
Tutorial

How to Design, Create, and Evaluate an Instruction-Tuning Dataset for Large Language Model Training in Health Care: Tutorial From a Clinical Perspective

Wojciech Nazar¹, MD, PhD; Grzegorz Nazar²; Aleksandra Kamińska²; Ludmila Danilowicz-Szymanowicz³, MD, PhD

¹Department of Allergology, Faculty of Medicine, Gdańsk Medical University, Gdansk, Poland

²Faculty of Medicine, Gdańsk Medical University, Gdansk, Poland

³Department of Cardiology and Electrotherapy, Faculty of Medicine, Gdańsk Medical University, Gdansk, Poland

Corresponding Author:

Wojciech Nazar, MD, PhD

Department of Allergology

Faculty of Medicine

Gdańsk Medical University

Smoluchowskiego 17

Gdansk, 80-214

Poland

Phone: 48 585844300

Email: wojciech.nazar@gumed.edu.pl

Abstract

High-quality data are critical in health care, forming the cornerstone for accurate diagnoses, effective treatment plans, and reliable conclusions. Similarly, high-quality datasets underpin the development and performance of large language models (LLMs). Among these, instruction-tuning datasets (ITDs) used for instruction fine-tuning have been pivotal in enhancing LLM performance and generalization capabilities across diverse tasks. This tutorial provides a comprehensive guide to designing, creating, and evaluating ITDs for health care applications. Written from a clinical perspective, it aims to make the concepts accessible to a broad audience, especially medical practitioners. Key topics include identifying useful data sources, defining the characteristics of well-designed datasets, and crafting high-quality instruction-input-output examples. We explore practical approaches to dataset construction, examining the advantages and limitations of 3 primary methods: fully manual preparation by expert annotators, fully synthetic generation using artificial intelligence (AI), and an innovative hybrid approach in which experts draft the initial dataset and AI generates additional data. Moreover, we discuss strategies for metadata selection and human evaluation to ensure the quality and effectiveness of ITDs. By integrating these elements, this tutorial provides a structured framework for establishing ITDs. It bridges technical and clinical domains, supporting the continued interdisciplinary advancement of AI in medicine. Additionally, we address the limitations of current practices and propose future directions, emphasizing the need for a global, unified framework for ITDs. We also argue that artificial general intelligence (AGI), if realized, will not replace empirical research in medicine. AGI will depend on human-curated datasets to process and apply medical knowledge. At the same time, ITDs will likely remain the most effective method of supplying this knowledge to AGI, positioning them as a critical tool in AI-driven health care.

(*J Med Internet Res* 2025;27:e70481) doi: [10.2196/70481](https://doi.org/10.2196/70481)

KEYWORDS

generative artificial intelligence; large language models; instruction-tuning datasets; tutorials; evaluation framework; health care

Introduction**Background**

Why is high-quality data the cornerstone of modern artificial intelligence (AI)-driven health care? Reliable data enable AI

algorithms to assist medical professionals in making evidence-based decisions, reducing the likelihood of errors and improving patient outcomes [1-5].

Accurate datasets are the foundation for developing large language models (LLMs) and deep learning models based on

the transformer architecture [1,6-12]. This technique enables the model to learn the structure, grammar, and nuances of language, as well as factual knowledge and patterns of reasoning [9-12]. Examples of state-of-the-art all-purpose transformer-based models include generative pretrained transformers (GPTs) from OpenAI, Gemini and Gemma models from Google DeepMind, and LLaMA LLMs from Meta [8,13,14].

Foundation LLMs are trained on vast amounts of data encompassing millions of samples from diverse sources such as books, studies, and websites [8,13,14]. The goal is to improve the model's generalization capabilities. This means enabling the model to apply learned patterns and knowledge to a wide range of unseen inputs, ensuring it produces accurate, meaningful, and contextually relevant outputs rather than merely memorizing the training data [8,13,14]. Among various types of datasets used to train LLMs, instruction-tuning datasets (ITDs) used for instruction fine-tuning (IFT) has emerged as a pivotal technique in enhancing LLM performance and generalization capabilities across diverse tasks [8,9,15]. For example, OpenAI researchers report that, even though the InstructGPT model with 1.3 billion parameters has more than 100 times fewer parameters than the original 175 billion-parameter GPT-3, its outputs were preferred over GPT-3 [8]. An ITD contains examples of task instructions paired with corresponding responses, enabling models to understand better and follow human-like directives [8,9,16]. This structured training approach not only improves task-specific accuracy but also enhances the LLM's ability to generalize knowledge across multiple domains, sometimes including even domains that were not extensively covered in the original dataset [8,9,15,16].

Given the importance of ITDs, preparing high-quality ITDs in health care is critical. Such datasets may facilitate the formation of robust LLMs capable of addressing the nuanced requirements of complex medical questions, where precision, adaptability, and context-specific understanding are essential. In the clinical setting, instruction-tuned LLMs may support clinical decision-making and reduce the risk of medical errors. It will ultimately improve clinically relevant outcomes like increased patient safety and reduced hospitalizations, intensive care unit admissions, or deaths [4,5,17-20].

Aim

This paper provides a guide on the key principles of designing, creating, and evaluating ITDs for training LLMs in health care applications.

What Is an ITD?

ITDs are used during the IFT of LLMs [8,9,12,21]. These datasets typically consist of instruction-input-output (IIO) triples, such as an appropriate instruction combined with a question and its corresponding answer [8,9,21]. The primary aim of IFT is to improve a model's ability to comprehend and execute various instructions, particularly those relevant to the medical domain, ultimately developing a specialized medical LLM [12,21,22].

ITDs differ significantly from the general datasets used for supervised fine-tuning (SFT). The key distinction between databases used for SFT and IFT lies in their objectives and methodologies. SFT primarily seeks to integrate domain-specific medical knowledge into a general LLM by continuing pretraining, enhancing the model's understanding of medical texts [8,12]. It creates a "medical foundation model" [12,21]. Conversely, IFT focuses on improving the model's responsiveness to instructions and tailoring its outputs to align with specific guidance and human-like responses rather than emphasizing token prediction accuracy as in ST [12,21]. Usually, the IFT follows the SFT [8,9,12,21].

