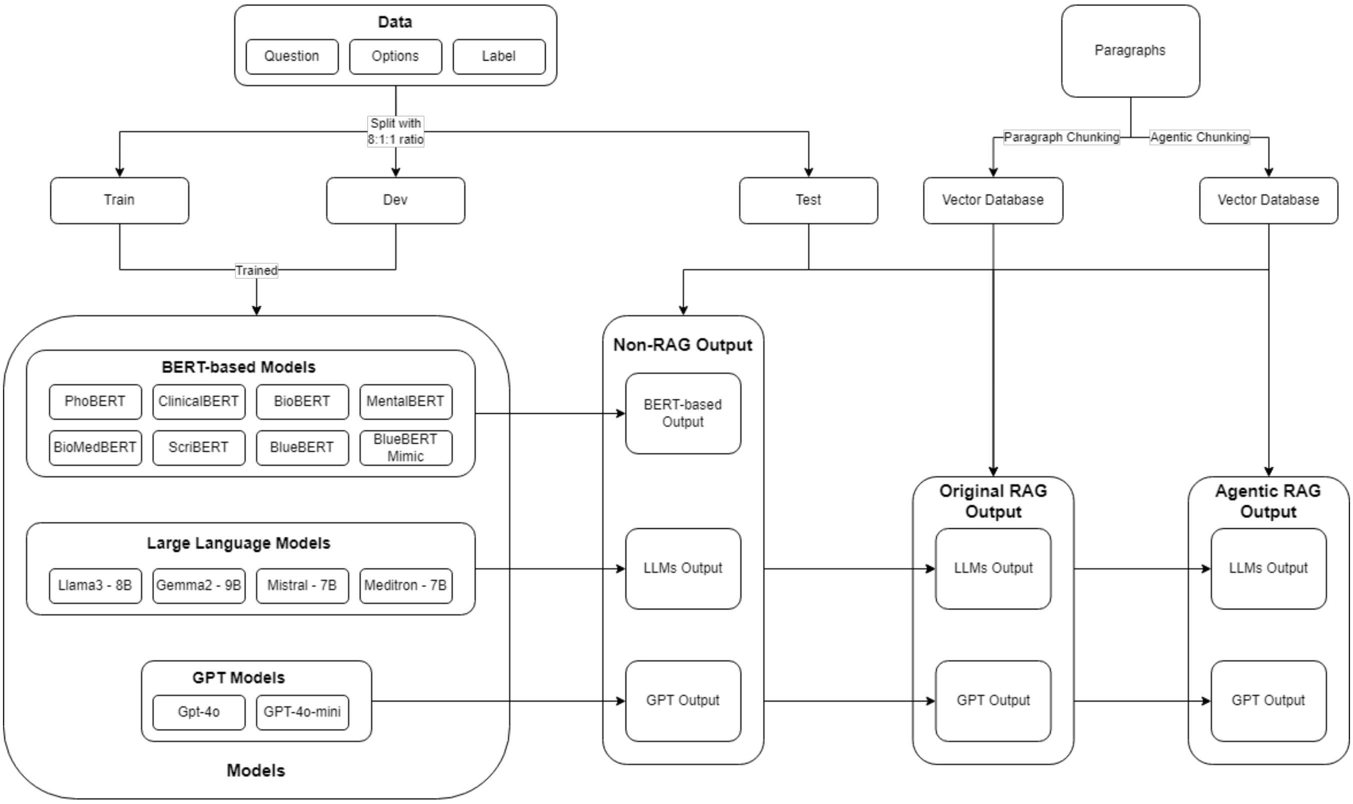


Figure 6. Overview of model architectures and data flow for VMHQA evaluation.

Mistral 7B [29] and Meditron 7B [30] were also selected for their specialised features. Mistral 7B, with its architectural innovations like Grouped-Query Attention and Sliding Window Attention, delivered high performance while maintaining computational efficiency. It performed very well on more complicated and lengthy tasks. Meditron 7B showed an advanced understanding of medical terms after training on medical datasets like clinical notes and exam questions. It successfully tackled the domain-specific issues in the VMHQA dataset because of its alignment with the mental health domain.

4.5. GPT models

When reviewing the open-access model, the VMHQA dataset also assessed GPT-4o and GPT-4o mini. These state-of-the-art architectures offer an ordinary way to determine open-access models like Llama 3.1 8B, Gemma 2-9B, Mistral 7B, and Meditron 7B. GPT-4o and GPT-4o mini were part of the test to see if costly devices with features would exceed open-access choices when reacting to particular questions.

The comparison focuses on accurate and linguistic understanding and their ability to manage the Vietnamese language and medical terminology challenges. Testing both models and those integrated with an RAG system provided insight into their effectiveness across configurations. This evaluation allows for understanding the trade-off between using the access model and a paid alternative, contributing to a decision on the approach for handling the VMHQA dataset.

