
Supplementary Material for UltraMedical

Contents

1	Dataset Details	2
1.1	Details of UltraMedical Instructions	2
1.2	Details of UltraMedical Preference	2
1.3	Details of Medical Reward Bench	2
1.4	Details of General Instructions	2
1.5	Details of Benchmarks	3
2	Dataset Analysis	3
2.1	Correlation of model-based scores	3
2.2	Complexity Evolving of Instructions	3
2.3	Instruction Distribution	4
2.4	Instruction Decontaminate	4
3	Dataset Release	4
3.1	Website	4
3.2	Limitations	5
4	The Medical RewardBench	6
4.1	Construction	6
4.2	Human Annotation	6
4.3	Data Statistics	7
5	Prompts	8
6	Examples	12
7	Croissant Document	23

1 Dataset Details

1.1 Details of UltraMedical Instructions

We display the composition of the UltraMedical collections in Figure 1a, where multi-choice question answering comprises about 50%, PubMed question answering accounts for about 20%, and the remaining 30% consists of open-ended instructions and dialogues. As displayed in Figure ??, we randomly selected 200K prompts from the UltraMedical collection and mapped them into vectors using Atlas Nomic.AI. We present the topic distribution in Figure ?? and the task distribution in Figure 1b, both of which validate the effectiveness of our diversity-driven process. Details about the map can be viewed through this [Nomic AI Atlas](#).

1.2 Details of UltraMedical Preference

We present the model’s accuracy for QA tasks in Figure 4a, the models’ win percentages in binarized preference in Figure 7, and the scores and rankings of all models across various tasks from GPT-4 in Figures 6 and 5.

1.3 Details of Medical Reward Bench

For the easy set, we selected gpt-4-1106-preview as the chosen model, while gpt-3.5-turbo-1106, Mixtral-8x22B-Instruct, and Mixtral-8x7B-Instruct were rejected. For the hard set, we selected models with the highest and nearly highest scores, including gpt-4-1106-preview, Meta-Llama-3-70B-Instruct, and Llama-3-8B-UltraMedical. For the set without length bias, we selected Meta-Llama-3-70B-Instruct and Meta-Llama-3-8B-Instruct as chosen and rejected, respectively, which have a significant gap in scores but the same answer length.

For the initially given 1,000 test pairs, we ultimately retained 777 pairs following human expert annotation. These include 238 easy, 196 hard, 180 length-based, and 163 human-judged pairs. Approximately 233 pairs were filtered out due to issues such as incorrect formulations, difficulty in answering, or both. The human category comprises pairs where preferences differ between human annotators and GPT-4, which is regarded as even hard for GPT-4 to recognize.

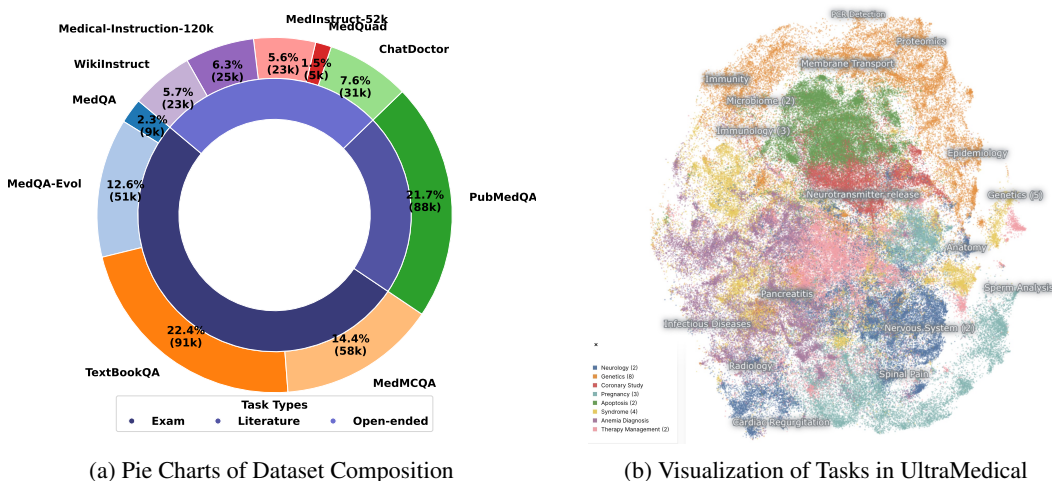


Figure 1: Statistics of UltraMedical dataset.

1.4 Details of General Instructions

To enhance the general instruction-following capabilities, we integrate the UltraMedical with high-quality prompts from various general domains, sourced from UltraChat, Dolphin, Wizard, Orca, and additional datasets included in 0-hero/Matter-0.1.

Table 1: Statistics of datasets in UltraMedical.

Type	Dataset	Count	Description
Exam	MedQA	9,343	General medical knowledge in US medical licensing exam
	MedMCQA	58,871	General medical knowledge in Indian medical entrance exams
	MedQA-Evol	51,809	Synthesized data evolved from the original MedQA data
	TextBookQA	91,684	Multiple-choice questions derived from medical books
Literature	PubMedQA	88,688	Closed-domain question answering given PubMed abstract
Open-End	MedQuad	5,957	Medical question-answer pairs created from 12 NIH websites
	MedInstruct-52k	23,032	Generated medical instruction-following data with self-instruct
	Medical-Instruction-120k	25,806	Various thoughts proposed by the people and synthetic responses
	ChatDoctor	31,115	Real conversations between patients and doctors from HealthCareMagic
	WikiInstruct	23,288	Detailed knowledge and instructions expanded from thousands of biomedical concepts from Wikipedia pages.

1.5 Details of Benchmarks

The number of evaluations and descriptions of the tasks are presented in Table 2.

Table 2: Statistics of datasets for evaluations.

Domain	Dataset	Count	Description
Medical	MedQA (UCMLE)	1273	General medical knowledge in US medical licensing exam
	MedMCQA	4183	General medical knowledge in Indian medical entrance exams
	PubMedQA	500	Closed-domain question answering given PubMed abstract
	MMLU-Clinical knowledge	265	Clinical knowledge multiple-choice questions
	MMLU-Medical genetics	100	Medical genetics multiple-choice questions
	MMLU-Anatomy	135	Anatomy multiple-choice questions
	MMLU-Professional medicine	272	Professional medicine multiple-choice questions
	MMLU-College biology	144	College biology multiple-choice questions
	MMLU-College medicine	173	College medicine multiple-choice questions
	K-QA	201	Real-world clinical questions with physician-curated answers (long-form answers)
	MultimedQA	140	Consumer medical question-answering data (long-form answers)
General	MT-Bench	80	Multi-turn question answering benchmark evaluating eight different abilities
	Alpaca-Eval 2	805	General world knowledge question-answering for chat-models
	Arena-Hard	500	Built from live data in the Chatbot Arena with challenging user queries
	MMLU	116k	Multi-choice questions for massive multitask language understanding
	GPQA	198	Very hard multiple-choice and question answering tasks in biology, physics, and chemistry
	GSM8K	1319	Grade school math word problems for question answering
MATH	5000	Challenging competition mathematics problems	

2 Dataset Analysis

2.1 Correlation of model-based scores

We have selected `gpt-3.5-turbo` as the evaluator for instruction scoring, as it remains highly competitive with mainstream open-source LLMs and offers scalability due to its lower cost. `gpt-3.5-turbo` demonstrates a high correlation and maintains stability across multiple evaluation iterations, as shown on the left side of Figure 2. Additionally, `gpt-3.5-turbo` exhibits a strong correlation with `gpt-4-turbo`, as depicted in the middle of Figure 2. The primary difference is that instructions typically receive slightly lower scores in `gpt-4-turbo` evaluations.

Beyond model-based scoring, previous studies have also attempted to rank instructions directly based on length. As illustrated on the right side of Figure 2, the correlation between model-based scores and lengths is very low, indicating that the evaluator prioritizes assessing instruction complexity rather than merely its length.

2.2 Complexity Evolving of Instructions

Complexity of instructions is a principal characteristic of high quality. For our synthetic datasets, we conduct two additional rounds of instruction evolution to increase complexity. As shown in Figure 3, the scores of instructions across the three datasets consistently increase. Within these datasets, instructions in TextBookQA are synthesized based on few-shot examples and paragraphs from textbooks, resulting in minor score changes. The WikiInstruct dataset, which includes various open-ended questions based on entities from Wikipedia, exhibits the highest complexity scores.