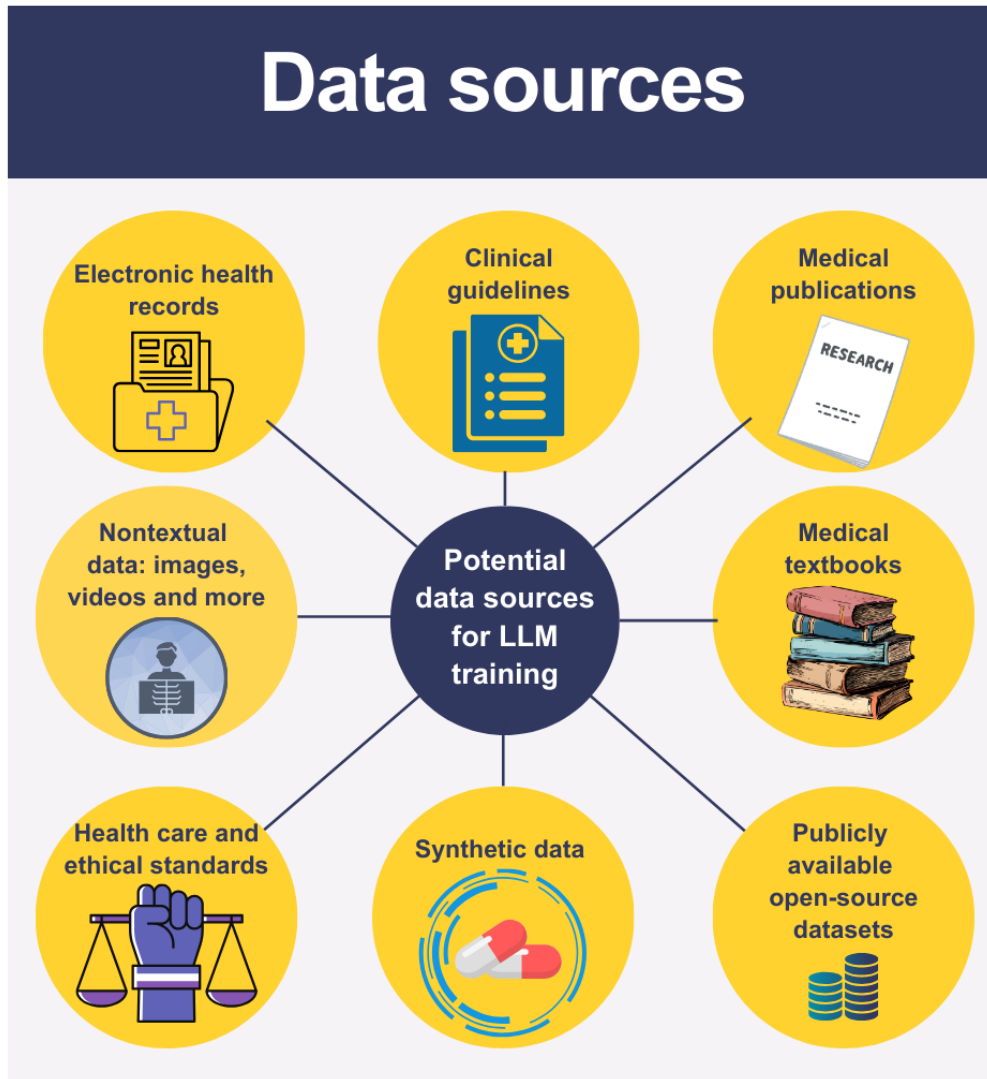
While SFT relies heavily on the volume of training data, IFT prioritizes the quality and diversity of the data. In general, IFT improves the performance of the baseline SFT model [7,16,21,22]. Recent research has explored combining these approaches to build robust medical LLMs, leveraging the strengths of both techniques for better overall accuracy [7,16,21,22].

Considerations for Designing ITDs in Health Care

Data Sources

Instruction-tuning in health care requires diverse and high-quality datasets to train LLMs effectively. Potential sources for such data are presented in [Figure 1](#) and described in detail in [Multimedia Appendix 1](#).

Figure 1. Potential data sources for instruction-tuning datasets. LLM: large language model.



Key Features of a Well-Designed ITD

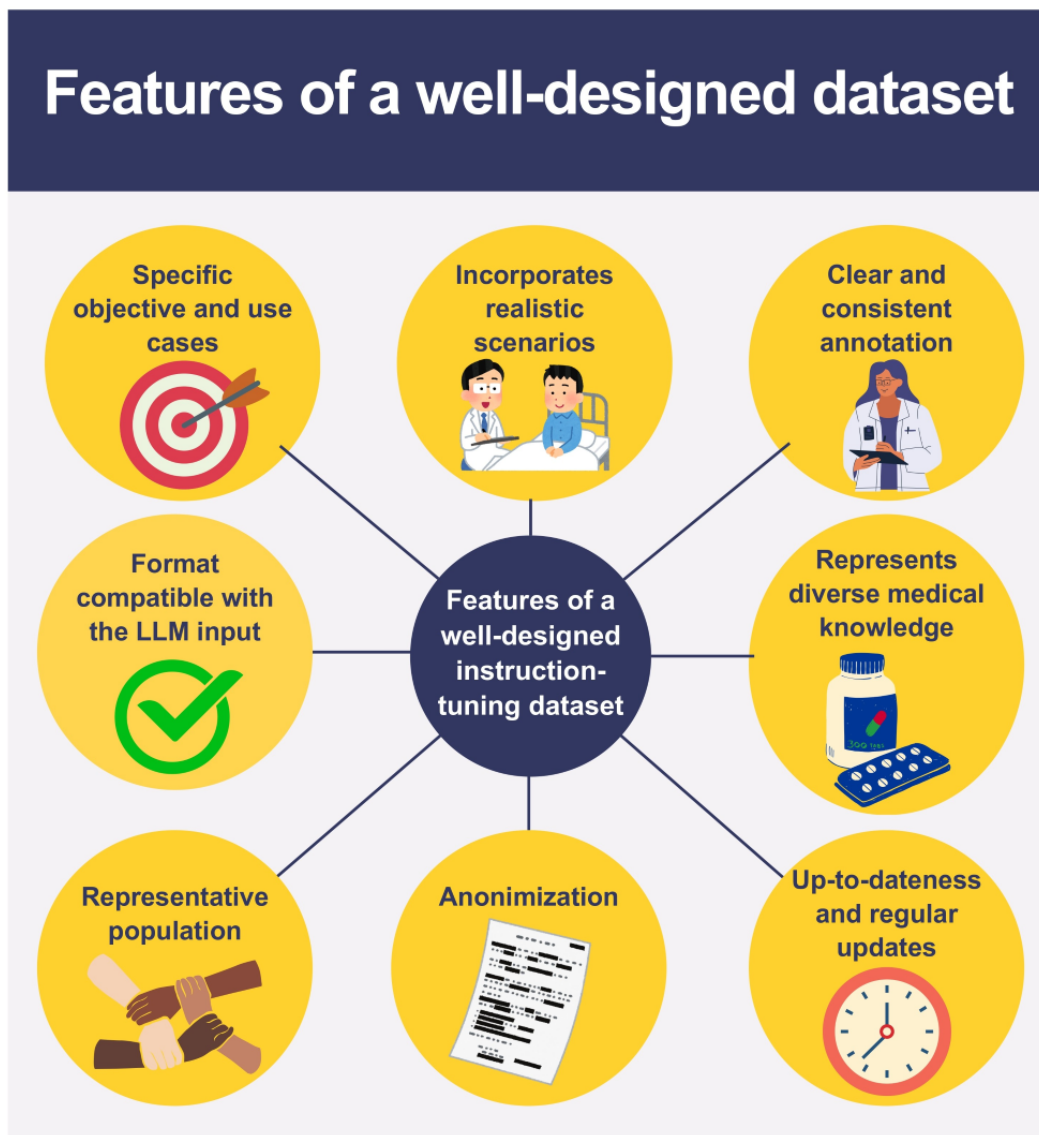
The dataset should align with the specific objectives and use cases for which the model is fine-tuned, whether clinical decision support, patient education, or administrative tasks (Figure 2) [11,21]. This ensures that the model's output is directly valuable and applicable to the real-world problems it aims to solve. Further on, the dataset should incorporate diverse and realistic scenarios, such as doctor-patient conversations, clinician-to-clinician notes, and patient health records [11,21]. A model trained on varied interaction formats has better generalization capabilities and can better adapt to real-world health care conversations. The samples should be clearly and consistently annotated. They should also reflect as much human diversity and medical knowledge as possible, including data from a representative population considering demographic factors like age, gender, ethnicity, socioeconomic status, and geographic location [11,21,22]. Moreover, incorporating

evidence-based medical information is crucial for patient safety [11,21,22]. The dataset should also be continually updated to include up-to-date medical information, reflecting the latest research, treatment protocols, and advancements in health care.

Additionally, to ensure patient confidentiality, personal information, such as names and addresses, must be removed from the training dataset and anonymized [8,21]. Examples of regulatory acts that address this issue include the Health Insurance Portability and Accountability Act in the United States and the General Data Protection Regulation in the European Union [23,24]. Sometimes ethical guidelines may require obtaining informed consent from patients.

Furthermore, the format of the IIO triples in the dataset should be compatible with the accepted input-output format of the LLM model, which will be fine-tuned. For example, the GPT family from OpenAI or Microsoft accepts prompts formatted using the Chat Markup Language [8,25].

Figure 2. Features of a well-designed dataset. LLM: large language model.



IIO Examples

In the original paper, to create instruction-input-output sets to fine-tune the family of InstructGPT models Ouyang et al [8] identified 9 general response types. Starting from the most popular, the dataset included the following scenarios: generation,

open question-answering, brainstorming, chat, rewrite, summarization, classification, closed question-answering, and extract (Figure 3). Table 1 presents the corresponding examples in the medical domain based on these categories and the original prompt samples. Additional examples are provided in Multimedia Appendix 1.

