# VM14K: FIRST VIETNAMESE MEDICAL BENCHMARK

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Medical benchmarks are indispensable for evaluating the capabilities of language models in healthcare for non-English-speaking communities,therefore help ensuring the quality of real-life applications. However, not every community has sufficient resources and standardized methods to effectively build and design such benchmark, and available non-English medical data is normally fragmented and difficult to verify. We developed an approach to tackle this problem and applied it to create the first Vietnamese medical question benchmark, featuring 14,000 multiple-choice questions across 34 medical specialties. Our benchmark was constructed using various verifiable sources, including carefully curated medical exams and clinical records, and eventually annotated by medical experts. The benchmark includes four difficulty levels, ranging from foundational biological knowledge commonly found in textbooks to typical clinical case studies that require advanced reasoning. This design enables assessment of both the breadth and depth of language models' medical understanding in the target language thanks to its extensive coverage and in-depth subject-specific expertise. We release the benchmark in three parts: a sample public set (4k questions), a full public set (10k questions), and a private set (2k questions) used for leaderboard evaluation. Each set contains all medical subfields and difficulty levels. Our approach is scalable to other languages, and we open-source our data construction pipeline to support the development of future multilingual benchmarks in the medical domain.

## 1 INTRODUCTION

The rise of Large Language Models (LLMs) in healthcare has become more prominent than ever, projecting tremendous impacts across various fields such as clinical research, diagnosis support systems and other different medical specialties (Liu et al., 2024; Al-Garadi et al., 2025; Shah et al., 2024). In such a critical field like medicine, medical benchmarks are indispensable for evaluating the quality of these models before applying to real-world application. However, the lack of diversity in terms of language - most medical resources including data, models, and benchmarks are primarily in English, or Chinese - implies the need for developing benchmarks in other non-popular languages to support local communities (Jin et al., 2023; Kasai et al., 2023; Kweon et al., 2024; Yu et al., 2024). Force prompting in target language or pure translate question could be an option, however it might suffer from information loss, hallucination due to inherent difference in language semantic (Gérardin et al., 2023; Jin et al., 2023). Additionally, while it is straightforward to develop and test LLMs on universal benchmarks, testing for highly-specialized fields like medical in local languages requires careful design with detailed standards (Jin et al., 2019; Ben Abacha et al., 2017; Hendrycks et al., 2021b; Pal et al., 2022). Several reports highlight how off-the-shelf models and benchmarks are not optimal for multilingual applications; common problems include missing concepts in specific languages or special characteristics in the medical language of particular regions (Gérardin et al., 2023; Jin et al., 2023).

Based on our knowledge, there is no universal process for designing local medical benchmarks for LLMs. Collection methods are often fragmented or even nonexistent in many languages, and the definition and level of detail sometimes differ across regions, posing difficulties in the testing process. Moreover, it poses significant challenges such as data verification, as medical data is normally hard to verify and heavily skewed toward common diseases in the distribution of questions(Wang et al., 2025; Qin & Tong, 2025; Park et al., 2024).

Virtually all countries have a corpus of medical textbooks and testing systems for their students and doctors. Even if not every community has a modernized medical education system, these materials are at least verified or standardized by experts to maintain educational quality. We consider this a valuable source of truth that is already verified by experts in the field and continuously maintained and updated. More importantly, most historical data is public and could be explored from different perspectives to serve the purposes of LLMs in the medical domain, from training to testing model quality (Zhang & Ni, 2024; Zhang et al., 2024; Qiu et al., 2024; Lamurias et al., 2020).

To advance the progress of local LLMs in the medical field for non-English speaking communities, we **(1) developed a scalable framework** that curate medical data from various sources (websites, text, PDF), **(2) defined a simple yet flexible standard** for designing medical benchmarks for local LLMs, and **(3) created a first medical benchmark for Vietnamese** with 14,000 questions spanning 34 medical specialties, carefully annotated by medical students across four difficulty levels. We evaluated robustness of our benchmark on multiple foundation and open source medical models. We also open-source our data infrastructure to contribute to advancements in the medical LLM field with broader impact.

## 2 RELATED WORKS

### 2.1 NON-ENGLISH MEDICAL BENCHMARK

Existing non-English benchmarks are still fragmented and limited in scope. For example, Chinese and Korean medical benchmarks have shown their evaluation feasibility (Liu et al., 2023; Kweon et al., 2024; Yu et al., 2024; Tian et al., 2019), but they lack a standardized structure for difficult levels, question formats, and annotation guidelines. Furthermore, they also limit themselves to a small range of specialties only. Our benchmark directly address these drawbacks by covering across 34 medical topics with a predefined schema.

Due to the lack of native language resources, some researches have tried to translate English benchmarks into other languages (Singhal et al., 2022; Kasai et al., 2023; Qiu et al., 2024). However, it often introduces terminology inconsistency, ambiguity, or cultural misalignment, which can reduce the quality of training and evaluation process. Even high-quality machine translations can lead to clinical nuance and misinformed content in the biomedical domain (Gérardin et al., 2023). Furthermore, medical education is different between regions and may not align with local structure or practice standards (Jin et al., 2023).