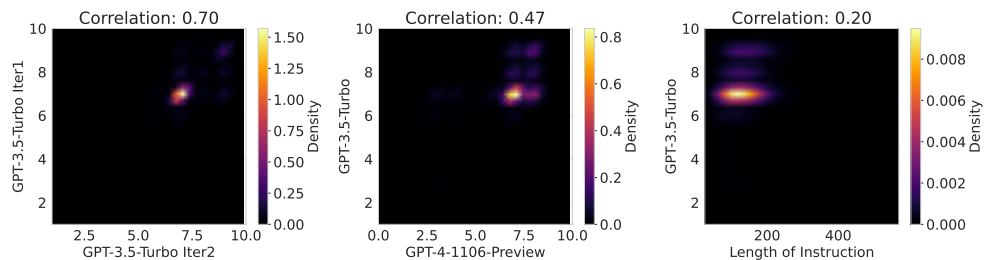


Figure 2: Correlation analysis of various scores, including those from different models and the length of instructions.

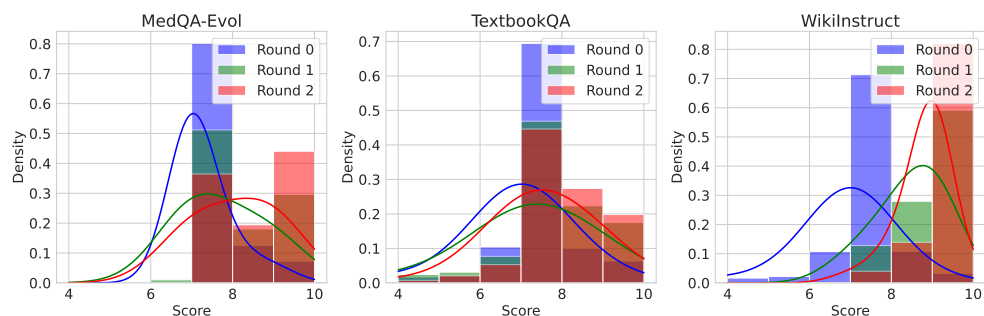


Figure 3: Distribution of model-based evaluation score progression across evolution rounds for our three synthetic datasets, illustrating how instruction evolution contributes to increased complexity.

2.3 Instruction Distribution

The UltraMedical collections contain three main task types and ten sub-tasks, as illustrated in Figure 1a. Questions derived from exams and textbooks account for approximately 50%, literature-based questions for about 20%, and open-ended instructions and questions for around 30%. We randomly sample 5,000 examples from each sub-task, embed them using `intfloat/e5-mistral-7b-instruct`[6], and subsequently project them into two dimensions with t-SNE. As depicted in Figure 1b, questions in the exam series exhibit broad and diverse topics, while instructions from literature and our synthetic instructions based on Wikipedia entities are complementary.

2.4 Instruction Decontaminate

Due to the use of large-scale synthetic data, we implement decontamination operations to prevent test set leakage, as described in the `bage1` project¹. Our approach involves clustering all training and test data based on dense vectors and then calculating the length correlation within the top K nearest samples. This method revealed no potential test data leakage in UltraMedical.

3 Dataset Release

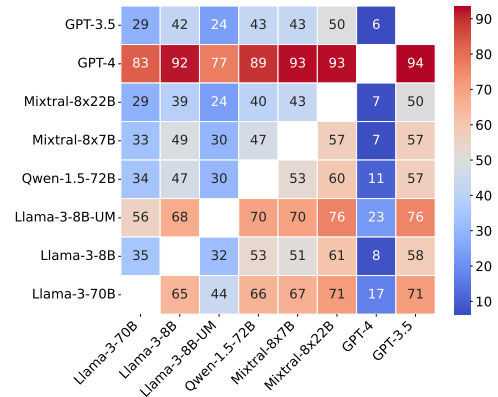
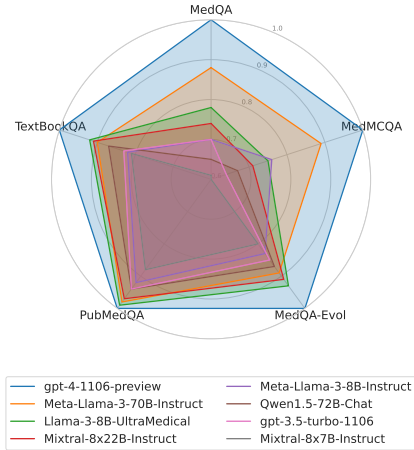
3.1 Website

We have released the datasets and models on GitHub and Huggingface, which ensure long-term availability.

GitHub: <https://github.com/TsinghuaC3I/UltraMedical>

Huggingface: <https://huggingface.co/TsinghuaC3I>

¹<https://github.com/jondurbin/bage1/tree/main>



(a) Models' accuracy on QA tasks. (b) Models's Win Percentage in Binarized Preference.

Figure 4: Annotation and preference of models statistic results.

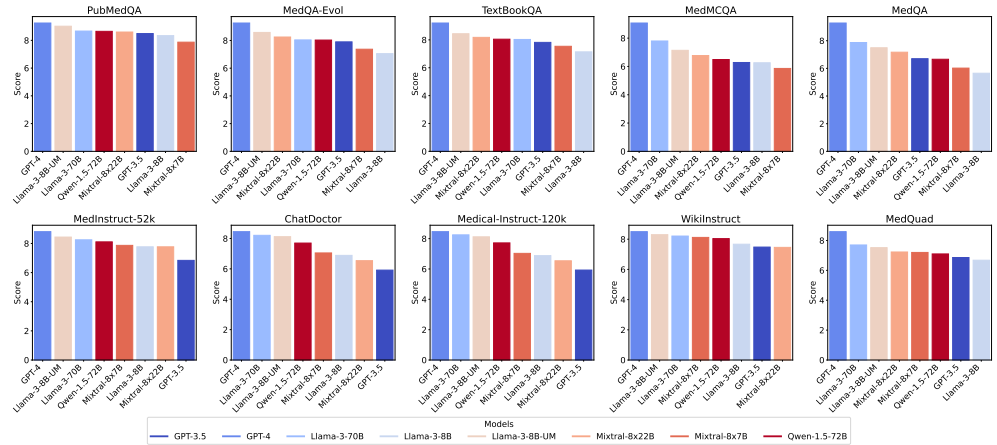


Figure 5: Scores of all models across various tasks from GPT-4 (higher is better).

We have also sampled 200,000 examples and displayed them on the [Nomic AI Atlas](#) for topic and task distribution exploration.

All model usage must adhere to the [Llama3 License](#)², and datasets derived from OpenAI GPT are not permitted for commercial use.

3.2 Limitations

While the UltraMedical suites show promising performance on several benchmarks, they still have limitations, such as hallucinations. Additionally, the outputs are synthesized from GPT-4, which may also exhibit bias. We plan to address these issues and verify the accuracy of facts in UltraMedical in future research.

²<https://llama.meta.com/llama3/license/>

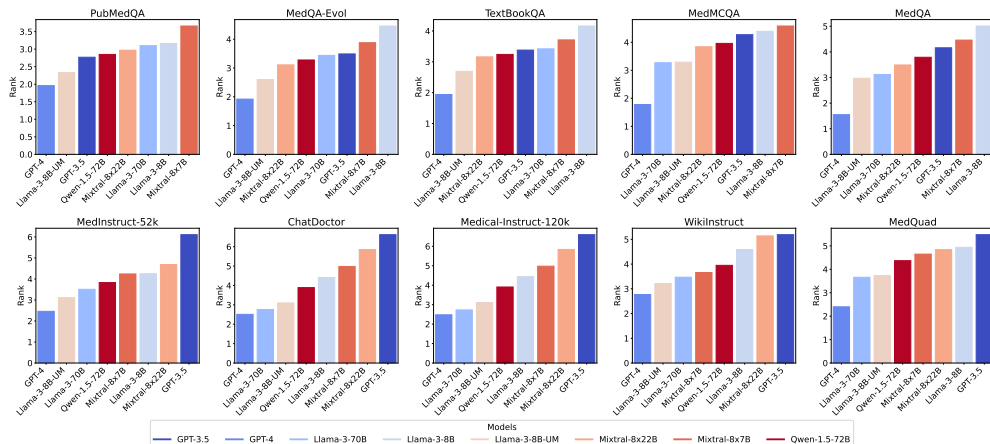


Figure 6: Ranking of all models across various tasks from GPT-4 (lower is better).