Figure 3. Generative AI response types. AI: artificial intelligence.



Table 1. Prompt categories with corresponding examples from the medical domain.

Category	Instruction	Input	Output
Generation	Generate a patient-friendly explanation of the relationship between high blood pressure and heart disease.	Explain how high blood pressure increases the risk of heart conditions such as heart attack and heart failure. Use simple language and include suggestions for lifestyle changes.	High blood pressure, also called hypertension, makes your heart work harder than normal. Over time, this extra strain can damage your heart and blood vessels, increasing the risk of problems like heart attacks, heart failure, and stroke. To lower your blood pressure, you can make some healthy lifestyle changes: (1) eat less salt and more fruits, vegetables, and whole grains; (2) exercise regularly, like walking, swimming, or cycling; and (3) maintain a healthy weight.
Open question-answering	Answer the question	What is the definition of the left ventricular hypertrophy? How is left ventricular hypertrophy typically diagnosed?	Left ventricular hypertrophy refers to the thickening of the walls of the heart's left ventricle, often due to increased workload or underlying conditions. Diagnosis typically involves several steps. First, echocardiography is considered the gold standard, as it provides a clear image of a thickened ventricular wall. Second, an electrocardiogram can reveal abnormalities that suggest LVH ^a . Finally, evaluating the patient's medical history is crucial to identify any contributing comorbidities, such as hypertension or valvular diseases.
Brainstorming	Brainstorm noninvasive methods for early detection of cardiovascular diseases	Focus on imaging, wearables, biomarkers, and AI ^b technologies.	Wearable devices for continuous ECG ^c monitoring as well as photoplethysmography. AI-enhanced portable ultrasound for cardiac imaging.
Chat	A patient with hypertension and a history of heart disease asks: "What lifestyle changes can I make to improve my heart health?" Respond with clear and concise evidence-based advice.	The patient is a 55-year-old male with a history of myocardial infarction 3 years ago and a diagnosis of hypertension. He is currently on beta-blockers and ACE inhibitors. His BMI is 28, and his lipid profile shows elevated LDL ^d cholesterol. He works a sedentary desk job and smokes.	Quitting smoking. Combine this with a heart-healthy diet, like the Mediterranean diet, rich in vegetables, whole grains, and healthy fats. Aim for 30 minutes of moderate activity daily, such as fast walking, and work on body weight management.
Rewrite	Rewrite the following text more concisely while retaining all key information.	"Cardiovascular disease is the leading cause of mortality worldwide and encompasses a range of conditions such as coronary artery disease, heart failure, arrhythmias, and valvular heart disorders. Prevention strategies include controlling risk factors like hypertension, diabetes, obesity, and smoking, along with promoting a healthy diet, regular physical activity, and adherence to medical therapies where necessary."	"Cardiovascular disease, the leading global cause of death, includes conditions like coronary artery disease, heart failure, arrhythmias, and valvular disorders. Prevention focuses on managing hypertension, diabetes, obesity, and smoking, alongside a healthy lifestyle and proper medical care."
Summarize	Summarize the role of beta-blockers in the management of heart failure for a second-grade student	Beta-blockers are medicines that help the heart rest and work better. They slow down the heartbeat, so the heart does not have to work as hard. Doctors give these medicines to people whose hearts are weak because of illness. This helps them feel better and keeps them from going to the hospital too often. Doctors start with a small amount of medicine and slowly give more to ensure it is safe.	Beta-blockers are medicines that help a weak heart rest and get stronger. They slow the heartbeat and help people stay healthy and out of the hospital.
Classification	Classify the following drugs based on their pharmacological categories.	Carvedilol and rivaroxaban	<ul style="list-style-type: none"> • Carvedilol: beta-blocker • Rivaroxaban: antithrombic
Closed question-answering	Answer the following closed question based on your knowledge of cardiovascular medicine.	Is metoprolol a beta-blocker or antiplatelet drug?	Beta-blocker

Category	Instruction	Input	Output
Extract	Extract the diagnoses of the patient from the electronic health records.	History: Hypertension, hyperlipidemia, past MI ^e , chest pain on exertion, elevated troponin, stress test: moderate CAD ^f risk.	Hypertension, hyperlipidemia, MI, and CAD (moderate risk)

^aLVH: left ventricular hypertrophy.

^bAI: artificial intelligence.

^cECG: electrocardiogram.

^dLDL: low-density lipoprotein.

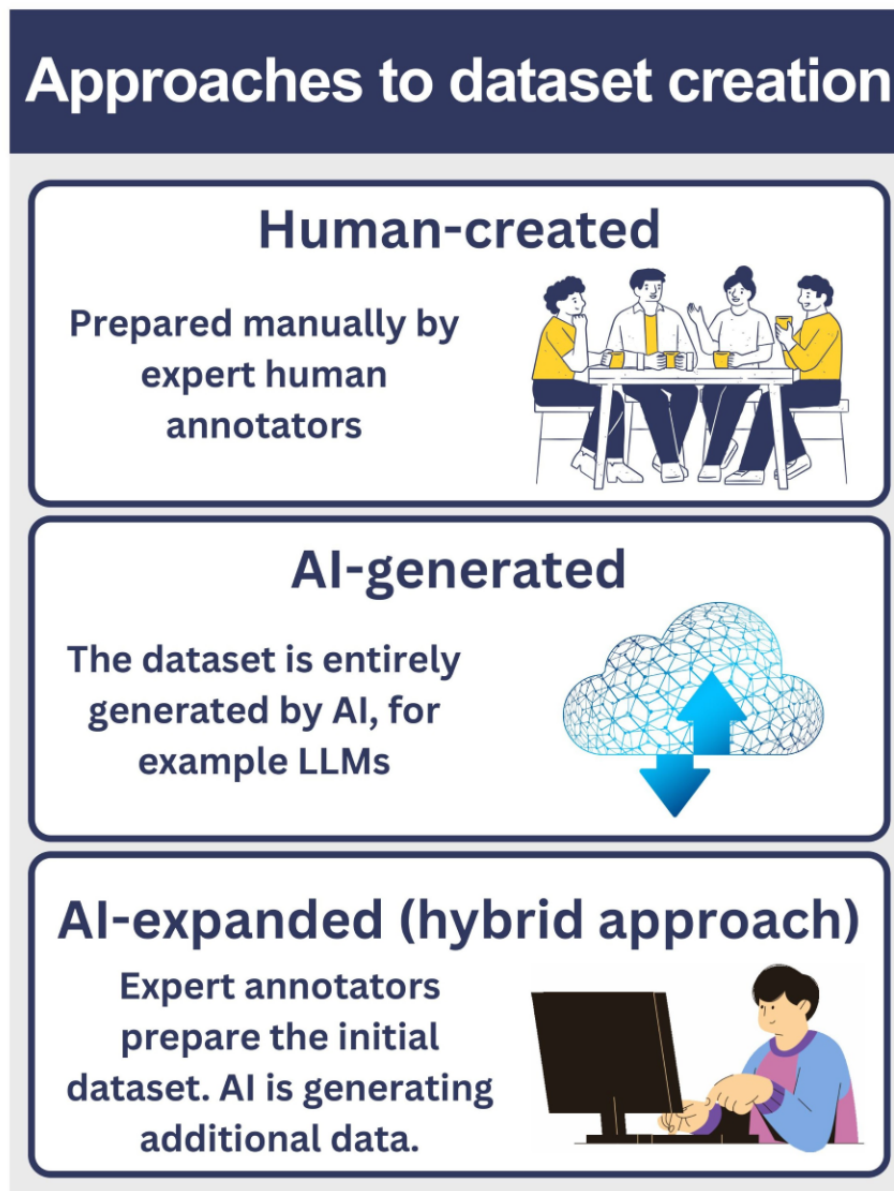
^eMI: myocardial infarction.

^fCAD: coronary artery disease.