5. Experiments

5.1. Experimental settings

The experiment evaluated eight BERT-based models on the Vietnamese mental health question-answering task using initial parameters derived from the MedMCQA dataset, including a batch size of 16, a maximum sequence length of 192, a learning rate of 2e-4, a hidden dropout probability of 0.4, and 5 training epochs. PhoBERT demonstrated the highest adaptability and performance during the initial evaluation and was subsequently fine-tuned with revised parameters, maintaining the batch size of 16 and 5 training epochs while updating the maximum sequence length to 256, the learning rate to 2e-5, and the hidden dropout probability to 0.3. These optimised settings were then sequentially applied to the other models, and if a model surpassed PhoBERT in performance in the dev dataset, it was further fine-tuned. This constant operation aims to find the most practical BERT-based model for Vietnamese mental health tasks. Figure 7 shows the layers and pipeline used in this setting.

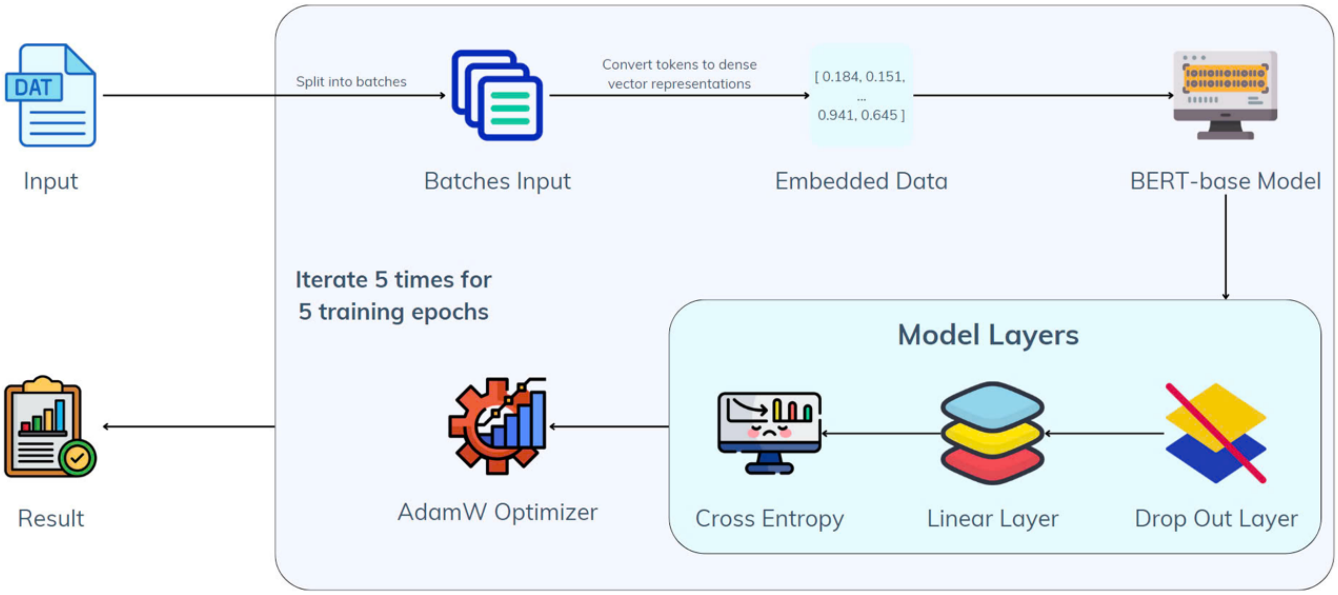


Figure 7. BERT-based model training pipeline in Experimental setup.

The experimental settings for fine-tuning the Large Language Models, including Mistral 7B, Llama 3.1 8B, Gemma 2-9B, and Meditron 7B, involve an iterative process to optimise performance on the Vietnamese mental health question-answering task. Each model was fine-tuned using the Low-Rank Adaptation (LoRA) method, with critical parameters set at r = 32, LoRA alpha = 16, and a learning rate 0.0002. The train used a batch size of 2 with gradient accumulation steps of 4, optimising with the AdamW 8-bit optimiser over two epochs, including five warm-up steps. The models utilised a weight decay of 0.01 and a linear learning rate scheduler. Mixed precision training was employed based on the system's bfloat16 support, and a random seed of 3407 ensured consistency across experiments. These parameters were applied uniformly to identify the most effective configuration for the task. Figure 8 shows the layers and pipeline used in this setting.

As for GPT models, using GPT-4o with GPT-4o mini was trained for the Vietnamese mental health question-answering task. The training and validation datasets consisted of structured data, including a system prompt, the question, four multiple-choice options, and the correct label. We were fine-tuning to enhance the model's ability to understand and generate precise responses for mental health-related multiple-choice questions.

5.2. Baseline results

BERT-based models. The baseline results for the BERT-based models are presented in Table 4 (referenced in Section 4). These models were evaluated on the Vietnamese Mental Health Question Answering dataset, focusing on their performance concerning precision and F1-score for development.

