LLM-Empowered Medical Patient Communication: A Data-Centric Survey From a Clinical Perspective

Anonymous ACL submission

Abstract

The integration of large language models (LLMs) into medical patient communication has shown promising potential for enhancing healthcare accessibility. Despite significant advancements in LLM capabilities, real-world clinical adoption remains challenging due to gaps between in-lab LLM training and the complexities of clinical practice. This survey provides a systematic and data-centric review of 21 text-based medical datasets that support LLM training and evaluation for patient communication. From a clinical perspective, we propose a novel taxonomy for classifying these datasets based on key clinical properties and upon which identify the training objectives they support. Additionally, we introduce a full lifecycle framework for optimizing the development of medical LLMs through alignment across dataset selection, fine-tuning methodologies, benchmark and evaluation metrics, highlighting the impact of alignment on model performance and training effectiveness. Finally, we provide guidance on enhancing medical datasets through clinically informed annotations and adaptive learning techniques to support the development of safe, clinically aligned LLMs for patient-centered communication in real-world healthcare settings.

1 Introduction

Augmenting medical patient communication with large language models (LLMs) has emerged as a promising solution to handle the growing demand for scalable and accessible healthcare services (Busch et al., 2025; Huo et al., 2025). By automating aspects of symptom consultation, treatment recommendations, and psychiatric support, these models can mitigate workforce shortages and improve patient outcomes (Omar et al., 2024). Recent advances, exemplified by MedPaLM2 (Singhal et al., 2025), Meditron (Chen et al., 2023b), Med42 (Christophe et al., 2024a,b), and GPT-4 (Nori et al., 2023), demonstrate that LLMs can achieve medical knowledge comparable to that of healthcare professionals, as evidenced by high performance on knowledge-based benchmarks such as MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and United States Medical Licensing Examination (USMLE). State-of-the-art medical LLMs have been shown to match or even surpass human experts in knowledge accuracy, response relevance, and social attributes such as empathy and supportiveness (Singhal et al., 2025; Paiola et al., 2024; Calle et al., 2024).

However, clinical feasibility studies of these LLM-powered tools in real-world healthcare settings have yielded mixed results, revealing inconsistencies in intervention effectiveness, human acceptance, and clinical applicability (Busch et al., 2025; Liu et al., 2024b). These findings underscore a persistent gap between in-lab LLM development and the complex, dynamic demands of clinical practice (Shi et al., 2024a,b). Notably, the performance of medical LLMs is primarily shaped by the training datasets, which ranges from standardized medical exams and scholarly articles to clinical documentation and real-world patient-provider interactions. Therefore, there is an urgent need to thoroughly understand the distinct clinical properties of these datasets that support LLM-empowered medical patient communication (Wu et al., 2024a).

Despite the urgency, the above-mentioned discrepancy has not been properly addressed due to the lack of clinical understanding of the diverse properties possessed by the medical datasets used in LLM training to support patient communication. These datasets vary significantly in their clinical properties—some, such as PubMedQA (Jin et al., 2019) and MedQA (Jin et al., 2021), are knowledge-based, consisting of scholarly content tailored for healthcare professionals, while others, such as medical dialogue datasets (e.g., NoteChat (Wang et al., 2023a), Psych8K (Liu et al., 2023), CMtMedQA

(Yang et al., 2024b)), focus on the real-world social and conversational dynamics in patient communication. Failing to systematically understand and leverage these distinctions risks misalignment between model training and real-world clinical application.

To address the current gap, this paper provides a systematic and data-centric review of 21 text-based medical datasets that support LLM training and evaluation in medical patient communication. From a clinical perspective, we *first* present a taxonomy for classifying and analyzing existing datasets based on key clinical properties such as inquiry types, communication dynamics, and target audiences, upon which we identify the training objectives supported by these datasets. *Second*, we propose a full lifecycle framework for optimizing the development of medical LLMs through alignment across critical steps, including the selection of training datasets, fine-tuning methodologies, benchmarks, and evaluation metrics. Based on a meta-analytical review of previous experiments, we demonstrate the fundamental impact of alignment and misalignment on model performance and training effectiveness. *Finally*, we provide guidance for data enhancement to bridge the gap between LLM training and clinical application by incorporating clinically informed and standardized data annotations and employing adaptive learning techniques to develop safe, clinically aligned medical LLMs that support patient-centered communication.

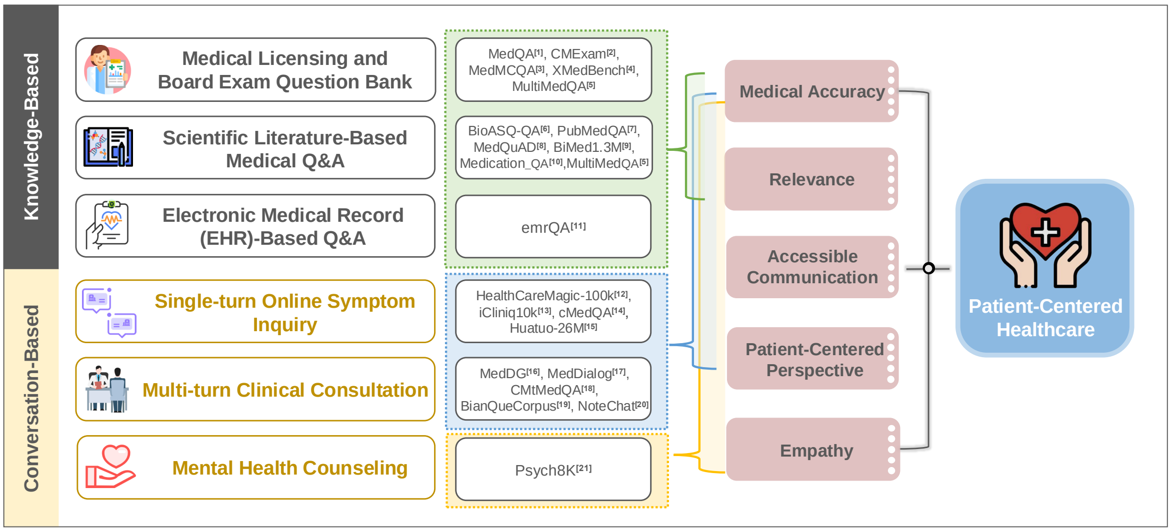
Our contributions are summarized as three-fold:

- **A Novel Data-Centric Taxonomy.** We introduce a taxonomy that categorizes medical patient communication datasets based on key clinical properties, providing a foundation for understanding their roles in developing LLMs.
- **Systematic Methodological Review.** We present a comprehensive review of the full lifecycle development of medical LLMs, emphasizing alignment across dataset selection, fine-tuning methodologies, and evaluation metrics.
- **Guidance for Dataset Enhancement.** We propose a framework for enhancing medical datasets through clinically informed annotations and adaptive learning techniques to ensure model alignment with clinical practices and support patient-centered communication.