Approaches to Dataset Creation

There are 3 main approaches to preparing an ITD (Figure 4) [7-9,15,16,21].

Figure 4. Approaches to dataset creation. AI: artificial intelligence; LLM: large language model.



Human-Created

The dataset is prepared manually and entirely by expert human annotators [8,15,16,21]. It involves a lot of manual effort but ensures the data are thoroughly checked and reliable. Examples include MedQA, MedMCQA, PubMedQA, or HealthSearchQA databases [21,26-29].

AI-Generated

The dataset is entirely generated by AI, for example, LLMs. Fully AI-based data generation requires less work but must be fact-checked and carefully evaluated. Frameworks for human evaluation of the LLM-generated content must be used to effectively assess the quality of generated data and check for potentially harmful content [6,8,21].

Automated AI generation was applied by Wu et al [15], who built a 400,000 instruction-following dataset, "MIMIC-Instr," from the publicly available MIMIC-IV electronic health record (EHR) database [15,30]. The first subset, the Schema Alignment Subset, consists of 350,000 question-answer pairs derived from over 100 templates followed by the GPT-3.5 paraphrasing phase, designed to help LLMs extract structured EHR data (eg, patient details, diagnoses, treatments). Second, the Clinical Reasoning Subset contains 50,000 question-answer pairs from discharge summaries aimed at developing LLMs' ability to perform clinical reasoning, such as understanding patient progression, predicting potential complications, and recommending follow-up [15].

AI-Expanded Using Human Seed (Hybrid Approach)

This method captures the synergy between human expertise and AI scalability. Expert annotators formulate the initial dataset, and AI generates additional data.

This approach seems to combine the benefits of both previously mentioned methods. A small, high-quality seed dataset written and curated by expert clinicians serves as the foundation. Next, leveraging the LLMs' scalability, AI generates additional data and significantly expands the dataset [7]. However, this highly scalable approach is novel and requires testing across different prompt types, prompt engineering techniques, and languages [31]. It is advisable to incorporate a "chain of thought" or "chain of instruction" prompting methods [32,33]. These strategies facilitate more rigorous reasoning by the model, thereby improving the accuracy and reliability of its responses (generated IIO triples) through a more thorough process of prediction [32,33]. Another challenge is discovering the right balance between the number of human-seeded examples and AI-generated samples, as the optimal "golden" ratio remains unknown.

Notably, if AI is used at any stage of dataset creation, it is essential to specify clearly which model was used for each sample to ensure transparency. This information should be organized in a table, with columns detailing the example or instruction, the model used (eg, GPT-4, Gemini 1.5, Claude 3), and the corresponding predicted response [14,34,35]. Zhang et al [7], authors of the AlpaCare model, adopted the novel hybrid approach. They also authored the MedInstruct-52k database, using GPT-4 to generate a diverse set of over 52,000 instructions

based on a high-quality expert-curated seed set encompassing 167 samples [7]. Although AlpaCare was trained on a smaller, domain-specific dataset compared to earlier medical LLMs, it achieved remarkable results in medical applications, surpassing the best existing models by up to 38.1% in medical free-form instruction evaluations [7]. Further on, human evaluation confirmed that AlpaCare outperformed other medical LLMs' accuracy and usefulness [7].

Nevertheless, many well-known datasets (used as benchmarks for the medical LLMs) like MedQA, MedMCQA, PubMedQA, MMLU clinical topics database, and HealthSearchQA were curated by human annotators or entirely created through manual effort [21,26-29]. Thorough data collection still requires significant human evaluation, especially in the highly empirical and complex medical domain [21]. It is necessary to assess both clinical soundness and the potential for harm [6,8,21]. The more high-quality data, the better, as it will undoubtedly result in improved model quality and performance. Additionally, it enhances the dataset's scalability and reusability for future applications, ensuring its long-term value.

Database-to-Model Compatibility

Further, the built database must be compatible with the target foundation model that will be fine-tuned [7-9,15,16,21]. Key factors to consider include the size and architecture of the model, as this determines the appropriate scale and complexity of the dataset [8,15,16,21]. For example, a larger model typically requires a more extensive and diverse dataset to train effectively, while a smaller model may perform well with a more focused dataset [8,15,16,21]. Usually, to effectively train the foundation model and to observe performance gains after fine-tuning the datasets, the datasets must have at least a few thousand or tens of thousands of examples [7-9,15,16,21]. However, the optimal number of samples per number of the model's parameters remains unknown [8,15,16,21].

Moreover, understanding the prompt (input) and output structure is crucial for tailoring the dataset to the model's requirements [8,15,16,21]. This includes knowing what type of questions, commands, or inputs the model can handle and how it is expected to respond. Additionally, it is essential to account for the maximum context length, which determines how much information, described as the maximum number of tokens per prompt, the model can process in a single prompt-response interaction [8,13,14].

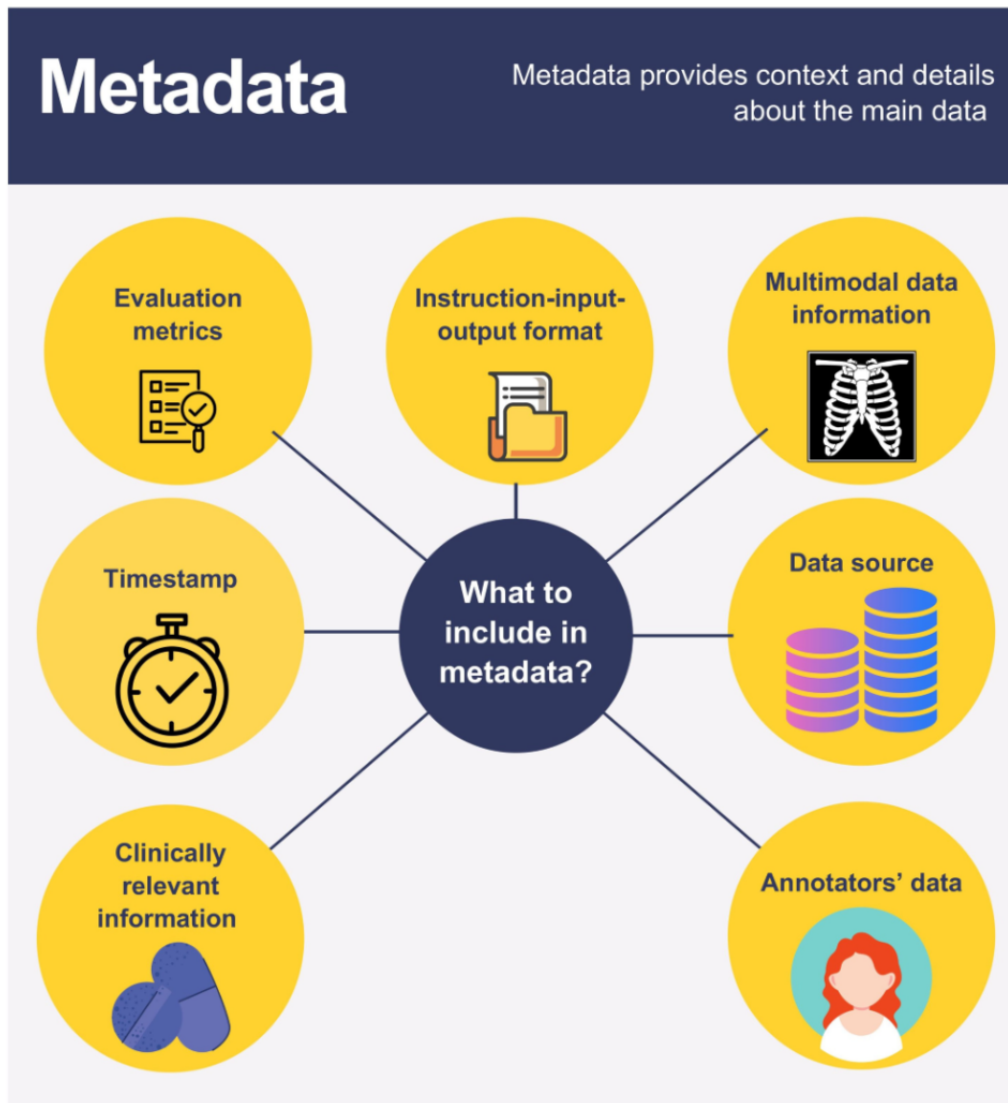
Metadata

Overview

Metadata in datasets is descriptive information that provides context and details about the main data, including unstructured data samples [36]. In the case of ITDs, the main data are IIO triples [7,8,15,22]. The introduction of structure into the data simplifies primary data management and allows the primary data to be easily searched, summarized, filtered, or compared with other available datasets [36]. Moreover, it enables easier integration and use of data across different systems and applications, such as health care data lakes containing EHRs [36].

