HCMC NATIONAL UNIVERSITY

**UNIVERSITY OF SCIENCE**

Information Technology Department

# MedicalQA – EDA Survey Report

Datasets: ViMQ, MedXpertQA, ViMedAQA

***Students:***

Trần Phúc Hải (22127096)

Nguyễn Văn Minh Thiện (22127398)

***Advisors:***

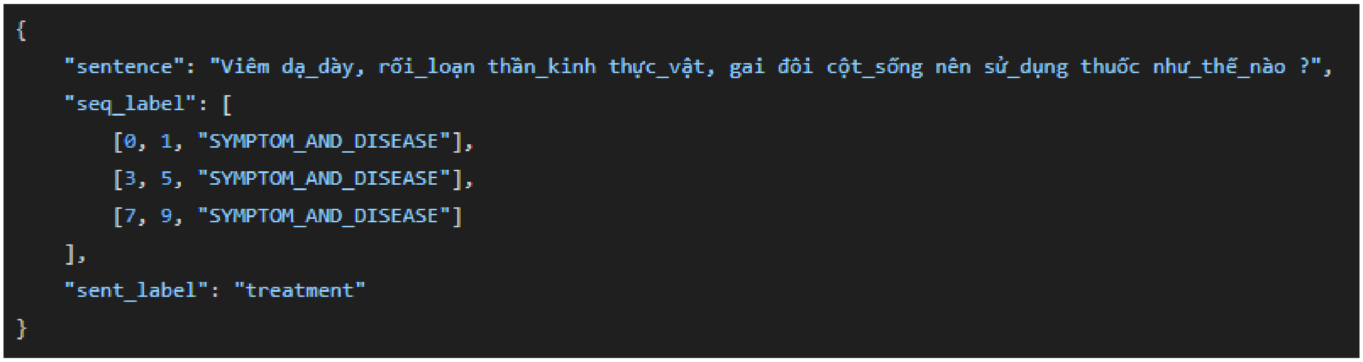
Dr. Lê Thanh Tùng

Dr. Nguyễn Tiến Huy

# Table of Contents

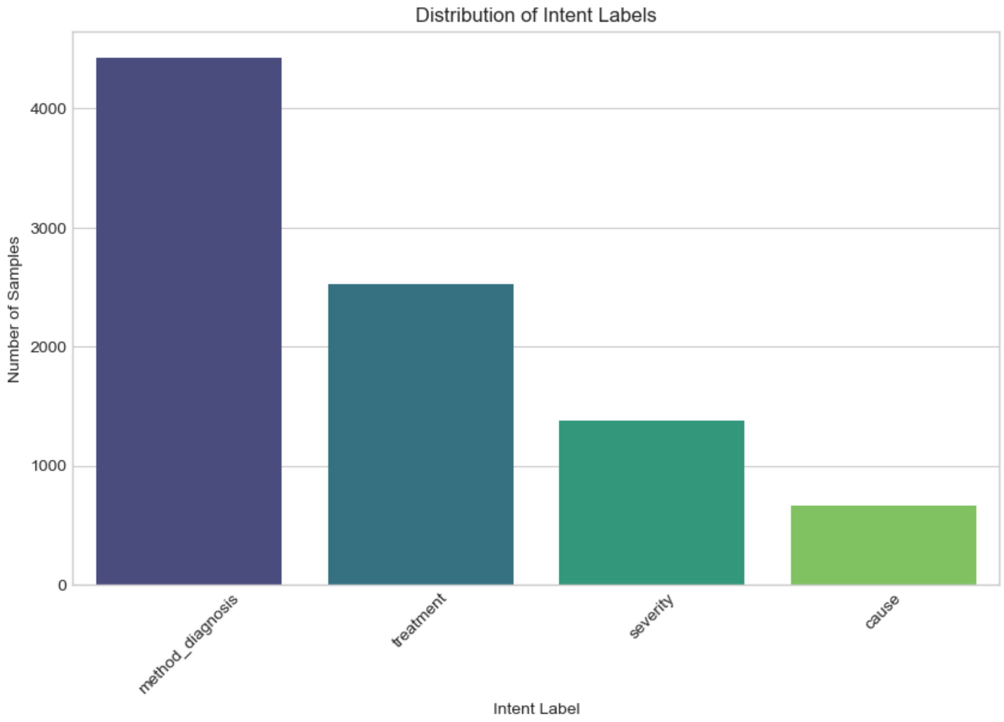
# ViMQ

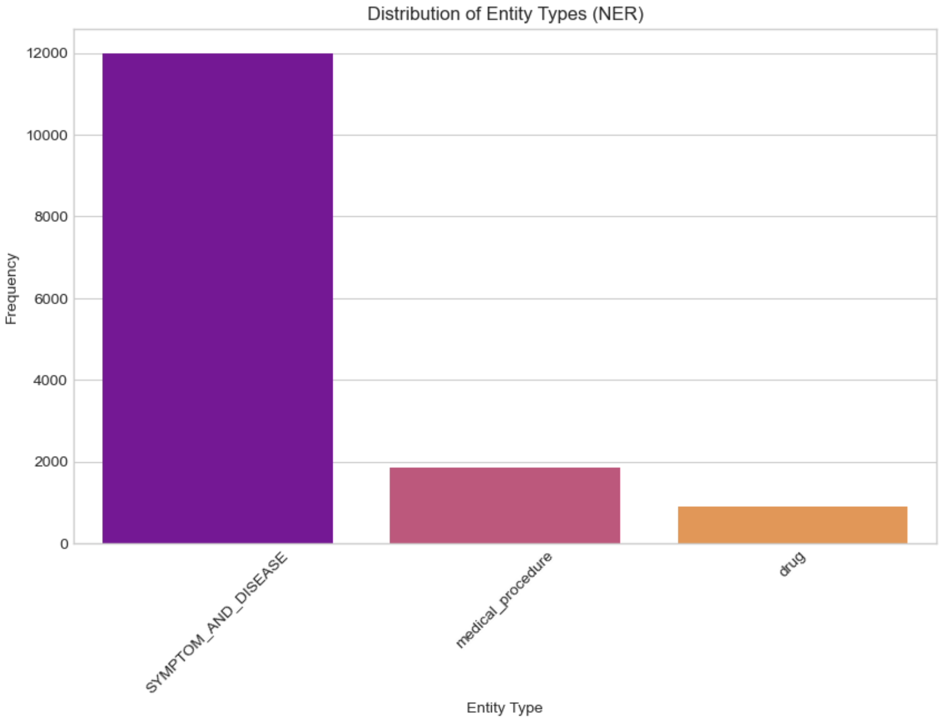
- Small-to-medium-sized collection of 9,000 Vietnamese medical questions designed for Intent Classification and Named Entity Recognition (NER) tasks.
- Sample data point:



```
{
    "sentence": "Viêm dạ_dày, rối_loạn thần_kinh thực_vật, gai đôi cột_sống nên sử_dụng thuốc như_thế_nào ?",
    "seq_label": [
        [0, 1, "SYMPTOM_AND_DISEASE"],
        [3, 5, "SYMPTOM_AND_DISEASE"],
        [7, 9, "SYMPTOM_AND_DISEASE"]
    ],
    "sent_label": "treatment"
}
```

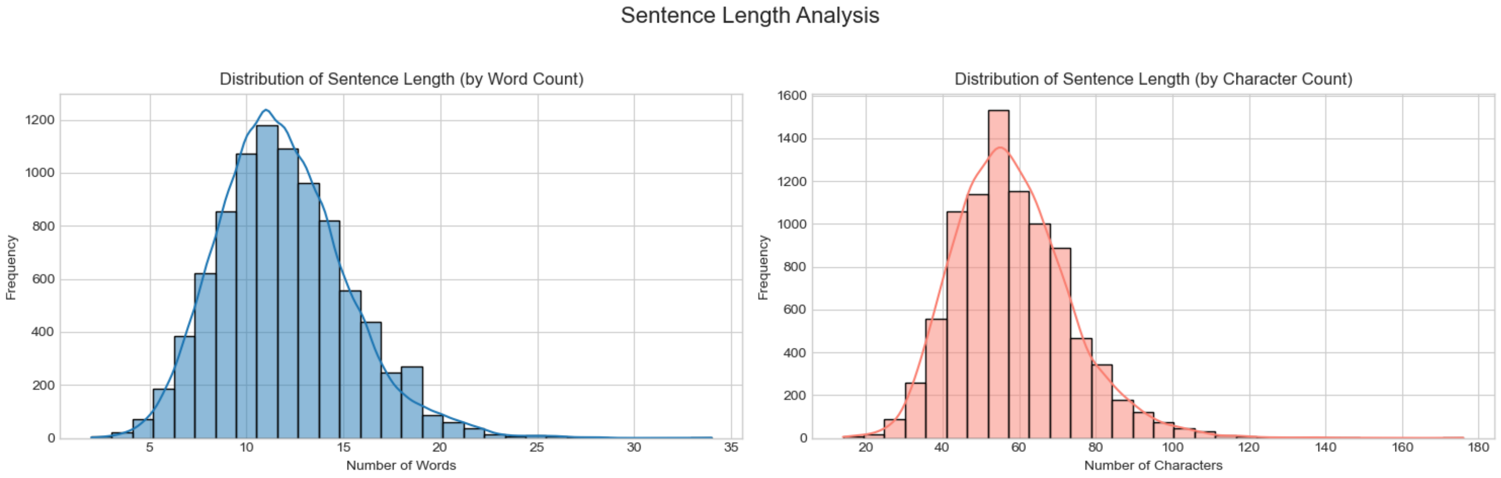
  - ○ "**sentence**": Contains the raw text of the user's medical question in Vietnamese.
    i. Need to use correct tokenizer to correctly handle the pre-combined multi-word tokens like "dạ_dày", "rối_loạn", etc.
  - ○ "**seq_label**": This is a list that pinpoints and labels specific medical terms (entities) within the sentence, used for NER. Each element has the format [start_token_index, end_token_index, entity_type].
  - ○ "**sent_label**": This field identifies the overall goal or intent of the entire sentence.
- Significant class imbalances in both intent labels and entity types, with a heavy skew towards `method_diagnosis` intent and `SYMPTOM_AND_DISEASE` entities:



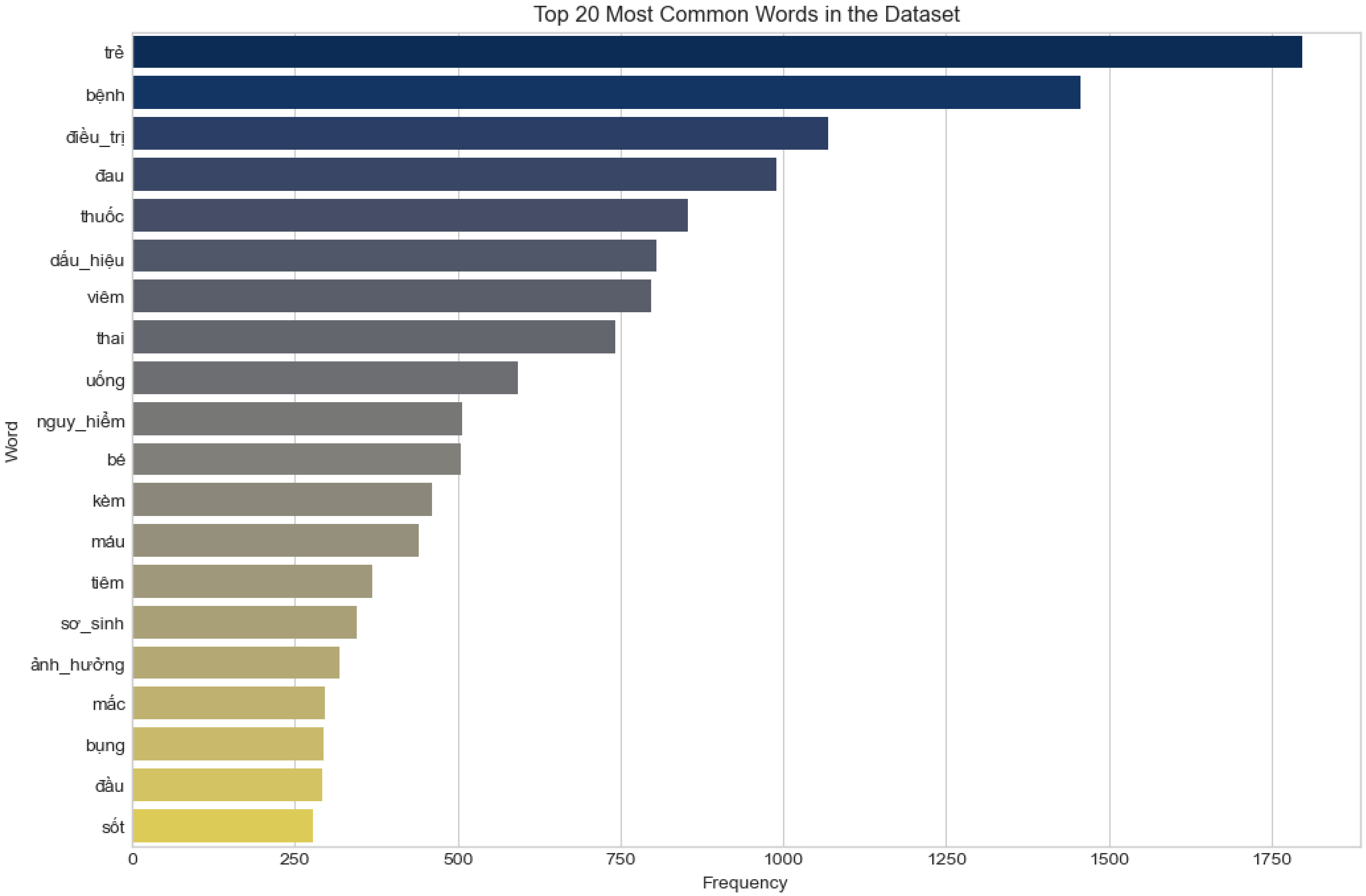


- Observations:
  1. A single class making up half the data is a classic sign of a heavy skew.
  2. We can observe a long tail, where one or two classes are very frequent and the rest become progressively rarer.
  3. A model trained on this data would be heavily biased towards predicting the majority class simply because it has seen so many more examples of it.
- Possible explanations:
  1. The primary reason for this skewness is that the data mirrors real-world patterns of how people communicate about their health. When individuals seek medical information or consult a doctor, their primary focus is usually on understanding their condition and its symptoms, for disease diagnosis.
  2. Questions about treatment, severity, and cause often follow after a diagnosis has been suggested or confirmed. Therefore, in a general collection of medical queries, the volume of diagnostic questions, along with symptoms and diseases, will naturally be higher.
  3. A single disease or a set of symptoms can be discussed at length. In contrast, the number of specific medical procedures or drugs mentioned in relation to a particular condition is often smaller. For example, a patient might describe five different symptoms of the flu, but the recommended treatment might only be one or two drugs.
  4. → **This is a problem often seen in medical datasets, not a specific problem with ViMQ**.