Structure of This Survey. In Section 2, we propose a taxonomy for 21 text-based medical datasets that support LLMs in patient communication, analyzing their clinical properties, data types, annotations, and communication qualities to identify the training objectives they support. Section 3 presents a full lifecycle framework for medical LLM development, covering dataset selection, fine-tuning methodologies, and evaluation metrics, and examining the impact of alignment and misalignment across these components on model performance and training effectiveness. Section 4 provides guidance on enhancing medical datasets to improve clinical alignment, emphasizing clinically informed annotations, adaptive learning techniques, and evaluation metrics that reflect real-world healthcare needs. Finally, Section 5 reviews existing surveys on medical datasets used for LLM benchmarking and highlights the novelty of this study.

2 Medical Patient Communication Datasets: A Taxonomy

Medical patient communication datasets are text-based datasets that capture, support, or simulate communication related to medical care between healthcare providers and patients. These datasets include medical Q&A, patient-provider dialogues, medical exam questions, mental health counseling transcripts, and other text forms relevant to patient education, diagnosis, treatment, and health management. Such datasets are employed to develop medical LLMs that support effective patient communication, which requires not only the accurate transmission of medical knowledge but also the contextualization and accessible communication of complex medical information (Ha and Longnecker, 2010). Evidence-based practices of patient communication, essential for patient-centered care and improved health outcomes, demand medical accuracy, communication accessibility, and a patient-centered approach. Reflecting the underlying communication processes, we categorize existing medical patient communication datasets into (i) *knowledge-based* and (ii) *conversation-based* ones. Knowledge-based datasets prioritize the accuracy of clinical tasks such as disease diagnosis and clinical reasoning, while conversation-based datasets capture the communication dynamics and clinical principles across diverse patient-provider interactions. This section provides a comprehensive review of 21 medical datasets (see Appendix A for details), analyzing their clinical properties, data type, annotation methods, and target audiences to assess their suitability for various LLM training



Knowledge-based Dataset: [1]MedQA (Jin et al., 2021),[2]CMExam (Liu et al., 2024a),[3]MedMCQA (Pal et al., 2022),[4]XMedBench (Wang et al., 2024d), [5]MultiMedQA (Singhal et al., 2022), [6]BioASQ-QA (Krithara et al., 2023), [7]PubMedQA (Jin et al., 2019), [8]MedQuAD (Abacha et al., 2019), [9]BiMed1.3M (Pieri et al., 2024), [10]Medication_QA (Abacha et al., 2019), [11]emrQA (Pampari et al., 2018);
Conversation-based Dataset: [12]HealthCareMagic-100k (Li et al., 2023c), [13]iCliniq10k (Li et al., 2023c), [14]CMedQA (Zhang et al., 2017), [15]Huatuo-26M (Li et al., 2023a), [16]MedDG (Liu et al., 2022), [17]MedDialog (Zeng et al., 2020), [18]CMtMedQA (Yang et al., 2024b), [19]BianQueCorpus (Chen et al., 2023a), [20]NoteChat (Wang et al., 2023a), [21]Psych8K (Liu et al., 2023).

Figure 1: The proposed taxonomy and a comprehensive overview of medical patient communication datasets.

objectives. An overview of the proposed taxonomy and datasets is presented in Figure 1. Below we review the two types of datasets by first introducing their mainstreams, followed by discussing their alignment with clinical practices and current limitations, respectively.

2.1 Knowledge-Based Datasets

Knowledge-based medical patient communication datasets are text-based corpora employed to ensure medical accuracy, technical precision, and clinical reasoning in LLMs. These datasets, formatted as medical Q&A to address patient queries, are curated from three primary sources. First, multiple-choice question answering (MCQA) datasets, derived from medical licensing and board exam question banks, present questions in a multiple-choice format to assess factual recall, critical thinking, and clinical reasoning. For example, the MedQA dataset (Jin et al., 2021), sourced from the United States Medical Licensing Examination (USMLE), Mainland China Medical Licensing Examination (MCMLE), and Taiwan Medical Licensing Examination (TWMLE), comprises 60K MCQA questions. Second, open-domain question answering (Open Q&A) datasets, such as BioASQ-QA (Krithara et al., 2023) and PubMedQA (Jin et al., 2019), leverage vast repositories such as MEDLINE and PubMed to generate scientific literature-based medical Q&A pairs. Third, EHR-based Q&A datasets are derived from electronic health records (EHRs), which are digital collections of patient information offering comprehensive, real-time health data, including medical history, diagnoses, medications, allergies, and laboratory results.

2.1.1 Clinical Properties and Supported Training Objectives

Knowledge-based datasets convey domain-specific knowledge and are often curated with expert annotations, such as question labels (e.g., disease groups, clinical departments, medical disciplines, areas of competency, and question difficulty levels, as in (Krithara et al., 2023)) or structured annotations (e.g., question type, concept, Q&A, supporting material, reference, as in (Liu et al., 2024a)). The rigorous annotation process ensures high-quality, validated medical content, making these datasets essential for training and evaluating LLMs in domain-specific language understanding, medical knowledge retrieval, structured problem-solving, clinical reasoning, and interpretability. State-of-the-art medical LLMs are predominantly trained and evaluated using knowledge-based datasets, often achieving clinical reasoning and medical accuracy comparable to or exceeding that of healthcare professionals (Singhal et al., 2025; Christophe et al., 2024b).

2.1.2 Gaps in Supporting Patient Communication

While knowledge-based datasets provide a structured and efficient foundation for model training and benchmarking, they deviate significantly from real-world clinical practice, limiting the applicability of medical LLMs in healthcare settings (Shi et al., 2024b). Effective patient communication requires not only accurate medical content but also contextualized, personalized, and accessible delivery (Kurtz, 2002; Matusitz and Spear, 2014). Current knowledge-based datasets fall short in supporting patient communication due to: 1) insufficient contextualized reasoning; and 2) inadequate accessibility in communication.

Insufficient Contextualized Reasoning. Exam-style datasets prioritize factual recall and structured reasoning, but fail to account for the socio-cultural, psychological, and structural barriers that influence medical decision-making. Real-world patient communication requires more than adherence to clinical guidelines; it involves understanding patient-centered factors such as health literacy, socio-cultural beliefs, and psychological barriers (Ha and Longnecker, 2010), which are absent in standardized MCQA and Open Q&A datasets. Consequently, LLMs trained on these datasets may generate generic, decontextualized responses that oversimplify complex diagnostic reasoning and lack the dynamic, evolving nature of shared decision-making and clinical interactions.

Inadequate Accessibility in Communication. Knowledge-based datasets are essentially tailored to healthcare professionals and prioritize technical precision over linguistic accessibility. Open Q&A corpora, such as BioASQ, PubMedQA, and MedQuAD, derive information from scientific literature, which is often dense, highly specialized, inaccessible to lay users, and lacking both readability and social attributes such as empathy and emotional support. Thus, LLMs trained on knowledge-based datasets are insufficient for delivering accessible communication tailored to non-expert audiences. Hence these models struggle to produce patient-friendly responses, limiting their effectiveness in patient communication (Christophe et al., 2024b).