Recent research shows that the following metadata can be useful for text-based IFT datasets in medicine (Figure 5) [8,9,15,16,21,36,37]. Multimedia Appendix 1 provides details about the metadata.

Figure 5. Metadata in instruction-tuning datasets.



Human Evaluation of ITDs

Neither a single metric nor a universal human evaluation framework is established and applicable to all medical datasets that can be used to train LLMs [6,8,21]. Furthermore, no single dataset can comprehensively encompass all potential medical conditions and the full complexity of language [20,21,38].

Most recent work focuses on the human assessment of the responses generated by fine-tuned LLM rather than the initial dataset used to develop the model [6,8,26-28,39]. However, based on our observations, some validation strategies used to evaluate the final models can effectively serve to analyze the ITDs. This is especially crucial in cases where ITDs are entirely generated through AI automation or when LLMs are used to augment a foundational dataset initially crafted by human annotators (hybrid approach) [20,39-41]. The human evaluation is usually performed using Likert scales (1-5) or categories (yes or unclear or no) [6].

Implementation of a Clear and Objective Evaluation Process

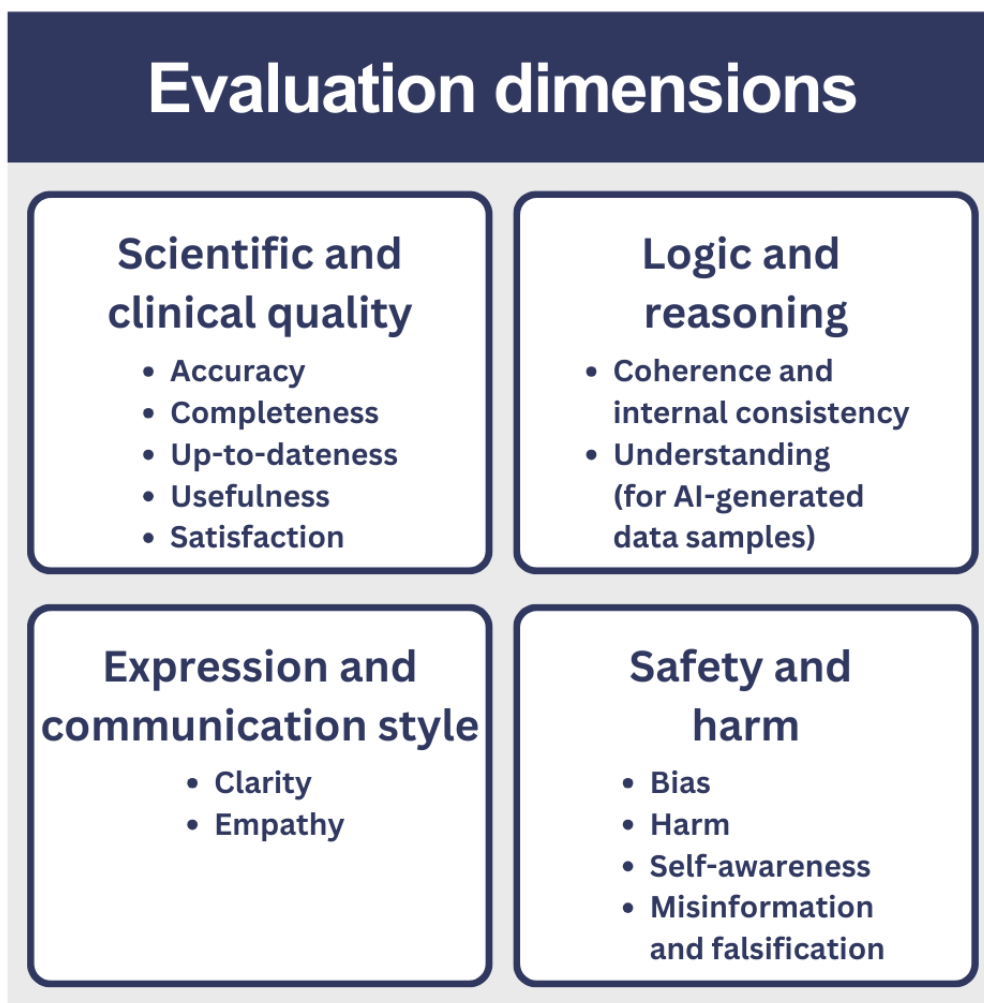
The implementation of objective evaluation involves 3 key phases: training, optimization, and final scoring. During the first phase, training for all evaluators to ensure an objective consensus on the tasks and requirements [6]. Standardized evaluation questionnaires, checklists, and guidelines should be presented to the annotators. Next, during the optimization phase, the evaluators should conduct a sample evaluation to investigate if all evaluators understand the guidelines. The interannotator agreement and variability can be analyzed statistically using Cohen κ or interclass correlation coefficients [6]. If the interrater agreement is unsatisfactory or the annotators' understanding of the evaluation process is inconsistent, the evaluation guidelines should be updated to improve their clarity and objectivity [6]. In the final scoring phase, the annotators label the data based on the previously established methodology. Final evaluation scores for each dimension are calculated.

Evaluation Dimensions

Four dimensions may be used to evaluate each sample (IIO

triple) in the ITDs (Figure 6) [6,8,26-28,39]. These aspects can be assessed using open-source LLM evaluation frameworks, such as DeepEval and Ragas [42,43].

Figure 6. Evaluation dimensions. AI: artificial intelligence.



Scientific and Clinical Quality

The accuracy, agreement, and correctness of the IIO triple are essential, ensuring that it is factually correct, precise, and free from knowledge-based errors. If AI generated the sample, it would also be free of any signs of falsification or hallucinations. Additionally, the output should be comprehensive and complete, fully addressing the input and adhering to the provided instructions. The sample must also reflect the most up-to-date knowledge available.

Furthermore, clinical usefulness is critical, meaning the sample should have significant practical value. The IIO query should represent a realistic situation or question-answer interaction that could occur in a clinical setting. Finally, overall user satisfaction is an important consideration. Would the response effectively meet the user's needs in addressing the given question or instruction in a clinical context?

Logic and Reasoning

Coherence, reasoning, and internal consistency involve ensuring that the instruction, input, and output are logically connected

and aligned. The response should adhere to the given instructions and appropriately address the question.

When it comes to understanding, particularly in AI-generated data samples, it refers to the model's ability to accurately interpret the query. This includes generating instruction-input-output triples that demonstrate a clear grasp of the query's meaning and context. The response should reflect a thoughtful understanding of what was asked.

Expression and Communication Style

Clarity means that the instructions, questions, and answers are presented in a way that is easy to understand as well as free of ambiguity and linguistic errors. Communication should be straightforward and concise.

Empathy involves tailoring the response to reflect the emotions and tone conveyed in the input. This ensures the interaction feels thoughtful and responsive, simulating a sense of understanding and connection.

Safety and Harm

“*Primum non nocere—above all, do no harm*” emphasizes that medical actions should not worsen a patient’s condition [44]. Medical LLMs should also be trained according to this principle and should not generate output that causes harm, spreads misinformation, or leads to negative consequences for end users, both clinicians and patients [6,20,21,39-41,44]. Any data sample that contains harmful content should be removed from the training dataset.