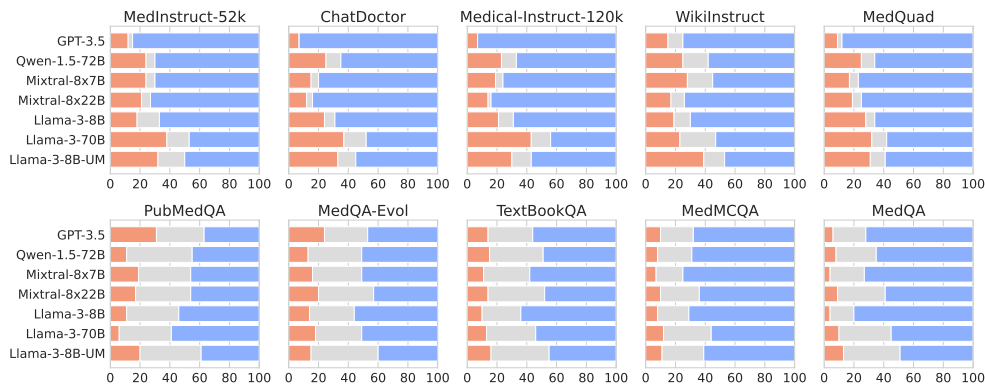


Figure 7: Ranking of models relative to GPT-4 (win/tie/loss) across various tasks, based on feedback from GPT-4. Green, gray, and red represent win, tie, and loss, respectively.

4 The Medical RewardBench

4.1 Construction

As we introduce in the main content, we firstly prompt GPT-4 for preference annotation of about 100K instructions with about 800K completions, which produce about 1.8 million (instruction, chosen, rejected) pairs. To control different level of gap between chosen and rejected completions, we further sample pairs based on easy, hard and length criteria from the 1.8 million pairs, resulting into about 100K pairs. We chose a subset pairs about 1,000 from the final 100K pairs as test set, where includes 100 for each sub-task.

4.2 Human Annotation

We requested a human expert to review and re-annotate 1,000 preference pairs. The web interface used for annotation is displayed in Figure 8.

For biomedical-related questions (including clinical, exam, and open research questions) along with responses from Model A and Model B, the task is to choose the best response (vote) based on the following criteria:

- “Honest”: The response is more accurate and verifiable with no factual errors.

- “Helpful”: The response is more useful, addresses the problem effectively, and does not contain platitudes.
- “Harmless”: The response is safe and ethically sound (generally always the case).
- “Length Bias”: A longer response is not necessarily better; avoid verbosity, prioritizing the above 3H.

Annotation Process:

- Enter the name of the annotator for later processing and filtering of invalid annotations.
- Review the question and the answers from the two models (for reference, see GPT-4’s explanation), and vote for the best response.
- After clicking to vote, the question will automatically refresh. To simplify the process, returning to a previous question is not supported!
- If a suitable response cannot be selected due to a lack of knowledge, errors in the question or answers, etc., click skip.
- Also, leave a comment in the comments section with the corresponding question number.

Welcome to the Voting System

Your Name
test

Start Voting

Question (MedQuad,8305):
What are the genetic changes related to fragile X-associated tremor/ataxia syndrome ?

Reference:
None

Answer from model A:
Fragile X-associated tremor/ataxia syndrome (FXTAS) is a genetic condition associated with expansions in the FMR1 (Fragile X Mental Retardation 1) gene located on the X chromosome. The genetic change specifically related to FXTAS involves a premutation in the number of CGG trinucleotide repeats in the FMR1 gene.

In the general population, the FMR1 gene typically contains between 5 to 44 CGG repeats. Individuals with FXTAS usually have a premutation range of approximately 55 to 200 CGG repeats. This is distinct from individuals with Fragile X syndrome (FXS), who have a full mutation with over 200 CGG repeats, which leads to hypermethylation of the gene and a reduction or absence of the gene's product, the FMRP protein.

The premutation associated with FXTAS does not cause the gene to be methylated or silenced. Instead, it leads to a different mechanism of disease. The increased number of CGG repeats in the premutation range results in elevated levels of FMR1 mRNA, which is thought to have a toxic gain-of-function effect on the cell. The exact mechanisms are not fully understood, but the elevated mRNA may lead to cellular stress and the formation of inclusions in neurons and astrocytes, contributing to the neurodegenerative process observed in FXTAS.

FXTAS typically affects older adults, with symptoms often starting after the age of 50. The syndrome is characterized by intention tremor, cerebellar ataxia, parkinsonism, cognitive decline, and general brain atrophy. It is more common in males than females, likely due to the presence of only one X chromosome in males, which means the premutation has a higher chance of manifesting symptoms.

It's important to note that not all individuals with the FMR1 premutation will develop FXTAS; the penetrance is incomplete, and other genetic or environmental factors may influence the development and severity of the condition. Genetic counseling is recommended for individuals with the FMR1 premutation to understand their risks and the potential implications for their family members.

GPT-4 Evaluation:
The response gives a clear explanation of the genetic changes in FXTAS, focusing on the premutation and its consequences. It discusses the role of elevated FMR1 mRNA and its potential toxic effects comprehensively. The response is well-structured and scientifically sound, providing a good level of detail.

GPT-4 Score	GPT-4 Rank
5	2

Vote A

Answer from model B:
Fragile X-associated tremor/ataxia syndrome (FXTAS) is a genetic disorder caused by a premutation in the FMR1 gene, which is located on the X chromosome. The FMR1 gene normally produces a protein called fragile X mental retardation protein (FMRP), which is important for neural development and function.

In individuals with FXTAS, the genetic change typically involves a CGG trinucleotide repeat expansion in the 5' untranslated region (5' UTR) of the FMR1 gene. While individuals with Fragile X syndrome (caused by a full mutation of the same gene) usually have more than 200 CGG repeats, those with FXTAS generally have an intermediate number of repeats, often between 55 to 200, which is referred to as a premutation.

The premutation can lead to the production of an abnormal RNA transcript that may sequester RNA-binding proteins, including neuronal nucleolin, which in turn affects the translation of specific mRNAs and disrupts neuronal function. The exact mechanisms by which the premutation causes the various symptoms of FXTAS are still being investigated, but it is believed that the expansion leads to cellular stress and damage, particularly in the brain, resulting in the clinical features of the syndrome.

FXTAS is associated with two types of symptoms: primary symptoms and secondary symptoms. Primary symptoms include intention tremor (a tremor that occurs during voluntary movements), action myoclonus (sudden, involuntary muscle jerks), parkinsonism (movement difficulties similar to Parkinson's disease), cognitive decline, and problems with memory and executive function. Secondary symptoms may include anxiety, depression, and sleep disturbances.

FXTAS typically affects older adults, with onset often in the sixth decade of life or later. It can be sporadic or familial, and there is a gender bias, with men being more severely affected than women, possibly due to the presence of only one X chromosome in males, which means that the premutation has a higher likelihood of being expressed.

It's important to note that the FMR1 gene premutation does not only cause FXTAS but is also associated with Fragile X-associated primary ovarian insufficiency (FXPOI) in females, and there may be a risk for certain individuals to develop Fragile X syndrome if the premutation expands to a full mutation in their offspring.

If you or someone you know is suspected of having FXTAS, genetic counseling and testing are recommended to confirm the diagnosis and discuss the implications for family members, as the premutation can be inherited from an affected parent.

GPT-4 Evaluation:
The response effectively outlines the genetic basis of FXTAS and includes detailed descriptions of the premutation effects and their biological consequences. However, it could improve by offering more direct information on how these genetic changes lead to the specific symptoms of FXTAS.

GPT-4 Score	GPT-4 Rank
4	6

Vote B

Skip

Figure 8: The WebUI used by human experts to annotate and correct preferences.