In many regions, including Southeast Asia, there is no standardized effort to benchmark local-language medical knowledge; rather than relying on datasets from medical licensing exams or student materials, there is no framework to support cross-specialty generalization and scalability (Zhang et al., 2024). Our work addresses this gap by proposing a systematic framework, firmly based on local medical education resources, to ensure high quality and relevance.

### 2.2 MEDICAL DATA CURATION

For medical QA, medical benchmarks are majorly curated from English resources such as USMLE practice questions, PubMed research abstracts (MMLU Medical (Hendrycks et al., 2021a;b), MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019)), U.S. National Library of Medicine(TREC-2017 LiveQA (Ben Abacha et al., 2017)). Occasionally, there are datasets collected from other non-English resources such as medical board exam in China (cMedQA2 (Zhang et al., 2018)), AIIMS and NEET PG from India (MedMCQA (Pal et al., 2022)). Besides, there exists datasets that were synthesized from other non-mainstream resources such as Reddit, StackExchange (BiQA (Lamurias et al., 2020)). These datasets usually collected and curated from resources that have the same structure of data or a subspace of medical knowledge space, which is not often verified and representative enough to be used for building and evaluating medical LLMs. In addition, manually synthesizing from different resources is not scalable and efficient due to the heterogeneous structure of data from different resources. Our dataset is collected and processed through a scalable pipeline that gather online Vietnamese medical resources (medical textbook, medical exams, online medical quizzes & flashcards). Based on these verifiable sources, selected data was carefully filtered, annotated to build a complete benchmark. This pipeline is applicable to any languages, allowing researchers to conduct data synthesizing on their desire language for pre-training, fine-tuning and downstream tasks.

# 3 METHODS

## 3.1 BENCHMARK DESIGN

Among several existing methodologies for benchmark development, we designed our approach with the primary criteria of medical comprehensiveness, as the benchmark must be broad enough to cover the full spectrum of medical knowledge while being clear, replicable, and extendable to other languages. The selection of medical specialties was intentionally comprehensive to ensure both breadth across the medical field and sufficient depth within each specialty. Our benchmark includes 34 distinct medical categories, carefully selected to provide a holistic evaluation of medical knowledge. This extensive categorization (Table 5) serves multiple critical purposes:

**(1) Complete Coverage of Medical Practice:** The categories span the entire medical field, from primary care to highly specialized domains, ensuring that no significant area of medical practice is overlooked.

**(2) Balanced Representation:** We deliberately included both common specialties (like Internal Medicine and Pediatrics) and more specialized fields (such as Nuclear Medicine and Palliative Care) to prevent bias toward frequently encountered medical concepts.

**(3) Integrated Medical Knowledge:** The selected categories not only represent universally recognized medical specialties practiced across different healthcare systems worldwide, but also have categories related to traditional medical systems that remain relevant in many communities, especially Eastern. This design making our framework adaptable to various cultural and linguistic contexts.

**(4) Public Health and Preventive Focus:** Categories such as Public Health and Preventive Healthcare were selected to ensure that the benchmark evaluates the understanding of population-level interventions and disease prevention strategies, not just treatment-oriented knowledge. By including questions on disease surveillance, vaccination strategies, health education campaigns, and environmental risk mitigation, our benchmark captures the full spectrum of understanding needed to anticipate and avert outbreaks, reduce health disparities, and promote social well-being. The benchmark emphasizes the importance of upstream interventions, such as policy development, community engagement, and behavior change frameworks, that can prevent illness before it occurs, complementing treatment-focused competencies and fostering a truly holistic approach to healthcare.

To comprehensively assess the capabilities of medical LLMs, we introduce a four-tier difficulty framework, ranging from fundamental knowledge to complex clinical reasoning. This framework is designed to evaluate LLMs across a spectrum of cognitive skills necessary for medical professionals. This definition of levels (Table 6) offered several advantages:

**(1) Comprehensive Evaluation:** It allows for a comprehensive assessment of LLMs by covering a wide range of cognitive skills, from basic recall to complex reasoning. It aligns with the cognitive skills expected of medical professionals, mirroring the progression from foundational knowledge to advanced clinical practice. This ensures that the evaluation is not limited to a single aspect of medical knowledge.

**(2) Nuanced Assessment:** By differentiating between levels, the framework provides a nuanced understanding of an LLM's strengths and weaknesses. For instance, an LLM might excel at recalling factual information but struggle with complex clinical scenarios.

**(3) Alignment with Medical Expertise:** The framework aligns with the cognitive skills expected of medical professionals, mirroring the progression from foundational knowledge to advanced clinical practice. The "Challenging" and "Hard" levels assess skills crucial for real-world medical practice, such as diagnostic reasoning and clinical decision-making, guiding the development of more effective training strategies and architectures.

## 3.2 DATA COLLECTION

### 3.2.1 DATA PIPELINE

**Extract data** The crawler service is the core component our systems, this service is implemented in Python to extract medical data from reliable resources including medical textbooks, medical exams/tests from medical institutions, and public medical records from research institutions. The

extracted data are stored in Amazon S3 (AWS) in a folder in its raw format including text, document files, PDF files, images, along with metadata.