Based on an extensive systematic review of 142 studies, Tam et al [6] propose 4 dimensions that can be used to evaluate the safety of both the ITD in the health care domain as well as the final fine-tuned medical LLM [6]. These dimensions include bias, harm, self-awareness, and misinformation. Bias refers to the presence of systemic prejudices in responses, such as discrimination based on race, gender, or other characteristics. Second, harm encompasses any potential negative outcomes

caused by responses, such as spreading misinformation, promoting offensive or harmful language, reinforcing stereotypes, encouraging illegal activities, or inciting violence. Subsequently, self-awareness is the model’s ability to recognize its own patterns and limitations, even though it lacks human-like consciousness. Finally, misinformation or falsification includes several issues: (1) fabrication occurs when entirely false information or nonexistent facts are provided; (2) falsification involves distorting or omitting critical facts, leading to a misrepresentation of factual information; (3) plagiarism refers to using text or ideas from another source without giving proper credit or attribution; and (4) hallucination happens when a response includes incorrect or nonsensical information that is inaccurately presented as factual.

Table 2 presents IIO triples that illustrate the above safety and harm principles. Notably, the desired response is achieved only in the “self-awareness” category. Additional examples are provided in the Multimedia Appendix 1.

Table 2. Examples of safety and harm principles.

Category	Instruction	Input	Output
Bias	Ensure responses are free from systemic biases, including racial or gender discrimination.	What are the risk factors for heart disease?	Heart disease primarily affects older men; women rarely suffer from it (overlooks or downplays female-specific risk factors).
Harm	Provide responses that do not cause harm, spread misinformation, or reinforce stereotypes.	Can I stop taking my blood pressure medication if I feel fine?	Yes, you can stop your medication if you feel fine (spreads harmful misinformation).
Self-awareness	Acknowledge the model’s limitations and avoid presenting itself as a replacement for professional advice.	Can you guarantee the accuracy of your medical advice?	I am a language model and cannot guarantee accuracy. In doubt, please consult a medical professional (correct answer).
Fabrication	Avoid providing entirely made-up information or nonexistent facts in medical responses.	What is the success rate of a new drug for heart failure?	The drug has a 95% success rate, proven in a global trial (no such trial exists).
Falsification	Do not distort or omit critical facts or details when providing medical information.	What are the key findings of the latest study on hypertension?	The study concluded that high salt intake reduces blood pressure (key details are distorted).
Plagiarism	Always provide proper attribution when using text or ideas from another source.	Explain the mechanism of action of beta-blockers.	Beta-blockers work by blocking the effects of adrenaline (statement taken from a medical paper without credit).
Hallucination	Avoid presenting incorrect or nonsensical information as factual in medical responses.	What is the normal range for ejection fraction in heart function?	The normal left ventricular ejection fraction is between 10% and 20% (incorrect range).

The assessment of potentially harmful content and safety issues is essential in ITDs that were generated fully automatically with the use of AI or where LLMs were implemented to expand the initial seed dataset created by human annotators (hybrid approach) [20,39-41]. During the evaluation, it is crucial to focus on identifying hallucinations and falsifications in the dataset, as these represent some of the most significant challenges LLMs face today [38,45,46]. Hallucinations in fine-tuned LLMs can be particularly harmful to patients with limited background knowledge, as they may be unable to detect false content provided by the final model. In contrast, health care professionals with extensive medical knowledge are better equipped to identify hallucinations and falsifications more easily

[40,41,47]. Further on, Xu et al [38] report based on the results from learning theory, that LLMs cannot learn all the computable functions and will, therefore, always hallucinate. Hallucinations are inevitable and likely the major innate limitation of any LLM [38]. Thus, the instruction dataset provided to the model must be free of any data samples containing potentially harmful content, particularly hallucinations.

Limitations and Future Directions

Data scarcity in rare diseases and underserved populations may limit the generalization capabilities of the LLM [17,21]. Synthetic AI-based generation of ITDs offers a potential solution by simulating patient scenarios and improving model

performance in underrepresented groups [7,15]. However, this approach carries risks and raises ethical concerns [40,41,48]. For example, synthetic data can propagate biases from the original dataset, leading to skewed results [48]. Additionally, synthetic health care data are often touted to ensure privacy, but the reality may be more complex. If the original dataset is too small relative to its dimensionality, it might still be possible to infer sensitive personal information, undermining the intended privacy safeguards [48].

Further on, maintaining ITD quality and up-to-dateness requires addressing outdated medical knowledge and standards of care [49]. Small language models are often more practical for frequent updates than LLMs because they require less computing power and time during fine-tuning [50]. Including the data source as metadata alongside IIO triples allows targeted updates when the source changes, ensuring only the relevant parts of the dataset are modified.

Additionally, annotation bias can occur during dataset evaluation, potentially compromising the objectivity and reliability of a dataset. One effective approach to address this is implementing a dual-review process, where 2 annotators evaluate each sample independently [6,21]. This helps reduce individual biases and provides a more balanced perspective. Ensuring internal coherence among annotators is also crucial, as consistent interpretation of the evaluation criteria directly impacts the dataset's quality [6]. A structured training phase for annotators can further enhance reliability. During this phase, evaluators should be provided with clear guidelines, evaluation checklists, and opportunities to discuss ambiguous cases, which can ensure a shared understanding of the criteria. Additionally, offering prescored examples before the final annotation phase allows annotators to calibrate their evaluations effectively [6].

Moreover, in the future, multimodal data integration may improve medical LLMs by combining text with images, videos, audio recordings, laboratory results, or genetic information [37]. Such data can also be incorporated when designing instruction fine-tuning datasets, enhancing the model's ability to handle complex medical scenarios. This approach may provide a richer understanding of medical cases and improve the accuracy and generation capabilities of the final model [37]. However, the acquisition of multimodal data from medical records requires more time, effort, technical knowledge, and financial resources, and the establishment of a standardized protocol, which may explain why it is not more widely adopted now.

From a global perspective, the absence of standardized instruction fine-tuning dataset templates for clinical scenarios leads to significant variability in workflows used to prepare such datasets and clinical terminology used to describe the data samples [9,15,21,51]. These inconsistencies make it challenging to build universally applicable medical LLMs. To address this, global initiatives are needed to establish a uniform health care

ITD framework. Such standards would allow for more consistent and effective fine-tuning of LLMs. Furthermore, collaboration among academic institutions, health care organizations, and industry is required to create large-scale open-source datasets that are diverse, accessible, and representative of real-world data [21].

Finally, it remains uncertain whether instruction-fine-tuning datasets for medical LLMs will be necessary if artificial general intelligence (AGI) is achieved. This question is still largely speculative [52-55]. Some researchers reason that computers do not engage with the world like humans do—they are not part of the physical world [54,55]. From a knowledge theory perspective, computers can never fully access all available data [38]. Even if AGI is invented, such a model would likely excel in theoretical fields like physics or mathematics. Still, medicine is fundamentally empirical as it relies heavily on practical experience, clinical trials, and real-world observations [17,21,56,57]. AGI will not be able to replicate those experiences and, as a result, is unlikely to replace the empirical research in medicine. Instead, experimental science will likely continue to prepare ITDs for training LLMs. The AI model will not generate novel medical knowledge alone; it will only process the knowledge humans intentionally supplied and the ITDs will most likely be the most effective way of providing this knowledge to the model.

Conclusions

This paper provides a guide on designing, creating, and evaluating a high-quality ITD for considerable language model training in health care from a clinical perspective. Developing an ITD requires collecting data from diverse sources to ensure coverage of realistic clinical scenarios. Moreover, an end user of the final model must be defined. ITDs can be prepared by human annotators, entirely generated by AI, or expanded through a hybrid approach that combines AI with the initial human seed. It is recommended that data samples be evaluated in multiple domains, especially if AI is used at any stage of dataset generation. Each IIO sample ought to be described using metadata. The datasets must comply with ethical standards of data privacy. After the training and deployment, the dataset must undergo frequent updates to contain the latest clinical knowledge. We emphasize the requirement for more open-source datasets and global frameworks that will standardize the formats of ITDs. Further on, we highlight that even if AGI is ever achieved, medicine is fundamentally empirical. Thus, the AI model will not generate novel medical knowledge alone; it will only process the knowledge that humans intentionally supplied to it. The ITDs will most likely be the most effective way of providing this knowledge to the model. Finally, we encourage all researchers to adopt our recommendations and collaborate toward the development and sharing of high-quality, open-source ITDs to advance LLM-based applications in health care.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary description of potential data sources and metadata. Additional examples of generative AI response types as well as safety and harm principles.