2.2 Conversation-Based Datasets

Conversation-based medical datasets capture real-world patient-provider interactions, reflecting how patients describe symptoms, express concerns, and seek reassurance. They emphasize naturalistic dialogue flow, accessible communication, and empathetic response. These datasets encompass various forms of medical dialogues: (1)single-turn online symptom inquiries, (2) multi-turn clinical consultations, and (3) mental health counseling. Single-turn online symptom inquiry datasets feature brief, one-question-one-answer exchanges where patients describe symptoms and doctors provide asynchronous responses. For example, HealthcareMagic-100k comprises 100K real-world medical inquiries from an online health platform (Li et al., 2023c). Multi-turn clinical consultation datasets involve extended dialogues with iterative exchanges between doctors and patients, including patient history gathering, follow-up questions, diagnostic or treatment recommendations, and shared decision making (e.g., BianQueCorpus (Chen et al., 2023a)). Mental health counseling datasets provide transcripts of therapeutic conversations, which capture counseling techniques such as active listening, cognitive behavioral therapy, and empathetic counseling. For example, Psych8K (Liu et al., 2023) contains 8K conversation fragments constructed from 260 real-world in-depth counseling sessions.

2.2.1 Clinical Properties and Supported Training Objectives

Conversation-based datasets contain multi-turn clinical consultations and entity-labeled exchanges across diverse medical cases and patient populations. For example, MedDG, an entity-centric medical dialogue dataset, provides expert annotations on disease, symptoms, medicine, examination, and attributes, facilitating the retrieval of medical knowledge and providing guidance on conversational flow (Liu et al., 2022). Similarly, Psych8K uses GPT-4 annotations on seven counseling metrics, such as approval & reassurance, direct guidance, and restatement, reflection & listening (Liu et al., 2023). These real-world doctor-patient conversations, along with AI or expert annotations, effectively train LLMs in patient communication skills, such as generating follow-up questions, clarifying symptoms, and tailoring explanations to patients' literacy levels, thereby significantly improving model performance on dialogue coherence, contextual adaptation, patient engagement, and adherence to clinical guidelines (Wang et al., 2023a; Chen et al., 2023a; He et al., 2024).

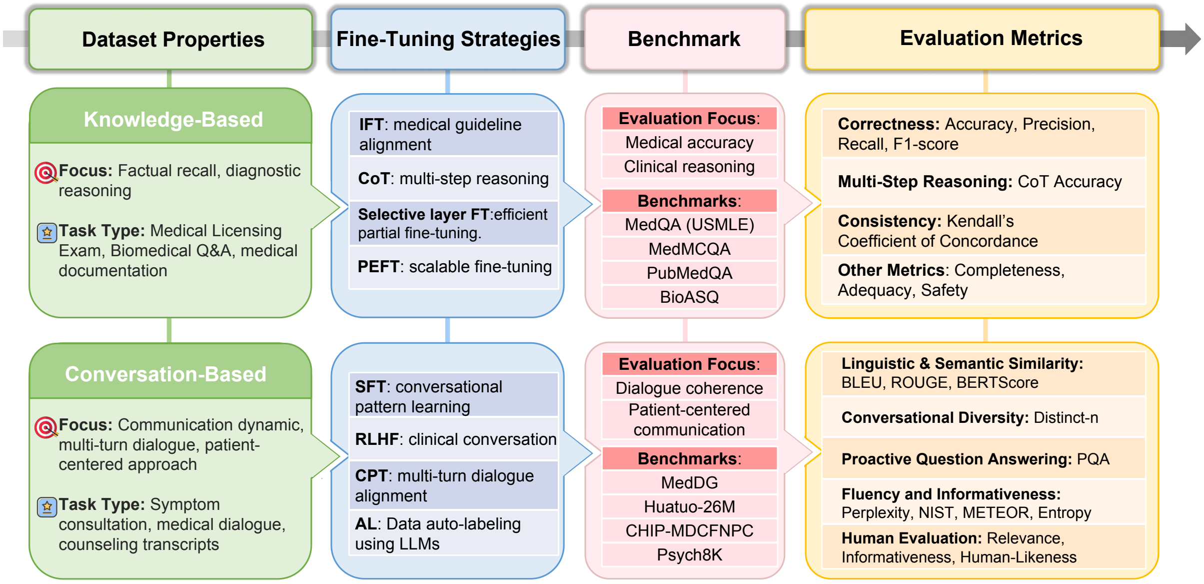


Figure 2: An overview for the full lifecycle development of LLMs for medical patient communication.

2.2.2 Gaps in Supporting Patient Communication

Existing conversation-based datasets often lack annotations grounded in key clinical principles that guide effective patient communication, such as empathy, cultural sensitivity, and shared decision-making. This poses significant challenges for LLM training. Unlike knowledge-based datasets, which rely on standardized clinical guidelines for clear-cut annotations, patient communication is highly contextualized, individualized, and socio-culturally constructed (Ha and Longnecker, 2010). Consequently, real-world medical dialogues vary significantly based on patient expectations, physician styles, and shared cultural norms. This inherently contextualized and personalized nature complicates the development of standardized annotation schema, as there are no absolute right or wrong responses. As a result, clinical principles are often inconsistently represented, inadequately annotated, and exhibit low inter-rater reliability in existing conversation-base datasets.

3 Development Lifecycle of Medical Patient Communication LLMs

Building clinically applicable LLMs requires a strategic alignment among training data, fine-tuning methodologies, and evaluation benchmarks. Here we introduce a data-centric lifecycle for developing medical patient communication LLMs, highlighting key stages such as training data, fine-tuning methodologies, benchmark and evaluation metrics, all grounded in the clinical properties of the employed datasets and their supported training objectives (refer to Figure 2 for an overview).

3.1 Data-Centric Strategies for Fine-Tuning

The fine-tuning strategies employed for LLMs differ significantly based on the properties of datasets used (Alghisi et al., 2024). Below we highlight the representative fine-tuning strategies for knowledge- and conversation-based datasets, respectively.

Fine-Tuning on Knowledge-Based Datasets. Knowledge-based datasets, such as MedQA, PubMedQA, BioASQ, and MedMCQA, primarily focus on improving factual recall and clinical reasoning. Models trained on these datasets frequently employ instruction fine-tuning (IFT) to enhance performance on medical Q&A tasks (Kamble and AlShikh, 2023; Singhal et al., 2025). For example, Med-PaLM and LLaMA 7B utilize IFT to align responses with medical guidelines (Singhal et al., 2025; Li et al., 2023c). Additionally, models like Med-PaLM leverage chain-of-thought (CoT) prompting, enabling step-by-step reasoning to handle multi-step medical queries effectively (Singhal et al., 2025). Advanced techniques such as selective layer fine-tuning and domain-specific vocabulary integration, as seen in MultiMedQA, help models adapt to complex diagnostic queries (Hamzah and Sulaiman). Efficient tuning approaches, such as parameter-efficient fine-tuning (PEFT) with quantized low-rank adaptation (QLoRA), are also utilized in models like ChatBode-7B to preserve general capabilities while improving performance (Paiola et al., 2024).