**Transformation** We leveraged Spark to efficiently transform large amounts of data in parallel. This transformation process standardized most of the text data to JSON format while image and PDF data were transferred to another flow. The transformed data are stored in S3 in the destination path, which we refer to as the bronze zone. Further advanced transformed data can be stored in the silver zone or the gold zone depending on the storage structure convention.

**Transfer data** We implemented an EL (Extract-Load) service from scratch, which helps transfer data from various source types to various destination types based on needs. This design allowed us to craft data effectively for different purposes besides benchmarking.

**Orchestral workflow** We used Apache Airflow for orchestration as our services would be triggered periodically to update and process new incoming data.

**Incremental extracting** We use PostgreSQL as a metadata database for data tracking to handle data duplication. We concatenate and hash strings created from fields to determine whether the data is new or not. The hash column serves as a unique identifier in the metadata. The crawler service uses metadata to detect which data have been extracted and to avoid duplicates.

**Data processing** The data was reformatted by removing extra index, HTML tags and then standardized to our JSON format while PDFs and documents were further extracted and deduplicated as needed.

#### 3.2.2 Structured data processing

We applied the method below for PDF and Microsoft Word files to extract and standardize questions from these sources.

**Extract content** We used docx2python [1] to extract text from Microsoft Words and pymupdf [2] to extract text from PDF. The raw text of questions (mostly multiple-choice questions and their options) is saved in the same output file.

**Extract questions** We utilized OpenAI's GPT-4o (OpenAI, 2024) and Gemini 2.0 Flash (Google, 2024) to extract questions to structured output (Table 1). Since our full raw dataset included 100,000 questions and produced large text files, we divided theses into chunks for parallel batch processing. We noticed that each question usually has 5 to 7 lines of text, so we set the chunk size to 100 lines (to ensure within the context length limit of GPT-4o) and overlap 5 lines to prevent losing questions. For the questions that do not have difficultLevel and medicalTopic, we used GPT-4o and Geimini Flash to label those based on our definition (Table 5, 6).

The extracted questions contain the following fields:

**Data verification** Figure 1 describes the pipeline that we built to remove redundant entries. This approach is effective for the Vietnamese language, where accents create many variances. Additionally, removal based on clustering and the Levenshtein distance is suitable when the questions and answers are slightly different at the character level, as is common in chemical-related topics. Afterwards, the answer verification step includes three sources of references: (1) the correct answers from the source data, (2) the answers from foundation LLMs, and (3) the answers from medical experts. We ranked questions for annotation based on mutual disagreement and level of difficulty. The first questions to be verified by human experts are the easiest ones, where the original source answers and the answers from foundation models differ. This approach is based on the assumption that all sources of truth should have mutual agreement on easy questions. Since the annotation process took a long time and the data distribution was heavily skewed toward common topics, we sampled 14,000

[1] https://github.com/ShayHill/docx2python

[2] https://github.com/pymupdf/PyMuPDF

Table 1: Extracted question field

| Field | Description |
|---|---|
| question | Medical multiple-choice question |
| answers | List of options for the answer |
| regularFormat | True if the question is multiple-choice. False otherwise |
| correctOption | The correct answer according to GPT-4o |
| difficultLevel | Level of difficulty of the question evaluated by GPT-4o |
| medicalTopic | List of relevant medical topics |

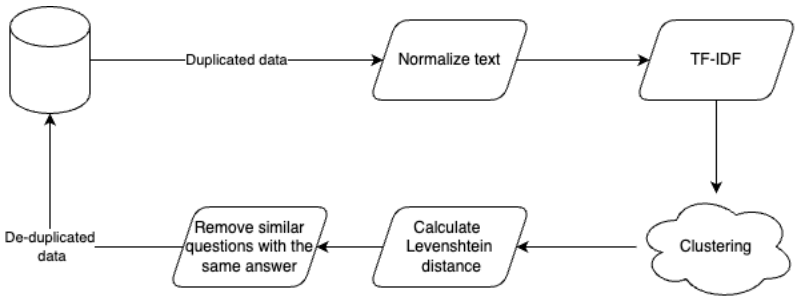


Figure 1: Duplication removal process.

questions from a total of 100,000, ensuring an even representation of the 34 medical categories included in our benchmark. This prevents common topics from dominating the dataset and allows for a more balanced evaluation of LLM performance across different areas of medical knowledge. This whole process offers early detection of potential problems, ensures representativeness, and provides a holistic verification of our benchmark.

## 4 EXPERIMENTS

In this section, we conduct a comprehensive evaluation of several LLMs including general models with multilingual capability and medical specialized models. We employ two evaluation metrics: pass@$k$ (with $k = 3$, $k = 1$ for non-reasoning models and $k = 1$ for reasoning models, except for Deepseeek-R1) and ensemble for all models. The ensemble setup involved running the model three times on the same set of questions, but choices were shuffled independently for each run, then voting is used for final aggregation. We analyze the results and provide insight about the potential and limitation of those LLMs in Vietnamese medical domain. The main results are shown in Table 2 and Table 3.