[\[DOCX File, 54 KB-Multimedia Appendix 1\]](#)

References

1. Schwabe D, Becker K, Seyferth M, Klauf A, Schaeffter T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *NPJ Digit Med*. 2024;7(1):203. [FREE Full text] [doi: [10.1038/s41746-024-01196-4](https://doi.org/10.1038/s41746-024-01196-4)] [Medline: [39097662](https://pubmed.ncbi.nlm.nih.gov/39097662/)]
2. Pipino LL, Lee YW, Wang RY. Data quality assessment. *Commun ACM*. 2002;45(4ve):211-218. [doi: [10.1145/505999.506010](https://doi.org/10.1145/505999.506010)]
3. Ehsani-Moghaddam B, Martin K, Queenan JA. Data quality in healthcare: a report of practical experience with the Canadian primary care sentinel surveillance network data. *Health Inf Manag*. 2021;50(1-2):88-92. [doi: [10.1177/1833358319887743](https://doi.org/10.1177/1833358319887743)] [Medline: [31805788](https://pubmed.ncbi.nlm.nih.gov/31805788/)]
4. Nazar W, Szymanowicz S, Nazar K, Kaufmann D, Wabich E, Braun-Dullaeus R, et al. Artificial intelligence models in prediction of response to cardiac resynchronization therapy: a systematic review. *Heart Fail Rev*. 2024;29(1):133-150. [FREE Full text] [doi: [10.1007/s10741-023-10357-8](https://doi.org/10.1007/s10741-023-10357-8)] [Medline: [37861853](https://pubmed.ncbi.nlm.nih.gov/37861853/)]
5. Nazar W, Romantowski J, Niedoszytko M, Daniłowicz-Szymanowicz L. Cardiac adverse drug reactions to COVID-19 vaccines. a cross-sectional study based on the Europe-wide data. *Eur Heart J Cardiovasc Pharmacother*. 2024;10(7):599-607. [doi: [10.1093/ehjcvp/pvae063](https://doi.org/10.1093/ehjcvp/pvae063)] [Medline: [39174484](https://pubmed.ncbi.nlm.nih.gov/39174484/)]
6. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digit Med*. 2024;7(1):258. [FREE Full text] [doi: [10.1038/s41746-024-01258-7](https://doi.org/10.1038/s41746-024-01258-7)] [Medline: [39333376](https://pubmed.ncbi.nlm.nih.gov/39333376/)]
7. Zhang X, Tian C, Yang X, Chen L, Li Z, Petzold LR. AlpaCare:Instruction-tuned large language models for medical application. *ArXiv*. Preprint posted on October 23, 2023. [FREE Full text]
8. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, editors. *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Red Hook, NY: Curran Associates Inc; 2023.
9. Zhang S, Dong L, Li X, Zhang S, Sun X, Wang S, et al. Instruction tuning for large language models: a survey. *ArXiv*. Preprint posted online on August 21, 2023. [FREE Full text]
10. Parthasarathy VB, Zafar A, Khan A, Shahid A. The ultimate guide to fine-tuning LLMs from basics to breakthroughs: an exhaustive review of technologies, research, best practices, applied research challenges and opportunities. *ArXiv*. Preprint posted online on August 23, 2024. [FREE Full text]
11. Li Q, Yang X, Wang H, Wang Q, Liu L, Wang J, et al. From beginner to expert: modeling medical knowledge into general LLMs. *ArXiv*. Preprint posted online on December 2, 2023. [FREE Full text]
12. Zhou H, Liu F, Gu B, Zou X, Huang J, Wu J, et al. A survey of large language models in medicine: progress, application, and challenge. *ArXiv*. Preprint posted online on November 9, 2023. [FREE Full text] [doi: [10.48550/arXiv.2311.05112](https://doi.org/10.48550/arXiv.2311.05112)]
13. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: Open and efficient foundation language models. *ArXiv*. Preprint posted online on February 27, 2023. [FREE Full text]
14. Gemini Team Google, Georgiev P, Lei V, Burnell R, Bai L, Gulati A, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*. Preprint posted online on March 8, 2024. [FREE Full text]
15. Wu Z, Dadu A, Nalls M, Faghri F, Sun J. Instruction tuning large language models to understand electronic health records. 2024. Presented at: 2024 Conference on Neural Information Processing Systems; December 10-15, 2024; Vancouver, BC. URL: <https://neurips.cc/virtual/2024/poster/97801>
16. Jiang Z, Sun Z, Shi W, Rodriguez P, Zhou C, Neubig G, et al. Instruction-tuned language models are better knowledge learners. *ArXiv*. Preprint posted online on May 26, 2024. [FREE Full text] [doi: [10.18653/v1/2024.acl-long.296](https://doi.org/10.18653/v1/2024.acl-long.296)]
17. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. [doi: [10.1038/s41591-018-0300-7](https://doi.org/10.1038/s41591-018-0300-7)] [Medline: [30617339](https://pubmed.ncbi.nlm.nih.gov/30617339/)]
18. Sufyan M, Shokat Z, Ashfaq UA. Artificial intelligence in cancer diagnosis and therapy: current status and future perspective. *Comput Biol Med*. 2023;165:107356. [doi: [10.1016/j.compbiomed.2023.107356](https://doi.org/10.1016/j.compbiomed.2023.107356)] [Medline: [37688994](https://pubmed.ncbi.nlm.nih.gov/37688994/)]
19. Wu J, Liu X, Li M, Li W, Su Z, Lin S, et al. Clinical text datasets for medical artificial intelligence and large language models—a systematic review. *NEJM AI*. 2024;1(6). [doi: [10.1056/aira2400012](https://doi.org/10.1056/aira2400012)]
20. Matheny M, Israni ST, Whicher D, Ahmed M. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril*. Washington, DC: National Academies Press; 2023.
21. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. [FREE Full text] [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
22. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. *Nat Med*. 2025. [doi: [10.1038/s41591-024-03423-7](https://doi.org/10.1038/s41591-024-03423-7)] [Medline: [39779926](https://pubmed.ncbi.nlm.nih.gov/39779926/)]