4.3 Data Statistics

For the initially given 1,000 test pairs, we ultimately retained 777 pairs following human expert annotation. These include 238 easy, 196 hard, 180 length-based, and 163 human-judged pairs. Approximately 233 pairs were filtered out due to issues such as incorrect formulations, difficulty in

answering, or both. The human category comprises pairs where preferences differ between human annotators and GPT-4, which is regraded as even hard for GPT-4 to recognize.

5 Prompts

QA Prompts: We use the prompts and question-answer format referenced in MedPrompt [5], where we regard all the questions in PubMedQA [1] as multiple-choice problems, including options for “yes,” “no,” or “maybe.” We use the same zero-shot CoT prompts for all models.

Prompts for Instruction Evaluation: We utilize GPT-3.5 to assess the complexity of each question across all datasets. Each instruction is rated on a scale from 1 to 10, where 1 represents the easiest and 10 denotes the most challenging, as described in [3].

Prompts for Preference Annotation: We employ GPT-4 to annotate preferences for all completions of instructions. For each annotation, we provide GPT-4 with all candidate model responses alongside the question and potential answers (primarily for multiple-choice QA), and then instruct GPT-4 to score each response on a scale from 1 to 5, where 1 is the worst and 5 is the best, based on a 5-level requirement system. Finally, GPT-4 ranks all models according to these scores. Our approach mainly references [2] to define the 5-level requirements from a biomedicine perspective.

Prompts for Instruction Evaluation: We conduct instruction evaluation on MedQA problems using GPT-4. The goal of this evaluation is to enhance the complexity of the questions using four base methods, as utilized in EvolInstruct [8, 4].

Prompts for TextBook Question Generation: We present three examples and a paragraph from a collection of 18 widely used medical textbooks, which serve as crucial references for students preparing for the United States Medical Licensing Examination (USMLE). These textbooks can be accessed at [MedRAG/textbooks](#)³.

Prompts for Wikipedia Instruction Generation: The process begins by crawling all topics from the BioMedicine page on Wikipedia, followed by prompting GPT-4 to generate sub-topics within this field. Subsequently, we instruct GPT-4 to create open-domain instructions for various applications, based on these sub-topics and a background introduction, akin to the approach in Self-Instruct [7].

We provide all above prompts in Table 3.

Table 3: This table displays the prompts used in our experiments.

<u>ZERO-SHOT PROMPTS FOR QA</u>
Question { { question } }
Task Answer the above question with format ‘So, the answer is’ after your explanation. For example, if the answer is A, write ‘So, the answer is A’.
Answer Let’s think step by step.
<u>PROMPTS FOR INSTRUCTIONS EVALUATION BY GPT-3.5</u>
Please evaluate the following question and rate its difficulty and complexity on a scale from 1 to 10, with 1 being the least difficult/complex and 10 being the most difficult/complex. Consider factors such as the breadth and depth of knowledge required, the number of concepts involved, the level of technical terminology, and the presence of quantitative or analytical components.
In addition to the numerical score, provide a brief justification (1-2 sentences) explaining your rationale for the assigned score. This will help us better understand the reasoning behind your evaluation.
Question {question}

³<https://huggingface.co/datasets/MedRAG/textbooks>

Evaluation
Justification:
Score: [1-10]

PROMPTS FOR PREFERENCE ANNOTATION BY GPT-4

Please evaluate the following user instruction and the proposed response within the context of biomedicine.

Evaluation Criteria

Use the following 5-point scale to assess how well the AI Assistant's response addresses the biomedical inquiry:

1: Inadequate - The response is incomplete, vague, off-topic, or controversial. It may lack necessary biomedical data, use incorrect terminology, or include irrelevant clinical examples. The perspective may be inappropriate, such as personal experiences from non-scientific blogs or resembling a forum answer, which is unsuitable given the precision required in biomedicine.

2: Partially Adequate - The response addresses most biomedical aspects requested but lacks direct engagement with the core scientific question. It might provide a general overview instead of detailed biomedical mechanisms or specific clinical applications.

3: Acceptable - The response is helpful, covering all basic biomedical queries. However, it may not adopt an AI Assistant's typical scientific voice, resembling content from general health blogs or web pages and could include personal opinions or generic information.

4: Good - The response is clearly from an AI Assistant, accurately focusing on the biomedical instruction. It is complete, clear, and comprehensive, presented in a clinically appropriate tone. Minor improvements could include adding more precise scientific details or a more formal presentation.

5: Excellent - The response perfectly represents an AI Assistant in biomedicine, addressing the user's scientific inquiry without any irrelevant content. It demonstrates in-depth knowledge, is scientifically accurate, logically structured, engaging, insightful, and impeccably written.

Question and Reference Answer

Question: {question}

Reference Answer: {answer}

Model Responses

{candidates}

Feedback and Rankings

Provide feedback and an overall score between 1 to 5 for each response based on the **Evaluation Criteria**. Then rank the model responses, even if they share the same score, based on criteria such as clarity of response logic, richness of information, and naturalness of language.

Format your feedback and rankings as follows:

```
---  
{  
  "feedback": {  
    "Model 1": {  
      "Evaluation": "",  
      "Score": ""  
    },  
    // Similar entries for other models  
  },  
}
```

```
"ranking": [  
  {"rank": 1, "model": "Model X"}],  
  // Subsequent rankings  
]  
}}  
...
```

PROMPTS FOR INSTRUCTIONS EVOLUTION BY GPT-4

Act as a Question Rewriter to make biomedical multiple-choice questions more challenging for AI systems like ChatGPT and GPT-4, while remaining reasonable for human experts to understand and answer.

Complicate the given question using one of these methods:

```
[METHOD 1] Add one more constraint or requirement.  
[METHOD 2] Replace general concepts with more specific ones.  
[METHOD 3] Make the choices hard to differentiate by adding more complex distractors.  
[METHOD 4] If solvable with simple thinking, request multi-step reasoning.
```

Limit additions to 10-20 words. Ensure a unique answer exists among the choices.

Question:
{question}

Output JSON format:

```
... {{  
  "question": "Rewritten question in the format: "xxx\nA. xxx\nB. xxx\nC. xxx\nD. xxx",  
  "answer": "A/B/C/D"  
}}  
...
```

PROMPTS FOR TEXTBOOK QUESTION GENERATION BY GPT-4

Paragraph from the medical textbook
{paragraph}

Example multi-choice questions

Example 1

Question: {example1}

Answer: {answer1}

Example 2

Question: {example2}

Answer: {answer2}

Example 3

Question: {example3}

Answer: {answer3}

Instructions

1. Evaluate the examination significance of the provided paragraph.
2. Assess whether the paragraph contains sufficient knowledge to evaluate a powerful AI like GPT-4. Consider factors such as:

- Depth and breadth of the medical concepts covered
- Specificity and technicality of the information provided
- Potential for testing higher-order thinking skills

3. If the paragraph is deemed significant and contains enough knowledge to evaluate GPT-4, generate a synthetic multi-choice question based on the paragraph's content and the provided examples. Ensure that the generated question has a single, unambiguous correct answer among the provided choices.

4. If the paragraph is not significant or lacks sufficient knowledge for AI evaluation, set the value of "generated_question" to an empty object ({}).
5. Provide the output in the specified JSON format.

Output Format (JSON)

```

{
  "examination_significance": boolean,
  "sufficient_knowledge_for_ai_evaluation": boolean,
  "generated_question": {
    "question": string,
    "answer_choices": [
      {
        "choice": string,
        "correct": boolean
      },
      {
        "choice": string,
        "correct": boolean
      },
      {
        "choice": string,
        "correct": boolean
      },
      {
        "choice": string,
        "correct": boolean
      }
    ]
  }
}

```

PROMPTS FOR WIKIPEDIA SUB-TOPICS GENERATION BY GPT-4

{entity}: {description} As an expert in the field of {entity}, I need you to do the following:

1. List {number} subfields within the realm of {entity} research.
2. Ensure that these subfields represent distinct areas of entity without any overlap.
3. Provide a brief description for each subfield, highlighting its main research focus and characteristics.
4. Aim for this list to comprehensively reflect the diversity and breadth of the biomedical field.
5. Present this list in an array of dictionaries format, with each dictionary containing two keys: 'name' (the name of the subfield) and 'description' (a brief description of the subfield).