Fine-Tuning on Conversation-Based Datasets. In contrast, conversation-based datasets, such as BianQueCorpus, MedDialog, Psych8K, and CMtMedQA, emphasize dialogue quality, multi-turn conversation handling, and patient-centered communication. Fine-tuning on these datasets often uses supervised fine-tuning (SFT) to help models learn appropriate conversational patterns, follow-up questioning, and context-aware responses (Ye et al., 2023; Zhang et al., 2023; Li et al., 2023b; Chen et al., 2023a). Additionally, models such as CMtMedQA and IIMedGPT utilize reinforcement learning with human feedback (RLHF) to improve conversation quality and alignment with clinical standards (Zhang et al., 2025; Zhao et al., 2024; Yang et al., 2024b). Some models, such as Ziya-LLaMA, implement conversational preference training (CPT) to align responses with clinical principles during multi-turn dialogues (Tian et al., 2023; Acikgoz et al., 2024; Yang et al., 2024b).

3.2 Strategies for Benchmark Selection

The alignment between fine-tuning datasets and evaluation benchmarks significantly impacts model performance. Models generally demonstrate optimal performance when their training objectives, as supported by fine-tuning datasets align with evaluation benchmarks. For example, models fine-tuned on knowledge-based datasets excel on similarly structured knowledge-intensive benchmarks (e.g., USMLE, PubMedQA, and MedMCQA) (Jin et al., 2019; Christophe et al., 2024a; Guo et al., 2023; Wang et al., 2024c; Kamble and AlShikh, 2023), as evidenced by MedPaLM2's high accuracy (86.5%) on MedQA (Singhal et al., 2025). Likewise, conversation-oriented models, such as BianQue (Chen et al., 2023a) and Zhongjing (Yang et al., 2024b), which are trained on multi-turn medical dialogues, perform well on medical dialogue benchmarks (e.g., MedDG, CMtMedQA, huatuo-26M) assessing coherence and patient engagement (Wang et al., 2023a; Zeng et al., 2020; Dou et al., 2023; He et al., 2024; Tian et al., 2023). However, model performance declines when there is a misalignment of clinical properties between training datasets and benchmarks. For instance, ChiMed-GPT, which is fine-tuned on dialogue data, underperforms on knowledge-based tasks like MEDQA due to minimal exposure to MCQA formats (Tian et al., 2023). Conversely, models trained exclusively on knowledge-centric datasets (e.g., Med42) struggle on dialogue-heavy benchmarks, evidenced by low BLEU and ROUGE scores, due to their inability to generate context-aware responses (Kim et al., 2024). The impact of misalignment on model performance underscores the importance of using benchmarks that match the training objectives.

3.3 Evaluation Strategies

The choice of evaluation metrics also depends on the properties of the fine-tuning datasets. Knowledge-based LLMs are often assessed upon metrics like accuracy, precision, recall, and F1-score, particularly for multiple-choice and open-ended medical Q&A benchmarks such as MedQA, PubMedQA, and USMLE (Liu et al., 2024a). Reasoning-intensive benchmarks sometimes employ CoT accuracy to evaluate LLM's logical consistency, as in TCMBench (Yue et al., 2024) and Med-PaLM's multi-step reasoning tasks (Singhal et al., 2025). By contrast, communication-oriented models are commonly evaluated using BLEU, ROUGE, BERTScore, and distinct n-gram that capture linguistic overlap and conversational diversity (Zhao et al., 2024; Tian et al., 2023; Dou et al., 2023). Some benchmarks also incorporate specialized metrics like proactive questioning ability (PQA) to assess LLM's capacity to encourage patient engagement (Chen et al., 2023a). Counseling-oriented datasets, such as Psych8K, rely on automatically generated annotations on counseling metrics (e.g., active listening, approval & reassurance), which do not necessarily map to factual accuracy. Misalignment occurs when knowledge-based metrics are applied to dialogue-oriented models (or vice versa), providing an incomplete picture of performance. For example, models fine-tuned for exam-style Q&A (e.g., Med42) excel in accuracy-based metrics but fare poorly when measured by BLEU or ROUGE, highlighting the mismatch between their training objectives, supported by the properties of fine-tuning datasets, and evaluation metrics (Kim et al., 2024).

3.4 Data-Centric Performance Analysis

The alignment among fine-tuning datasets, benchmarks, and evaluation metrics directly impacts model performance. Models achieve optimal performance when the properties of their training datasets align with both benchmarks and evaluation metrics (Li et al., 2023c; Zeng et al., 2020;

Dou et al., 2023; Jin et al., 2019; Singhal et al., 2025). Previous studies have observed a performance drop when there is a misalignment between training objectives and evaluation. Aqulia-Med, for instance, achieves only moderate accuracy on knowledge-based benchmarks because it is training on conversation-orientated data does not translate to multiple-choice Q&A formats (Zhao et al., 2024). Similarly, Med42, trained exclusively on knowledge-based datasets, scores poorly on conversational benchmarks due to a lack of exposure to multi-turn dialogue (Kim et al., 2024). Misalignment between evaluation metrics and training objectives also distorts performance assessments. Psych8K, which evaluates conversational skills using GPT-4-generated counseling metrics, fails to demonstrate superior performance on medical knowledge benchmarks, as its training objective emphasizes counseling skills rather than medical knowledge (Liu et al., 2023).

4 Data-Centric Future Opportunities

Advancing LLMs to support medical patient communication with higher quality requires a comprehensive approach that aligns dataset curation, fine-tuning methodologies, and model evaluation with clinical principles. Despite the recent progress in this domain, current translational efforts face significant limitations due to the gaps between in-lab LLM training and real-world clinical applications (Kim et al., 2025; Hager et al., 2024). Addressing these gaps requires interdisciplinary collaboration to ensure that models are not only accurate but also clinically applicable, providing patient-centered and accessible communication (Bajwa et al., 2021; Alowais et al., 2023). This section proposes key strategies to bridge LLM training with clinical practice, emphasizing the need for high-quality annotations, contextual integration, adaptive learning approaches, and data-centric fine-tuning strategies.

4.1 Enhancing Knowledge-Based Datasets

Enhancing knowledge-based datasets to support LLMs in contextualized reasoning and accuracy involves several critical strategies.