23. HIPAA for professionals. US Department of Health and Human Services. URL: <https://www.hhs.gov/hipaa/for-professionals/index.html> [accessed 2024-11-25]
24. The general data protection regulation. European Council and Council of the European Union. URL: <https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/> [accessed 2024-11-25]
25. Chat Markup Language ChatML (Preview). Microsoft. URL: <https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/chat-markup-language> [accessed 2024-11-25]
26. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring massive multitask language understanding. ArXiv. Preprint posted online on September 7, 2020. [FREE Full text]
27. Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: a dataset for biomedical research question answering. 2019. Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November 3-7, 2019:2567-2577; Hong Kong, China. [doi: [10.18653/v1/d19-1259](https://doi.org/10.18653/v1/d19-1259)]
28. Pal A, Umaphathi LK, Sankarasubbu M. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. Proc Conf Health Inference Learn. 2022:248-260. [FREE Full text]
29. Jin D, Pan E, Oufattole N, Weng W, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci. 2021;11(14):6421. [doi: [10.3390/app11146421](https://doi.org/10.3390/app11146421)]
30. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10(1):1. [FREE Full text] [doi: [10.1038/s41597-022-01899-x](https://doi.org/10.1038/s41597-022-01899-x)] [Medline: [36596836](https://pubmed.ncbi.nlm.nih.gov/36596836/)]
31. Schulhoff S, Ilie M, Balepur N, Kahadze K, Liu A, Si C, et al. The Prompt Report: a systematic survey of prompting techniques. ArXiv. Preprint posted online on June 6, 2024. [FREE Full text]
32. Wang X, Zhou D, Deepmind G. Chain-of-thought reasoning without prompting. ArXiv. Preprint posted online on February 15, 2024. [FREE Full text]
33. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. ArXiv. Preprint posted online on January 28, 2022. [FREE Full text]
34. Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. Anthropic. 2024. URL: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf [accessed 2025-01-26]
35. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 technical report. ArXiv. Preprint posted online on March 15, 2023. [FREE Full text]
36. Piantella D, Reali P, Kumar P, Tanca L. A minimum metadataset for data lakes supporting healthcare research. 2024. Presented at: SEBD 2024: 32nd Symposium on Advanced Database Systems; June 23-26, 2024; Villasimius, Sardinia, Italy. URL: <https://sebd2024.unica.it/papers/paper47.pdf>
37. AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault M, Damseh R, et al. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res. 2024;26:e59505. [FREE Full text] [doi: [10.2196/59505](https://doi.org/10.2196/59505)] [Medline: [39321458](https://pubmed.ncbi.nlm.nih.gov/39321458/)]
38. Xu Z, Jain S, Kankanhalli M. Hallucination is inevitable: an innate limitation of large language models. ArXiv. Preprint posted online on January 22, 2024. [FREE Full text]
39. Bommasani R, Liang P, Lee T. Holistic evaluation of language models. Ann N Y Acad Sci. 2023;1525(1):140-146. [doi: [10.1111/nyas.15007](https://doi.org/10.1111/nyas.15007)] [Medline: [37230490](https://pubmed.ncbi.nlm.nih.gov/37230490/)]
40. Ong JCL, Chang SYH, William W, Butte AJ, Shah NH, Chew LST, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health. 2024;6(6):e428-e432. [FREE Full text] [doi: [10.1016/S2589-7500\(24\)00061-X](https://doi.org/10.1016/S2589-7500(24)00061-X)] [Medline: [38658283](https://pubmed.ncbi.nlm.nih.gov/38658283/)]
41. Ethics and governance of artificial intelligence for health. World Health Organization. URL: <https://www.who.int/publications/i/item/9789240029200> [accessed 2021-06-28]
42. Ragas. URL: <https://docs.ragas.io/en/stable/> [accessed 2025-01-26]
43. DeepEval. URL: <https://docs.confident-ai.com/> [accessed 2025-01-26]
44. Smith CM. Origin and uses of primum non nocere—above all, do no harm! J Clin Pharm. 2013;45(4):371-377. [doi: [10.1177/0091270004273680](https://doi.org/10.1177/0091270004273680)]
45. Sun Y, Sheng D, Zhou Z, Wu Y. AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. Humanit Soc Sci Commun. 2024;11(1):1-14. [doi: [10.1057/s41599-024-03811-x](https://doi.org/10.1057/s41599-024-03811-x)]
46. Maleki N, Padmanabhan B, Dutta K. AI hallucinations: a misnomer worth clarifying. ArXiv. Preprint posted online on January 9, 2024. [FREE Full text] [doi: [10.1109/cai59869.2024.00033](https://doi.org/10.1109/cai59869.2024.00033)]
47. Hatem R, Simmons B, Thornton JE. A call to address AI 'hallucinations' and how healthcare professionals can mitigate their risks. Cureus. 2023;15(9):e44720. [FREE Full text] [doi: [10.7759/cureus.44720](https://doi.org/10.7759/cureus.44720)] [Medline: [37809168](https://pubmed.ncbi.nlm.nih.gov/37809168/)]
48. Chauhan P, Bongo LA, Pedersen E. Ethical challenges of using synthetic data. 2023. Presented at: Proceedings of the AAAI Symposium Series AAAI Press; July 17-19, 2023:133-134; Washington, DC. [doi: [10.1609/aaais.v1i1.27490](https://doi.org/10.1609/aaais.v1i1.27490)]
49. Adane K, Gizachew M, Kendie S. The role of medical data in efficient patient care delivery: a review. Risk Manag Healthc Policy. 2019;12:67-73. [FREE Full text] [doi: [10.2147/RMHP.S179259](https://doi.org/10.2147/RMHP.S179259)] [Medline: [31114410](https://pubmed.ncbi.nlm.nih.gov/31114410/)]

50. Wang F, Zhang Z, Zhang X, Wu Z, Mo T, Lu Q, et al. A comprehensive survey of small language models in the era of large language models: techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. ArXiv. Preprint posted online on November 4, 2024. [[FREE Full text](#)]
51. Chute CG. Clinical classification and terminology: some history and current observations. *J Am Med Inform Assoc*. 2000;7(3):298-303. [[FREE Full text](#)] [doi: [10.1136/jamia.2000.0070298](https://doi.org/10.1136/jamia.2000.0070298)] [Medline: [10833167](https://pubmed.ncbi.nlm.nih.gov/10833167/)]
52. Buttazzo G. Rise of artificial general intelligence: risks and opportunities. *Front Artif Intell*. 2023;6:1226990. [[FREE Full text](#)] [doi: [10.3389/frai.2023.1226990](https://doi.org/10.3389/frai.2023.1226990)] [Medline: [37693010](https://pubmed.ncbi.nlm.nih.gov/37693010/)]
53. Fei N, Lu Z, Gao Y, Yang G, Huo Y, Wen J, et al. Towards artificial general intelligence via a multimodal foundation model. *Nat Commun*. 2022;13(1):3094. [[FREE Full text](#)] [doi: [10.1038/s41467-022-30761-2](https://doi.org/10.1038/s41467-022-30761-2)] [Medline: [35655064](https://pubmed.ncbi.nlm.nih.gov/35655064/)]
54. van Rooij I, Guest O, Adolfi F, de Haan R, Kolokolova A, Rich P. Reclaiming AI as a theoretical tool for cognitive science. *Comput Brain Behav*. 2024;7(4):616-636. [doi: [10.1007/s42113-024-00217-5](https://doi.org/10.1007/s42113-024-00217-5)]
55. Fjelland R. Why general artificial intelligence will not be realized. *Humanit Soc Sci Commun*. 2020;7(1):1-9. [doi: [10.1057/s41599-020-0494-4](https://doi.org/10.1057/s41599-020-0494-4)]
56. Zhong T, Liu Z, Pan Y, Zhang Y, Zhou Y, Liang S, et al. Evaluation of OpenAI o1: opportunities and challenges of AGI. ArXiv. Preprint posted online on September 27, 2024. [[FREE Full text](#)]
57. Schwartz DA. Physician-scientists: the bridge between medicine and science. *Am J Respir Crit Care Med*. 2012;185(6):595-596. [doi: [10.1164/rccm.201110-1806ED](https://doi.org/10.1164/rccm.201110-1806ED)] [Medline: [22422900](https://pubmed.ncbi.nlm.nih.gov/22422900/)]

Abbreviations

AGI: artificial general intelligence
AI: artificial intelligence
EHR: electronic health record
IIO: instruction-input-output
IFT: instruction fine-tuning
ITD: instruction-tuning dataset
LLM: large language model
SFT: supervised fine-tuning

Edited by A Mavragani; submitted 23.12.24; peer-reviewed by FTJ Faria, LP Gorrepati; comments to author 20.01.25; revised version received 28.01.25; accepted 07.02.25; published 18.03.25

Please cite as:

Nazar W, Nazar G, Kamińska A, Danilowicz-Szymanowicz L

How to Design, Create, and Evaluate an Instruction-Tuning Dataset for Large Language Model Training in Health Care: Tutorial From a Clinical Perspective

J Med Internet Res 2025;27:e70481

URL: <https://www.jmir.org/2025/1/e70481>

doi: [10.2196/70481](https://doi.org/10.2196/70481)

PMID:

©Wojciech Nazar, Grzegorz Nazar, Aleksandra Kamińska, Ludmila Danilowicz-Szymanowicz. Originally published in the Journal of Medical Internet Research (<https://www.jmir.org>), 18.03.2025. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.