Example output format:

```

[
  {
    "name": "Gene Editing",
    "description": "Gene editing involves altering the genetic material of organisms to study gene functions or treat genetic diseases."
  },
  {
    "name": "Neuroscience",
    "description": "Neuroscience focuses on the study of the structure, function, and diseases of the nervous system."
  },
  // ... 18 more subfields
]

```

PROMPTS FOR WIKIPEDIA INSTRUCTIONS GENERATION BY GPT-4

{topic}: {description} As an expert in the field of {topic}, please devise {number} {topic}-related questions or instructions, formatted as an array of dictionaries, each with two keys: 'instruction' and 'context'. Follow these guidelines:

1. **Verb Diversity**: Incorporate a broad spectrum of verbs to diversify and enrich the instructions set.
2. **Language Style Variability**: Blend both interrogative and imperative sentence structures to enhance the dynamism of instructions.

3. **Range of Task Types**: Ensure the tasks span a variety of categories such as explanations, analyses, comparisons, and more. 1. **Difficulty levels** should vary from elementary concepts to complex scientific inquiries and extend to addressing novel, challenging scenarios.
4. **Exclusivity to Text-Based Tasks**: Frame all instructions in a text-only format. Refrain from incorporating tasks that require physical execution or laboratory experimentation.
5. **Conciseness and Precision**: Articulate each instruction in English with utmost precision, limiting it to 1 or 2 sentences for clarity and brevity.
6. **Background Information Accuracy**: For tasks necessitating supplementary context, provide succinct yet comprehensive descriptions (restricted to 100 words). For basic queries, simply state "None" in the context section.
7. **JSON Format Adherence**: Format the output as an array of dictionaries. Each dictionary should have two keys: 'instruction' for the task description and 'context' for the relevant background information.

Example output format:

```


{
  "instruction": "Explain the structure of liposomes and their role in drug delivery.",
  "context": "Liposomes are nanoscale carriers used in drug delivery, where their structure and function significantly impact efficiency."
},
{
  "instruction": "List three common cardiovascular diseases.",
  "context": "None"
},
// ... 18 more instructions


```

6 Examples

We provide examples for each sub-task within UltraMedical in Table 4, including the dataset name, question, answer from GPT-4 and the original dataset, as well as the score on the instruction by ChatGPT. For the WikiInstruct task, we construct the instructions from scratch, resulting in no answer provided.

Table 4: Examples of datasets in UltraMedical.

Dataset	Question	GPT4-Answer	Answer	Score
MedQA	<p>"A 23-year-old pregnant woman at 22 weeks gestation presents with burning upon urination. She states it started 1 day ago and has been worsening despite drinking more water and taking cranberry extract. She otherwise feels well and is followed by a doctor for her pregnancy. Her temperature is 97.7F (36.5C), blood pressure is 122/77 mmHg, pulse is 80/min, respirations are 19/min, and oxygen saturation is 98% on room air. Physical exam is notable for an absence of costovertebral angle tenderness and a gravid uterus. Which of the following is the best treatment for this patient?</p> <p>A. Ampicillin B. Ceftriaxone C. Doxycycline D. Nitrofurantoin"</p>	<p>"The clinical scenario suggests that the pregnant woman is likely experiencing a urinary tract infection (UTI), given her symptoms of burning upon urination. During pregnancy, the risk of UTIs increases due to physiological changes, and it is important to treat them to prevent complications such as pyelonephritis or preterm labor.</p> <p>When considering antibiotic therapy for a UTI in a pregnant woman, it's crucial to choose an agent that is safe for both the mother and the fetus. Let's evaluate the options:</p> <p>A. Ampicillin - This is a penicillin antibiotic that is generally considered safe during pregnancy and can be used to treat UTIs. However, it may not be the first choice due to resistance patterns.</p> <p>B. Ceftriaxone - This is a cephalosporin antibiotic that is also generally safe in pregnancy and effective against a broad range of bacteria. It is typically used for more complicated infections or when other antibiotics are not suitable.</p> <p>C. Doxycycline - This is a tetracycline antibiotic that is contraindicated during pregnancy because it can affect fetal bone growth and discoloration of teeth.</p> <p>D. Nitrofurantoin - This antibiotic is commonly used to treat uncomplicated UTIs and is considered safe during most of the pregnancy. However, it is not recommended at the very end of pregnancy (after 38 weeks) due to the risk of hemolytic anemia in the newborn.</p> <p>Given that the patient is at 22 weeks gestation and has an uncomplicated UTI, the best choice would be an antibiotic that is safe and commonly used for UTIs in pregnancy.</p> <p>So, the answer is D. Nitrofurantoin."</p>	D	<p>This question would likely be rated around a 7 or 8. It requires a strong understanding of obstetrics and gynecology, as well as knowledge of common urinary tract infections and their treatment. Additionally, the question involves the consideration of pregnancy-specific treatment options and the potential risks to the fetus.</p>

Dataset	Question	GPT4-Answer	Answer	Score
MedMCQA	"Chronic urethral obstruction due to benign prismatic hyperplasia can lead to the following change in kidney parenchyma A. Hyperplasia B. Hyperophy C. Atrophy D. Dyplasia"	"Chronic urethral obstruction, often caused by benign prostatic hyperplasia (BPH), can lead to a series of complications in the urinary tract, especially affecting the kidneys. The obstruction hinders the flow of urine, leading to urinary retention and increased pressure back up into the kidneys. This increased pressure can impair kidney function and eventually lead to changes in the kidney tissue. The correct term here is ""atrophy."" When there is chronic obstruction, such as from BPH, the kidneys can suffer from prolonged pressure, which leads to the deterioration of kidney tissue over time. This process, where the kidney tissue becomes thinner and weaker, is known as atrophy. Unlike hyperplasia, which is an increase in the number of cells, atrophy refers to the loss or shrinkage of tissue. Hyperophy is not a recognized medical term related to kidney changes. Dysplasia, on the other hand, refers to the presence of abnormal cells within a tissue or organ but does not directly result from obstructive processes like BPH. So, the answer is C."	C	"Score: 7 Justification: This question requires a deep understanding of medical terminology, anatomy, and the pathophysiology of urinary tract obstructions. It also involves knowledge of the specific changes that occur in kidney parenchyma due to chronic urethral obstruction."
MedQA-Evol	"In a 23-year-old pregnant female at 22 weeks gestation with a confirmed urinary tract infection experiencing dysuria, which condition could potentially develop as a direct sequela of this infection if left untreated, and is particularly associated with the immunological adaptations during pregnancy? A. Gestational diabetes B. Preterm labor C. Intrauterine growth restriction (IUGR) D. Hyperemesis gravidarum"	"In pregnancy, a woman's immune system undergoes various adaptations, making her more susceptible to certain infections, including urinary tract infections (UTIs). If a UTI during pregnancy is left untreated, it can ascend to the kidneys, leading to a more severe infection known as pyelonephritis. This condition can further complicate the pregnancy. A significant risk associated with untreated UTIs in pregnancy is preterm labor. The inflammation and infection can stimulate the production of prostaglandins, which are chemicals that can induce labor. This can result in preterm birth, where the baby is born before 37 weeks of gestation. Preterm labor is a well-recognized complication of untreated UTIs during pregnancy and is distinguished from the other options because: - Gestational diabetes is related to hormonal changes affecting insulin, not directly linked to urinary tract infections. - Intrauterine growth restriction (IUGR) refers to poor growth of a fetus while in the mother's womb during pregnancy and is not typically a direct consequence of UTIs. - Hyperemesis gravidarum, characterized by severe nausea and vomiting, is not directly caused by UTIs. So, the answer is B."	B	"Score: 7 Justification: This question requires a deep understanding of both urinary tract infections and the immunological adaptations during pregnancy. It also involves understanding potential complications of untreated infections during pregnancy. The question is complex and requires a high level of medical knowledge."

