Evaluation of Medical Large Language Models Taxonomy, Review, and Directions

Anisio Lacerda[1] , **Gisele Lobo Pappa**[1] , **Adriano César Machado Pereira**[1] , **Wagner Meira Jr**[1] , **Alexandre Guimarães de Almeida Barros**[2]
[1]Computer Science Department, Federal University of Minas Gerais (UFMG).
[2]Internal Medicine Department, INCT-NeuroTec-R - UFMG Faculty of Medicine
{anisio, glpappa, adrianoc, meira}@dcc.ufmg.br, xandebarros@gmail.com

Abstract

The integration of Large Language Models (LLMs) into medicine presents both great opportunities and significant challenges, particularly in ensuring these models are accurate, reliable, and safe. While LLMs have shown impressive capabilities in understanding and generating human language, their application in the medical domain requires careful evaluation due to the critical nature of medical applications which are inherently linked to patient life and health. Current evaluations of LLMs in medicine are often fragmented and insufficient, with a lack of standardized performance metrics, limited use of real patient data, and insufficient attention to important applications, such as documentation, education, and research. Furthermore, traditional NLP-based evaluations are often inadequate for assessing the text generated by LLMs. Therefore, a robust evaluation is essential to ensure the responsible and effective use of LLMs in medical settings, and to address the inherent challenges associated with their implementation. This paper explores the various dimensions of LLM evaluation in the medical domain, proposes a new taxonomy for categorizing medical applications, and discusses directions for future research in this critical area.

1 Introduction

Large Language Models have shown promise in several medical applications, but their evaluation requires a nuanced approach that goes beyond simple accuracy metrics. Other papers in the literature have already compared and evaluated how different LLM models perform in general and healthcare applications [Chang *et al.*, 2024; Bedi *et al.*, 2025]. This paper, in contrast, assumes that a Large Language Model (LLM) is given and we want to evaluate how effectively this model addresses the medical application at hand, which can be clinical diagnosis, aiding physicians in treatment choice, or helping students understand a specific medical topic.

The assessment of LLM results may follow an automatic or a manual, human-driven approach [Tam *et al.*, 2024]. The protocols and metrics used in these approaches differ, but there are at least three well-defined dimensions one wants to evaluate in any LLM application, namely: (i) model accuracy, (ii) robustness, and (iii) fairness and biases. Model accuracy measures how precise a model is on a given application, and the way it is measured is highly dependent on the application. For example, accuracy can be measured by traditional machine learning predictive measures, such as F1-score, in diagnosis problems, or by the ROUGE score in document summarization tasks. Robustness evaluates how the response of the model changes when data distribution changes, noise is added to data or other challenging inputs are presented to the model. Fairness and bias assess whether the model treats different data groups consistently, where groups can be defined by age, race, gender, etc.

The rigorous evaluation of Large Language Models (LLMs) is paramount for their responsible integration into the medical domain. The potential benefits of LLMs are substantial, but their use without thorough assessment carries risks of unreliability, inaccuracy, and harm. Given the dependence evaluation metrics have on different medical applications, this paper proposes a taxonomy of medical applications and then reviews how different model types (discriminative versus generative) used in different applications imply different evaluation metrics. This paper outlines the key considerations that underscore the importance of evaluating LLMs in this critical domain, emphasizing commonly used datasets and metrics most appropriate to different applications. Additionally, we identify the limitations and challenges of applying LLMs in medical tasks, highlighting why meticulous assessment is not optional, but a fundamental requirement for their safe and beneficial adoption.

This paper fills this gap by introducing a taxonomy of medical tasks, explicitly linking them to LLM evaluation metrics. Unlike previous works that either broadly cover healthcare applications or focus on specific tasks, this study provides a structured approach to medical applications with a strong LLM evaluation perspective. Figure 1 details the structure of this paper. We summarize our contributions to the Medical LLMs evaluation as:

- A comprehensive survey of the relevant literature.
- A novel taxonomy for the evaluation of medical LLMs in different applications.
- A discussion on promising lines for future research addressing current challenges.

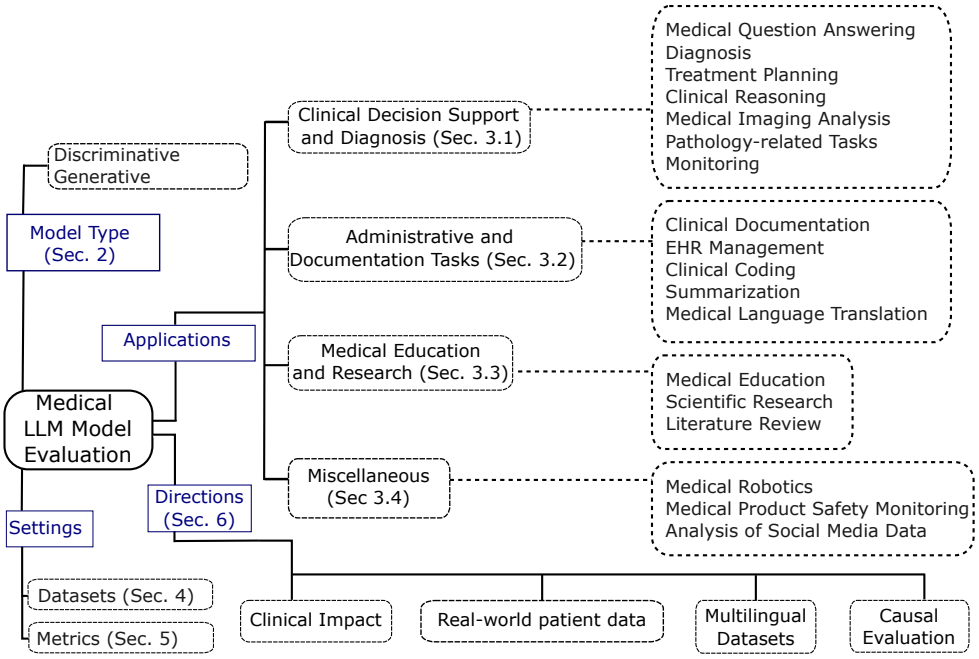

Figure 1: Structure of this paper.