Standardizing Annotation Protocols. Implementing standardized annotation protocols enhances dataset quality and minimizes the risk of embedding biases into LLMs. Current knowledge-based datasets contain expert annotations on medical context, question types, and clinical activities. Additional annotations such as institutional and sociocultural factors should be included to capture variations in clinical guidelines across different regions and healthcare settings. This comprehensive approach ensures that LLMs are trained on data reflective of diverse clinical practices, thereby improving their generalizability and fairness. Moreover, cross-institutional data integration necessitates protocols to address potential noise and formatting inconsistencies, ensuring the synthesized dataset maintains high quality and uniformity.

Enhancing Health Equity. The composition of training datasets profoundly influences an LLM's performance, especially concerning health equity and patient-centered care. Under-representation of certain patient demographics, medical institutions, and regional clinical practices can introduce biases, limiting the model's applicability across diverse healthcare settings. To mitigate this, it is imperative to actively include data from underrepresented populations and regions. Techniques such as oversampling or stratified curation can be employed to balance the dataset. Subsequently, appropriate fine-tuning strategies can be adopted to assign appropriate weights to the sampled subsets, promoting equitable performance across various patient groups and clinical scenarios.

Facilitating Context-Aware Reasoning. Facilitating contextualized reasoning in LLMs requires synthesizing multimodal data to create contextually rich datasets that capture the full spectrum of the clinical reasoning process. Integrating EHR with patient-centric data such as demographics, medical history, and psychosocial factors provides essential context about individual patients.

4.2 Enhancing Conversation-Based Datasets

Enhancing conversation-based datasets is crucial for training LLMs to effectively support patient-centered communication. Incorporating expert annotations and linguistic indicators of patient engagement can improve the alignment of LLM outputs with clinical principles.

Incorporating Conversational-Level Annotations. Developing datasets that reflect evidence-based practices in patient-centered communication necessitates collaborative effort among healthcare professionals, health communication researchers, and patients. This evidence-based and patient-centered approach ensures that annotations capture both clinical guidelines and patient experience (Alowais et al., 2023). In particular, future datasets should incorporate: (1) clinical perspectives that evaluate conversation adherence to clinical guidelines; and (2) patient perspectives that assess communication accessibility and adaptation to patient needs. The high-quality conversational-level annotations can be further employed to train reward models, facilitating LLM alignment with clinical principles and a patient-centered approach.

Integrating Linguistic Indicators of Patient Engagement. Previous research identified linguistic features in doctor-patient communication indicative of patient engagement, such as linguistic synchrony, word usage patterns, and dialogue coherence (Khaleghzadegan et al., 2024; Falkenstein et al., 2016). For example, research has shown that physicians' linguistic adaptation to patients' health literacy significantly improves communication effectiveness (Schillinger et al., 2021). Automated extraction of these linguistic metrics enables the creation of embedding representations that quantify patient engagement. These embeddings could further serve as essential inputs for training reward models that guide LLM fine-tuning, enhancing patient engagement and conversational adaptability (Coppolillo et al., 2024; Tennenholtz et al., 2025).

5 Related Work

Recent surveys and systematic reviews have explored medical datasets used in training and evaluating large language models (LLMs). For instance, (Yan et al., 2024) reviewed medical benchmarks across multiple modalities, including text (e.g., electronic health records, doctor-patient dialogues), images (e.g., X-rays, MRIs), and multimodal data (e.g., audio, video, ECG, omics), emphasizing their significance, data structures, and applications in clinical tasks such as diagnosis, medical report generation, and clinical summarization. Similarly, (Zhang et al., 2024) categorized medical datasets based on data sources (e.g., EHR, scientific literature, web data), structures (e.g., conversational text, multimodal data), and their roles in LLM pre-training, fine-tuning, and evaluation. (Wang et al., 2024b) provided a comprehensive survey of training corpora and evaluation benchmarks in medical LLMs, covering corpus sources (e.g., medical Q&A, knowledge graphs, clinical guidelines), data preparation (e.g., cleaning, augmentation, translation), training paradigms (e.g., instruction fine-tuning, PEFT, RLHF), and evaluation methods (e.g., machine and human-centric evaluations). Additionally, (Spasic and Nenadic, 2020) and (Wu et al., 2024a) reviewed clinical text data, such as clinical notes, pathology reports, and discharge summaries, identifying key obstacles in clinical NLP, including data scarcity, lack of synthetic data, and insufficient annotations.

While previous surveys have laid a strong foundation, our work distinguishes itself by offering a clinical perspective on medical patient communication datasets for LLM training, emphasizing their clinical properties and alignment with real-world healthcare practices. We introduce a novel taxonomy that categorizes datasets based on clinical properties, data type and annotations, and target audiences, bridging the gap between medical LLM training and clinical applicability. Our review extends beyond dataset classification by providing a comprehensive analysis of fine-tuning strategies, critically examining how dataset properties influence the selection of training objectives, fine-tuning methodologies, benchmarks, and, more importantly, how their alignment affects model performance and training effectiveness. Grounded in a patient-centered approach, our work aims to advance the development of medical LLMs that are not only technically proficient but also clinically aligned and effectively augment patient communication in real-world healthcare settings, addressing critical gaps left by previous research.

6 Conclusion

The integration of LLMs into medical patient communication necessitates a data-centric approach to ensure clinical applicability. This survey makes three key contributions: (1) we introduce a novel taxonomy for classifying medical patient communication datasets based on key clinical properties that determine their supported training objectives; (2) we propose a full lifecycle framework for developing medical LLMs, encompassing dataset selection, fine-tuning methodologies, and evaluation metrics, while highlighting the critical impact of alignment across these stages on model performance and clinical applicability; and(3) we provide guidance on enhancing medical datasets to support model alignment with clinical practices, emphasizing the importance of clinically informed annotations, standardized data curation, and adaptive learning techniques. This survey lays the foundation for developing safe, clinically aligned, and patient-centered LLM-powered medical communication systems.

7 Limitations

This survey is subject to several limitations. *First*, the current review is limited to a meta-analytical approach and could be strengthened by incorporating empirical experiments that examine the impact of alignment and misalignment among fine-tuning datasets, benchmarks, and evaluation metrics on LLM performance. *Additionally*, the proposed taxonomy could further benefit from iterative refinement through interviews with healthcare practitioners and researchers specializing in LLM training. Incorporating feedback from both clinical practice and LLM development communities would enhance the taxonomy's applicability and relevance across diverse medical contexts. *Lastly*, the proposed annotation framework, while foundational, could be further detailed into domain-specific annotation protocols. Tailoring these protocols to diverse medical contexts and clinical settings would ensure more precise and contextually appropriate expert annotation.

References

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 25–29. IOS Press.

Emre Can Acikgoz, Osman Batur İnce, Rayene Bench, Arda Anıl Boz, İlker Kesen, Aykut Erdem, and Erkut Erdem. 2024. Hippocrates: An open-source framework for advancing large language models in healthcare. *arXiv preprint arXiv:2404.16621*.

Simone Alghisi, Massimo Rizzoli, Gabriel Roccabruna, Seyed Mahed Mousavi, and Giuseppe Riccardi. 2024. Should we fine-tune or RAG? evaluating different techniques to adapt LLMs for dialogue. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 180–197, Tokyo, Japan. Association for Computational Linguistics.