PhoBERT emerged as the best-performing model, achieving an accuracy of 0.67 on the development set and 0.64 on the test set, corresponding to F1 Scores of 0.66 and 0.64. This highlights the adaptability of a language model fine-tuned for Vietnamese tasks.

MentalBERT also demonstrated competitive results, with an accuracy of 0.64 on both the development and test sets and similar F1-scores of 0.64. This is noteworthy, as MentalBERT is explicitly designed for mental health applications, which aligns well with the focus of the VMHQA dataset.

Other models like BioMedBERT and SciBERT, tailored for broader scientific or biomedical texts, performed moderately well, with accuracies ranging from 0.61 to 0.64 and F1-scores from 0.61 to 0.64. This indicates that while general biomedical or scientific models can handle mental health-related questions to some extent, they are outperformed by models such as PhoBERT and MentalBERT.

Notably, BlueBERT and BlueBERT + MIMIC, while focused on biomedical data, showed performance with accuracy scores of 0.60 to 0.61 and F1 scores of 0.59 to 0.60. This underperformance is likely due to the outdated nature of the data this model was trained on, limiting their ability to generalise to more recent and nuanced mental health contexts.

Large language models. The assessment of LLMs for the VMHQA task has successfully underlined the potential of open-source models, such as Gemma 2-9B, giving avenues of trailblazing ideas for cost-effective AI

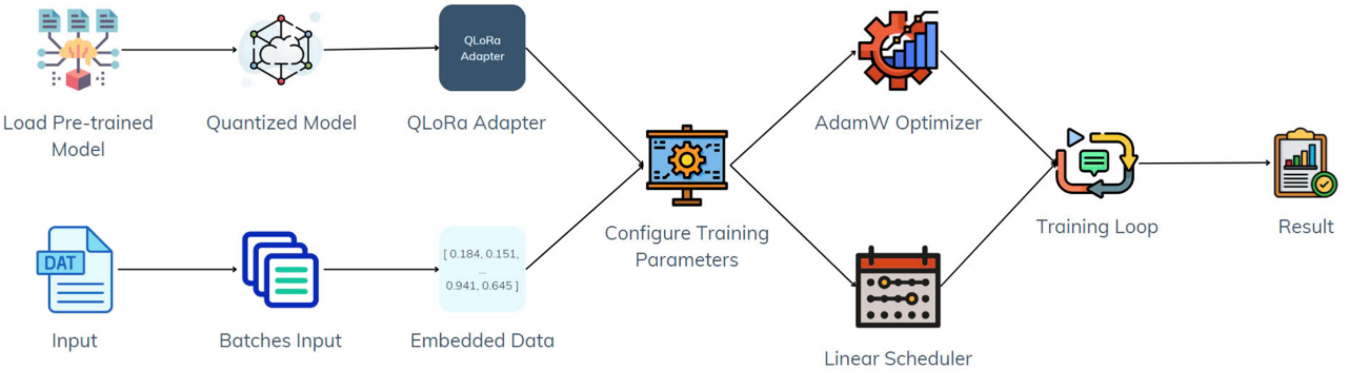


Figure 8. LLM fine-tuning and training pipeline for 7-9B parameter models.

Table 4. Performance of all baseline models in accuracy (percentage) and F1-score on VMHQA test and dev set.

Model	Accuracy		F1-Score	
	Dev	Test	Dev	Test
PhoBERT	**0.67**	**0.64**	**0.66**	**0.64**
ClinicalBERT	0.62	0.62	0.62	0.62
BioBERT	0.61	0.62	0.61	0.62
BlueBERT	0.63	0.60	0.63	0.60
BlueBERT + MIMIC	0.61	0.60	0.61	0.59
MentalBERT	0.64	0.62	0.64	0.62
BioMedBERT	0.64	0.61	0.64	0.61
SciBERT	0.63	0.62	0.63	0.62

Table 5. The performance of LLMs in the test file.

Models	Origin	Fine-tuned
Llama 3.1 8B	77.40%	89.10%
Mistral 7B	70.40%	84.40%
Gemma 2 9B	**83.50%**	**90.70%**
Meditron 7B	55.10%	80.00%
GPT 4o	72.10%	90.10%
GPT 4o mini	69.60%	87.60%

applications. Gemma 2-9B has surpassed GPT-4o with an impressive accuracy of 90.70%

Other open-access models (such as Mistral 7B and Llama 3.1 8B) also showed positive results, achieving fine-tuned accuracies of 84.40% and 89.10%, respectively. These models highlight the potential for cost-effective solutions, particularly in resource-constrained environments. In contrast, Meditron 7B, trained on medical datasets, lagged with 80.00% accuracy.

The standout performance of Gemma 2-9B suggests a bright future for low-cost, scalable applications. These include mental health chatbots, virtual assistants, and early diagnostic tools. Moreover, these models provide opportunities for localised solutions, handling multilingual tasks, and empowering organisations to develop affordably.