### 4.1 GENERAL MODEL WITH MULTILINGUAL CAPABILITY

**Results** As shown in Table 2, Deepseek R1 outperforms all other models significantly in both Pass@1 and Ensemble. Apart from Deepseek R1 outstanding performance, other reasoning models (o3-mini, Qwen 3) are not better than non-reasoning models such as Claude 3.5 Sonnet, GPT-4o, Gemini 2.0 Flash. This is in direct conflict with the performance on MedQA (Jin et al., 2020), an English medical benchmark, where o3-mini and Qwen 3 performed better than or at the same level as Deepseek R1 and much better than non-reasoning models (ValsAI, 2025). These observations suggest that for our benchmark, other factors, such as Vietnamese capability, are better predictors of performance than medical ability in English.

The ensemble method with shuffled choices shows an improvement across the board for all models. GPT-4o saw the highest improvement in ensemble performance over Pass@1. Interestingly, the lowest improvement from ensembling is observed for Deepseek R1, the best-performing model.

Table 2: Model performance on VM14K with pass@$k$ metric. Values are accuracies in percentages. Challenging and Hard questions are combined under "Hard".

| Models | | Overall | Easy | Medium | Hard | Overall | Easy | Medium | Hard |
|---|---|---|---|---|---|---|---|---|---|
| | | pass@3 | | | | pass@1 | | | |
| *Closed-source models* | | | | | | | | | |
| GPT-4o | | **80.40** | **82.78** | **79.57** | **77.95** | 72.74 | 75.11 | 72.06 | 69.19 |
| o3-mini | | – | – | – | – | 71.42 | 73.21 | 70.92 | 68.58 |
| Claude Sonnet | 3.5 | 75.62 | 77.50 | 74.81 | 74.43 | 71.46 | 73.12 | 70.77 | 70.58 |
| Gemini Flash | 2.0 | 77.92 | 80.16 | 77.34 | 74.77 | **75.67** | **78.03** | **75.18** | **71.75** |
| *Open-source models* | | | | | | | | | |
| Deepseek-R1 | | **83.02** | **84.68** | **82.32** | **82.31** | **78.17** | **79.89** | **77.54** | **77.03** |
| Qwen3-32B | | – | – | – | – | 72.47 | 74.37 | 72.10 | 68.79 |
| Qwen3-30B-A3B | | – | – | – | – | 71.32 | 73.69 | 70.28 | 65.08 |
| Llama 4 Maverick | | 73.13 | 74.68 | 72.66 | 71.17 | 71.76 | 72.95 | 71.54 | 69.66 |
| Phi-4 | | 69.18 | 71.57 | 68.77 | 60.99 | 51.15 | 53.82 | 50.44 | 45.52 |
| Gemma3-27B-Instruct | | 64.45 | 67.36 | 63.65 | 55.65 | 63.38 | 66.46 | 62.50 | 53.99 |
| Gemma3-12B-Instruct | | 60.03 | 62.86 | 59.13 | 53.16 | 58.05 | 60.60 | 57.39 | 50.84 |
| Llama-3.1-8B-Instruct | | 60.88 | 63.55 | 59.88 | 55.28 | 48.73 | 52.33 | 47.36 | 42.08 |
| *Open-source medical specialized models* | | | | | | | | | |
| Llama-3-70B-UltraMedical | | **67.34** | **70.53** | **65.64** | **65.07** | **54.96** | **58.34** | **53.55** | 51.07 |
| Meditron3-70B | | 59.55 | 62.86 | 58.17 | 45.42 | 42.27 | 45.57 | 40.86 | 37.17 |
| HuatuoGPT-o1-8B | | – | – | – | – | 54.52 | 57.93 | 53.14 | **52.19** |
| Llama-3.1-8B-UltraMedical | | 48.01 | 52.26 | 46.85 | 38.36 | 41.86 | 46.13 | 40.66 | 31.06 |
| Meditron3-8B | | 30.04 | 31.25 | 28.82 | 29.54 | 23.05 | 23.81 | 22.08 | 23.20 |

**Difficulty Levels** As shown in Figure 2, the performance of most models decreases as difficulty increases. However, the difference in accuracy between difficulty levels is not as large as in smaller open-source models. In particular, Claude 3.5 Sonnet and Llama 4 Maverick maintain relatively consistent accuracy across difficulty levels, although there is still a slight downward trend.

**Medical Topics** Table 4 shows that these foundation models have different levels of accuracy for different medical topics. Among the top 10 most popular medical topics, Oncology and Endocrinology are topics with higher accuracy while most models struggle with Pediatrics. These differences in performance are distributed quite evenly among foundation models, i.e. there are no clear signs any models have strengths in some certain topics and weaknesses in others.