Dataset	Question	GPT4-Answer	Answer	Score
TextBookQA	<p>"In a comprehensive regional anatomical analysis of the thorax emphasizing the respiratory system's components, which structure would NOT be expected to be scrutinized, considering the respiratory focus?</p> <p>A. External intercostal muscles B. Costovertebral joints C. Visceral pleura D. Coronary vessels"</p>	<p>"The thorax, or chest region, contains several structures, including those related to the respiratory system. When focusing on the respiratory components, one would typically examine structures directly involved in the process of breathing and gas exchange.</p> <p>A. External intercostal muscles - These muscles are significant in the respiratory system because they are involved in the mechanics of breathing, particularly in the elevation of the ribs during inspiration.</p> <p>B. Costovertebral joints - These joints are where the ribs articulate with the vertebrae. They are relevant to the respiratory system because they allow the movement of the ribs during respiration.</p> <p>C. Visceral pleura - This is the membrane that directly covers the lungs. It is definitely a part of the respiratory system, playing a role in the lubrication and protection of the lungs.</p> <p>D. Coronary vessels - These are the blood vessels of the heart (coronary arteries and veins) that provide the myocardium (heart muscle) with oxygen and nutrients. They are not directly involved in the respiratory system but are part of the circulatory system.</p> <p>Given that the coronary vessels are not components of the respiratory system but rather of the circulatory system, they would not be expected to be scrutinized in a comprehensive regional anatomical analysis of the thorax emphasizing the respiratory system's components.</p> <p>So, the answer is D."</p>	D	<p>"Score: 7 Justification: This question requires a deep understanding of regional anatomy of the thorax, specifically focusing on the respiratory system. It involves knowledge of anatomical structures, their relationships, and their functions. The question also requires the ability to identify which structure is not directly related to the respiratory system, adding a level of complexity."</p>

Dataset	Question	GPT4-Answer	Answer	Score
PubMedQA	<p>"Context: Phosphatidylethanolamine N-methyltransferase (PEMT), a liver enriched enzyme, is responsible for approximately one third of hepatic phosphatidylcholine biosynthesis. When fed a high-fat diet (HFD), <i>Pemt(-/-)</i> mice are protected from HF-induced obesity; however, they develop steatohepatitis. The vagus nerve relays signals between liver and brain that regulate peripheral adiposity and pancreas function. Here we explore a possible role of the hepatic branch of the vagus nerve in the development of diet induced obesity and steatohepatitis in <i>Pemt(-/-)</i> mice. Context: 8-week old <i>Pemt(-/-)</i> and <i>Pemt(+/+)</i> mice were subjected to hepatic vagotomy (HV) or capsaicin treatment, which selectively disrupts afferent nerves, and were compared to sham-operated or vehicle-treatment, respectively. After surgery, mice were fed a HFD for 10 weeks. Context: HV abolished the protection against the HFD-induced obesity and glucose intolerance in <i>Pemt(-/-)</i> mice. HV normalized phospholipid content and prevented steatohepatitis in <i>Pemt(-/-)</i> mice. Moreover, HV increased the hepatic anti-inflammatory cytokine interleukin-10, reduced chemokine monocyte chemoattractant protein-1 and the ER stress marker C/EBP homologous protein. Furthermore, HV normalized the expression of mitochondrial electron transport chain proteins and of proteins involved in fatty acid synthesis, acetyl-CoA carboxylase and fatty acid synthase in <i>Pemt(-/-)</i> mice. However, disruption of the hepatic afferent vagus nerve by capsaicin failed to reverse either the protection against the HFD-induced obesity or the development of HF-induced steatohepatitis in <i>Pemt(-/-)</i> mice. Does vagus nerve contribute to the development of steatohepatitis and obesity in phosphatidylethanolamine N-methyltransferase deficient mice? A. maybe B. yes C. no"</p>	<p>"The evidence from the context given indicates that the hepatic branch of the vagus nerve does play a role in the development of diet-induced obesity and steatohepatitis in <i>Pemt(-/-)</i> mice. This is supported by the observation that hepatic vagotomy (HV) abolished the protection against HFD-induced obesity and glucose intolerance in <i>Pemt(-/-)</i> mice. HV also normalized phospholipid content and prevented steatohepatitis, indicating a clear connection between the hepatic vagus nerve and the development of these conditions. However, the disruption of the hepatic afferent vagus nerve by capsaicin treatment did not reverse the protection against HFD-induced obesity or the development of steatohepatitis. This suggests that while the hepatic vagus nerve has a role, it may specifically involve the efferent fibers, which are responsible for sending signals from the brain to the liver, rather than the afferent fibers, which carry signals from the liver to the brain. Based on the provided information, the hepatic branch of the vagus nerve contributes to the development of steatohepatitis and obesity in <i>Pemt(-/-)</i> mice, as evidenced by the effects of HV. However, the specific pathways and mechanisms, whether efferent or afferent, require further clarification. So, the answer is B."</p>	B	<p>This question would be rated as a 9. It requires a deep understanding of hepatic physiology, neurology, and molecular biology, as well as familiarity with the specific roles of phosphatidylethanolamine N-methyltransferase and the vagus nerve in the development of obesity and steatohepatitis. Additionally, the question involves complex experimental design and interpretation of results.</p>

Dataset	Question	GPT4-Answer	Answer	Score
MedQuad	How to diagnose Lymphocytic Choriomeningitis (LCM) ?	<p>"Diagnosing Lymphocytic Choriomeningitis (LCM) typically involves a combination of clinical assessment, laboratory testing, and, in some cases, imaging studies. Here,Äôs a detailed approach:</p> <ol style="list-style-type: none"> Clinical Assessment: The initial step is a thorough medical history and physical examination. LCM should be suspected in individuals presenting with symptoms like fever, headache, muscle aches, and meningitis or encephalitis signs, especially if there's a history of exposure to rodents or their droppings, as the LCM virus is primarily transmitted through rodent contact. Laboratory Testing: - Serologic Testing: This is the primary method for diagnosing LCM. Blood tests can detect antibodies against the LCM virus. IgM antibodies suggest recent infection, while IgG antibodies indicate past exposure. - Polymerase Chain Reaction (PCR): This test can detect LCM virus RNA in blood or cerebrospinal fluid (CSF) and is useful in the early stages of the disease. - CSF Analysis: Examination of cerebrospinal fluid obtained via lumbar puncture can show a lymphocytic pleocytosis (increased lymphocytes), which is indicative of meningitis. While not specific to LCM, it can support the diagnosis in the appropriate clinical context. Imaging Studies: While not diagnostic for LCM specifically, imaging studies like MRI or CT scans of the brain may be performed to rule out other causes of neurological symptoms. Differential Diagnosis: It,Äôs important to consider and rule out other causes of meningitis or encephalitis, such as other viral, bacterial, or fungal infections, and non-infectious causes like autoimmune diseases. Epidemiological Links: In cases where LCM is suspected, identifying a potential source of infection (like contact with rodents) can support the diagnosis. <p>Remember, the diagnosis of LCM should be made based on a combination of clinical findings, laboratory results, imaging studies (when necessary), and epidemiological links. It,Äôs also important to consult with healthcare professionals for accurate diagnosis and management."</p>	<p>"During the first phase of the disease, the most common laboratory abnormalities are a low white blood cell count (leukopenia) and a low platelet count (thrombocytopenia). Liver enzymes in the serum may also be mildly elevated. After the onset of neurological disease during the second phase, an increase in protein levels, an increase in the number of white blood cells or a decrease in the glucose levels in the cerebrospinal fluid (CSF) is usually found. Laboratory diagnosis is usually made by detecting IgM and IgG antibodies in the CSF and serum. Virus can be detected by PCR or virus isolation in the CSF at during the acute stage of illness."</p>	<p>This question would likely be rated a 7. It requires a deep understanding of medical terminology and diagnostic procedures, as well as knowledge of the specific symptoms and characteristics of LCM. Additionally, it may involve knowledge of laboratory tests and diagnostic imaging techniques.</p>

Dataset	Question	GPT4-Answer	Answer	Score
MedInstruct-52k	<p>"Rewrite the following passage into simpler language to help pre-med students understand.</p> <p>Hypertrophic cardiomyopathy is a condition where the myocardium becomes hypertrophied without an obvious cause, resulting in impaired ventricular filling and sometimes outflow obstruction. The most common presenting symptom is dyspnea on exertion"</p>	<p>Hypertrophic cardiomyopathy is when the heart muscle gets abnormally thick for no clear reason. This thickening can make it hard for the heart to fill with blood properly, and sometimes it can also block blood from flowing out of the heart easily. The symptom people usually notice first is getting short of breath when they do physical activities.</p>	<p>"Hypertrophic cardiomyopathy is a condition where the muscle in the heart becomes too thick without a clear reason, leading to problems with the filling of the heart's chambers and sometimes blocking the flow of blood. The main symptom that most people experience is shortness of breath when they do physical activities.</p> <p>The Answer is: ""Hypertrophic cardiomyopathy means the heart muscle getting larger for no clear reason, causing trouble with filling the heart and sometimes blocking blood flow. The main symptom is feeling out of breath during exercise.""</p>	<p>"I would rate this question a 7. Justification: This question requires a deep understanding of medical terminology and concepts related to cardiology. It also requires the ability to simplify complex medical information for a specific audience, adding to its difficulty and complexity."</p>