In other words, such open-source models discussed above have proven effective in presenting an affordable alternative to premium models, significantly reducing costs for mental health consultation services supported by AI. Such an innovation promises a paradigm shift towards AI-backed services for mental health, paving the way for future innovations.

Performance with full-context chunking and agentic chunking RAG. When integrated with RAG systems, the models demonstrate further improvement in handling long-context and retrieval-based queries, as reflected in the tables assessing full-context and agentic chunking RAG.

Gemma 2-9B consistently outperforms other models across different RAG configurations, achieving the highest Top 1 accuracy of 85.90% in full-context chunking and 86.00% in agentic chunking, indicating its ability.

Trained GPT-4o and Trained GPT-4o mini also deliver competitive performances, with Top 1 accuracies of 87.20% and 88.30% in full-context chunking, and 87.90% and 88.00% in agentic chunking, respectively. These results highlight the effectiveness of combining pre-training data with a retrieval mechanism.

Llama 3.1 8B and Mistral 7B performed reasonably well but fell behind Gemma 2-9B and GPT-4o, with Top 1 accurate in the 72.80% to 80.20% range across both RAG configurations. This suggests that while they are powerful language models, their performance benefits more from fine-tuning than from RAG integration compared to the more robust performance of Gemma 2-9B and GPT-4o.

In our study, although Gemma 2-9B in a non-RAG configuration achieved the highest Top-1 accuracy (90.70%), RAG-integrated models like trained GPT-4o mini (Top-1: 88.60%) still demonstrated competitive performance, especially in chunking strategies that promote focused retrieval. This supports the argument that RAG methods remain promising when accuracy is balanced with traceability, modularity, and generalizability in deployment contexts.

In summary, the experimental result indicated that models fine-tuned with domain-specific data, such as PhoBERT and MentalBERT for BERT-based models and Gemma 2-9B and GPT-4o for LLMs, are better suited for handling the challenge of mental health-related question-answering tasks. Moreover, the incorporation of RAG further enhances performance.

Table 6. The performance of LLMs with Full-Context Chunking RAG.

Model	Top 1	Top 3	Top 5	Top 10
Llama 3.1 8B	72.80%	55.10%	58.20%	55.90%
Mistral 7B	80.00%	79.50%	79.30%	76.50%
Gemma 2 9B	85.90%	86.30%	86.40%	85.90%
Meditron 7B	45.90%	37.40%	35.20%	35.10%
GPT 4o	67.00%	69.10%	69.50%	69.70%
GPT 4o mini	60.10%	61.20%	63.80%	64.40%
Trained GPT 4o	87.20%	84.80%	83.20%	83.50%
Trained GPT 4o mini	**88.30%**	**88.60%**	**88.40%**	**88.50%**

Table 7. The performance of LLMs with Agentic Chunking RAG.

Model	Top 1	Top 3	Top 5	Top 10
Llama 3.1 8B	72.80%	54.00%	58.80%	53.40%
Mistral 7B	80.20%	79.50%	78.50%	77.50%
Gemma 2 9B	86.00%	87.00%	87.30%	86.60%
Meditron 7B	48.20%	40.50%	36.40%	35.80%
GPT 4o	64.40%	70.30%	70.30%	68.70%
GPT 4o mini	59.50%	62.60%	63.40%	63.30%
Trained GPT 4o	87.90%	83.80%	83.80%	84.60%
Trained GPT 4o mini	**88.00%**	**88.50%**	**87.90%**	**88.60%**

6. Discussion

The baseline experiments show how well LLMs and BERT-based models perform on the VMHQA task. The strong performance was demonstrated by models like PhoBERT and MentalBERT, illustrating the importance of domain-specific fine-tuning, especially for understanding mental health terminology and the linguistic complexity of Vietnamese. The higher accuracy supported the need for specialised pre-training, and the F1-scores these models got compared to general models such as BioBERT and SciBERT. On the other hand, models trained on out-of-date biomedical data, such as BlueBERT and BlueBERT + MIMIC, performed poorly, highlighting the necessity of current, domain-specific datasets.

Gemma 2-9B's superior performance in VMHQA can be attributed to the diverse and comprehensive dataset utilised in its training. Initially trained on various English-language web text documents, Gemma 2 developed a deep understanding of complex linguistic structures through exposure to multiple styles, topics, and vocabularies. Its training also included programming languages and mathematical texts, enhancing its logical reasoning and precision, which is essential for handling structured queries in mental health applications. With 9 billion parameters, Gemma 2-9B offers greater capacity than models like Mistral 7B and Llama 3.1 8B, allowing for more effective retention and processing of large volumes of context-rich information. Combined with specialised fine-tuning for the linguistic characteristics of Vietnamese, Gemma 2-9B excels in managing the tonal and grammatical complexities of the language. Its flexible chunking strategies, full-context, and agentic enable it to process lengthy, multi-turn queries in mental health dialogues while maintaining critical contextual information. These attributes position Gemma 2-9B as a highly effective and scalable solution, outperforming open-access and proprietary models in this specialised domain.