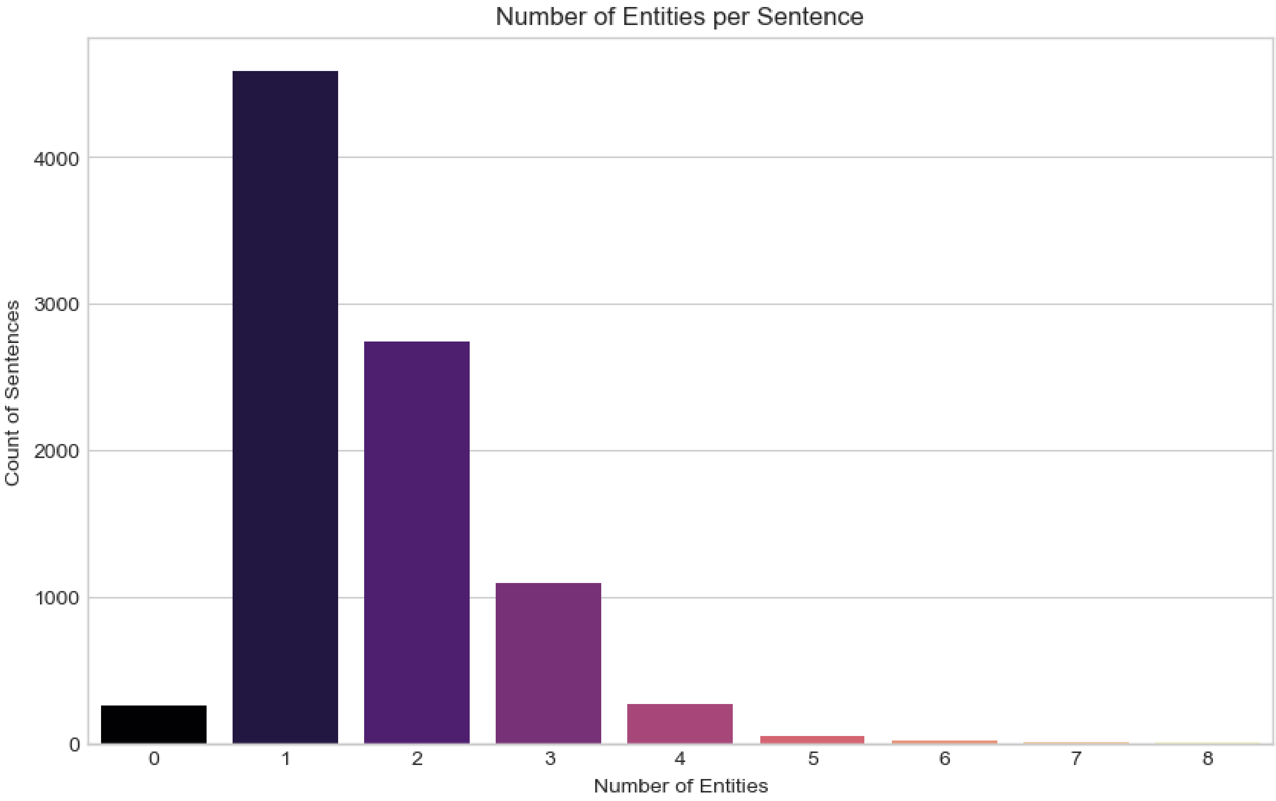
- The data is characterized by short sentences, averaging 12 words:



- Highly domain-specific, limited vocabulary of 3,272 unique words, not counting stop words → The model may not generalize well to unseen medical terminology. Medical terminology is one of the largest specialized terminologies and is estimated to contain over 250,000 items.[1]
    - Below are the top 20 most common words in the dataset:



  - The most frequent words match the intent label and entity type imbalances:
    1. For intents:
       1. điều_trị (treatment): The 3rd most common word.
       2. nguy_hiểm (dangerous): The 10th most common word, directly mapping to the severity intent.
       3. Keywords like dấu_hiệu (symptom) and bệnh (disease) are fundamental to the method_diagnosis intent.
    2. For entities:
       1. bệnh (disease), dấu_hiệu (symptom), đau (pain), viêm (inflammation): These are all prime examples of the SYMPTOM_AND_DISEASE entity.
       2. thuốc (drug): This directly maps to the drug entity.
  - The word đau (pain) is the 4th most common word. Its high frequency, combined with body part words on the list like bụng (stomach) and đầu (head), suggests that many queries are about describing pain in specific locations. Understanding questions about pain is a key capability the model will need to develop.
  - Words like điều_trị (treatment), thuốc (drug), and uống (to take/drink medicine) highlights that users are not just describing problems; they are actively seeking solutions and instructions. This reinforces that the dataset is well-suited for building a practical, action-oriented Q&A system, rather than just a descriptive text-analysis tool.
  - **A focus on pediatrics (nhi khoa) and obstetrics (sản khoa)**: trẻ (child), thai (pregnancy), bé (baby), and sơ sinh (newborn).
    1. The words viêm (inflammation/infection), sốt (fever), and đi ngoài (diarrhea) are very common reasons parents seek medical advice for their children. Respiratory infections and diarrheal diseases are major causes of childhood morbidity worldwide.
- The entity density is low, with an average of 1.6 entities per sentence, and 75% of sentences have 2 or fewer entities.

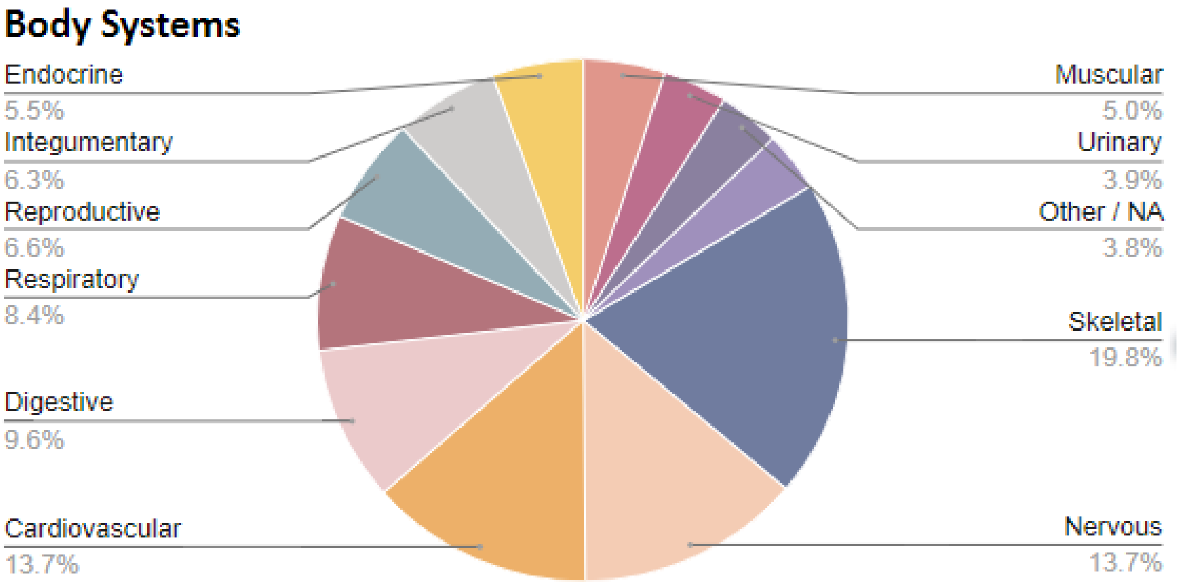


- **Advantages** of few entities per sentence: It is easier to understand the relationships between entities and words in a sentence, when there are fewer entities.
- **Disadvantages**:
  1. Inability to learn complex relationships.
  2. Poor generalization to noisy real-world data*:* The dataset is clean and simple. Human language is not. People often combine multiple ideas, provide background context, and ask multipart questions in a single sentence.
     a. Training data example: "How to treat a skin rash?" (2 entities)
     b. Real-world query example: "My son has a skin rash after we switched to a new laundry detergent, and he also has a history of eczema; what's the best treatment?"
  3. Difficulty with ambiguity resolution*:* Context is key to resolving ambiguity. Often, the presence of other entities in a sentence helps clarify the meaning of a specific word. With few entities, there is less context to learn from.
  4. → **Major challenge if the goal is to build more sophisticated applications beyond simpler Q&A**.

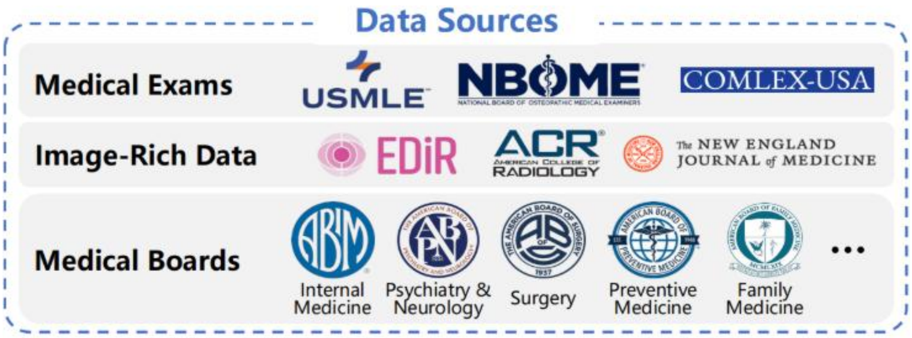
# MedXpertQA

## Overview

- The benchmark includes 4,460 multiple-choice questions that cover 17 medical specialties and 11 body systems in English.

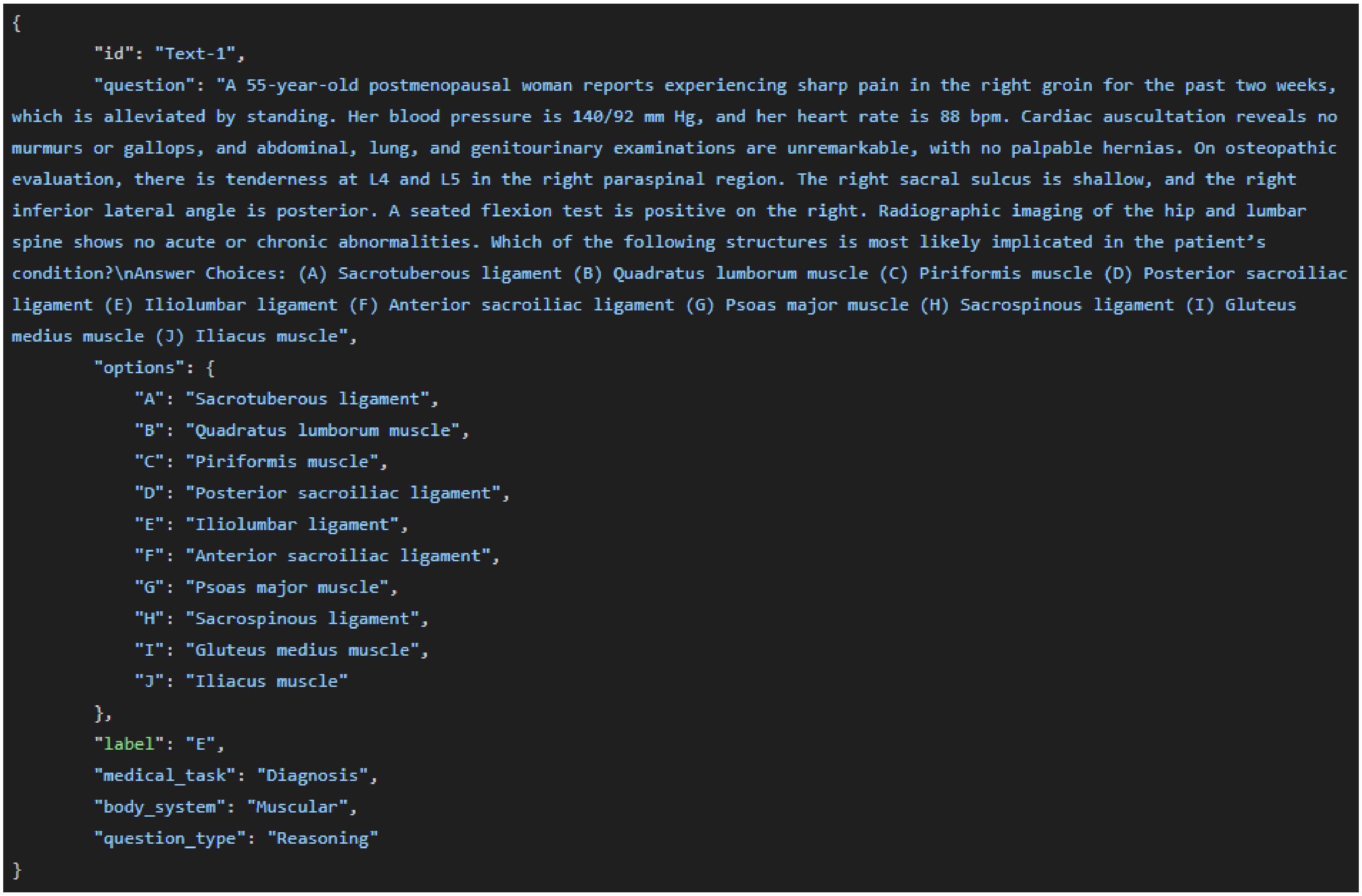


- Existing benchmarks lack clinical relevance. To overcome these limitations, MedXpertQA incorporates questions from specialty board exams *(professional medical examinations designed to assess expert-level knowledge and advanced reasoning in specific medical fields)* and uses a rigorous process of filtering and augmentation. The creators also implemented data synthesis and multiple rounds of expert reviews to minimize data leakage and ensure accuracy.