**Cost-Performance analysis** Figure 3 shows the cost analysis of the foundation models, which can inform the use of these models as a service for Vietnamese medical use cases. Deepseek R1 is the best model at the cost of $ 3.59 inference cost to answers 1000 questions, in contrast with the other reasoning model, OpenAI's o3-mini, which is the second-most expensive model but also the

Table 3: Model performance on VM14K with ensemble metric. Values are accuracies in percentages. Challenging and Hard questions are combined under "Hard". *We included Pass@$k$ column for reference.

| Models | Pass@$k$* | Overall | Easy | Medium | Hard |
|---|---|---|---|---|---|
| *Closed-source models* | | | | | |
| GPT-4o | **80.40** | 73.80 (-6.60) | 76.36 | 72.79 | 71.87 |
| GPT-o3-mini | 71.42 | 72.02 (+0.60) | 74.11 | 71.35 | 69.33 |
| Claude 3.5 Sonnet | 75.62 | 72.31 (-3.31) | 73.81 | 71.71 | 71.67 |
| Gemini 2.0 Flash | 77.29 | **76.30 (-0.99)** | **78.89** | **75.58** | **72.84** |
| *Open-source models* | | | | | |
| Deepseek-R1 | **83.02** | 78.42 (-4.60) | 80.59 | 77.60 | 76.78 |
| Qwen3-32B | 72.47 | **84.96 (+12.49)** | **86.60** | **84.51** | **82.23** |
| Qwen3-30B-A3B | 71.32 | 84.04 (+12.72) | 85.48 | 83.63 | 81.42 |
| Llama 4 Maverick | 73.13 | 72.46 (-1.33) | 73.71 | 72.06 | 71.42 |
| Phi-4 | 69.18 | 72.25 (+3.07) | 74.59 | 71.89 | 67.25 |
| Gemma3-27B-Instruct | 64.45 | 77.11 (+12.66) | 79.16 | 76.33 | 72.64 |
| Gemma3-12B-Instruct | 60.03 | 74.09 (+14.06) | 76.12 | 73.57 | 67.68 |
| Llama-3-8B-Instruct | 60.88 | 69.92 (+9.04) | 72.57 | 68.84 | 68.06 |
| *Open-source medical specialized models* | | | | | |
| Llama-3-70B-UltraMedical | **67.34** | 72.57 (+5.23) | 75.10 | 71.30 | 72.33 |
| Meditron3-70B | 59.55 | 56.10 (-3.45) | 59.31 | 54.98 | 45.42 |
| HuatuoGPT-o1-8B | 54.52 | **73.89 (+19.37)** | **76.41** | **72.99** | **73.03** |
| Llama-3.1-8B-UltraMedical | 48.01 | 56.24 (+8.23) | 60.19 | 55.45 | 44.27 |
| Meditron3-8B | 30.04 | 30.09 (+0.05) | 31.69 | 28.80 | 28.57 |

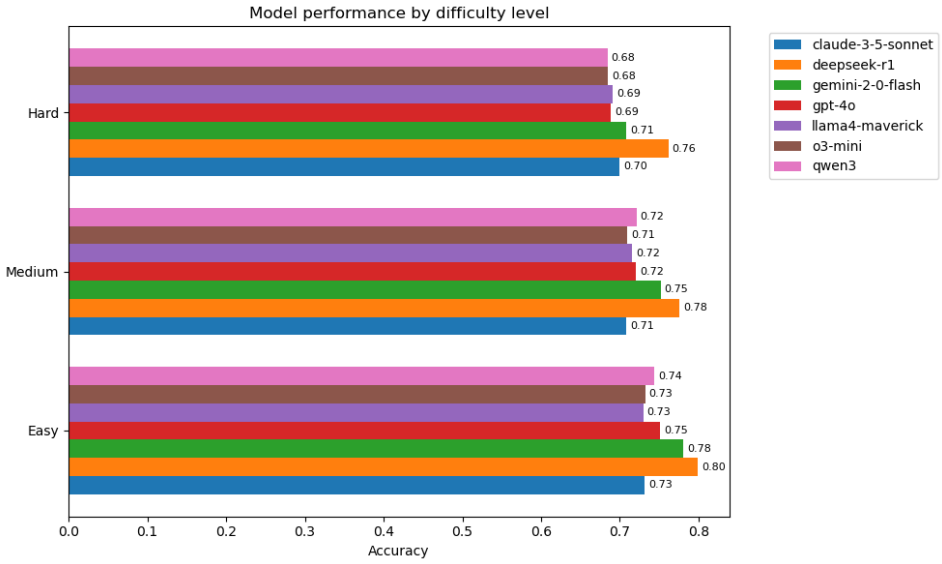


Figure 2: Performance of general models on the benchmark by difficulty levels. Challenging and Hard questions are combined under "Hard".

most inaccurate. Gemini 2.0 Flash is the best value-for-money model, only trailing Deepseek R1 in performance and is the cheapest among the 7 models.

Table 4: Pass@1 by Top 10 popular medical topics. Values are accuracies in percentages.

| Medical Topic | GPT 4o | o3 mini | Llama 4 Maverick | Claude 3.5 Sonnet | Qwen3 32B | Gemini 2.0 Flash | Deepseek R1 |
|---|---|---|---|---|---|---|---|
| Gastroenterology | 72.07 | 73.19 | 73.31 | 72.43 | 74.02 | 77.01 | **81.47** |
| Obstetrics and Gynecology | 74.61 | 69.80 | 72.29 | 74.12 | 71.55 | 77.18 | **79.71** |
| Pulmonology | 70.87 | 70.82 | 68.12 | 68.66 | 70.23 | 74.00 | **77.35** |
| Infectious Diseases | 73.23 | 72.24 | 71.74 | 71.63 | 71.74 | 76.20 | **77.79** |
| Endocrinology | 76.92 | 76.79 | 76.34 | 75.63 | 78.01 | 78.27 | **81.24** |
| Oncology | 75.86 | 75.74 | 76.58 | 75.15 | 77.41 | 79.67 | **82.64** |
| Pathology | 71.23 | 67.61 | 69.16 | 70.97 | 73.03 | 74.19 | **78.19** |
| Pediatrics | 67.74 | 67.48 | 64.52 | 68.65 | 67.23 | 68.39 | **73.16** |
| Radiology | 70.81 | 72.05 | 69.57 | 69.25 | 70.19 | 74.84 | **77.48** |
| Surgery | 68.77 | 66.83 | 68.77 | 68.61 | 68.77 | 71.36 | **75.89** |