Although GPT-4o and GPT-4o mini showed good performance, especially after fine-tuning, their capabilities were further increased by integrating RAG systems. Models can handle complex and context-rich queries thanks to RAG's ability to retrieve external knowledge from massive datasets. This capability is crucial for mental health applications requiring specialised terminology and nuanced language. RAG-based models might, however, perform poorly in VMHQA scenarios despite these developments. The propensity of RAG systems to produce confused or irrelevant responses-often referred to as 'hallucinations'-when processing substantial volumes of external data is a significant problem. This happens because the retrieved content might not precisely match the requirements of the query, especially for domain-specific tasks like mental health, where precision and contextual awareness are crucial.

Moreover, RAG systems' embedding and searching strategies can struggle with subtle emotional nuances and sensitive contextual clues in mental health dialogues. Queries often involve implicit meanings or expressions of distress, which cannot be addressed solely by retrieving information from external sources. The embedding processes, with a limit of 1536 dimensions, may fail to capture the emotional and psychological depth of the query, while search strategies like cosine similarity may produce overly factual or fragmented responses, missing the patient's underlying intent. This limitation has significant implications for the practical application of RAG in mental health care, where accurate language interpretation is vital for compassionate and practical support.

Two approaches for managing long-context queries in RAG systems have been examined: full and agentic chunking. Agentic chunking separates the context into smaller, easier-to-manage pieces, allowing the model to concentrate on essential components and facilitating more efficient processing. In contrast, full-context chunking gives the model the entire context at once, improving its preservation of long-term dependencies. The way RAG and non-RAG pipelines respond to queries is a difference between them. RAG pipelines process context-rich queries by dividing them into manageable components and using chunking strategies to retrieve external knowledge. While agentic chunking concentrates on smaller segments, full-context chunking manages the entire query simultaneously. However, non-RAG pipelines, like Gemma 2-9B, often outperform RAG-based pipelines, achieving a fine-tuned accuracy of 90.70% compared to RAG-configured pipelines like GPT-4o. This advantage is due to the non-RAG pipeline's reliance on internal, pre-trained knowledge, which reduces latency and avoids introducing noise or irrelevant information from external retrieval. In domain-specific tasks like VMHQA, where language nuances are critical, non-RAG pipelines provide more consistent and accurate responses by focusing solely on internal knowledge. Thus, while RAG pipelines offer advantages in handling complex queries, non-RAG pipelines, with their streamlined, fine-tuned architecture, often prove more effective in specialised tasks.

According to the results, Gemma 2–9B consistently performed better than other models in full-context and agentic chunking RAG configurations. It has successfully managed global and focused retrieval-based tasks, achieving the highest Top 1 accuracy of 85.90% in full-context chunking and 86.00% in agentic chunking. Gemma 2-9B's performance in both RAG configurations highlights its adaptability and resilience in managing intricate, domain-specific content, especially when handling questions about mental health. With Top 1 accuracies of 87.20% and 88.30% in full-context chunking and 87.90% and 88.00% in agentic chunking, respectively, GPT-4o and GPT-4o mini, on the other hand, perform competitively but fell short of Gemma 2-9B. This means open-source models can be as effective as proprietary models like GPT-4o.

The contrast between agentic and full-context chunks highlights how important chunking methods are in retrieval-based tasks. Full-context chunking is more helpful for queries that need an extensive understanding of more extended text sequences, like those found in conversations about mental health. On the other hand, agentic chunking has more effect when particular data must be retrieved and processed separately, making it appropriate for some educational or diagnostic applications. In summary, the findings show that when fine-tuned, LLMs combined with RAG systems offer an effective strategy for managing difficulties. A bright future for reasonable, highly effective AI solutions in specialised fields is shown by the ability of open-source models like Gemma 2-9B to surpass premium models like GPT-4o in various RAG configurations.

In addition to the technical considerations, addressing the challenges of handling sensitive mental health information while maintaining scientific and ethical integrity is crucial. One of the significant challenges in developing the VMHQA task is the potential inconsistency in the dataset due to its collection by psychology students rather than senior professionals. While the students were trained and followed a standardised process, the lack of extensive clinical experience may lead to variability in the quality and depth of the annotations. This can result in a dataset that is not entirely consistent, as more nuanced cases might not be captured with the same accuracy as they would be by seasoned experts. Another issue is the overlap of symptoms between different mental health conditions, such as depression and anxiety. These disorders often share common signs, which can make it difficult to formulate questions that are both specific and diagnostically conclusive. Some questions in the dataset may provide only general information, which might not wholly indicate the complexity required for real-world diagnosis, potentially leading to answers that are not entirely convincing or definitive. Additionally, the complexity of mental health symptoms, which can present differently across individuals, adds another layer of difficulty in creating a dataset that is both comprehensive and nuanced. The challenge of ensuring diagnostic accuracy and the ethical handling of sensitive mental health information underscores the need for continuous refinement of the dataset and careful consideration of these factors in model development.