S. A. Alowais, S. M. Alghamdi, F. D. Alqahtani, N. M. Alzahrani, W. R. Alsharif, A. M. Alotaibi, A. S. Alqahtani, A. I. Alzahrani, A. A. Alghamdi, and A. S. Alzahrani. 2023. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, 23(1):1–9.

Junaid Bajwa, Usman Munir, Aditya Nori, and Bryan Williams. 2021. Artificial intelligence in healthcare: transforming the practice of medicine. *Future healthcare journal*, 8(2):e188–e194.

Kuluhan Binici, Abhinav Ramesh Kashyap, Viktor Schlegel, Andy T Liu, Vijay Prakash Dwivedi, Thanh-Tung Nguyen, Xiaoxue Gao, Nancy F Chen, and Stefan Winkler. 2024. Medsage: Enhancing robustness of medical dialogue summarization to asr errors with llm-generated synthetic dialogues. *arXiv preprint arXiv:2408.14418*.

Felix Busch, Lena Hoffmann, Christopher Rueger, Elon HC van Dijk, Rawen Kader, Esteban Ortiz-Prado, Marcus R Makowski, Luca Saba, Martin Hadamitzky, Jakob Nikolas Kather, et al. 2025. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5(1):26.

Paul Calle, Ruosi Shao, Yunlong Liu, Emily T Hébert, Darla Kendzor, Jordan Neil, Michael Businelle, and Chongle Pan. 2024. Towards ai-driven healthcare: Systematic optimization, linguistic analysis, and clinicians' evaluation of large language models for smoking cessation interventions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, Xiangmin Xu, et al. 2023a. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *arXiv preprint arXiv:2310.15896*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. 2024a. Med42–evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024b. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*.

Erica Coppolillo, Marco Minici, Federico Cinus, Francesco Bonchi, and Giuseppe Manco. 2024. Engagement-driven content generation with large language models. *arXiv preprint arXiv:2411.13187*.

Chengfeng Dou, Zhi Jin, Wenping Jiao, Haiyan Zhao, Zhenwei Tao, and Yongqiang Zhao. 2023. Plugmed: Improving specificity in patient-centered medical dialogue generation using in-context learning. *arXiv preprint arXiv:2305.11508*.

Angelica Falkenstein, Brandon Tran, Daniel Ludi, Afshin Molkara, Henry Nguyen, Arnold Tabuenca, and Kate Sweeny. 2016. Characteristics and correlates of word use in physician-patient communication. *Annals of Behavioral Medicine*, 50(5):664–677.

Yichen Gao, Licheng Zong, and Yu Li. 2024. Enhancing biomedical question answering with parameter-efficient fine-tuning and hierarchical retrieval augmented generation. *CLEF Working Notes*.

Quan Guo, Shuai Cao, and Zhang Yi. 2022. A medical question answering system using large language models and knowledge graphs. *International Journal of Intelligent Systems*, 37(11):8548–8564.

Z Guo, P Wang, Y Wang, and S Yu. 2023. Improving small language models on pubmedqa via generative data augmentation. arxiv.

Jennifer Fong Ha and Nancy Longnecker. 2010. Doctor-patient communication: a review. *Ochsner journal*, 10(1):38–43.

Philipp Hager, Florian Jungmann, Robert Holland, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9):2613–2622.

Farizal Hamzah and Nuraini Sulaiman. Optimizing llama 7b for medical question answering: A study on fine-tuning strategies and performance on the multimedqa dataset.

Yuhong He, Yongqi Zhang, Shizhu He, and Jun Wan. 2024. Bp4er: Bootstrap prompting for explicit reasoning in medical dialogue generation. *arXiv preprint arXiv:2403.19414*.

Bright Huo, Amy Boyle, Nana Marfo, Wimonchat Tangamornsuksan, Jeremy P Steen, Tyler McKechnie, Yung Lee, Julio Mayol, Stavros A Antoniou, Arun James Thirunavukarasu, et al. 2025. Large language models for chatbot health advice studies: A systematic review. *JAMA Network Open*, 8(2):e2457879–e2457879.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

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

Kiran Kamble and Waseem AlShikh. 2023. Palmyra-med: Instruction-based fine-tuning of llms enhancing medical domain performance.

Sogol Khaleghzadegan, Mitchell Rosen, Allison Links, Aisha Ahmad, Megan Kilcullen, Emily Boss, Mary Catherine Beach, and Somnath Saha. 2024. Validating computer-generated measures of linguistic style matching and accommodation in patient-clinician communication. *Patient Education and Counseling*, 119:108074.

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2025. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *arXiv preprint arXiv:2502.04381*.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. Medexqa: Medical question answering benchmark with multiple explanations. *arXiv preprint arXiv:2406.06331*.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Suzanne M Kurtz. 2002. Doctor-patient communication: principles and practices. *Canadian Journal of Neurological Sciences*, 29(S2):S23–S29.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023a. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*.

Wenqiang Li, Lina Yu, Min Wu, Jingyi Liu, Meilan Hao, and Yanjie Li. 2023b. Doctorgpt: A large language model with chinese medical question-answering capabilities. In *2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, pages 186–193. IEEE.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023c. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024a. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. 2024b. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2024c. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1114.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.

Jonathan Matusitz and Jennifer Spear. 2014. Effective doctor–patient communication: an updated examination. *Social work in public health*, 29(3):252–266.

Andreea Maria Monea and Anca Nicoleta Marginean. 2021. Medical question entailment based on textual inference and fine-tuned biomed-roberta. In *2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 319–326. IEEE.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Mahmud Omar, Girish N Nadkarni, Eyal Klang, and Benjamin S Glicksberg. 2024. Large language models in medicine: A review of current clinical trials across healthcare applications. *PLOS Digital Health*, 3(11):e0000662.

Pedro Henrique Paiola, Gabriel Lino Garcia, Joao Renato Ribeiro Manesco, Mateus Roder, Douglas Rodrigues, and João Paulo Papa. 2024. Adapting llms for the medical domain in portuguese: A study on fine-tuning and model evaluation. *arXiv preprint arXiv:2410.00163*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

Keivalya Pandya. 2023. Peft-medaware: Large language model for medical awareness. *arXiv preprint arXiv:2311.10697*.

Sara Pieri, Sahal Shaji Mullappilly, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. Bimedix: Bilingual medical mixture of experts llm. *arXiv preprint arXiv:2402.13253*.

Víctor Ponce-López. 2024. Large language models for multi-choice question classification of medical subjects. *arXiv preprint arXiv:2403.14582*.

Dean Schillinger, Nicholas D. Duran, Danielle S. McNamara, Scott A. Crossley, Renu Balyan, and Andrew J. Karter. 2021. Precision communication: Physicians' linguistic adaptation to patients' health literacy. *Science Advances*, 7(51):eabj2836.