- A publically-available repository of metrics, datasets, and an annotated taxonomy (see https://github.com/MALL-DCC-UFMG/medical-llms-evaluation).

2 Preliminaries

The type of model used to address the medical application, i.e., whether it follows a discriminative or generative model, provides a useful framework for understanding how LLMs process information and the kind of output they provide. It is also key to the type of evaluation that should be performed.

Discriminative Models Discriminative models focus on understanding and categorizing existing data. These tasks are about making distinctions, classifying information, and extracting relevant details from input. In a medical context, this often means analyzing patient records [Karabacak and Margetis, 2023], or medical literature to answer specific questions [Chen and Li, 2023], or categorize information. The goal is to find patterns, relationships, and categories within the given input, and provide structured outputs such as labels, classifications, or extracted data points. For instance, a discriminative task might involve identifying all instances of "diabetes" mentioned in a patient's medical history or classifying a clinical note as either a "diagnosis" or "treatment" record.

Generative Models Generative models, conversely, involve generating text or images based on the given input. These tasks are not about categorizing or extracting information, but rather about producing content that is supposed to be relevant and coherent. In a medical context, generative models might include summarizing a lengthy medical report, generating a simplified explanation of a complex medical condition for a patient [Gao *et al.*, 2023], or translating medical information into another language [Genovese *et al.*, 2024]. The emphasis is on the model's ability to synthesize new text that adheres to specific instructions and accurately conveys the necessary information.

Evaluation Discriminate and generative models require different approaches and evaluation metrics. Although both will be evaluated according to accuracy, robustness, and fairness and biases, their metrics differ. Discriminative models are often evaluated based on metrics such as accuracy, precision, and recall, which measure how well the model identifies and categorizes information. On the other hand, generative models are evaluated based on the quality of the generated text using metrics such as ROUGE scores, which assess the similarity between the generated text and reference summaries or human-created content.

Besides the three aforementioned evaluation dimensions, generative model evaluation should also consider other dimensions related to text generation, such as factuality and metrics referring to text quality. Factuality evaluates how information or answers provided by the model are consistent with real-world knowledge and is paramount for model trust. It is a way to detect model hallucinations.

3 LLMs Medical Applications Taxonomy

To better understand the capabilities and limitations of LLMs in medicine, we categorize their applications into four classes: (i) clinical decision support and diagnosis (CDSD) (§3.1), (ii) administrative and documentation tasks (ADT) (§3.2), (iii) medical education and research (MER) (§3.3), (iv) miscellaneous applications (§3.4).

3.1 Clinical Decision Support and Diagnosis

Clinical decision support and diagnosis (CDSD) encompass the use of LLMs to aid medical professionals in making in-

formed decisions about patient care [Goh *et al.*, 2024]. We further categorize the decision and diagnosis-based clinical data according to their specific purpose, each one with different evaluation needs.

Medical Question Answering (MQA). This application involves LLMs providing precise answers to clinical questions based on medical knowledge, literature, or patient data. MQA can assist clinicians in quickly finding information about symptoms, treatment options, and drug interactions [Singhal *et al.*, 2023]. Specifically, the task can be open-domain, in which the model answers without any specific reference text being provided, or closed-domain, in which answers are inferred from a supporting text, such as a research abstract. MQA evaluation often involves both automatic – metric usage to assess the accuracy of answers, and manual methods – real clinicians assessing the quality of responses. Multiple datasets are used to evaluate MQA models, such as MedQA [Jin *et al.*, 2021], PubMedQA [Jin *et al.*, 2019], and MedMCQA [Pal *et al.*, 2022]. These datasets were organized into one of the few benchmarks for medical LLMs, namely MultiMedQA [Singhal *et al.*, 2023]. HealthSearchQA [Singhal *et al.*, 2023], which focuses on real-world health queries, is also a popular dataset choice.

Diagnosis. LLMs can analyze patient data to suggest potential diagnoses. Diagnosis model evaluations have primarily been conducted in simplified medical settings, such as standardized hypothetical patient scenarios or structured clinical case assessments [Eriksen *et al.*, 2024; Kanjee *et al.*, 2023]. In both instances, all necessary diagnostic information is presented upfront, and the task involves selecting a single correct answer from a predefined list of options. However, LLMs deployed in a high-stakes clinical setting must not only demonstrate exceptional accuracy but also comply with diagnostic and treatment guidelines, exhibit robustness, and effectively follow instructions. These criteria have yet to be thoroughly evaluated in prior medical assessments. Tackling these limitations, in [Rao *et al.*, 2023] the authors evaluated an LLM across the diagnostic clinical workflow by using curated lists of potential answers and analyzing hypothetical clinical vignettes. Since it is unrealistic to assume that all necessary information is readily available in real-world clinical settings, efforts are being made to assess LLMs' ability to autonomously gather information and perform open-ended diagnoses [Hager *et al.*, 2024b].

Treatment Planning. LLMs can support treatment planning by suggesting personalized recommendations based on the latest clinical evidence and patient-specific factors. They have also been used to predict patient care trajectories and identify potential treatment complications [Andrew, 2024]. For instance, in [Pagano *et al.*, 2023], the authors propose the usage of LLMs in the orthopaedics recommendation treatment based on medical records from patients presenting hip or knee disorders. They measured agreement between the model and specialists through Cohen's Kappa coefficient. Furthermore, the evaluation of LLM-based treatment planning may also be dependent on the medical speciality. For instance, in mental health, certain evaluation criteria hold greater significance than others. A structured assessment framework for clinical LLMs – placing primary emphasis on risk and safety, followed by feasibility, acceptability, and effectiveness – is consistent with established guidelines for evaluating digital mental health applications [Stade *et al.*, 2024].