- MedXpertQA is divided into two subsets:
  - Text (2455 samples): for text-based evaluation.
  - MM (2005 samples): for multimodal evaluation, which includes questions with diverse images and rich clinical information like patient records and examination results.

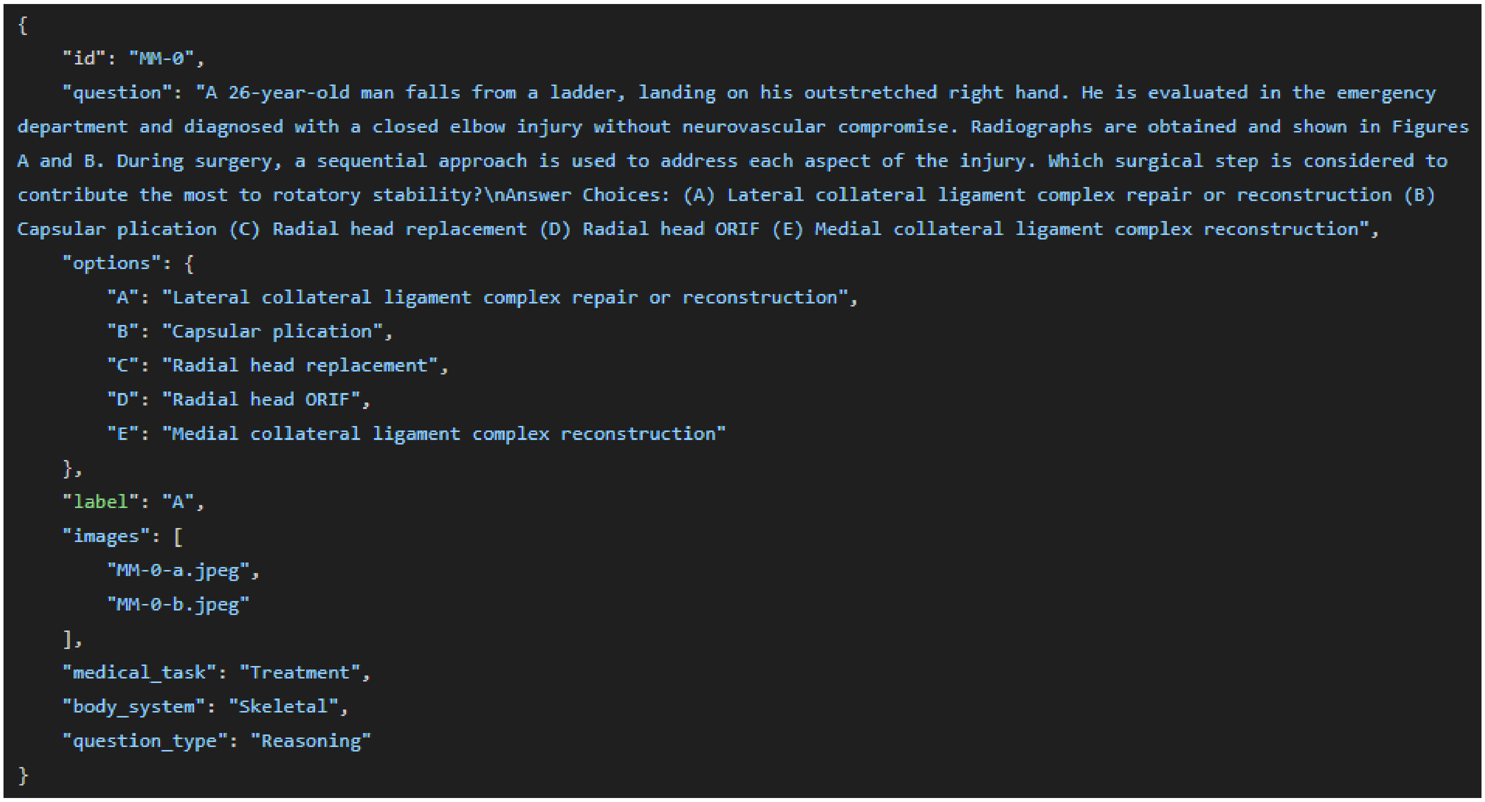
- Text data example:



```
{
        "id": "Text-1",
        "question": "A 55-year-old postmenopausal woman reports experiencing sharp pain in the right groin for the past two weeks, which is alleviated by standing. Her blood pressure is 140/92 mm Hg, and her heart rate is 88 bpm. Cardiac auscultation reveals no murmurs or gallops, and abdominal, lung, and genitourinary examinations are unremarkable, with no palpable hernias. On osteopathic evaluation, there is tenderness at L4 and L5 in the right paraspinal region. The right sacral sulcus is shallow, and the right inferior lateral angle is posterior. A seated flexion test is positive on the right. Radiographic imaging of the hip and lumbar spine shows no acute or chronic abnormalities. Which of the following structures is most likely implicated in the patient’s condition?\nAnswer Choices: (A) Sacrotuberous ligament (B) Quadratus lumborum muscle (C) Piriformis muscle (D) Posterior sacroiliac ligament (E) Iliolumbar ligament (F) Anterior sacroiliac ligament (G) Psoas major muscle (H) Sacrospinous ligament (I) Gluteus medius muscle (J) Iliacus muscle",
        "options": {
            "A": "Sacrotuberous ligament",
            "B": "Quadratus lumborum muscle",
            "C": "Piriformis muscle",
            "D": "Posterior sacroiliac ligament",
            "E": "Iliolumbar ligament",
            "F": "Anterior sacroiliac ligament",
            "G": "Psoas major muscle",
            "H": "Sacrospinous ligament",
            "I": "Gluteus medius muscle",
            "J": "Iliacus muscle"
        },
        "label": "E",
        "medical_task": "Diagnosis",
        "body_system": "Muscular",
        "question_type": "Reasoning"
}
```

  - “**id**”: A unique identifier for each question-answer pair.
  - “**question**”: The full text of the medical problem, presented as a clinical case followed by a multiple-choice question.
  - “**options**”: Contains the set of possible answers for the multiple-choice question.
  - “**label**”: The key that identifies the correct answer from the options dictionary.
  - “**medical_task**”: Categorizes the primary medical task the question is about.
  - “**body_system**”: Categorizes the body system of the question out of the 11 types in the “Body Systems” image above.
  - “**question_type**”: Categorizes the question as “Reasoning” or “Understanding” (requires less reasoning, more about medical facts).

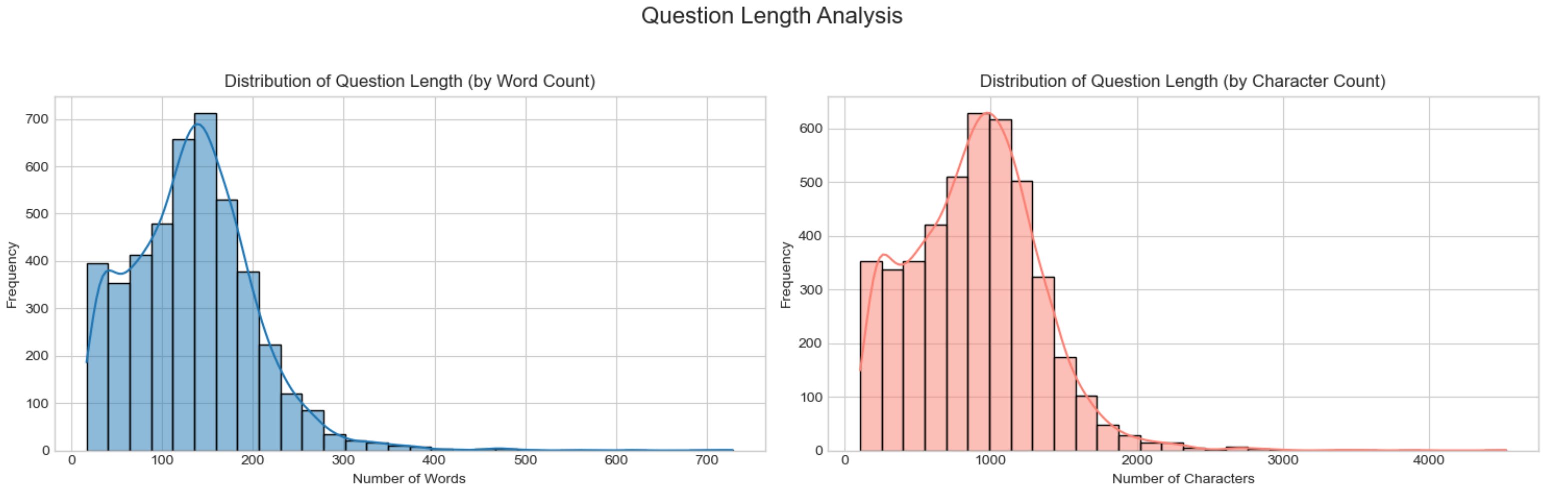
- MM data example: Like text data but with added “images” field.



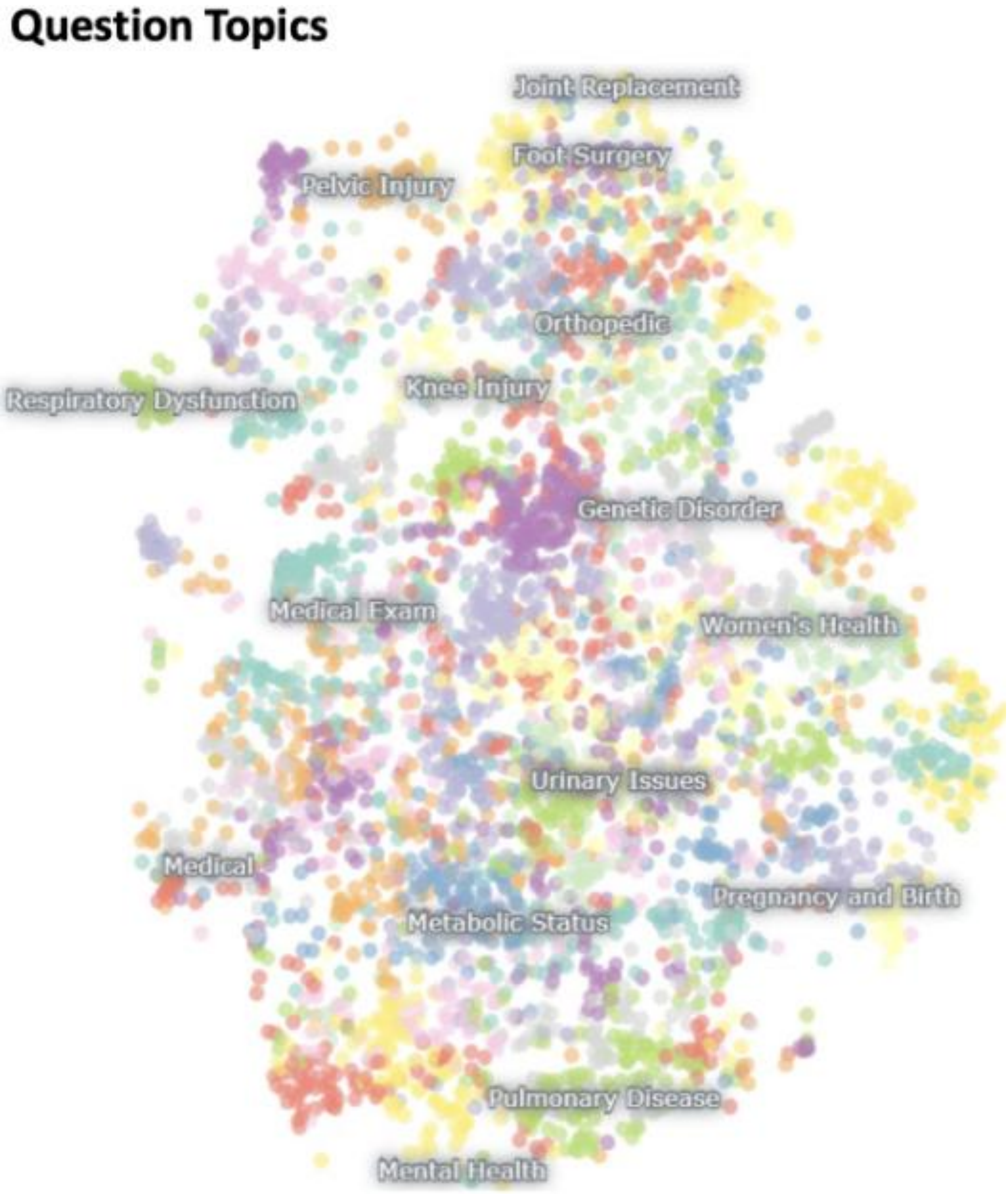
```
{
    "id": "MM-0",
    "question": "A 26-year-old man falls from a ladder, landing on his outstretched right hand. He is evaluated in the emergency department and diagnosed with a closed elbow injury without neurovascular compromise. Radiographs are obtained and shown in Figures A and B. During surgery, a sequential approach is used to address each aspect of the injury. Which surgical step is considered to contribute the most to rotatory stability?\nAnswer Choices: (A) Lateral collateral ligament complex repair or reconstruction (B) Capsular plication (C) Radial head replacement (D) Radial head ORIF (E) Medial collateral ligament complex reconstruction",
    "options": {
        "A": "Lateral collateral ligament complex repair or reconstruction",
        "B": "Capsular plication",
        "C": "Radial head replacement",
        "D": "Radial head ORIF",
        "E": "Medial collateral ligament complex reconstruction"
    },
    "label": "A",
    "images": [
        "MM-0-a.jpeg",
        "MM-0-b.jpeg"
    ],
    "medical_task": "Treatment",
    "body_system": "Skeletal",
    "question_type": "Reasoning"
}
```