Tongyue Shi, Jun Ma, Zihan Yu, Haowei Xu, Minqi Xiong, Meirong Xiao, Yilin Li, Huiying Zhao, and Guilan Kong. 2024a. Stochastic parrots or icu experts? large language models in critical care medicine: A scoping review. *arXiv preprint arXiv:2407.19256*.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024b. Medical dialogue system: A survey of categories, methods, evaluation and challenges. *Findings of the Association for Computational Linguistics ACL 2024*, pages 2840–2861.

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. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Sarvesh Soni and Kirk Roberts. 2020. Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering. In *Proceedings of the twelfth language resources and evaluation conference*, pages 5532–5538.

Irena Spasic and Goran Nenadic. 2020. Clinical text data in machine learning: systematic review. *JMIR Medical Informatics*, 8(3):e17984.

Guy Tennenholtz, Yinlam Chow, Chih-Wei Hsu, Lior Shani, Yi Liang, and Craig Boutilier. 2025. Embedding-aligned language models. In *Advances in Neural Information Processing Systems 37*, pages 15893–15946.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2023. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences. *arXiv preprint arXiv:2311.06025*.

Anan Wang, Yunong Wu, Xiaojian Ji, Xiangyang Wang, Jiawen Hu, Fazhan Zhang, Zhanchao Zhang, Dong Pu, Lulu Tang, Shikui Ma, et al. 2024a. Assessing and optimizing large language models on spondyloarthritis multi-choice question answering: Protocol for enhancement and assessment. *JMIR Research Protocols*, 13(1):e57001.

Jinqiang Wang, Huansheng Ning, Yi Peng, Qikai Wei, Daniel Tesfai, Wenwei Mao, Tao Zhu, and Runhe Huang. 2024b. A survey on large language models from general purpose to medical applications: Datasets, methodologies, and evaluations. *arXiv preprint arXiv:2406.10303*.

Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024c. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2023a. Notechat: a dataset of synthetic doctor-patient conversations conditioned on clinical notes. *arXiv preprint arXiv:2310.15959*.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. 2023b. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024d. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Jiageng Wu, Shihui Liu, Wanxin Li, et al. 2024a. Clinical text datasets for medical artificial intelligence and large language models — a systematic review. *NEJM AI*.

Jiageng Wu, Xian Wu, and Jie Yang. 2024b. Guiding clinical reasoning with large language models via knowledge seeds. *arXiv preprint arXiv:2403.06609*.

Jiageng Wu, Xian Wu, Yefeng Zheng, and Jie Yang. 2024c. Medkp: Medical dialogue with knowledge enhancement and clinical pathway encoding. *arXiv preprint arXiv:2403.06611*.

Yunfei Xie, Juncheng Wu, Haoqin Tu, Siwei Yang, Bingchen Zhao, Yongshuo Zong, Qiao Jin, Cihang Xie, and Yuyin Zhou. 2024. A preliminary study of o1 in medicine: Are we closer to an ai doctor? *arXiv preprint arXiv:2409.15277*.

Niraj Yagnik, Jay Jhaveri, Vivek Sharma, Gabriel Pila, Asma Ben, and Jingbo Shang. 2024. Medlm: Exploring language models for medical question answering systems. *arXiv preprint arXiv:2401.11389*.

Lawrence KQ Yan, Qian Niu, Ming Li, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Benji Peng, Ziqian Bi, Pohsun Feng, Keyu Chen, et al. 2024. Large language model benchmarks in medical tasks. *arXiv preprint arXiv:2410.21348*.

Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yu He Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, et al. 2024a. Kg-rank: Enhancing large language models for medical qa with knowledge graphs and ranking techniques. *arXiv preprint arXiv:2403.05881*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19368–19376.

Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, Fenglin Liu, Meng Cao, Ziming Wang, Xuxin Cheng, Zhu Lei, et al. 2023. Qilin-med: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089*.

Wenjing Yue, Xiaoling Wang, Wei Zhu, Ming Guan, Huanran Zheng, Pengfei Wang, Changzhi Sun, and Xin Ma. 2024. Tcmbench: A comprehensive benchmark for evaluating large language models in traditional chinese medicine. *arXiv preprint arXiv:2406.01126*.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: a thorough analysis of the emrqa dataset. *arXiv preprint arXiv:2005.00574*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9241–9250.

Deshiwei Zhang, Xiaojuan Xue, Peng Gao, Zhijuan Jin, Menghan Hu, Yue Wu, and Xiayang Ying. 2024. A survey of datasets in medicine for large language models. *Intelligence & Robotics*, 4(4):457–478.

H Zhang, J Chen, F Jiang, et al. 2023. Huatuogpt, towards taming language model to be a doctor. arxiv preprint arxiv: 230515075.

Haitao Zhang and Bo An. 2024. Medgo: A chinese medical large language model. *arXiv preprint arXiv:2410.20428*.

Sheng Zhang, Xin Zhang, Hui Wang, Jiajun Cheng, Pei Li, and Zhaoyun Ding. 2017. Chinese medical question answer matching using end-to-end character-level multi-scale cnns. *Applied Sciences*, 7(8):767.

Yiming Zhang, Zheng Chang, Wentao Cai, MengXing Ren, Kang Yuan, Yining Sun, and Zenghui Ding. 2025. Iimedgpt: Promoting large language model capabilities of medical tasks by efficient human preference alignment. *arXiv preprint arXiv:2501.02869*.

Lulu Zhao, Weihao Zeng, Xiaofeng Shi, Hua Zhou, Donglin Hao, and Yonghua Lin. 2024. Aqulia-med llm: Pioneering full-process open-source medical language models. *arXiv preprint arXiv:2406.12182*.

A Overview of Medical Patient Communication Datasets

In this appendix, we present **Table 1**, which provides a structured summary of medical datasets utilized for training and evaluating Large Language Models (LLMs) in clinical communication tasks. This table serves as a reference for understanding the characteristics and applications of these datasets in medical AI research.

A.1 Construction of Table 1 and Objectives

Table 1 is compiled from a systematic survey of open-access medical datasets referenced throughout this paper. The primary objective is to offer a **data-centric taxonomy** that differentiates knowledge-based datasets, which focus on medical accuracy and structured reasoning, from conversation-based datasets, which emphasize interactive, patient-centered communication.

The inclusion criteria for datasets in Table 1 are:

- Publicly available or well-documented.
- Explicit focus on medical patient communication, including diagnostic Q&A, doctor-patient dialogues, and medical literature-based queries.
- Prior adoption in research for benchmarking medical LLMs.

Each dataset entry is sourced from **peer-reviewed publications, dataset repositories, or official documentation**, ensuring reliability and relevance.

A.2 Structure and Organization of the Table

Table 1 consists of multiple columns capturing essential details of each dataset. The rows represent individual datasets, categorized into two groups:

1. **Knowledge-Based Datasets**: These datasets primarily support factual medical knowledge extraction and diagnostic reasoning.
2. **Conversation-Based Datasets**: These datasets primarily focus on patient communication, interactive dialogue dynamics, and empathetic medical consultation.