Clinical Reasoning. This task uses LLMs to simulate a diagnostic process, providing interpretable predictions or diagnoses. Traditional evaluation settings for clinical reasoning focused on single-turn evaluations, in which the model is provided with complete clinical information and asked to select the correct answer [Jin *et al.*, 2021]. However, this evaluation setting fails to measure proactive reasoning, such as asking follow-up questions when necessary. To better simulate real-world scenarios multi-turn evaluations were introduced. The authors in [Li *et al.*, 2024b] propose a Patient System – simulating incomplete medical records, and an Expert System – an LLM that must seek additional information before diagnosis. They evaluated the Patient system with factuality and relevance metrics while investigating LLM accuracy performance for varying levels of input information.

Medical Imaging Analysis (MIA). LLMs have been used to analyze several types of medical images, including X-rays, CT scans, and MRIs to assist in diagnosis and treatments. This includes discriminative tasks such as image classification, segmentation, and object detection. MIA tasks are often used in diagnostic and treatment workflows. For example, in the evaluation of LLM-based radiology report generation models, the MIMIC-III [Johnson *et al.*, 2016] and MIMIC-IV [Johnson *et al.*, 2023] datasets are predominantly used for training and assessment, as they represent the largest publicly accessible free-text electronic health records (EHRs). Standard automatic evaluation metrics encompass lexical approaches, including BLEU [Papineni *et al.*, 2002], ROUGE [Lin, 2004], and METEOR [Banerjee and Lavie, 2005], alongside semantic-based methods such as BERTScore [Zhang *et al.*, 2019]. To enhance domain-specific assessment, methods to provide automatic labeling of exams such as CheXbert [Smit *et al.*, 2020] and RadGraph [Jain *et al.*, 2021] have been introduced. Other works, such as RadCliQ [Yu *et al.*, 2023], propose application-driven metrics adapted from standard accuracy metrics, offering a more nuanced evaluation of report accuracy and clinical relevance within radiology.

Pathology-related Applications. Pathology data refers to structured and unstructured medical data derived from tissue samples examination, cells, and genetic information to diagnose diseases, assess disease progression, and guide treatment decisions. Given a pathology image, the task consists of answering questions about the clinical findings contained in the image, i.e., a visual question answering task [Delbrouck *et al.*, 2022]. Standard metrics for evaluating LLM-generated responses include accuracy [Malinowski and Fritz, 2014] and BLEU [Papineni *et al.*, 2002]. While these metrics assess agreement with a gold standard, they do not adequately capture the factual accuracy of generated outputs, which remains a critical challenge[Zhang *et al.*, 2024]. To address this gap, recent research has introduced domain-specific metrics, such as the F_1-RadGraph score [Delbrouck *et al.*, 2022], which

Table 1: Representative medical-domain datasets grouped by medical applications and detailed in terms of data origin, scale, modality; and machine learning task. Meanings: CDSD – Clinical Decision Support and Diagnosis, ADT – Administrative and Documentation Tasks, MER – Medical Education and Research. Data modalities: (T) Text, (I) Image, (A) Tabular, (S) Speech, (Q) Question-answering pairs, (P) Physics-based simulation data. Machine Learning model: (D) Discriminative, (G) Generative.

Applications	Dataset	Source	Granularity	Modality	ML Model
CDSD	MIMIC-III [Johnson *et al.*, 2016]	Patients admitted to ICU at Beth Israel Deaconess Medical Center in Boston, Massachusetts.	53,423 hospital admissions with 38,597 unique patients, and 53,423 distinct ICU stays.	T I A	D G
	MIMIC-IV [Johnson *et al.*, 2023]	Patients admitted to ICU at Beth Israel Deaconess Medical Center in Boston, Massachusetts.	431,231 hospital admissions with 180,733 unique patients, and 73,181 ICU admissions with 50,920 unique patients.	T I A	D G
	MIMI-CDM [Hager *et al.*, 2024a]	Created using the MIMIC-IV database.	2,400 unique patients presented to the emergency department with acute abdominal pain.	T I A	D G
	PathVQA [He *et al.*, 2020]	Images and captions extracted from textbooks and online digital libraries.	Over 30,000 questions for medical visual question answering.	I Q	D
	SLAKE [Liu *et al.*, 2021]	Curated using open-source medical datasets.	642 annotated images referring to 12 diseases and 39 organs.	I Q	D G
ADT	MEDALIGN [Fleming *et al.*,]	Curated from real instruction collection provided by clinicians.	983 natural language instructions for EHR data, curated by clinicians from 7 specialities.	T	G
	ProbSum [Gao *et al.*, 2023]	Extracted from the MIMIC-III dataset.	Progress notes with an average of over 1k tokens, and unlabeled numerical data (e.g test results).	T	G
	ICD billing codes [Soroush *et al.*, 2024]	Disease and procedural codes extracted from EHRs.	23,647 disease and 3,673 procedural codes.	T	D
	ACI-Bench [Yim *et al.*, 2023]	Medical notes were created using a seq2seq model, then reviewed and corrected by medical scribes and physicians.	207 doctor-patient conversations and corresponding patient visit notes.	T	G
	MTS-Dialog [Abacha *et al.*, 2023]	Conversations created using clinical note sections.	1,700 doctor-patient conversations (16k turns, 18k sentences) and summarized clinical notes (6k sentences).	T	G
	PriMock57 [Papadopoulos Korfiatis *et al.*, 2022]	Mock consultation recordings with 7 clinicians and 57 actors posing as patients.	57 simulated primary care visits, including recordings, transcripts, and notes.	T S	G
MER	MedQA-USMLE [Kung *et al.*, 2023]	Derived from the US Medical Licensing Examination (USMLE) containing medical cases and queries.	376 Publicly-available tests.	T	D
	Medical Flash Cards [Han *et al.*, 2023]	Rephrased Q&A pairs from the front/back of medical flashcards.	33,955 Q&A pairs.	T	G
	PubMedQA [Jin *et al.*, 2019]	Derived from PubMed abstracts and curated as a set of yes/no/maybe research questions.	1,000 expert-annotated, 61,200 unlabeled, and 211,300 artificially generated Q&A instances.	T	D
	NCBI Disease [Doğan *et al.*, 2014]	Manual curation.	793 PubMed abstracts annotated, 6,892 disease mentions, 790 uniq. disease concepts.	T	D
	S2ORC [Lo *et al.*, 2020]	English academic papers.	4.9M papers with 75B+ tokens.	T	D
Other	ORBIT-Surgical [Yu *et al.*, 2024]	Robotic surgical tasks.	14 simulation tasks.	P	D
	SMM4H [Weissenbacher *et al.*, 2019]	Sourced from Twitter	19,699 Twitter annotated posts.	T	D
	COMETA [Basaldella *et al.*, 2020]	Curated from health-themed subreddits.	20k biomedical entity mentions with links to SNOMED-CT medical knowledge graph.	T	D