## Observations

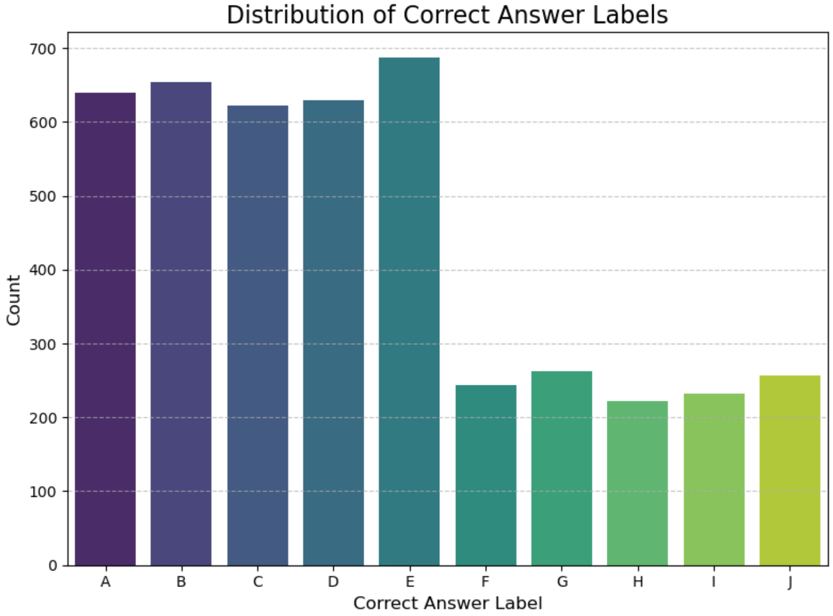
- “question” varies widely in length—mean around 134 words and 893 characters, with a sizeable right-tailed distribution (max up to 728 words, 4527 chars).
  - The median length (133 words) aligns with typical educational medical exam questions.
  - Compared to the average of 12 words per sentence of ViMQ, this is significantly more complex, and it mimics the real-word usage better.



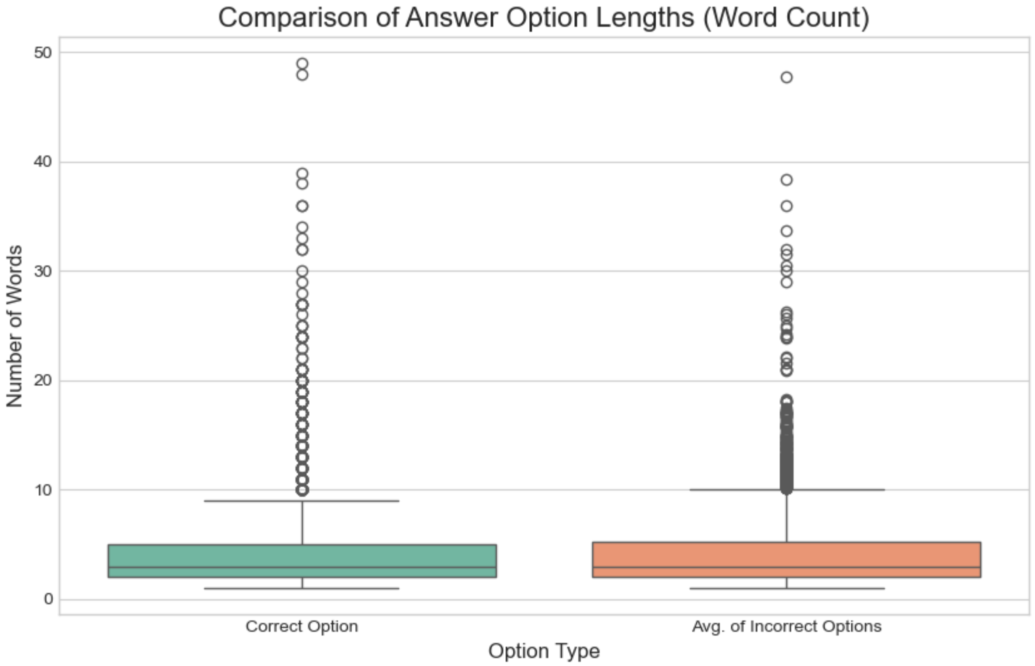
- Extreme length outliers (up to 728 words) may cause model training inefficiencies or require truncation strategies.
- Major question topics:



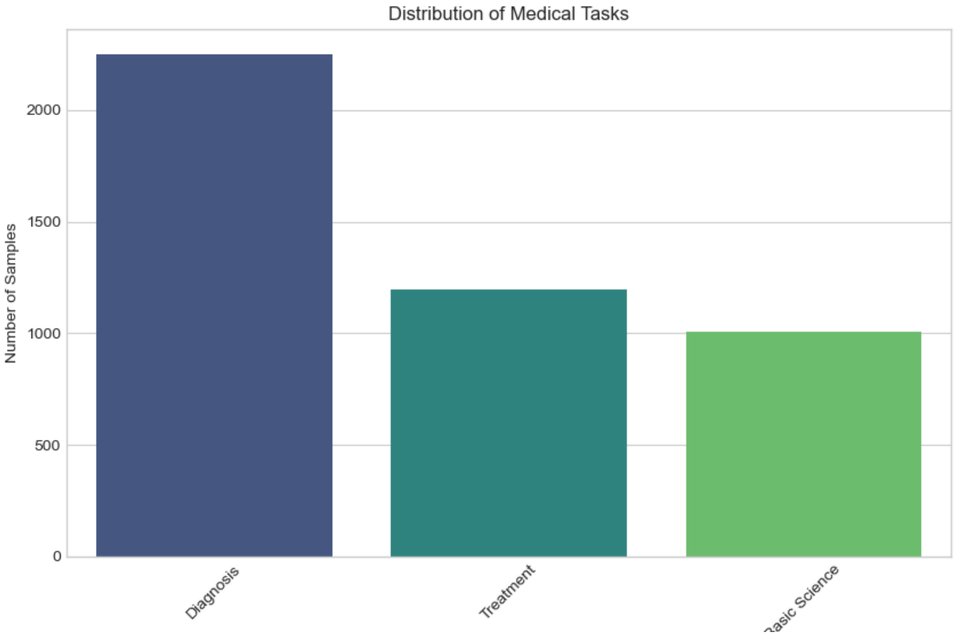
- “options”:

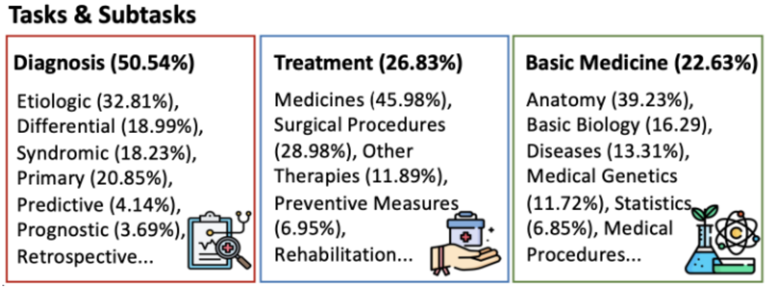


  - The counts for the first five options (A, B, C, D, E) are all relatively similar, hovering in the 600s. This is a very good sign of a well-designed dataset.
  - The counts for options F through J are significantly lower (in the 200s). This suggests that not all questions have 10 options.
  - Average options per question ~7.75, with a minimum of 5 and a maximum of 10.
  - → There are two types: questions with 5 options and questions with 10 options, with the 5 options being more popular.
  - Correct and incorrect options have very similar length distributions (mean ~4.16 words). This will prevent the model from associating option length with correctness:



- “medical_task” composition: Dominated by Diagnosis (2249 samples, ~50.5%), followed by Treatment (1194, ~26.8%) and Basic Science (1007, ~22.6%).

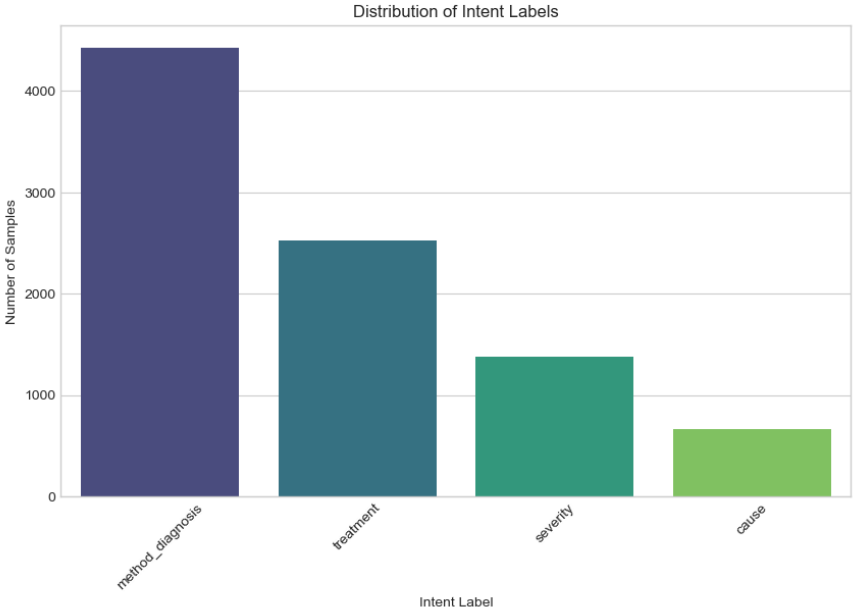




**Tasks & Subtasks**

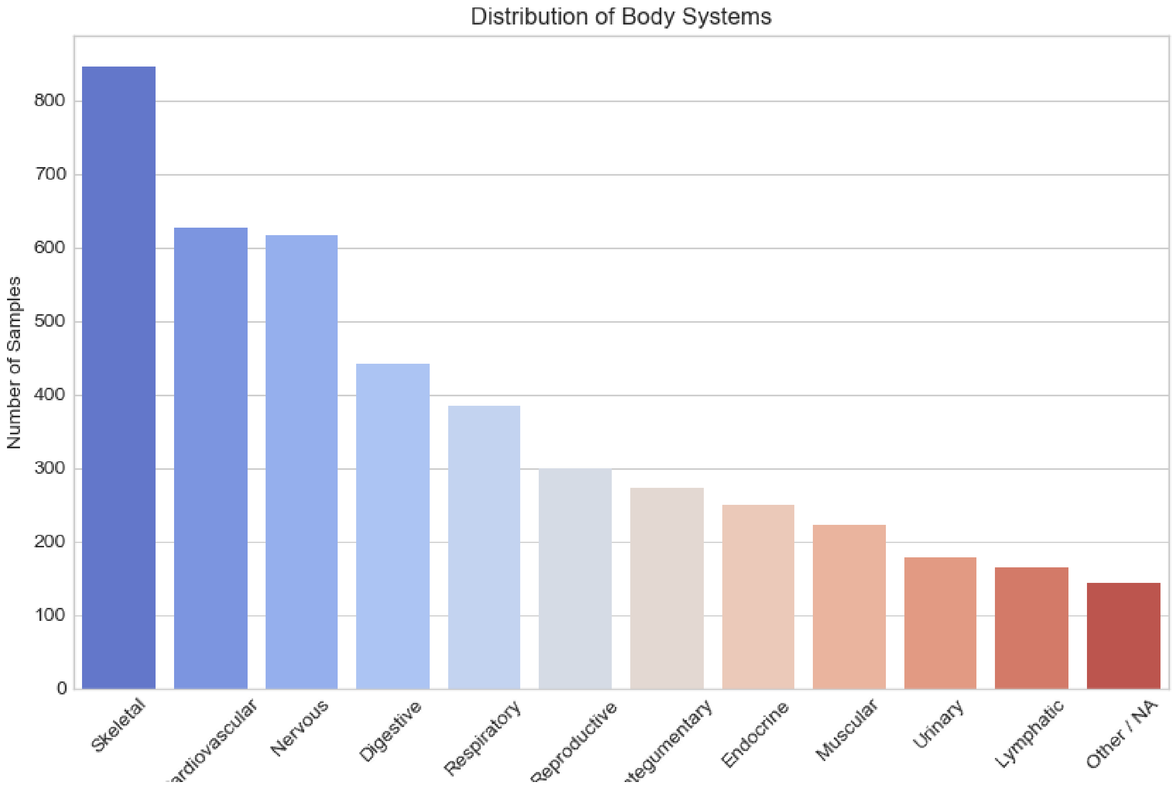
| Diagnosis (50.54%) | Treatment (26.83%) | Basic Medicine (22.63%) |
|---|---|---|
| Etiologic (32.81%), Differential (18.99%), Syndromic (18.23%), Primary (20.85%), Predictive (4.14%), Prognostic (3.69%), Retrospective... | Medicines (45.98%), Surgical Procedures (28.98%), Other Therapies (11.89%), Preventive Measures (6.95%), Rehabilitation... | Anatomy (39.23%), Basic Biology (16.29), Diseases (13.31%), Medical Genetics (11.72%), Statistics (6.85%), Medical Procedures... |

  - → Medical Tasks are uneven, skewed toward Diagnosis. This is the problem encountered with ViMQ, due to the nature of medical questions.