Ethical and Practical Implications. VMHQA is best applied in pre-consultation contexts, such as delivering reliable, accessible information about mental health conditions to users before they engage with a healthcare professional. Its structured format and medical alignment can assist in raising awareness and guiding individuals to seek help. However, it must not be used to diagnose, treat, or replace the role of licensed therapists or medical professionals. Any application built on VMHQA should include clear disclaimers and safeguards to ensure users understand its purpose is informational, not clinical.

7. Error Analysis

To estimate the precision of the dataset, a validation test was organised with a team of psychology students across different academic levels, including first-year students, sophomores, juniors, and seniors. A random sample of 100 records was selected from the test set and categorised by question type. The students were given 100 minutes to complete the task. The average accuracy achieved by the students was 58.3%. Several challenges were identified during this process. The time constraint imposed on the participants likely affected their performance, as many struggled to read and thoroughly analyse the lengthy question-option pairs.

Furthermore, an average similarity score of 52.42% across the dataset and 50.73% in the human-evaluated portion suggested that the high similarity between answer options (e.g., 'A: Strongly Agree' and 'B: Agree') led to confusion. This similarity increased cognitive load and caused participants to second-guess their choices, resulting in lower accuracy. Despite their knowledge of the subject, students often hesitated or overanalyzed subtle differences between the options, contributing to the overall performance drop.

Table 8. Accuracy of AI model by question type and record count.

Question Type	Number of records	Accuracy
What	357	0.91
Yes No	182	0.95
Which	140	0.89
Complex	136	0.91
How	104	0.81
Why	39	0.92
When	20	0.95
Who	16	0.88
Where	6	1

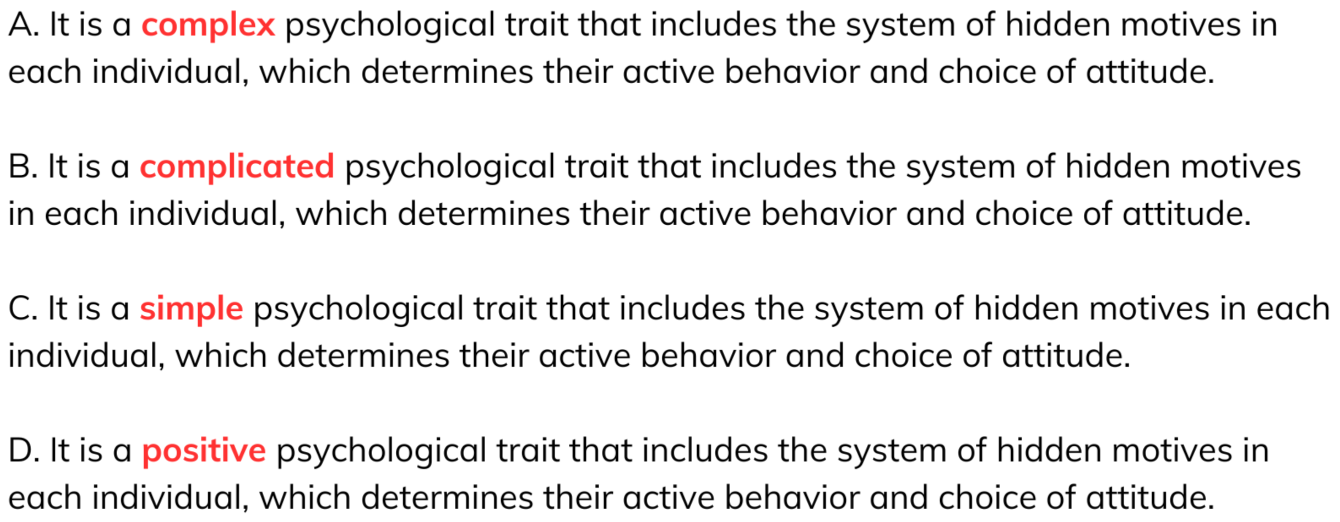
The error analysis of the model's performance through diverse question types, as shown in Table 8, offers further observation into the advantages and drawbacks of the AI model. The model performed exceptionally well on Yes/No and When questions, achieving an accuracy of 0.95%, highlighting its proficiency in handling straightforward factual queries. Similarly, it performed well on 'What' and 'Complex' questions, with accuracies of 0.91%, indicating that the model can effectively address fact-based inquiries. However, the model struggled with How questions, where the accuracy dropped to 0.81%, suggesting that it is challenging to generate accurate responses for procedural or explanatory queries requiring deeper reasoning. Additionally, the model achieved an accuracy of 0.88% on Who questions, indicating difficulty recognising or recalling specific entities. Despite a perfect accuracy of 1.0% for 'Where' questions, the small number of records (six) limits the significance of this result. The model's performance on 'Why' questions was respectable, with an accuracy of 0.92%, suggesting moderate capability in handling reasoning-based queries. While the model excels at direct, fact-based questions, it requires further fine-tuning to improve its handling of more interpretive or complex questions.