evaluates the factual correctness and completeness of the generated clinical text.

Monitoring. LLMs have been used to monitor patients in real-time, analyzing data from wearable devices and other sources to identify health issues [Pilowsky *et al.*, 2024]. This can include monitoring vital signs, symptoms, and other health-related data to help improve accuracy and reliability in real-time patient care. However, there exist few medical LLMs that can process physiological time-series data, such as electrocardiograms (ECGs) [Li *et al.*, 2024a] and sphygmomanometers (PPGs) [Liu *et al.*, 2024].In [Li *et al.*, 2024a], the authors propose a multi-modal self-supervised learning model that uses electrocardiograms (ECGs) and automatically generates clinical reports. They evaluate the model with PTB-XL – for training and testing, and the MIT-BIH – for external validation. Given the discriminative nature of their task (i.e. classification), they used accuracy and F1-score for evaluation. In [Liu *et al.*, 2024], the authors investigate the use of LLMs for estimating blood pressure based on physiological features from electrocardiogram (ECG) and photoplethysmogram (PPG) signals using an instruction-tuned approach. As validation, they used the CAS-BP dataset and regression-related metrics.

3.2 Administrative and Documentation Tasks

The evaluation of LLMs for administrative and documentation tasks is essential due to the direct impact they have on clinical workflows, physician well-being, and patient safety. It has the potential to alleviate the administrative burden on clinicians, improve the accuracy of medical records, and enhance healthcare efficiency, all while mitigating the risk of errors and potential harm. The main challenges of these tasks are ensuring that LLMs capture essential clinical details

while handling contextual understanding and interpreting nuanced medical language and terminology effectively.

Clinical Documentation. Generating clinical documentation is a time-consuming task, and LLMs can assist professionals in drafting medical documents, streamlining research processes, and improving the efficiency of medical workflows. They can also help with report generation, such as radiology reports or discharge summaries. However, challenges persist, such as accuracy, maintaining the integrity of biomedical research, and addressing the potential for errors. Additionally, the authors in [Van Veen *et al.*, 2023] show that LLM-generated reports tend to be less concise compared to human-written ones.

Electronic Health Record(EHR) Management. Effective management of EHR is essential for organizing and managing patient data. LLMs can automatically identify and categorize critical information, such as symptoms, medications, diseases, and lab results, from patient EHRs. However, the complexity of medical language and the diversity of medical contexts can make it difficult for LLMs to accurately capture the nuances of clinical practice, and they may be prone to errors in interpretation [Karabacak and Margetis, 2023].

Clinical Coding. Clinical coding, such as the International Classification of Diseases (ICD), is a key task for standardizing diagnostics, procedures, treatments, medical billing and reimbursement. LLMs can assist in clinical coding and formatting to improve efficiency and accuracy in this process. However, there are challenges in implementing LLMs for clinical coding, including the specialized nature of the content, and the need to ensure that they are up to date with the latest standards. Clinical coding, which is frequently framed as a multi-label classification task, is often developed and evaluated using the MIMIC dataset. The performance of models is measured using metrics such as F1 score, Area Under the Curve (AUC), and precision@k, which take into account either the top k most frequent labels or the entire label set [Wang *et al.*, 2024].

Summarization. Summarization of medical texts is an important task that can reduce the burden of analyzing large volumes of clinical information. LLMs have shown promise in medical text summarization, including radiology reports, progress notes, and doctor-patient dialogue [Gao *et al.*, 2023]. However, there are challenges in evaluating the quality of summaries and ensuring that they are complete, correct, and concise. Furthermore, current automatic evaluation methods for clinical report generation primarily focus on lexical metrics, which can lead to biased and inaccurate assessment of the contextual information present in the reports.

Medical Language Translation. Medical language translation can be divided into two key areas: translating medical terminology between languages and adapting medical dialogue for easier interpretation by non-professionals [Herrera-Espejel and Rach, 2023; Genovese *et al.*, 2024]. Both are essential for effective communication across different groups. Ensuring accuracy, preserving clinical meaning, and maintaining consistency across languages are critical. Human evaluation by bilingual medical experts is necessary to validate nuanced medical concepts, while comprehension tests with laypersons help assess the effectiveness of simplifying medical jargon for patients.