(ViMQ intent distribution)

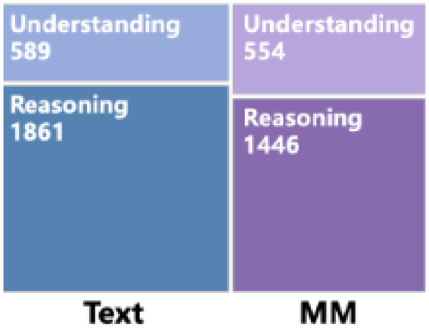
- “body_system” focus:



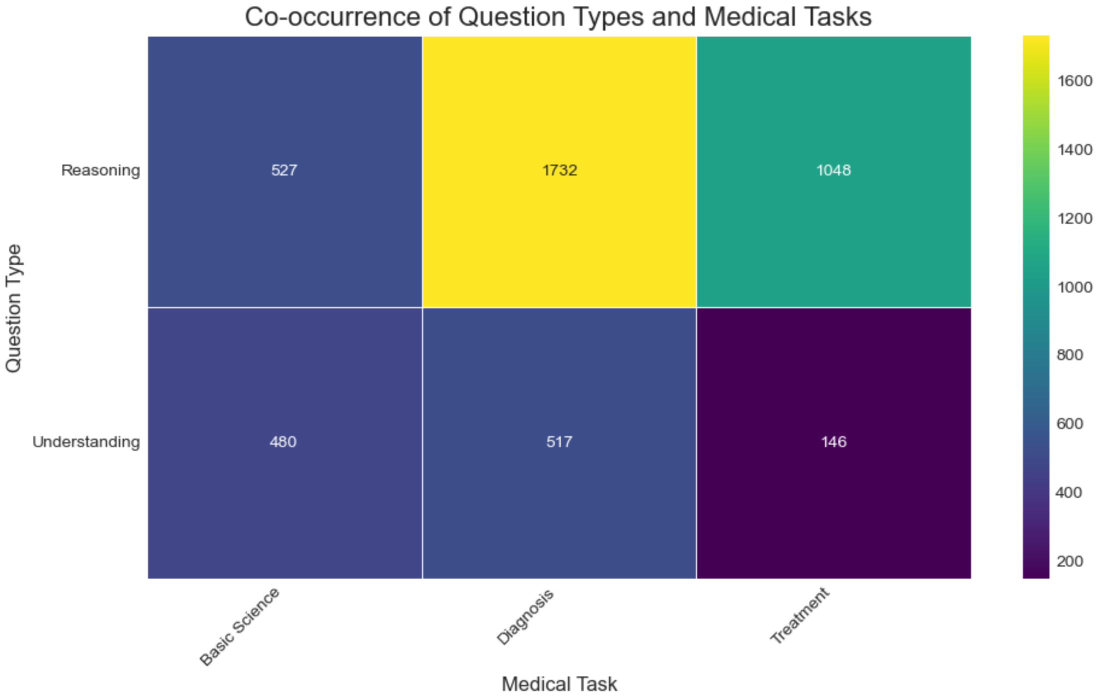
*(Cropped texts are Cardiovascular and Integumentary)*

  - Most questions relate to the Skeletal system (846 samples), Cardiovascular (626), and Nervous system (617). This is influenced by the paper’s data sources:
    i. Multimodal questions from ACR, EDiR, NEJM: A huge portion of diagnostic imaging (X-rays, CT scans, MRIs) is dedicated to the skeletal system (fractures, arthritis), the nervous system (brain MRIs/CTs for strokes, tumors), and the cardiovascular system (chest X-rays, CT angiograms).
    ii. Questions from COMLEX, a licensing exam that has a unique focus on osteopathic (nắn xương) principles, musculoskeletal (hệ cơ xương) medicine.
  - Least represented are Lymphatic (Hệ bạch huyết) (166) and Other/NA (144). This might affect generalizability across all medical domains.

- “question_type”: Vast majority are Reasoning questions (3307, ~74%), with the rest being Understanding (1143, ~26%).

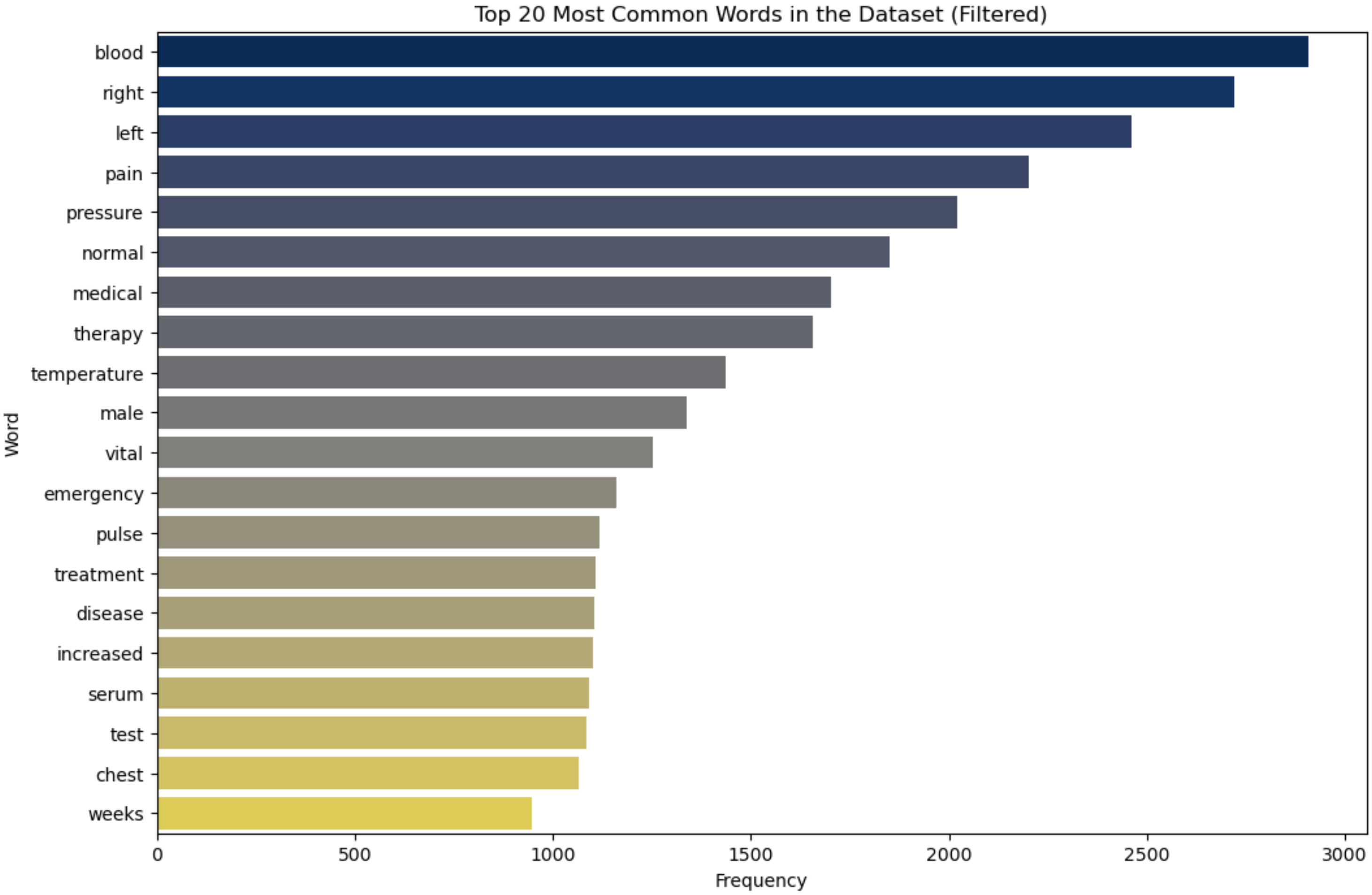


  - This imbalance is not an accidental bias but a deliberate and core design choice of the benchmark. Acccording to the paper, older benchmarks, such as MedQA, contain questions assessing medical knowledge only, hence are suboptimal for isolating the model's medical reasoning ability.
  - Heatmap of question types vs. medical tasks:



    i. Reasoning questions dominate Diagnosis and Treatment tasks; it makes sense as clinical decision-making emphasizes reasoning.
    ii. Understanding questions are more common in Basic Science, reflecting testing of foundational knowledge.

- Vocabulary:
  - Large vocabulary (24,343 unique words) condensed with filtering to 17,115 filtering out stop words and some uninformative words.
    i. Compared to 3,272 unique words of ViMQ, this is an improvement to cover more medical terms.
  - Top words include clinically relevant terms: blood, pain, pressure, therapy, temperature.



- Comparisons with ViMQ:
  i. Key difference 1: Focus
     a. MedXpertQA appears to be a general, broad-spectrum medical dataset. The vocabulary (blood, pressure, temperature, emergency, chest) is universal to adult medicine, surgery, and emergency care. The word male points to a focus on adult patient demographics.
     b. ViMQ has a clear and overwhelming focus on pediatrics and obstetrics. This is immediately obvious from the high frequency of words like trẻ (child), thai (pregnancy), bé (baby), and sơ_sinh (newborn).

  ii. Key difference 2: Language
     a. MedXpertQA uses the language of objective, clinical measurement and assessment. Words like right, left (laterality), pressure, temperature, pulse, vital (vital signs), serum, test, and normal vs. increased are all staples of a formal clinical workup. This is the language a doctor uses in a chart or a formal exam.
     b. ViMQ uses the language of symptoms and practical treatment actions. Words like sốt (fever), viêm (inflammation), đau (pain), and dấu_hiệu (sign/symptom) are prominent. Furthermore, the focus is on the method

of treatment: thuốc (medicine), uống (taking orally), tiêm (injecting). This is closer to the language used when discussing a condition and its management, perhaps with a patient.

iii. Key difference 3: Source

   a. MedXpertQA's vocabulary strongly suggests it is sourced from expert-level materials like board exams, case studies, and clinical vignettes. The language is precise, formal, and diagnostic, focusing on the data needed for a clinician to make a decision.

   b. ViMQ's vocabulary suggests it may be sourced from patient-facing materials, such as a medical Q&A forum, public health articles, or a different type of clinical resource. Words like nguy_hiểm (dangerous) and ảnh_hưởng (affect/influence) are more descriptive and evaluative, common in patient education, compared to the objective language in MedXpertQA.