In particular, the lower performance on the 'How' question type can be attributed to several factors. Among the questions that were mispredicted, 60% involved options related to quantities, such as 'How much,' 'How many,' or 'How long.' This is problematic because the dataset primarily consists of context-based word responses, confusing the model in predicting the numbers precisely. Additionally, 20% of the incorrect predictions occurred due to the high similarity between answer options. In some cases, such as a question where all four options were almost identical semantically and syntactically, the model struggled to differentiate between the choices, leading to confusion. Lastly, another 20% of the wrong predictions appear to be a result of randomness, which is inherent in multiple-choice questions. In these instances, the model may have guessed the answer due to a lack of apparent distinguishing features between the options. These specific issues illustrate why the 'How' question type presents more challenges for the model and why targeted improvements are necessary to address them.

Table 9. Accuracy of AI Model by Answer Type and Record Count.

Answer Type	Number of records	Accuracy
A	352	0.93
B	239	0.89
C	233	0.88
D	176	0.92

Moreover, as presented in Table 9, the model's performance by answer type demonstrates a lack of significant bias despite a notable disparity in the number of records between different answer options. For instance, option A has 352 records compared to only 176 for option D. Yet, the model shows comparable accuracy for both, with 0.93 for A and 0.92 for D. Similarly, answer types B and C show slight variations in accuracy, with 0.89 and 0.88, respectively. These findings suggest that the model has not overfitted to the uneven distribution of answer types and maintains consistent accuracy across all options. The model's robustness in handling various answer types without bias further supports its general reliability in making predictions, ensuring fair performance even with an imbalanced dataset. This consistency highlights the model's capacity to handle different answer options with similar levels of accuracy, reinforcing the validity of its predictions across a wide range of inputs.



A. It is a **complex** psychological trait that includes the system of hidden motives in each individual, which determines their active behavior and choice of attitude.

B. It is a **complicated** psychological trait that includes the system of hidden motives in each individual, which determines their active behavior and choice of attitude.

C. It is a **simple** psychological trait that includes the system of hidden motives in each individual, which determines their active behavior and choice of attitude.

D. It is a **positive** psychological trait that includes the system of hidden motives in each individual, which determines their active behavior and choice of attitude.

Figure 9. An example about ambiguity in answer options (translated to English).

8. Conclusion

The research results show significant developments in using AI models to address issues related to mental health in Vietnam. The dataset about mental health offers a valuable tool for improving Vietnamese NLP. The thorough examination of many models shows that domain-specific adaptation and fine-tuning are crucial for achieving high accuracy.

The open-source model Gemma 2-9B demonstrates the potential for cost-effective AI solutions in domains like mental health, where scalability and accessibility are critical. However, the integration of RAG does not improve performance as expected. Sometimes, the RAG system underperformed compared to non-RAG models; this means that RAG is not always suitable for handling complex, context-rich queries in this domain. The use of RAG, even with chunking strategies like Agentic Chunking and Full Context Chunking, did not deliver the anticipated improvements. This suggests that current RAG strategies may not be well-suited for specific, specialised tasks. Future work will address the limitation of the current Retrieval-Augmented Generation system cause its performance has not matched that of the non-RAG model in this task. RAG may be helpful by integrating outside knowledge, but it does not always give accurate or valuable data. Future strategies should improve the problem by replacing or refining RAG mechanisms with more efficient retrieval techniques. The alignment between the query and the retrieved knowledge can be improved, for instance, by using more target retrieval techniques or improving the embedding models used for retrieval. Furthermore, adopting a strategy that emphasises more effective and specific retrieval, using a hybrid approach, can assist models in better managing the complex demands of questions related to mental health. Future research attempts to optimise information retrieval and enhance answer accuracy in this area by changing the retrieval strategy and investigating alternative approaches.

Acknowledgement. This research is funded by International University, VNU-HCM under grant number SV2023-02-IT. We also would like to express our earnest gratitude to the National Mental Hospital 2 team for their invaluable support in consultation on dataset design and validating its quality, which have made this work possible. Moreover, this research is supported by the Central Interdisciplinary Laboratory in Electronics and Information Technology, 'AI and Cooperation Robot,' International University-Vietnam National University of Ho Chi Minh City.

References

[1] R.M., B., Vardhan, K.B., Nidhish, M., Kiran C., S., Nahid Shameem, D. and Sai Charan, V. (2024) Eye Disease Detection Using Deep Learning Models with Transfer Learning Techniques. *ICST Transactions on Scalable Information Systems* **11**. doi:10.4108/eetsis.5971.