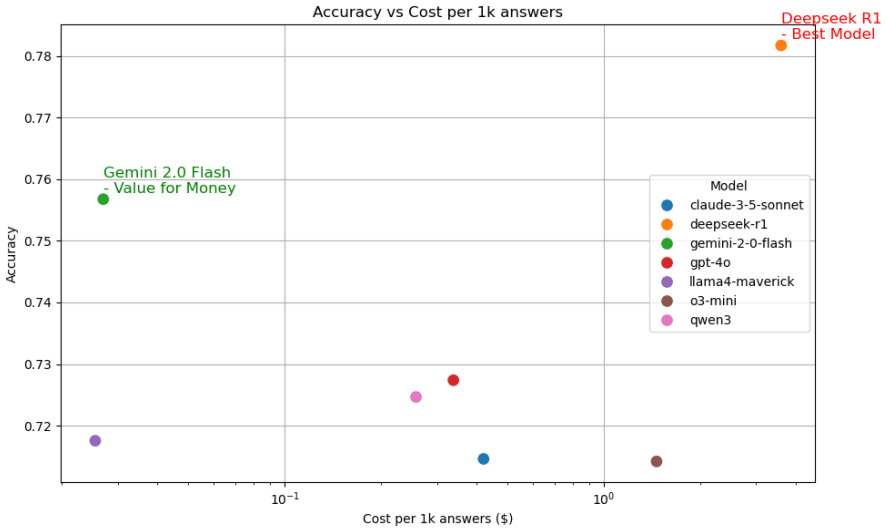


Figure 3: Cost versus performance analysis.

### 4.2 Medical Specialized Models

**Results** In general, larger models (70B) have more capacity to learn complex medical knowledge and reasoning patterns, leading to higher accuracy. This trend is clear across all difficulty levels. They can also maintain their performance better as question difficulty increases, which means scale helps with complex, multi-step reasoning required for harder questions. The drop in accuracy from Easy to Hard is less severe for larger models, except for HuatuoGPT-o1-8B: despite being 8B, it performs at the top, likely due to specialized training or data focused on medical reasoning. However, the rest of the 8B models are consistently the lowest performers; and the best results are achieved when both scale and specialization are present (for example Llama-3-70B-UltraMedical).

**Ensemble** The ensemble results are higher than pass@1 and often close to or better than pass@3, showing that combining multiple runs or shuffles helps models "cover" more correct answers, especially for difficult questions. We also observe a large shift in performance between Llama-3-70B-UltraMedical and HuatonGPT-o1-8B as the evaluation method changed from Pass@$k$ to Ensemble where the latter is the better. We argue that the Ensemble metric is more holistic to evaluate of

the model as some randomness can easily affect Pass@$k$. That said, a small well-trained reasoning model model could reach similar performance of a larger one.

Also, it is important to note that the number of open-source medical language models is relatively low, and most of the publicly available ones are small-scale. The majority of large-scale medical language models are currently proprietary. This is understandable, since training medical LLMs is costly and requires careful alignment and adherence to regulations.

## 5 CONCLUSION

In this work, we address the gap in linguistic diversity for medical benchmark by introducing an complete approach to collect, process reliable data sources and design benchmark. We applied this method to build the first Vietnamese medical benchmark (14,000 questions, 34 specialties, four difficulty levels). This benchmark, featuring expert-annotated questions, is designed for a specialized, non-English setting and sets a precedent for culturally relevant evaluation frameworks. Our open-source, scalable data pipeline, integrating a complete process from data collecting to correctness verification, facilitates the creation of this benchmark from diverse local medical resources. The benchmark and pipeline aim to foster collaboration, accelerate advancements in evaluating medical LLMs for underserved linguistic communities, and support the development of future multilingual benchmarks.

This effort holds the potential to significantly impact the field of medical AI by enabling more accurate and reliable evaluation of LLMs in languages other than English. The availability of such benchmarks will empower researchers and developers to create and refine LLMs that are better suited to the specific needs and contexts of diverse populations. Ultimately, this can lead to improved healthcare outcomes, enhanced medical education, and more equitable access to the benefits of AI in medicine across the globe. Our data pipeline and benchmark were publicly shared for the replication and advancement of the communities.