## Multimodal

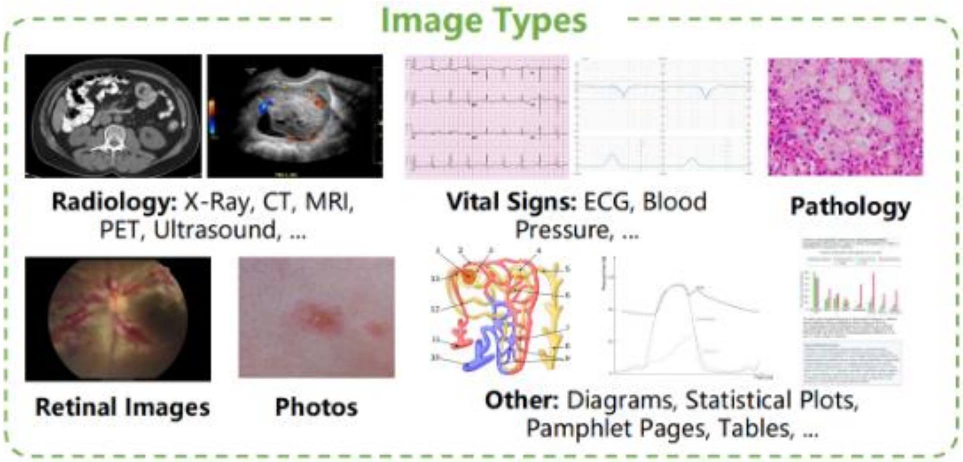
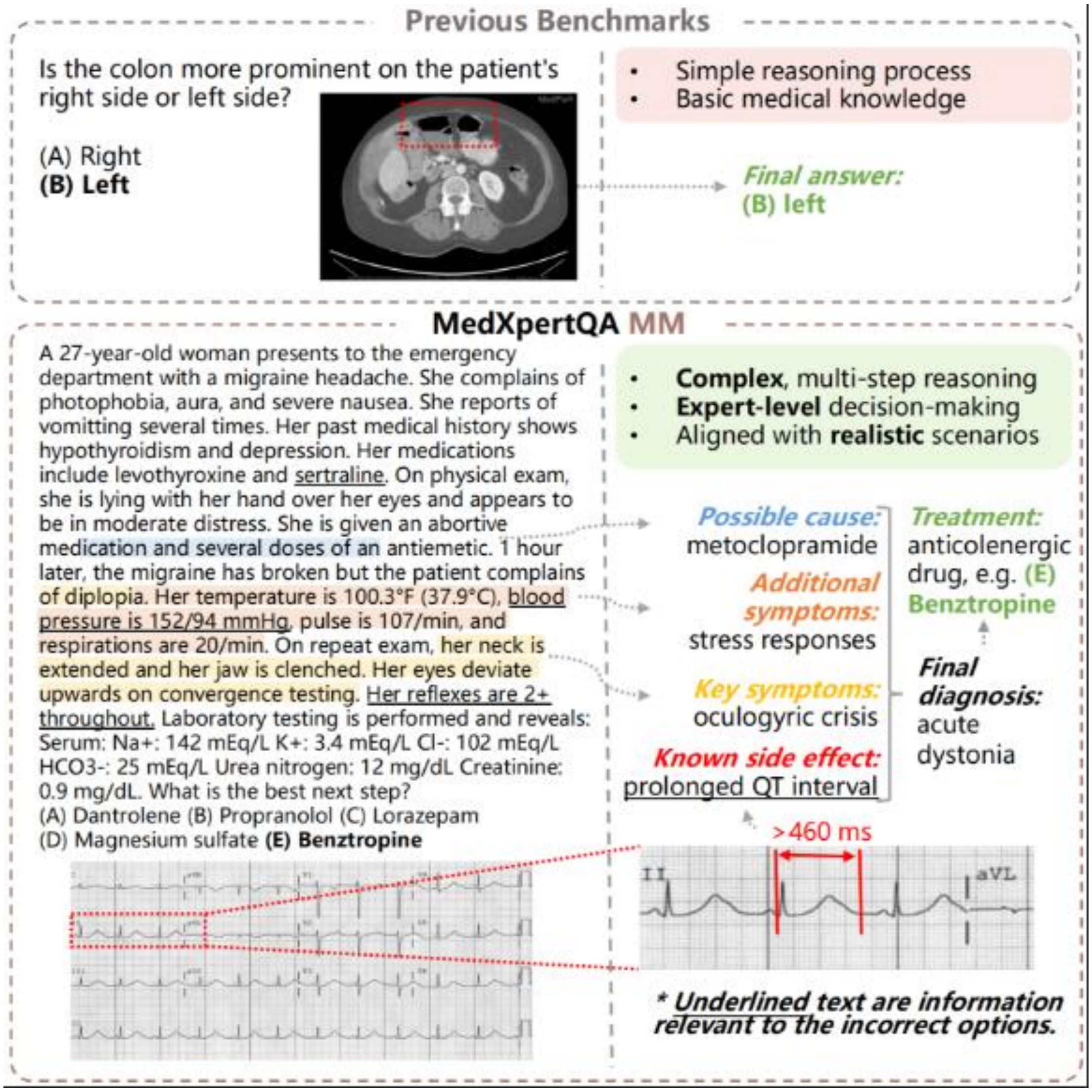


Image Types
Radiology: X-Ray, CT, MRI, PET, Ultrasound, ...
Vital Signs: ECG, Blood Pressure, ...
Pathology
Retinal Images
Photos
Other: Diagrams, Statistical Plots, Pamphlet Pages, Tables, ...

- The multimodal subset is a significant advancement from traditional benchmarks that often use simple question-answer pairs generated from image captions.



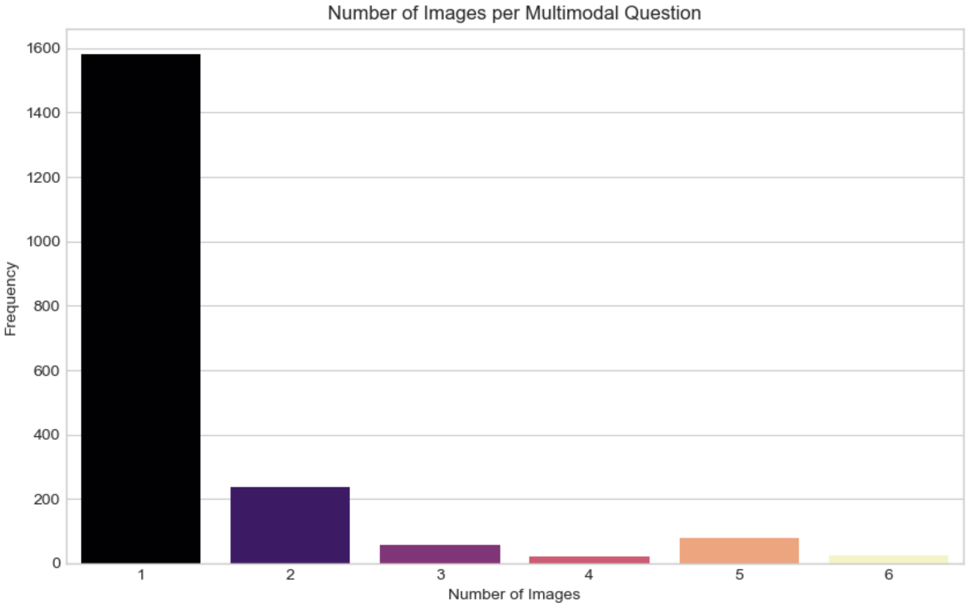
Is the colon more prominent on the patient's right side or left side?

(A) Right
**(B) Left**

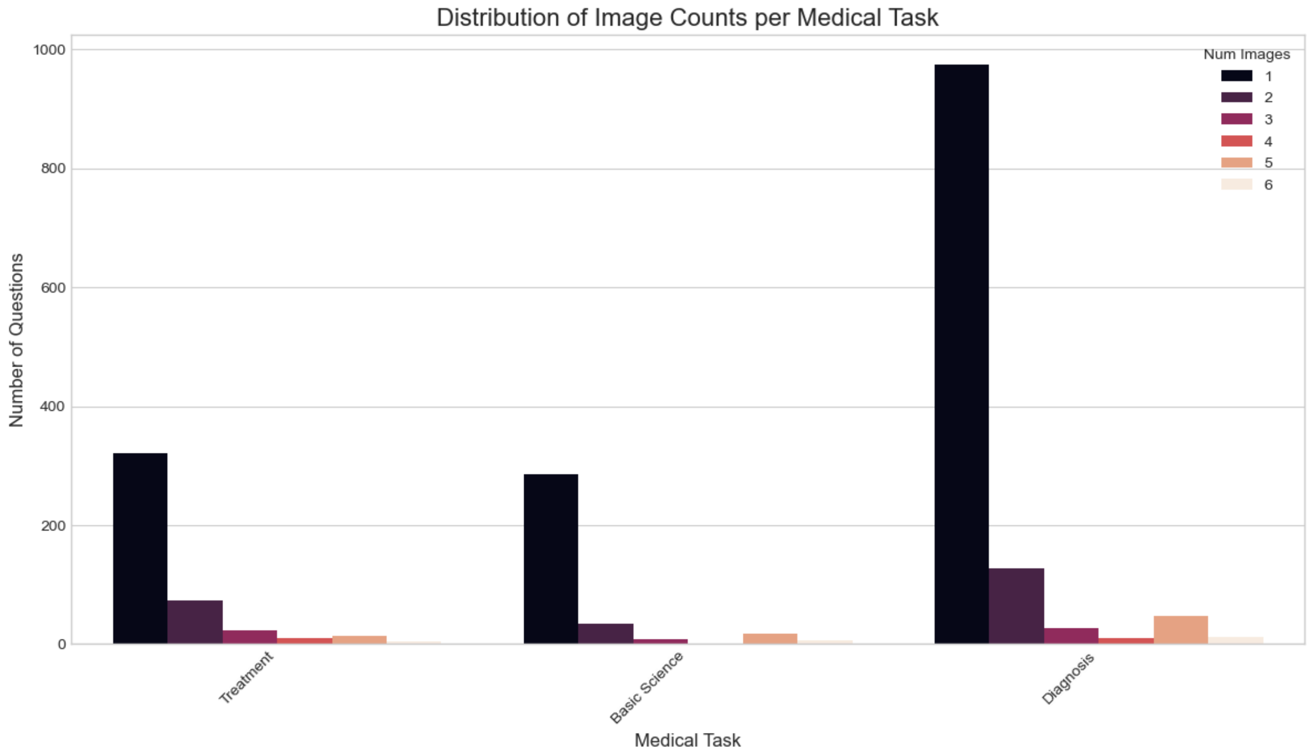
A 27-year-old woman presents to the emergency department with a migraine headache. She complains of photophobia, aura, and severe nausea. She reports of vomitting several times. Her past medical history shows hypothyroidism and depression. Her medications include levothyroxine and sertraline. On physical exam, she is lying with her hand over her eyes and appears to be in moderate distress. She is given an abortive medication and several doses of an antiemetic. 1 hour later, the migraine has broken but the patient complains of diplopia. Her temperature is 100.3°F (37.9°C), blood pressure is 152/94 mmHg, pulse is 107/min, and respirations are 20/min. On repeat exam, her neck is extended and her jaw is clenched. Her eyes deviate upwards on convergence testing. Her reflexes are 2+ throughout. Laboratory testing is performed and reveals: Serum: Na+: 142 mEq/L K+: 3.4 mEq/L Cl-: 102 mEq/L HCO3-: 25 mEq/L Urea nitrogen: 12 mg/dL Creatinine: 0.9 mg/dL. What is the best next step?
(A) Dantrolene (B) Propranolol (C) Lorazepam
(D) Magnesium sulfate **(E) Benztropine**

*** *Underlined text are information relevant to the incorrect options.***

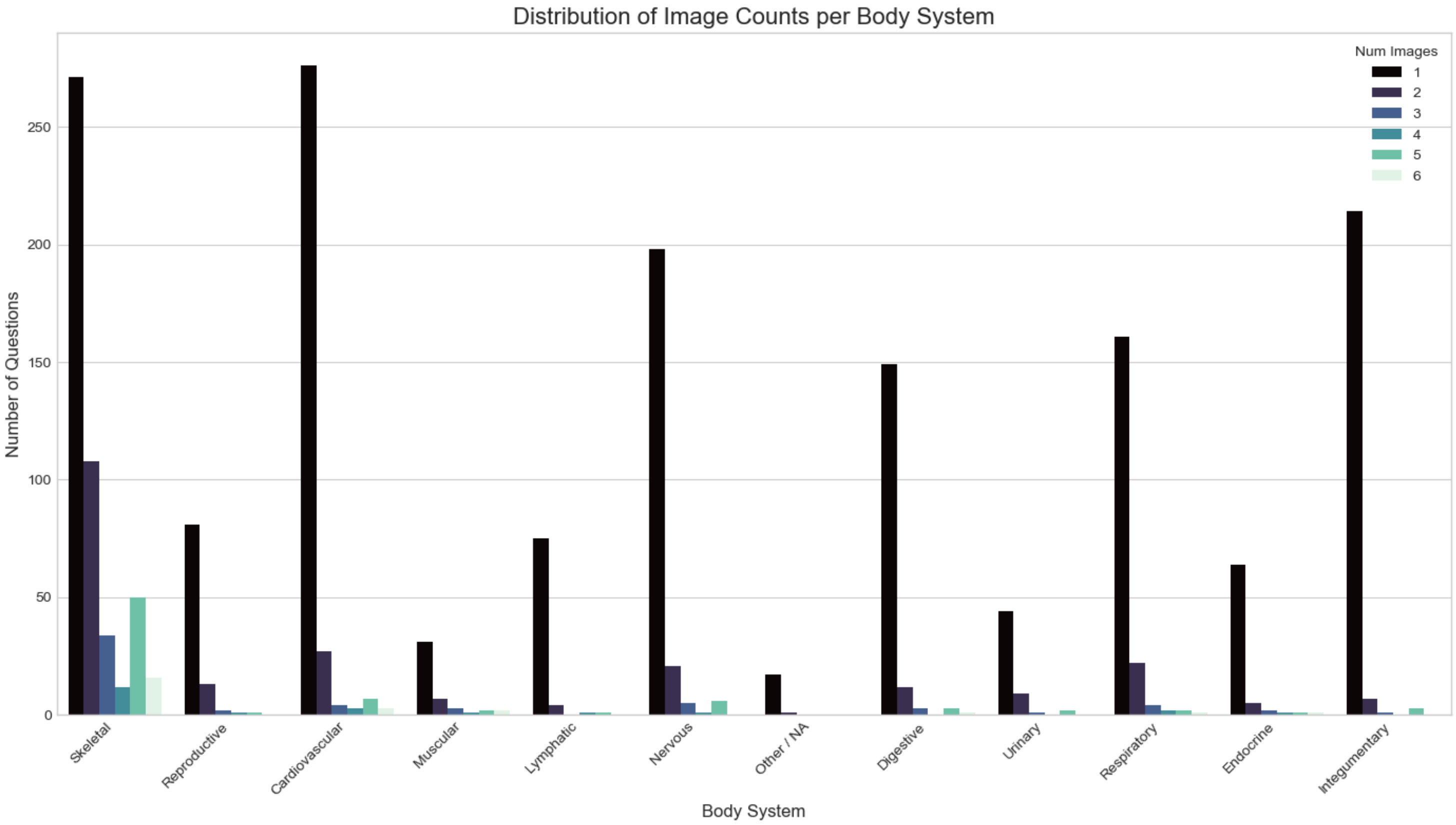
- Multimodal questions predominantly contain a single image (mean ~1.43 images/question), but some contain up to 6 images.