3.3 Medical Education and Research.

The evaluation of LLMs in Medical Education and Research is essential, due to the unique challenges within these domains, requiring a high degree of accuracy, reliability, and ethical considerations [Preiksaitis and Rose, 2023]. It is necessary to ensure that LLMs enhance, rather than hinder, learning and discovery in medical contexts, while also addressing the specific demands of these tasks. In medical education, the primary challenges include the risk of diminished critical thinking and reasoning skills in students, the potential for misinformation and bias, a lack of emotional understanding in LLM interactions, and the importance of balancing personalized learning with human guidance.

Medical Education. The use of LLMs in medical education is a rapidly growing area with the potential to enhance learning and teaching. LLMs can provide summaries, presentations, translations, explanations, and step-by-step guides on many topics, with customizable depth and style [Preiksaitis and Rose, 2023]. They can also be used to create interactive learning simulations, for instance allowing students to practice conversations with virtual patients [Preiksaitis and Rose, 2023]. Here, ensuring the accuracy of the information, mitigating potential biases, and preventing over-reliance on the output of LLMs continue to be challenges.

Scientific Research. LLMs can significantly enhance the efficiency and effectiveness of scientific research in the medical domain by analyzing large scientific datasets, generating research hypotheses, and summarizing complex research findings [Cascella *et al.*, 2023]. However, there is a need for experimental validation to ensure scientific accuracy and the risk of plagiarism or copyright issues [Mishra *et al.*, 2024]. It is also important to address issues of transparency and explainability to ensure the integrity of medical research.

Literature Review. LLMs can be useful tools for literature review, aiding in the summarization of vast amounts of information into a concise, readable format. They can assist in identifying relevant papers, extracting key findings, and generating comprehensive reviews of specific topics [Chen and Li, 2023]. Despite these potential benefits, it is important to acknowledge challenges such as the possibility of fabricated information and the lack of sophisticated evaluation methods that can account for the nuances of clinical information.

Medication Management. Medication management is an area where LLMs can support patients by providing information on medication dosages, side effects, and interactions, as well as helping them adhere to their medication schedules. These systems can also monitor drug safety and provide alerts about potential adverse effects. A crucial challenge here is the need for LLMs to provide up-to-date information and advice in a clear, concise manner, avoiding the generation of any information that might be inaccurate, incomplete, or misleading.

3.4 Miscellaneous Applications.

While some medical applications, such as Q&A, have been assessed using metrics like accuracy [Kanjee *et al.*, 2023; Samaan *et al.*, 2023], many applications lack well-defined evaluation methods that capture the nuances of human judgment and the complexity of real-world scenarios. For example, LLMs have been used in medical robotics and for medical product safety monitoring, but the means of evaluating their performance is not always clear. Moreover, the use of LLMs for analyzing social media data introduces further evaluation difficulties, considering the subjective nature of the content.

Medical Robotics. The integration of LLMs with medical robotics has the potential to transform surgical procedures and patient care. LLMs can provide real-time surgical navigation information and analyze physiological parameters, offering intraoperative support to surgeons [Cheng *et al.*, 2023]. This capability is especially important in complex procedures, potentially enhancing surgical precision and efficiency. A primary challenge lies in ensuring the accuracy and reliability of LLM outputs, as errors in a surgical context can have serious consequences. Further, the real-time processing demands of surgical robotics require LLMs to have low latency and robust performance in dynamic environments.

Medical Product Safety Monitoring. LLMs can analyze vast amounts of data from various sources, such as clinical trials, patient reports, and social media to monitor the safety of medical products. This is especially relevant in the post-market surveillance phase where real-world data becomes available [Raval *et al.*, 2021]. Identifying adverse events, drug interactions, and potential safety risks are key advantages of LLMs in this space. The challenge is ensuring that LLMs can distinguish between actual safety signals and background noise in the data, as well as address bias' issues in the datasets. Furthermore, the need to adhere to regulatory requirements and ethical guidelines in data handling adds complexity to the application of LLMs for medical product safety monitoring.

Analysis of Social Media Data. LLMs can be used to analyze social medical data, offering insights into public health trends, patient experiences, and emerging health concerns. This capability allows for real-time monitoring of health-related discussions, which can be valuable for identifying potential outbreaks, understanding public perception of health interventions, and assessing patient needs. The use of social media data brings with it many challenges, including the need to filter out misinformation, handle inherent biases in social media data, and ensure that patient privacy is not compromised. Moreover, interpreting the informal language and diverse expressions on social media requires sophisticated natural language processing techniques.

4 Datasets and Benchmarks

When analyzing medical datasets (and benchmarks, defined as sets of datasets used for model evaluation) for evaluating LLMs, several important dimensions need to be considered to ensure the models have a set of desired properties. Here, we present the datasets following the taxonomy defined for medical applications and also categorize them according to the data source, granularity, modality, and type of modeling.

By understanding the medical application (e.g., diagnosis, medical education), the data source (e.g. EHRs, medical literature), and the type of machine learning model (i.e., discriminative vs generative) being addressed, we gain insights into the specific skills an LLM is being used and tested for. Among other possibilities, these skills include knowledge recall, reasoning, text generation, and image analysis. Additionally, categorizing by data granularity (e.g., # patients) and modality (e.g., text, images) allow nuanced evaluation, revealing how LLMs perform across different granularity of medical data and with various types of input. Table 1 shows 19 datasets (and/or benchmarks) widely used by researchers in the medical domain and that have already been processed by LLMs in the four domains of medical application previously defined [1].