Dataset	Question	GPT4-Answer	Answer	Score
Medical-Instruction-120k	Hi.My names Ahdieh.I m from a small city in Iran.My father had a heart attack on sunday as doctor said cpr condition.after 40 minuts he came back to life. now he isnot conscious.he is in ICU.his doctor said his conscious rate is 5.what do you think about his condition?If it is needed I can send you his cardiograf and blood test and the drugs they are using for him in ICU.Kindly tell me can we carry him to a better hospital in Tehran(It takes 4 hours to reach there)?	<p>"Hello Ahdieh, I'm deeply sorry to hear about your father's situation. It must be a very stressful and difficult time for you and your family.</p> <p>Regarding your father's condition, a Glasgow Coma Scale (GCS) score of 5 indicates severe impairment of consciousness. This scale is commonly used to assess the level of consciousness in individuals who have suffered brain injury or are in a coma. The scale ranges from 3 to 15, with lower scores indicating more severe impairment. A score of 5 suggests that your father is in a critical condition and needs intensive medical care.</p> <p>I am not a medical doctor, but I can provide some general information. The decision to transfer a patient from one hospital to another, especially one who is critically ill, involves weighing the risks and benefits. Transporting a critically ill patient over a long distance can be risky, and the medical team must be prepared to manage any complications that might arise during the transfer. The potential benefits of being treated at a facility with more advanced resources must be carefully considered against these risks.</p> <p>Here are some factors to consider:</p> <ol style="list-style-type: none"> **Current Level of Care**: Is the current hospital equipped to provide the necessary level of care for your father's condition? If not, a transfer might be justified. **Stability for Transfer**: Is your father stable enough for the transfer? Patients must often be stable before they can be transported, especially over long distances. **Medical Transport Services**: If a transfer is necessary, it should be done with appropriate medical transport services equipped to handle critical patients, including the presence of a medical team that can provide care en route. **Receiving Hospital**: Does the hospital in Tehran have the capability to provide better or more specialized care for your father's condition? Have you confirmed that they can accept your father and that they are prepared to provide the necessary treatment? **Consultation with Medical Professionals**: The most important step is to consult with the doctors and medical team currently treating your father. They can provide specific advice based on his condition and the resources available at both the current hospital and the potential receiving hospital in Tehran. <p>Regarding the cardiogram, blood tests, and medication list, I am not a physician and cannot interpret these medical documents or provide medical advice. It's essential that you discuss these details with your father's medical team. They can provide you with the most accurate information and guidance.</p> <p>Remember to take care of yourself during this challenging time, and don't hesitate to reach out to family, friends, or support groups for emotional support. Your well-being is also important.</p> <p>I hope your father's condition improves, and I wish you strength and courage as you navigate this difficult situation."</p>	" Hello Addie! Welcome and thank you for asking on Chat Doctor! I passed carefully through your question and would explain that your father clinical situation could be related to prolonged brain hypoxia, due to the cardiac arrest. The time (40 minutes) needed for his resuscitation is long and may have caused irreversible damage to his brain, due to low blood flow (also called hypoxic encephalopathy). For this reason, I recommend consulting with a neurologist and performing a brain MRI and EEG to examine his brain function. Hope you will find this answer helpful! Kind regards, "	This question would be rated a 9 in difficulty and complexity. It requires a deep understanding of medical terminology, knowledge of CPR and its effects, interpretation of medical tests, and the ability to make a critical decision about transferring a patient to another hospital. Additionally, it involves emotional and ethical considerations.

Dataset	Question	GPT4-Answer	Answer	Score
ChatDoctor	gyno problemsfor the past few months, I have been having issues with my vagina. there always seems to be something wrong with me. its either an infection or a yeast infection from the medication used to treat the previous infection or a herpes outbreak as a result of a yeast infection. most recently, I had a uti. I was treated for that and everything seemed fine, until after I finished the medication. it still hurt when I had sex and still is uncomfortable to pee. I dont know whats going on and this has been going on for months.	<p>"I'm sorry to hear that you've been experiencing these issues. It's important to remember that I'm an AI and not a medical professional, but I can offer some general advice that might help guide you towards getting the appropriate care.</p> <p>Recurrent infections and discomfort can be caused by a variety of factors, and it's crucial to address these with a healthcare provider. Here are some steps you might consider:</p> <ol style="list-style-type: none"> Follow-Up with Your Healthcare Provider: Since you're still experiencing discomfort after treatment, it's important to return to your healthcare provider for a follow-up. They might need to conduct further tests to understand why you're having recurrent issues. Get Tested for STIs: If you haven't already, make sure to get a full screening for sexually transmitted infections (STIs), including herpes, which you mentioned. This will help to rule out or confirm any infections that may require specific treatment. Consider Other Causes: Recurrent urinary tract infections (UTIs) and yeast infections can sometimes be caused by factors like sexual activity, personal hygiene products, antibiotics, hormonal changes, or even underlying health conditions. Review Your Medications: If you find that you're getting yeast infections after antibiotic treatments, this could be due to the antibiotics disrupting the natural balance of bacteria and yeast in your vagina. Your doctor might suggest probiotics or other treatments to help maintain this balance. Discuss Sexual Health: Pain during intercourse (dyspareunia) can be caused by infections, but it can also be due to other conditions like endometriosis, vaginal dryness, or pelvic inflammatory disease. It's important to discuss this with your healthcare provider. Lifestyle Adjustments: Consider lifestyle changes that can reduce the risk of infections, such as wearing breathable cotton underwear, avoiding douches and scented hygiene products, urinating before and after sex, and ensuring good hydration. Specialist Referral: If your primary care provider is unable to determine the cause or provide effective treatment, you may need to see a specialist such as a gynecologist or a urologist. Pain Management: If you're experiencing ongoing pain, your healthcare provider can also help you with pain management strategies. Immune System Evaluation: Since recurrent infections can sometimes be a sign of a weakened immune system, it might be worth evaluating your overall health and immune function. <p>It's important to address these issues sooner rather than later, as untreated infections can lead to more serious complications. Be open with your healthcare provider about all of your symptoms and concerns so they can provide the most effective care.</p> <p>Remember, this is general advice and should not replace professional medical consultation. It's essential to work closely with a healthcare provider to get to the root of the problem and find a treatment plan that works for you."</p>	<p>Dear Friend. Welcome to Chat Doctor. I am Chat Doctor. I understand your concern. Recurring yeast / final infection occur due to:1. Improper treatment.2. Inadequate treatment3. Wrong diagnosis.4. Not treating the partner. I would advise you to get a Urine examination done for bacteria and fungus. When the reports are available, get treated. Also, treat your partner. This is my personal opinion for you based on available details. If you have any further query please ask me. Stay Healthy-ChatDoctor, MD</p>	This question would be rated as a 9. It requires a deep understanding of gynecological issues, including infections, medication interactions, and the potential for underlying conditions such as herpes. It also involves analyzing the patient's symptoms and medical history to make an accurate diagnosis.