- Image counts vary by task and system, with Skeletal and Cardiovascular systems having the highest coverage:



i. The relatively low multimodal presence in Basic Science aligns with more text-based conceptual knowledge.



  ii. Skeletal system questions often contain multiple images (significant amount of 2,3,5 images per question) compared to other types. These often require multiple radiographs or scans.

- Images predominantly JPEG format (2683), some PNG (161), few JPG (8). This may limit image format diversity.

# ViMedAQA

- **ViMedAQA** is a Vietnamese Medical Abstractive Question Answering dataset collected from https://youmed.vn/. Its main goal is to support research and development of automated question-answering systems in Vietnamese, with a focus on the medical domain.
- **Scale**
  - Total samples: 44,313 question-answer-context triplets.
  - Topics: Body part – 0, Disease – 1, Drug – 2, Medicine – 3.

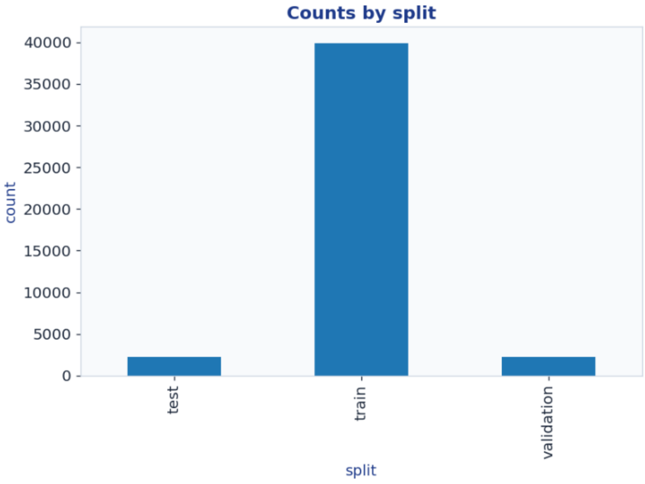
    Sorted by counts:

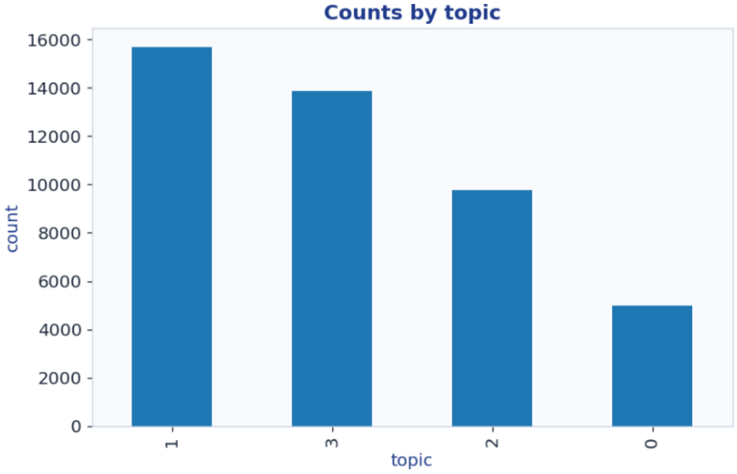
    i. Disease – 1 (15,690)

    ii. Medicine – 3 (13,873)

    iii. Drug – 2 (9,780)

    iv. Body part – 0 (4,970)
  - Splits: Train (39,881), Validation (2,215), Test (2,217)



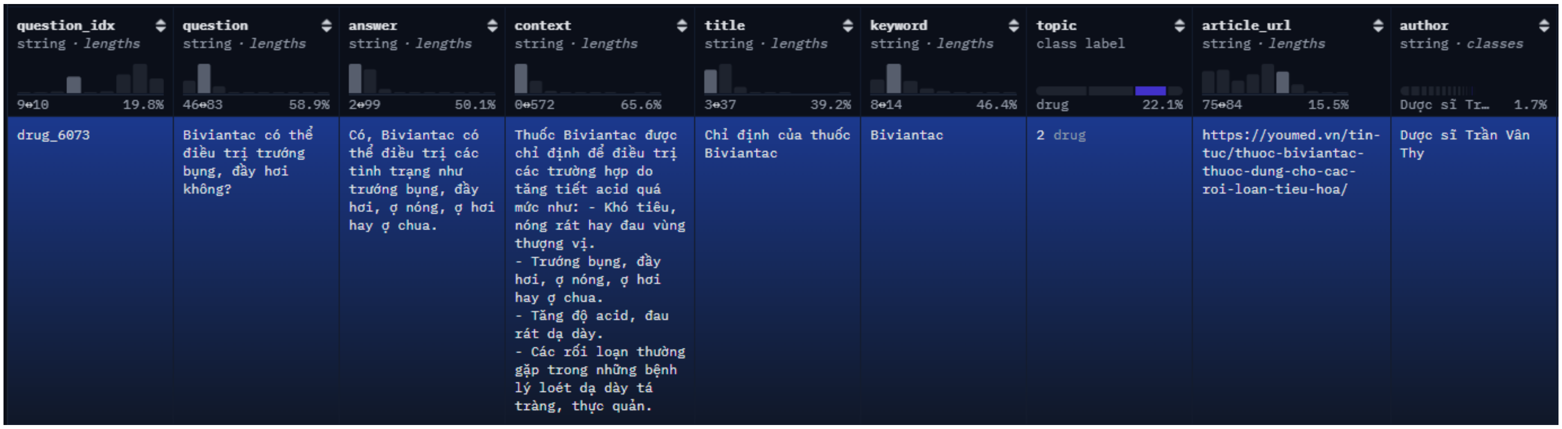


- **Structure of each record**
  - *'question_idx':* The index of the sample.
  - *'question':* The question to be answered.
  - *'answer':* The answer to the question.
  - *'context':* The context or pargraph that contains the infromation to answer the question.
  - *'title':* The title of the corresponding article from which the context was taken.
  - *'keyword':* The key disease/drug/body_part in the question.

- *'topic':* The topic of the question/context. It can be one of the following: Body part, Disease, Drug, and Medicine.
- *'article_url':* The URL of the original article.

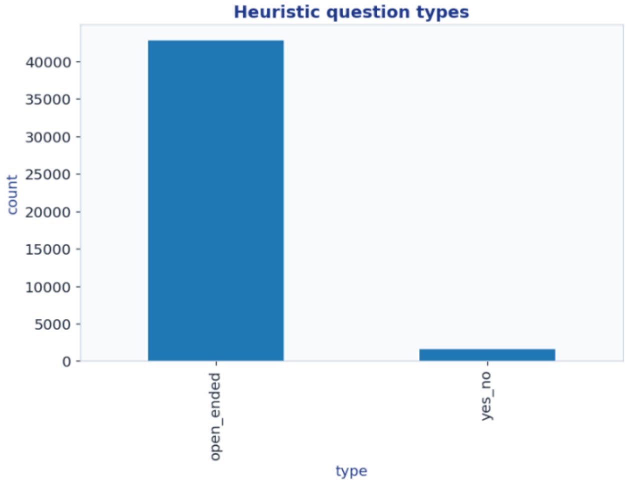
***Note***: Each topic in ViMedAQA originates from a single expert-authored source article.

***E.g. data point***



| question_idx<br>string · *lengths* | question<br>string · *lengths* | answer<br>string · *lengths* | context<br>string · *lengths* | title<br>string · *lengths* | keyword<br>string · *lengths* | topic<br>class label | article_url<br>string · *lengths* | author<br>string · *classes* |
|---|---|---|---|---|---|---|---|---|
| 9↔10 19.8% | 46↔83 58.9% | 2↔99 50.1% | 0↔572 65.6% | 3↔37 39.2% | 8↔14 46.4% | drug 22.1% | 75↔84 15.5% | Dược sĩ Tr… 1.7% |
| drug_6073 | Biviantac có thể điều trị trướng bụng, đầy hơi không? | Có, Biviantac có thể điều trị các tình trạng như trướng bụng, đầy hơi, ợ nóng, ợ hơi hay ợ chua. | Thuốc Biviantac được chỉ định để điều trị các trường hợp do tăng tiết acid quá mức như: - Khó tiêu, nóng rát hay đau vùng thượng vị.<br>- Trướng bụng, đầy hơi, ợ nóng, ợ hơi hay ợ chua.<br>- Tăng độ acid, đau rát dạ dày.<br>- Các rối loạn thường gặp trong những bệnh lý loét dạ dày tá tràng, thực quản. | Chỉ định của thuốc Biviantac | Biviantac | 2 drug | https://youmed.vn/tin-tuc/thuoc-biviantac-thuoc-dung-cho-cac-roi-loan-tieu-hoa/ | Dược sĩ Trần Vân Thy |

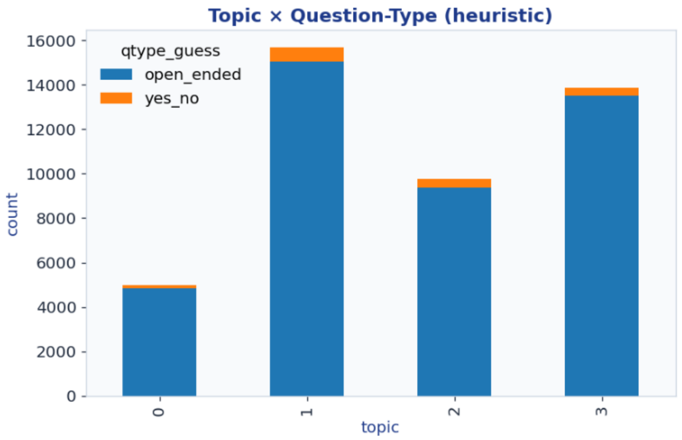
- **Question Types**



- The dataset contains **44,313** questions, divided into **Yes-No** (3,740) and **Open-Ended** (40,443) types. Open-Ended questions include subtypes:

| **Subtype** | What | How | When | Where | Which | Why | How much/many | Who | Others |
|---|---|---|---|---|---|---|---|---|---|
| **#Questions** | 27,403 | 5,546 | 2,385 | 1,294 | 1,205 | 1,204 | 721 | 685 | 130 |

- **Topic x Question Type**



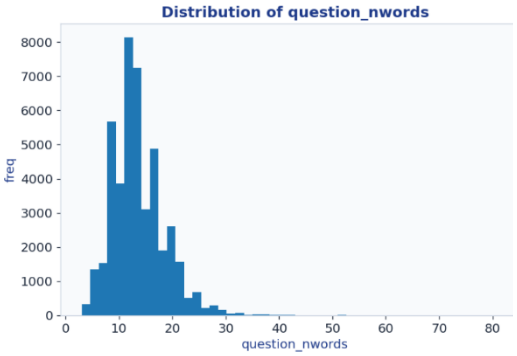
Across four topics, there are 40,759 open-ended and 1,554 yes-no questions:

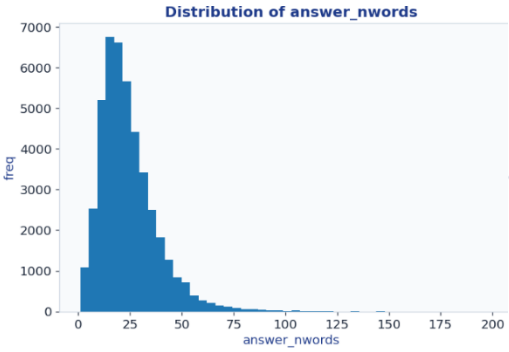
| | #Open-Ended questions | #Yes-No questions |
|---|---|---|
| **Body part** | 4,838 | 132 |
| **Disease** | 15,037 | 653 |
| **Drug** | 9,375 | 405 |
| **Medicine** | 13,509 | 364 |

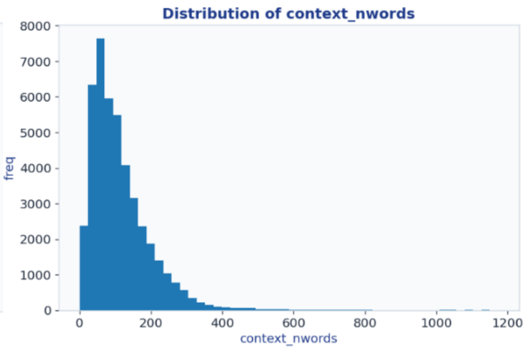
Most questions are **open-ended**, especially in **Disease** and **Medicine**, while **Body part** has the fewest. **Drug** and **Disease** show a relatively higher share of **yes-no** questions compared to other topics

- **Length Analysis of Questions, Answers, and Contexts**

The dataset shows a clear distinction in their lengths, reflecting its design for abstractive QA.