Among the 19 datasets, most focus on textual data, with exceptions being those that use the MIMIC and its derivations and PriMock57, which has audio recordings. For ADT, most works follow a generative approach, as these applications are usually interested in generating patient notes, documents or summaries [Abacha *et al.*, 2023]. For medical education, most datasets are made of Q&A pairs, and most applications follow a discriminative model –whether the answer to the question is correct or not [Preiksaitis and Rose, 2023]. We also find datasets from social media, where discriminative tasks such as [Weissenbacher *et al.*, 2019; Basaldella *et al.*, 2020] have been previously addressed. For clinical decision-making, data from MIMIC and its derivations are those with more modalities explored with both discriminative and generative approaches [Johnson *et al.*, 2016; Johnson *et al.*, 2023; He *et al.*, 2020].

5 Evaluation Measures

The proper choice of measures/metrics ensures that LLMs' performance is assessed accurately and in alignment with specific medical application requirements. As previously shown, a significant portion of evaluations found in the literature rely solely on accuracy as the primary metric [Samaan *et al.*, 2023; Kanjee *et al.*, 2023], especially in tasks involving medical knowledge assessment. This includes multiple-choice Q&A formats, often mirroring medical licensing exams. However, LLM evaluation in medicine is moving toward a more comprehensive, multi-faceted approach that combines traditional metrics with qualitative assessments and addresses issues such as bias and safety. There is an agreement among most researchers that evaluations must move beyond simple benchmarks and examine the real-world clinical impact of these models [Hager *et al.*, 2024a; Hager *et al.*, 2024b].

Concerning the three dimensions we define as important for the evaluation of LLMs, only accuracy is accounted for in most studies. Regarding robustness, we did not find any study directly evaluating it automatically. This is done mainly

[1] This is not an exhaustive list. Please, find an extended list of datasets in our repository.

Table 2: Summary of evaluation metrics for language models in the medical domain.

Metric	Description	Text Quality			Application-specific
		Syntactic	Semantic	Factuality	
BLEU	Measures the overlap of n-grams between generated and reference text, useful for tasks like machine translation or text generation.	✓			
ROUGE-L	Evaluates similarity based on the longest common subsequence, considering both precision and recall. It is useful for evaluating text summarization.	✓			
chrF++	A character n-gram based metric that is useful for evaluating text generation and translation tasks.	✓			
BERTScore	Uses contextual BERT embeddings to evaluate semantic similarity between generated and reference texts. It is appropriate for assessing overall meaning and context.		✓		
METEOR	Evaluates text similarity, but unlike BLEU, it also considers stemming and synonyms.		✓		
MPNetv2	A transformer-based metric that assesses the similarity between two texts based on their semantic representation. This metric is useful for measuring the consistency of LLM responses over time.		✓		
CIDEr	This metric evaluates the consensus between generated content and a gold standard in image captioning tasks.		✓		
MEDCON	It gauges the consistency of medical concepts by using QuickUMLS to extract biomedical concepts via string-matching algorithms. It is useful in evaluating the medical terminology usage.		✓		
F1-RadGraph score	Qualitatively assesses the factual correctness and completeness of generated radiology reports.			✓	✓
Exact String Match	Measures if the generated output is an exact match to a reference answer, useful in question-answering tasks where precise answers are required.				✓
Token-level F1	Measures the overlap of tokens between generated and reference text, used for tasks such as question answering.				✓
Word Error Rate (WER)	Measures the accuracy of automatic speech recognition in clinical dialogue transcription tasks.				✓
UniEval	A source-augmented metric that uses T5-large to evaluate model output source.				✓
COMET	A source augmented metric that uses XLM-RoBERTa to evaluate model output considering source.				✓

by human evaluators [Tam *et al.*, 2024]. Regarding fairness and biases, although there are metrics that can be adapted to consider these dimensions in medical data, specially in discriminative tasks, they are not used. Among these metrics we mention demographic parity difference (DPD), which measures if model predictions vary across different population groups, and equalized odds differences (EOD), which guarantees the model presents equal error rates across different populations [Chang *et al.*, 2024].

Table 2 provides a comprehensive catalog of these standard evaluation criteria currently used, which serves as a valuable reference. Note the metrics are defined only for generative models, because discriminative models follow traditional metrics, such as accuracy and f-measure [Chang *et al.*, 2024]. For generative models, we classify metrics according to which way they evaluate text quality, i.e., considering syntactic or semantic features, and whether an evaluation of factuality – one of the paramount dimensions – is evaluated. Note that measures for semantic quality are currently applied on top of purely syntactic ones. Also observe that domain-specific metrics are difficult to avoid in certain applications, as they may be the only way to automatically evaluate these methods.

6 Future Directions

Knowing the evaluation of LLMs in medical applications is still far from established, we discuss future research directions to improve evaluation metrics and methodologies.

Clinical Impact. The evaluation of LLMs in medicine should go beyond (e.g., question-answering) accuracy and focus on real-world clinical impact. Randomized controlled trials (RCTs) can assess their effects on mortality, morbidity, and other outcomes like efficiency and satisfaction. Observational studies can further explore their benefits and risks across diverse clinical settings. Additionally, developing evaluation frameworks for accuracy, robustness, fairness, and equity is essential, incorporating participatory methods to align with patient values and specific clinical use cases.

Real-world patient data. Current evaluations of LLMs often rely on question answering related to medical exams, which may not reflect the complexities of real patient data and interaction. Additionally, even studies that used datasets such as MIMIC are observational investigations focused on a specific patient population, which will have challenges to generalize to distinct scenarios of application, including diverse ethnic groups, and clinical protocols.

Multilingual datasets. Current datasets used to train medical LLMs are primarily in English and Chinese languages. The development of multilingual medical datasets is of primary importance to enable the LLMs to capture the domain-specific nuances of medical terms across other languages. Further, expanding clinical datasets in languages other than English and Chinese is needed to ensure the gains of AI are shared equitably among communities.