[2] Bhuvanya, R., Kujani, T. and Sivakumar, K. (2024) Fusing Attention and Convolution: A Hybrid Model for Brain Stroke Prediction. *ICST Transactions on Scalable Information Systems* **11**. doi:10.4108/eetsis.7022.

[3] Alvi, A.M., Khan, M.J., Manami, N.T., Miazi, Z.A., Wang, K., Siuly, S. and Wang, H. (2024) XCR-Net: A Computer Aided Framework to Detect COVID-19. *IEEE Transactions on Consumer Electronics* **70**(4): 7551–7561. doi:10.1109/TCE.2024.3446793.

[4] Tawhid, M.N.A., Siuly, S., Wang, K. and Wang, H. (2024) GENet: A Generic Neural Network for Detecting Various Neurological Disorders From EEG. *IEEE Transactions on Cognitive and Developmental Systems* **16**(5): 1829–1842. doi:10.1109/TCDS.2024.3386364.

[5] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. (2016), SQuAD: 100,000+ Questions for Machine Comprehension of Text. 1606.05250.

[6] Lai, G., Xie, Q., Liu, H., Yang, Y. and Hovy, E. (2017), RACE: Large-scale ReAding Comprehension Dataset From Examinations. 1704.04683.

[7] Vilares, D. and Gómez-Rodríguez, C. (2019), HEAD-QA: A Healthcare Dataset for Complex Reasoning. 1906.04701.

[8] Jin, D., Pan, E., Oufattole, N., Weng, W.H., Fang, H. and Szolovits, P. (2020), What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. 2009.13081.

[9] Pal, A., Umapathi, L.K. and Sankarasubbu, M. (2022), MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. 2203.14371.

[10] Huy, T.D., Tu, N.A., Vu, T.H., Minh, N.P., Phan, N., Bui, T.H. and Truong, S.Q.H. (2021) ViMQ: A Vietnamese Medical Question Dataset for Healthcare Dialogue System Development. **1517**, 657–664. doi:10.1007/978-3-030-92310-5_76. 2304.14405.

[11] Le, K., Nguyen, H., Le Thanh, T. and Nguyen, M. (2022) VIMQA: A Vietnamese dataset for advanced reasoning and explainable multi-hop question answering. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S. *et al.* [eds.] *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (Marseille, France: European Language Resources Association): 6521–6529.

[12] Nguyen, K.V., Nguyen, D.V., Nguyen, A.G.T. and Nguyen, N.L.T. (2020), A Vietnamese Dataset for Evaluating Machine Reading Comprehension. 2009.14725.

[13] American Psychiatric Association (2013) *Diagnostic and Statistical Manual of Mental Disorders* (American Psychiatric Association), fifth edition ed. doi:10.1176/appi.books.9780890425596.

[14] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H. *et al.* (2021), Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2005.11401.

[15] Chen, T., Wang, H., Chen, S., Yu, W., Ma, K., Zhao, X., Zhang, H. *et al.* (2024), Dense X Retrieval: What Retrieval Granularity Should We Use? 2312.06648.

[16] (2024), About the USMLE | USMLE, https://www.usmle.org/about-usmle.

[17] Vu Anh (2024), Undertheseanlp/underthesea, Under The Sea.

[18] OpenAI (2024), GPT-4 Technical Report. 2303.08774.

[19] Gao, T., Yao, X. and Chen, D. (2022), SimCSE: Simple Contrastive Learning of Sentence Embeddings. 2104.08821.

[20] Nguyen, D.Q. and Nguyen, A.T. (2020), PhoBERT: Pre-trained language models for Vietnamese. 2003.00744.

[21] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J. (2020) BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**(4): 1234–1240. doi:10.1093/bioinformatics/btz682. 1901.08746.

[22] Huang, K., Altosaar, J. and Ranganath, R. (2020), ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. 1904.05342.

[23] Peng, Y., Yan, S. and Lu, Z. (2019), Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. 1906.05474.

[24] Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P. and Cambria, E. (2021), MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. 2110.15621.

[25] Chakraborty, S., Bisong, E., Bhatt, S., Wagner, T., Elliott, R. and Mosconi, F. (2020) BioMedBERT: A Pre-trained Biomedical Language Model for QA and IR. In *Proceedings of the 28th International Conference on Computational Linguistics* (Barcelona, Spain (Online): International Committee on Computational Linguistics): 669–679. doi:10.18653/v1/2020.coling-main.59.

[26] Beltagy, I., Lo, K. and Cohan, A. (2019), SciBERT: A Pretrained Language Model for Scientific Text. 1903.10676.

[27] Dubey, A., Jauhri, A., Pandey, A., Kadian, A. and Al-Dahle (2024), The Llama 3 Herd of Models. 2407.21783.

[28] Team, G. (2024), Gemma 2: Improving Open Language Models at a Practical Size. 2408.00118.

[29] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F. *et al.* (2023), Mistral 7B. 2310.06825.

[30] Chen, Z., Cano, A.H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M. *et al.* (2023), MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. 2311.16079.