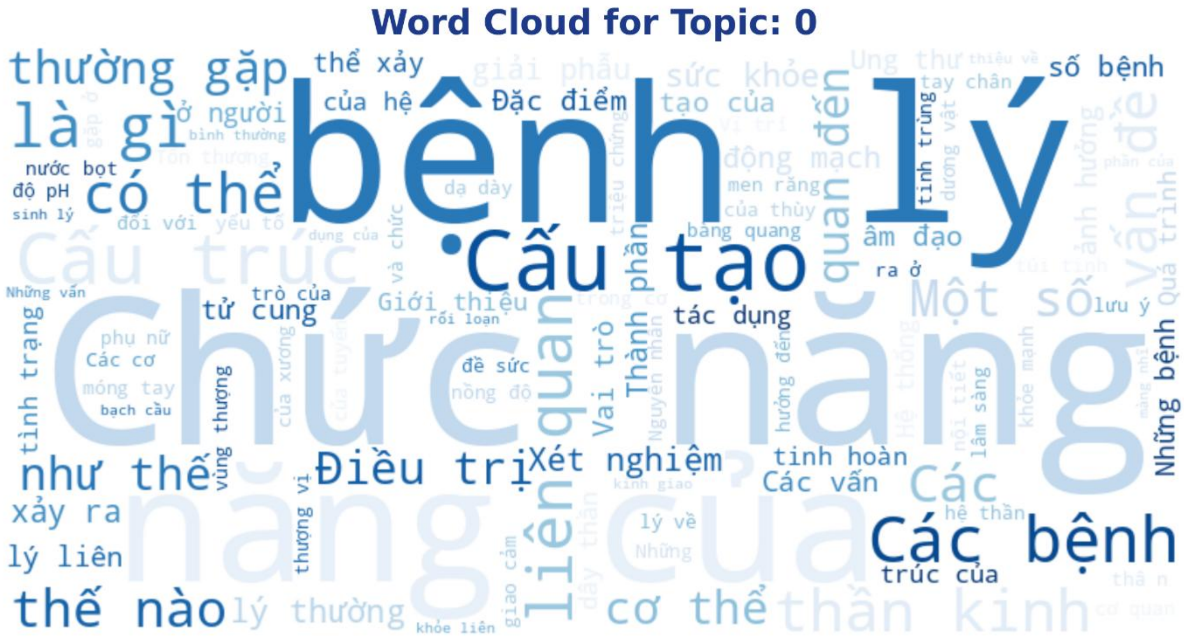
- Questions
  1. Mean: ~13.5 words, Median: 13 words, Range: 3 to 83 words → most questions are short, natural, and well-suited for both retrieval-based and generative QA.

        ii. 90% ≤ 20 words, 95% ≤ 22 words → very few long or noisy questions; models can expect concise inputs.
        iii. Some questions up to 83 words → introduces diversity for evaluating reasoning and summarization capabilities.
    - Answers
        i. Mean: ~17.5 words, Median: ~15 words, Range: ~3 to 200 words → answers are concise yet informative, suitable for abstractive QA.
        ii. 90% ≤ 29 words, 95% ≤ 36 words → most answers are short summaries, promoting efficient evaluation.
        iii. Few long answers (up to ~200 words) → provide edge cases for testing model summarization and reasoning.
    - Contexts
        i. Mean: ~111 words, Median: ~92 words, Range: ~10 to 1,171 words → moderately long passages, rich in detail for QA.
        ii. 90% ≤ 211 words, 95% ≤ 281 words → most contexts are manageable, with only a small fraction being long.
        iii. Long-tail contexts (up to ~1,171 words) → introduce challenging comprehension tasks for advanced models.

- **Duplicate Analysis**
    - **206 exact duplicate questions** were found, likely caused by inconsistencies during **LLM-based question generation and annotation**, which may introduce minor labeling noise.
    - **Only 4 duplicate question–context pairs** exist, showing low redundancy and strong dataset diversity.
- **Word Cloud Title Analysis by Topic**

  To better understand the content focus of each topic, we generated word clouds based on sample's title. These visualizations highlight the most frequent terms and provide quick insight into thematic emphasis. Distinct patterns across topics confirm the dataset's clear topical boundaries and curated domain-specific coverage.

- Topic 0 – Body Part

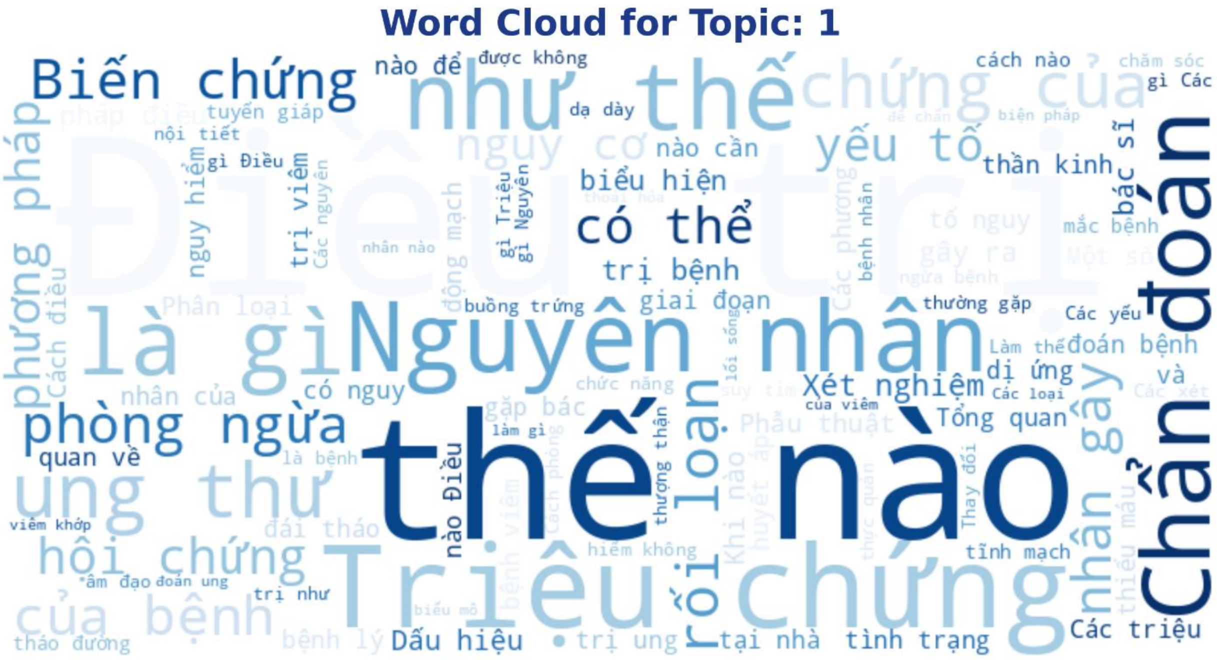


Keywords like *structure, function, body, organ, treatment* suggest that samples in this topic:

i. Describe **anatomy and physiology** of body parts and systems.

ii. Provide foundational knowledge on organ roles and links to medical conditions.

- Topic 1 – Disease

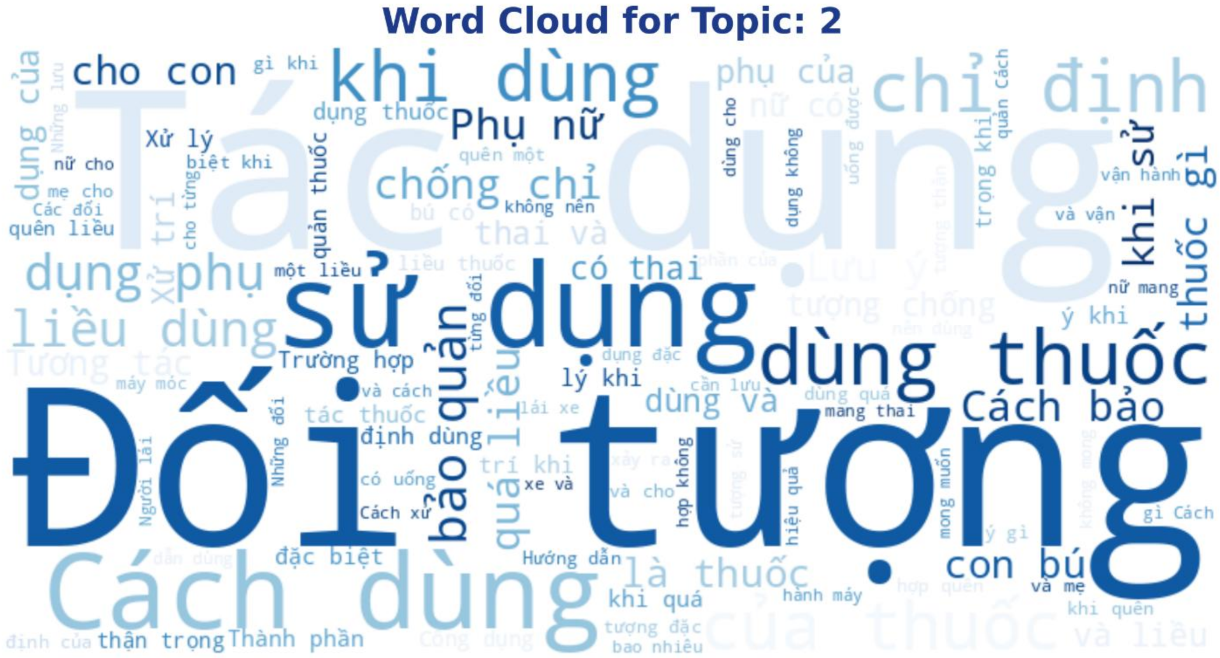


Frequent terms include *cause, symptom, diagnosis, prevention, complication:*

i. Samples focus on **disease definitions, pathogenesis, and clinical features**.

ii. Many titles emphasize **diagnostic methods, treatment, and prevention strategies**.

- Topic 2 – Drug

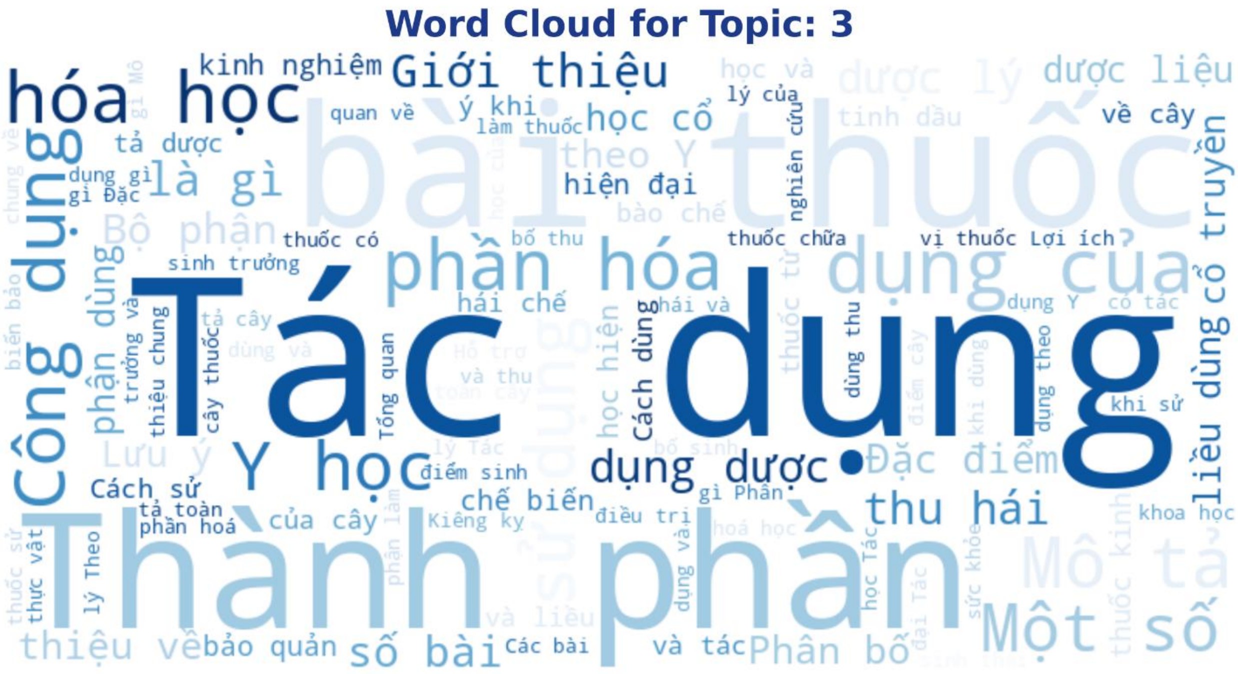


Prominent words such as *ussage, indication, dosage, side effect*:

i. Content details drug information, safe use, dosage, and contraindications.

   ii. Provides essential context for pharmaceutical safety and prescription guidelines.

- Topic 3 – Medicine



Keywords like *ingredient, effect, preparation, medicinal herb:*

   i. Covers traditional medicine, herbal remedies, ingredients, and preparation methods.

   ii. Adds diversity to the dataset by integrating **holistic and traditional knowledge**.

- **Conclusion**
  - ViMedAQA is a **44,313-sample Vietnamese Medical Abstractive QA dataset** covering Body Part, Disease, Drug, and Traditional & Herbal Medicine. EDA shows:

     i. Questions are short (~13 words), answers concise (~17 words), contexts moderate (~111 words).

     ii. Open-ended questions dominate (~92%), promoting reasoning and comprehension tasks.

     iii. Minimal duplicates and clear topic separation confirm good quality.

  - This dataset is ideal for **abstractive QA, retrieval-augmented generation, information retrieval, and topic classification**.
  - **Strengths:** expert-authored, well-structured, diverse coverage of modern and traditional medicine.
  - **Limitations:** single-source per topic, minor annotation noise, some long contexts, and risk of outdated medical knowledge.

[1] PubMed - Medical terminology: Its size and typology