Causal Evaluation. Medical data complexity can lead to unstable causal identification and erroneous correlation inferences when using traditional inductive bias for learning models. Also, since most studies are based on observational studies, it is key to avoid confounding and selection bias impacts. Hence, the use of causal inference is critical in the medical domain for determining the true causal effects of medical interventions (e.g., treatments) through data analysis.

References

[Abacha *et al.*, 2023] A B Abacha, W Yim, Y Fan, and T Lin. An empirical study of clinical note generation from doctor-patient encounters. In *Proc. of ACL*, pages 2291–2302, 2023.

[Andrew, 2024] A Andrew. Potential applications and implications of large language models in primary care. *Family Medicine and Community Health*, 12, 2024.

[Banerjee and Lavie, 2005] S Banerjee and A Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[Basaldella *et al.*, 2020] M Basaldella, F Liu, E Shareghi, and N Collier. COMETA: A corpus for medical entity linking in the social media. In *Proc. of EMNLP*, pages 3122–3137, 2020.

[Bedi *et al.*, 2025] S Bedi, Y Liu, L Orr-Ewing, D Dash, S Koyejo, A Callahan, JA Fries, M Wornow, et al. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*, 333(4):319–328, 01 2025.

[Cascella *et al.*, 2023] Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of medical systems*, 47(1):33, 2023.

[Chang *et al.*, 2024] Y Chang, X Wang, J Wang, Y Wu, LZ Yang, H Kand Chen, X Yi, C Wang, Y Wang, W Ye, Y Zhang, Y Chang, PS. Yu, Q Yang, and X Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), 2024.

[Chen and Li, 2023] Minyu Chen and Guoqiang Li. Chatgpt for mechanobiology and medicine: A perspective. *Mechanobiology in Medicine*, 1(1):100005, 2023.

[Cheng *et al.*, 2023] K Cheng, Z Sun, Y He, S Gu, and H Wu. The potential impact of chatgpt/gpt-4 on surgery: will it topple the profession of surgeons? *Int. Journal of Surgery*, 109(5):1545–1547, 2023.

[Delbrouck *et al.*, 2022] J Delbrouck, P Chambon, C Bluethgen, E Tsai, O Almusa, and CP Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. *arXiv preprint arXiv:2210.12186*, 2022.

[Doğan *et al.*, 2014] RI Doğan, R Leaman, and Z Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Jour. of biomedical informatics*, 47:1–10, 2014.

[Eriksen *et al.*, 2024] Alexander V Eriksen, Sören Möller, and Jesper Ryg. Use of gpt-4 to diagnose complex clinical cases, 2024.

[Fleming *et al.*,] S L Fleming, A Lozano, W J Haberkorn, JA Jindal, E Reis, E Thapa, L Blankemeier, JZ Genkins, E Steinberg, A Nayak, et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *AAAI'24*.

[Gao *et al.*, 2023] Y Gao, D Dligach, T Miller, and M Afshar. Overview of the problem list summarization (ProbSum) 2023 shared task on summarizing patients' active diagnoses and problems from electronic health record progress notes. In *Workshop on Biomedical NLP and BioNLP Shared Tasks*, 2023.

[Genovese *et al.*, 2024] A Genovese, S Borna, C Gomez-Cabello, S Haider, S Prabha, A Forte, and B Veenstra. Artificial intelligence in clinical settings: a systematic review of its role in language translation and interpretation. *A. of Trans. Medicine*, 2024.

[Goh *et al.*, 2024] E Goh, R Gallo, J Hom, E Strong, Y Weng, H Kerman, JA Cool, Z Kanjee, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969, 2024.

[Hager *et al.*, 2024a] P Hager, F Jungmann, R Holland, K Bhagat, I Hubrecht, M Knauer, J Vielhauer, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

[Hager *et al.*, 2024b] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

[Han *et al.*, 2023] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

[He *et al.*, 2020] X He, Y Zhang, L Mou, E Xing, and P Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[Herrera-Espejel and Rach, 2023] PS Herrera-Espejel and S Rach. The use of machine translation for outreach and health communication in epidemiology and public health: Scoping review. *JMIR Public Health and Surveillance*, 9(1):e50814, 2023.

[Jain *et al.*, 2021] S Jain, A Agrawal, A Saporta, SQH Truong, DN Duong, T Bui, P Chambon, Y Zhang, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

[Jin *et al.*, 2019] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[Jin *et al.*, 2021] D Jin, E Pan, N Oufattole, W Weng, H Fang, and P Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[Johnson *et al.*, 2016] AEW Johnson, TJ Pollard, L Shen, LH Lehman, M Feng, M Ghassemi, B Moody, P Szolovits, L Anthony Celi, and RG Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[Johnson *et al.*, 2023] AEW Johnson, L Bulgarelli, L Shen, A Gayles, A Shammout, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[Kanjee *et al.*, 2023] Zahir Kanjee, Byron Crowe, and Adam Rodman. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *Jama*, 330(1):78–80, 2023.

[Karabacak and Margetis, 2023] Mert Karabacak and Konstantinos Margetis. Embracing large language models for medical applications: opportunities and challenges. *Cureus*, 15(5), 2023.

[Kung *et al.*, 2023] TH Kung, M Cheatham, A Medenilla, C Sillos, L De Leon, C Elepaño, et al. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2), 2023.

[Li *et al.*, 2024a] Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ecg zero-shot learning. In *Medical Imaging with Deep Learning*, pages 402–415. PMLR, 2024.

[Li *et al.*, 2024b] SS Li, V Balachandran, S Feng, JS Ilgen, E Pierson, PW Koh, and Y Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *NeurIPS*, 2024.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004.