## 6 LIMITATIONS

Our study has several limitations. First, while our benchmark covers a wide range of medical specialties and difficulty levels, the published question format is limited to multiple-choice questions, which may not fully capture the complexities of clinical reasoning in real-world scenarios. Future work could explore incorporating other question formats, such as open-ended questions or clinical case simulations. Second, the benchmark is designed for Vietnamese, and while our data pipeline is scalable to other languages, the process of adapting and validating the benchmark for new languages may present unique challenges. These challenges include the availability of high-quality medical resources, the need for expert annotation, and the potential for cultural and linguistic variations in medical knowledge. Third, while having 100,000 entries, we were just able to verify 14,000 multiple-choice questions within 3 months. The rest of the data is valuable as it also includes clinical questions and medical records and should be exploited not just for evaluating but also for synthesizing training data. Finally, the evaluation of LLMs using our benchmark focuses primarily on their ability to answer medical questions. While this is a crucial aspect of medical LLM performance, it does not encompass other important factors such as the model's ability to generate accurate and comprehensive medical reports, engage in effective doctor-patient communication, or adhere to ethical and legal guidelines. Future research should consider a more holistic evaluation framework that assesses these additional dimensions of LLM performance in the medical domain.

## 7 ACKNOWLEDGE

This work was generously supported by computing credits from Azure, AWS, and Google Cloud Platform. We also gratefully acknowledge the technical support provided by Venera AI. Additionally, we would like to express our deep gratitude to the individuals who offered invaluable support and guidance throughout the development of this project: Quynh-Anh Pham, Viet-Cuong Nguyen, and Manh-Cuong Nguyen from Thai Nguyen University of Medicine and Pharmacy; Huy-Hoang Ha, Phuc Phan, Hop Bui and Diep Duong from VietAI.

## REFERENCES

Mohammed Al-Garadi, Tushar Mungle, Abdulaziz Ahmed, Abeed Sarker, Zhuqi Miao, and Michael E. Matheny. Large language models in healthcare, 2025. URL https://arxiv.org/abs/2503.04748.

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. Overview of the medical question answering task at trec 2017 liveqa. In *TREC 2017*, 2017.

Google. Introducing gemini 2.0: our new ai model for the agentic era, 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Christel Gérardin, Yuhan Xiong, Perceval Wajsbürt, Fabrice Carrat, and Xavier Tannier. Impact of translation on biomedical information extraction from real-life clinical notes, 2023. URL https://arxiv.org/abs/2306.02042.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries, 2023. URL https://arxiv.org/abs/2310.13132.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations, 2023. URL https://arxiv.org/abs/2303.18027.

Sunjun Kweon, Byungjin Choi, Gyouk Chu, Junyeong Song, Daeun Hyeon, Sujin Gan, Jueon Kim, Minkyu Kim, Rae Woong Park, and Edward Choi. Kormedmcqa: Multi-choice question answering benchmark for korean healthcare professional licensing examinations, 2024. URL https://arxiv.org/abs/2403.01469.

Andre Lamurias, Diana Sousa, and Francisco M. Couto. Generating biomedical question answering corpora from q&a forums. *IEEE Access*, 8:161042–161051, 2020. doi: 10.1109/ACCESS.2020.3020868.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. Benchmarking large language models on cmexam – a comprehensive chinese medical exam dataset, 2023. URL https://arxiv.org/abs/2306.03030.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. A survey on medical large language models: Technology, application, trustworthiness, and future directions, 2024. URL https://arxiv.org/abs/2406.03712.

OpenAI. Hello GPT-4o, 2024. URL https://openai.com/index/hello-gpt-4o/.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Yu Jin Park, Aishwarya Pillai, Jing Deng, and et al. Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Medical Informatics and Decision Making*, 24(72), 2024. doi: 10.1186/s12911-024-02459-6. URL https://doi.org/10.1186/s12911-024-02459-6.

Hai Qin and Yi Tong. Opportunities and challenges for large language models in primary health care. *J Prim Care Community Health*, 16:21501319241312571, Jan-Dec 2025. doi: 10.1177/21501319241312571.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards building multilingual language model for medicine, 2024.

Nigam Shah, Mike Pfeffer, and Percy Liang. Holistic evaluation of large language models for medical applications. https://hai.stanford.edu/news/holistic-evaluation-of-large-language-models-for-medical-applications, Oct 2024. Accessed: 2025-05-15.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022. URL https://arxiv.org/abs/2212.13138.

Yuanhe Tian, Weicheng Ma, Fei Xia, and Yan Song. ChiMed: A Chinese medical corpus for question answering. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii (eds.), *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 250–260, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5027. URL https://aclanthology.org/W19-5027/.

ValsAI. MedQA benchmark, 2025. URL https://www.vals.ai/benchmarks/medqa-05-09-2025. Accessed: 2025-05-09.

Xiaoye Wang, Nicole Xi Zhang, Hongyu He, Trang Nguyen, Kun-Hsing Yu, Hao Deng, Cynthia Brandt, Danielle S. Bitterman, Ling Pan, Ching-Yu Cheng, James Zou, and Dianbo Liu. Safety challenges of ai in medicine in the era of large language models, 2025. URL https://arxiv.org/abs/2409.18968.

Ping Yu, Kaitao Song, Fengchen He, Ming Chen, and Jianfeng Lu. Tcmd: A traditional chinese medicine qa dataset for evaluating large language models, 2024. URL https://arxiv.org/abs/2406.04941.