Dataset	Question	GPT4-Answer	Answer	Score
WikiInstruct	Evaluate the survival curve graph illustrating the differential survival rates of organ transplants when augmented with regulatory T cell therapy compared to a control group lacking this immunomodulatory intervention, including an analysis of the variance in outcomes. Delve into the prognostic significance of these findings for future graft viability, considering underlying immune response mechanisms.	<p>Evaluating a survival curve that illustrates the differential survival rates of organ transplants augmented with regulatory T cell therapy compared to a control group without this intervention requires considering several key aspects:</p> <ol style="list-style-type: none"> Survival Curve Analysis: The survival curve, often plotted as a Kaplan-Meier curve, would show the proportion of grafts surviving over time for both groups. A higher curve for the group receiving regulatory T cell therapy would indicate better survival rates. It's important to look at the separation between the two curves; a significant gap suggests a meaningful impact of the therapy. Statistical Significance: To analyze variance in outcomes, the log-rank test or Cox proportional hazards model might be used to determine if the differences in survival rates are statistically significant. This analysis would help confirm whether the observed differences are likely due to the therapy rather than chance. Prognostic Significance: If the survival curve demonstrates significantly better outcomes with regulatory T cell therapy, this has profound implications for future graft viability. It suggests that modulating the immune response via regulatory T cells can effectively prolong graft survival, potentially reducing the need for long-term immunosuppression, which has its own side effects. Immune Response Mechanisms: The underlying mechanism involves regulatory T cells' role in maintaining immune tolerance. By augmenting with regulatory T cell therapy, the immune system's attack on the transplanted organ can be minimized. This therapeutic approach targets the immune system's fundamental response to foreign antigens, aiming to establish a more tolerant environment for the graft. Future Implications: The success of regulatory T cell therapy in improving graft survival rates could revolutionize transplant medicine by offering a more targeted approach to prevent graft rejection. This could lead to protocols that improve the quality of life for transplant recipients, reduce the incidence of graft-versus-host disease, and possibly decrease the reliance on general immunosuppressants, which have broad and often detrimental effects on the patient's overall immune function. <p>In summary, a survival curve showing improved outcomes with regulatory T cell therapy would highlight the therapy's potential as a game-changer in transplant medicine. It underlines the importance of immune modulation in transplant success and paves the way for research into similar strategies that could further improve graft survival and patient outcomes.</p>		I would rate this question a 9. The question requires a deep understanding of organ transplantation, immunology, and statistical analysis. It involves a high level of technical terminology and requires a comprehensive understanding of immune response mechanisms and their impact on graft viability.

Dataset	Question	GPT4-Answer	Answer	Score
----------------	-----------------	--------------------	---------------	--------------

7 Croissant Document

We have uploaded our dataset on Huggingface, which automatically generates the following cross-section. For details, please refer to <https://huggingface.co/api/datasets/TsinghuaC3I/UltraMedical/croissant>.

```
1 {
2   "@context": {
3     "@language": "en",
4     "@vocab": "https://schema.org/",
5     "citeAs": "cr:citeAs",
6     "column": "cr:column",
7     "conformsTo": "dct:conformsTo",
8     "cr": "http://mlcommons.org/croissant/",
9     "data": {
10      "@id": "cr:data",
11      "@type": "@json"
12    },
13    "dataBiases": "cr:dataBiases",
14    "dataCollection": "cr:dataCollection",
15    "dataType": {
16      "@id": "cr:dataType",
17      "@type": "@vocab"
18    },
19    "dct": "http://purl.org/dc/terms/",
20    "extract": "cr:extract",
21    "field": "cr:field",
22    "fileProperty": "cr:fileProperty",
23    "fileObject": "cr:fileObject",
24    "fileSet": "cr:fileSet",
25    "format": "cr:format",
26    "includes": "cr:includes",
27    "isLiveDataset": "cr:isLiveDataset",
28    "jsonPath": "cr:jsonPath",
29    "key": "cr:key",
30    "md5": "cr:md5",
31    "parentField": "cr:parentField",
32    "path": "cr:path",
33    "personalSensitiveInformation": "cr:personalSensitiveInformation",
34    "recordSet": "cr:recordSet",
35    "references": "cr:references",
36    "regex": "cr:regex",
37    "repeated": "cr:repeated",
38    "replace": "cr:replace",
39    "sc": "https://schema.org/",
40    "separator": "cr:separator",
41    "source": "cr:source",
42    "subField": "cr:subField",
43    "transform": "cr:transform"
44  },
45  "@type": "sc:Dataset",
46  "distribution": [
47    {
48      "@type": "cr:FileObject",
49      "@id": "repo",
50      "name": "repo",
51      "description": "The Hugging Face git repository.",
52      "contentUrl": "https://huggingface.co/datasets/TsinghuaC3I/
53      UltraMedical/tree/refs%2Fconvert%2Fparquet",
54      "encodingFormat": "git+https",
55      "sha256": "https://github.com/mlcommons/croissant/issues/80"
56    },
57    {
58      "@type": "cr:FileSet",
59      "@id": "parquet-files-for-config-default",
```

```

59     "name": "parquet-files-for-config-default",
60     "description": "The underlying Parquet files as converted by
Hugging Face (see: https://huggingface.co/docs/datasets-server/
parquet).",
61     "containedIn": {
62         "@id": "repo"
63     },
64     "encodingFormat": "application/x-parquet",
65     "includes": "default/**/*.parquet"
66 }
67 ],
68 "recordSet": [
69     {
70         "@type": "cr:RecordSet",
71         "@id": "default",
72         "name": "default",
73         "description": "TsinghuaC3I/UltraMedical - 'default' subset\n\
nAdditional information:\n- 1 skipped column: conversations",
74         "field": [
75             {
76                 "@type": "cr:Field",
77                 "@id": "default/id",
78                 "name": "default/id",
79                 "description": "Column 'id' from the Hugging Face parquet
file.",
80                 "dataType": "sc:Text",
81                 "source": {
82                     "fileSet": {
83                         "@id": "parquet-files-for-config-default"
84                     },
85                     "extract": {
86                         "column": "id"
87                     }
88                 }
89             },
90             {
91                 "@type": "cr:Field",
92                 "@id": "default/type",
93                 "name": "default/type",
94                 "description": "Column 'type' from the Hugging Face parquet
file.",
95                 "dataType": "sc:Text",
96                 "source": {
97                     "fileSet": {
98                         "@id": "parquet-files-for-config-default"
99                     },
100                    "extract": {
101                        "column": "type"
102                    }
103                }
104            },
105            {
106                "@type": "cr:Field",
107                "@id": "default/answer",
108                "name": "default/answer",
109                "description": "Column 'answer' from the Hugging Face
parquet file.",
110                "dataType": "sc:Text",
111                "source": {
112                    "fileSet": {
113                        "@id": "parquet-files-for-config-default"
114                    },
115                    "extract": {
116                        "column": "answer"
117                    }

```

```

118     }
119     },
120     {
121         "@type": "cr:Field",
122         "@id": "default/score",
123         "name": "default/score",
124         "description": "Column 'score' from the Hugging Face parquet
file.",
125         "dataType": "sc:Text",
126         "source": {
127             "fileSet": {
128                 "@id": "parquet-files-for-config-default"
129             },
130             "extract": {
131                 "column": "score"
132             }
133         }
134     }
135 ]
136 }
137 ],
138 "conformsTo": "http://mlcommons.org/croissant/1.0",
139 "name": "UltraMedical",
140 "description": "The UltraMedical Collections\n\t\n\nThe UltraMedical
Collections is a large-scale, high-quality dataset of biomedical
instructions, comprising 410,000 synthetic and manually curated
samples.\n\nThe construction of UltraMedical adheres to principles
of diversity and complexity.\n\nDetailed information can be found in
our GitHub repository: UltraMedical.\n\nStatistics of datasets in
the UltraMedical collections is shown in following table, where...
See the full description on the dataset page: https://huggingface
.co/datasets/TsinghuaC3I/UltraMedical.",
141 "alternateName": [
142     "TsinghuaC3I/UltraMedical",
143     "UltraMedical"
144 ],
145 "creator": {
146     "@type": "Organization",
147     "name": "TsinghuaC3I",
148     "url": "https://huggingface.co/TsinghuaC3I"
149 },
150 "keywords": [
151     "100K\u003Cn\u003C1M",
152     "English",
153     "mit",
154     "Croissant",
155     "Region: US"
156 ],
157 "license": "https://choosealicense.com/licenses/mit/",
158 "url": "https://huggingface.co/datasets/TsinghuaC3I/UltraMedical"
159 }

```

References

- [1] 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), pages 2567–2577, 2019.
- [2] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation, 2024.
- [3] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning, 2024.

- [4] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. Wizardcoder: Empowering code large language models with evol-instruct, 2023.
- [5] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. [arXiv preprint arXiv:2311.16452](#), 2023.
- [6] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. [arXiv preprint arXiv:2401.00368](#), 2023.
- [7] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022.
- [8] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. [arXiv preprint arXiv:2304.12244](#), 2023.