[Liu *et al.*, 2021] B Liu, L Zhan, L Xu, L Ma, Y Yang, and X Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *Int. Symposium on Biomedical Imaging*, pages 1650–1654, 2021.

[Liu *et al.*, 2024] Z Liu, C Chen, J Cao, M Pan, J Liu, N Li, F Miao, and Y Li. Large language models for cuffless blood pressure measurement from wearable biosignals. In *Proc of the Int. Conf. on Bioinformatics, Computational Biology and Health Informatics*, pages 1–11, 2024.

[Lo *et al.*, 2020] K Lo, LL Wang, M Neumann, R Kinney, and D Weld. S2ORC: The semantic scholar open research corpus. In *Proc. of ACL*, pages 4969–4983, 2020.

[Malinowski and Fritz, 2014] M Malinowski and M Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *NeurIPS*, 27, 2014.

[Mishra *et al.*, 2024] T Mishra, E Sutanto, R Rossanti, N Pant, A Ashraf, A Raut, G Uwabareze, A Oluwatomiwa, and B Zeeshan. Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers. *Scientific Reports*, 14(1):31672, 2024.

[Pagano *et al.*, 2023] D Pagano, S Holzapfel, T Kappenschneider, M Meyer, G Maderbacher, J Grifka, and DE Holzapfel. Arthrosis diagnosis and treatment recommendations in clinical practice: an exploratory investigation with the generative ai model gpt-4. *Journal of Orthopaedics and Traumatology*, 24(1):61, 2023.

[Pal *et al.*, 2022] A Pal, LK Umapathi, and M Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conf. on health, inference, and learning*, pages 248–260. PMLR, 2022.

[Papadopoulos Korfiatis *et al.*, 2022] A Papadopoulos Korfiatis, F Moramarco, R Sarac, and A Savkov. PriMock57: A dataset of primary care mock consultations. In *Proc. of ACL (Volume 2: Short Papers)*, pages 588–598, 2022.

[Papineni *et al.*, 2002] K Papineni, S Roukos, T Ward, and W Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, 2002.

[Pilowsky *et al.*, 2024] Julia K Pilowsky, Jae-Won Choi, Aldo Saavedra, Maysaa Daher, Nhi Nguyen, Linda Williams, and Sarah L Jones. Natural language processing in the intensive care unit: A scoping review. *Critical Care and Resuscitation*, 2024.

[Preiksaitis and Rose, 2023] Carl Preiksaitis and Christian Rose. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR medical education*, 9:e48785, 2023.

[Rao *et al.*, 2023] A Rao, M Pang, J Kim, M Kamineni, W Lie, A K Prasad, A Landman, K Dreyer, and MD Succi. Assessing the utility of chatgpt throughout the entire clinical workflow: development and usability study. *JMIR*, 25:e48659, 2023.

[Raval *et al.*, 2021] S Raval, H Sedghamiz, E Santus, T Alhanai, M Ghassemi, and E Chersoni. Exploring a unified sequence-to-sequence transformer for medical product safety monitoring in social media. In *Conf. on Empirical Methods in NLP*, pages 3534–3546, 2021.

[Samaan *et al.*, 2023] JS Samaan, YH Yeo, N Rajeev, L Hawley, S Abel, WH Ng, N Srinivasan, J Park, M Burch, R Watson, et al. Assessing the accuracy of responses by the language model chatgpt to questions regarding bariatric surgery. *Obesity surgery*, 33(6):1790–1796, 2023.

[Singhal *et al.*, 2023] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[Smit *et al.*, 2020] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

[Soroush *et al.*, 2024] Ali Soroush, Benjamin S Glicksberg, Eyal Zimlichman, Yiftach Barash, Robert Freeman, Alexander W Charney, Girish N Nadkarni, and Eyal Klang. Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI*, 1(5):AIdbp2300040, 2024.

[Stade *et al.*, 2024] EC Stade, SW Stirman, LH Ungar, CL Boland, HA Schwartz, DB Yaden, J Sedoc, RJ DeRubeis, R Willer, and JC Eichstaedt. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Res.*, 3(1), 2024.

[Tam *et al.*, 2024] TYC Tam, S Sivarajkumar, S Kapoor, AV Stolyar, K Polanska, KR McCarthy, et al. A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digital Medicine*, 7(1):258, 2024.

[Van Veen *et al.*, 2023] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Clinical text summarization: adapting large language models can outperform human experts. *Research Square*, 2023.

[Wang *et al.*, 2024] Hanyin Wang, Chufan Gao, Christopher Dantona, Bryan Hull, and Jimeng Sun. Drg-llama: tuning llama model to predict diagnosis-related group for hospitalized patients. *npj Digital Medicine*, 7(1):16, 2024.

[Weissenbacher *et al.*, 2019] D Weissenbacher, A Sarker, A Magge, A Daughton, K O'Connor, M Paul, and G Gonzalez. Overview of the 4th social media mining for health shared tasks at acl 2019. In *Proc. of social media mining for health applications workshop & shared task*, pages 21–30, 2019.

[Yim *et al.*, 2023] W Yim, Y Fu, Asma B Abacha, N Snider, T Lin, and M Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586, 2023.

[Yu *et al.*, 2023] F Yu, M Endo, R Krishnan, I Pan, A Tsai, EP Reis, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.

[Yu *et al.*, 2024] Q Yu, M Moghani, K Dharmarajan, V Schorp, WC Panitch, J Liu, K Hari, H Huang, M Mittal, et al. Orbit-surgical: An open-simulation framework for learning surgical augmented dexterity. *arXiv preprint arXiv:2404.16027*, 2024.

[Zhang *et al.*, 2019] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[Zhang *et al.*, 2024] Kai Zhang, R Zhou, E Adhikarla, Z Yan, Y Liu, J Yu, Z Liu, X Chen, BD Davison, H Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13, 2024.