Deshiwei Zhang, Xiaojuan Xue, Peng Gao, Zhijuan Jin, Menghan Hu, Yue Wu, and Xiayang Ying. A survey of datasets in medicine for large language models. *Intelligence & Robotics*, 4(4), 2024. ISSN 2770-3541. doi: 10.20517/ir.2024.27. URL https://www.oaepublish.com/articles/ir.2024.27.

S. Zhang, X. Zhang, H. Wang, L. Guo, and S. Liu. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2883637.

Z Zhang and H Ni. Critical care studies using large language models based on electronic healthcare records: A technical note. *J Intensive Med*, 5(2):137–150, 2024. doi: 10.1016/j.jointm.2024.09.002.

# A DETAIL OF DATASET

Table 5: Medical categories and descriptions

| Category | Description |
|---|---|
| Allergy and Immunology | Diagnosis and treatment of immune system disorders, including allergies, asthma, and autoimmune diseases. |
| Anesthesiology | Anesthesia and perioperative care to manage pain and vital functions during and after surgery. |
| Cardiology | Diseases related to the heart and blood vessels. |
| Dermatology | Conditions of the skin, hair, and nails, including cosmetic issues. |
| Endocrinology | Hormonal disorders, including diabetes and thyroid diseases. |
| Gastroenterology | Disorders of the digestive system: stomach, intestines, liver. |
| Geriatrics | Health care for the elderly, disease prevention and treatment in aging adults. |
| Hematology | Blood disorders, including anemia, clotting disorders, and blood cancers. |
| Infectious Diseases | Diseases caused by bacteria, viruses, fungi, or parasites. |
| Internal Medicine | General medical care for adults across a broad range of conditions. |
| Nephrology | Kidney-related disorders, including dialysis and electrolyte issues. |
| Neurology | Disorders of the nervous system: brain, spinal cord, nerves. |
| Nuclear Medicine | Use of radioactive substances for diagnosis and treatment, especially in cancer. |
| Obstetrics and Gynecology | Women's reproductive health, pregnancy, and childbirth. |
| Oncology | Diagnosis and treatment of cancer and tumors. |
| Ophthalmology | Medical and surgical care of eye diseases. |
| Orthopedics | Musculoskeletal conditions: bones, joints, muscles. |
| Otolaryngology | Disorders of the ear, nose, throat, head, and neck. |
| Palliative Medicine | Relief from serious illness symptoms to improve quality of life. |
| Pathology | Study of disease through tissue, cell, and fluid analysis. |
| Pediatrics | Medical care for infants, children, and adolescents. |
| Physical Medicine and Rehabilitation | Restoring function for people with physical impairments or disabilities. |
| Psychiatry | Mental health, including emotional and behavioral disorders. |
| Pulmonology | Disorders of the lungs and respiratory system. |
| Radiology | Diagnostic imaging like X-rays, MRI, CT scans. |
| Rheumatology | Autoimmune and inflammatory joint/muscle diseases. |
| Sports Medicine | Injuries and prevention related to sports and exercise. |
| Surgery | Operative procedures to treat diseases and injuries. |
| Urology | Disorders of the urinary tract and male reproductive system. |
| General Medicine | Broad general care and routine health maintenance. |
| Eastern Medicine | Traditional practices: acupuncture, herbal therapy, holistic care. |
| Public Health | Population health, prevention, and policy-based approaches. |
| Preventive Healthcare | Preventing illness via vaccines, screenings, and lifestyle. |
| Emergency Medicine | Acute care for urgent and life-threatening conditions. |

Table 6: Difficulty level definitions

| Level | Type | Description |
|---|---|---|
| Easy | Factual / Definitional | Questions about medical terms, basic definitions, or universally known concepts. *Keywords:* Definitions, terminology, recall. |
| Moderate | Mechanistic / Pharmacologic | Focus on disease mechanisms, drug properties, or pathophysiology. *Keywords:* Mechanisms, interactions, clinical application. |
| Challenging | Patient-Centered / Scenario-Based | Case studies requiring diagnostic reasoning or moderate clinical insight. *Keywords:* Diagnostic steps, patient context, moderate complexity. |
| Hard | Advanced Clinical Reasoning | Complex multi-symptom cases requiring integration and critical thinking. *Keywords:* Integrative reasoning, synthesis, high complexity. |

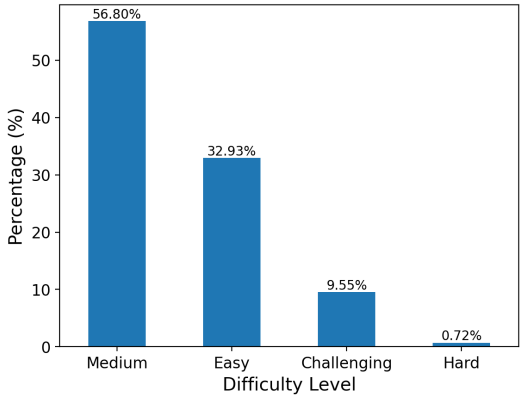


Figure 4: Difficulty level distribution of VM14K.

## B PROMPT DETAILS

We used the same prompt for all models:

```
{question}
Choose the correct option from these answers:
A. {optionA}
B. {optionB}
...

Only respond with 1 character
Example:
A
```

Figure 5: Prompt used in the experiments.