Columns in the Table:

- **Dataset Name**: The name of the dataset, along with references to primary sources.
- **Clinical Properties**: The primary medical communication focus (e.g., symptom inquiry, clinical consultation).
- **Data Type**: The nature of data collected, such as multiple-choice questions (MCQA), doctor-patient Q&A, or multi-turn dialogues.
- **Annotation**: The level of annotation provided, including question labels, structured metadata, or conversational tags.
- **Scale**: Dataset size, measured in number of examples, interactions, or Q&A pairs.
- **Application Papers**: Some of the key research papers that have used this dataset for model fine-tuning or evaluation.

A.3 Dataset Grouping and Distribution

The datasets are categorized into two types:

(A) Knowledge-Based Datasets

- **Medical Licensing and Board Exam Datasets**: Standardized MCQA datasets sourced from medical board exams (e.g., *MedQA, CMExam, MedMCQA*).
- **Scientific Literature-Based Q&A**: Datasets such as *PubMedQA, BioASQ, MedQuAD*, extracting knowledge from academic sources.
- **Electronic Health Record (EHR)-Based Q&A**: Structured datasets like *emrQA* that utilize clinical records.

(B) Conversation-Based Datasets

- **Single-Turn Symptom Inquiry Datasets**: Datasets such as *HealthCareMagic-100k, iCliniq10k, Huatuo-26M* provide doctor responses to patient symptom descriptions.
- **Multi-Turn Doctor-Patient Consultation Datasets**: Including *MedDG, BianQueCorpus, CMtMedQA*, these datasets capture extended interactions between doctors and patients.
- **Mental Health Counseling Transcripts**: The *Psych8K* dataset focuses on counseling conversations.

Table 1: Summary of Medical Patient Communication Datasets.

Dataset	Clinical Properties	Data Type	Annotation	Scale	Application Paper
Knowledge-Based					
MedQA (Jin et al., 2021)	Medical Licensing and Board Exam Question Bank	MCQA medical licensing exam	N/A	~60K	(Christophe et al., 2024a) (Wang et al., 2024c) (Paiola et al., 2024) (Kamble and AlShikh, 2023)
MedMCQA (Pal et al., 2022)		MCQA medical exam and mocked tests created by human experts	Explanations provided	~193K	(Ponce-López, 2024) (Christophe et al., 2024a) (Wang et al., 2024c) (Singhal et al., 2025) (Pal et al., 2022)
MultiMedQA (Singhal et al., 2022)		MCQA and Open QA synthesized from 7 medical Q&A datasets (MedQA, MedMCQA, PubMedQA, MMLU, LiveQA, MedicationQA, HealthSearchQA)	N/A	~474K development set and 9K test set	(Hamzah and Sulaiman) (Singhal et al., 2025)
CMExam (Liu et al., 2024a)		MCQA Medical Licensing Exam	Question labels: disease groups, clinical departments, medical disciplines, areas of competency, and question difficulty levels	~60K	(Liu et al., 2024a) (Wang et al., 2024a) (Wang et al., 2023b) (Ye et al., 2023) (Wu et al., 2024b) (Yue et al., 2024)
XMedBench (Wang et al., 2024d)		MCQA synthesized from multilingual medical Q&A datasets	N/A	N/A	(Xie et al., 2024)
BioASQ (Krithara et al., 2023)	Scientific Literature-Based Medical Q&A	Biomedical Q&A (including both exact answer and ideal answer) from scientific literature with reference and supporting material	Structured Q&A labels (e.g., question type, concept, answer, reference, supporting material.)	~5K	(Gao et al., 2024)
PubMedQA (Jin et al., 2019)		Biomedical Q&A collected from PubMed abstracts	Each Q&A instance labeled: Question + Context + Long Answer + Final Answer (yes/no/maybe)	~1k expert-annotated, 211.3k artificially generated QA, and 61.2k unlabeled	(Christophe et al., 2024a) (Jin et al., 2019) (Guo et al., 2023) (Zhao et al., 2024)
MedQuAD (Abacha et al., 2019)		Medical Q&A sourced from NIH websites	Each Q&A instance labeled: Question + Answer + Source + Focus Area	~47K	(Yagnik et al., 2024) (Pandya, 2023) (Monea and Marginean, 2021)
BiMed1.3M (Pieri et al., 2024)		MCQA, Q&A, synthesized multi-turn doctor-patient communication simulated with ChatGPT	N/A	~1.3M samples (423.8K Q&A, 638.1K MCQA, 249.7K chat)	(Pieri et al., 2024)
Medication QA (Abacha et al., 2019)		Medication Q&A	Each Q&A instance labeled: Focus (Drug) + Question Type + Answer + Section Title + URL	674	(Yang et al., 2024a)
emrQA (Pampari et al., 2018)	Electronic Medical Record (EHR)-Based Q&A	EHR-based Q&A, including both question-logical form pairs and Q&A pairs	EHR documents annotated with Q&A (Q&A and Question-Logical Form-Answer Evidence)	~1M question-logical form pairs, 400K Q&A	(Yue et al., 2020) (Soni and Roberts, 2020)
Conversation-Based					
HealthCareMagic-100K (Li et al., 2023c)	Single-turn Online Symptom Inquiry	Real-world user queries with doctor responses on an online health platform	N/A	~100K	(Li et al., 2023c) (Paiola et al., 2024)
iCliniq10K (Li et al., 2023c)		Real-world user queries with doctor responses on an online health platform	N/A	~10K	(Acikgoz et al., 2024)
cMedQA (Zhang et al., 2017)		Real-world patient queries answered by doctors from online medical Q&A forum	Question with a pair of ground truth answer and an incorrect answer	Total Questions: Q (54K) & A (102K) Training: Q (50K) & A (94K), Dev: Q (2K) & A (4K), Test: Q (2K) & A (4K)	(Guo et al., 2022)
Huatuo-26M (Li et al., 2023a)		Real-world patient queries answered by doctors from online medical Q&A forum; Medical Q&A collected from medical encyclopedia; Medical Q&A collected from knowledge graph	N/A	~26M	(Li et al., 2023a) (Li et al., 2023b) (Zhang et al., 2023) (Ye et al., 2023)
BianQue Corpus (Chen et al., 2023a)	Multi-turn Doctor-Patient Consultation Dialogues	Real-world multi-turn doctor-patient communication		~2.4M conversation samples	(Chen et al., 2023a)
MedDG (Liu et al., 2022)		Real-world multi-turn doctor-patient conversations	Each sentence labeled: Role (Doctor/Patient) + Symptom + Medicine + Examination + Attribute + Disease	18K	(Wu et al., 2024c) (Liu et al., 2024c) (Zhang and An, 2024) (He et al., 2024)
MedDialog (Zeng et al., 2020)		Real-world multi-turn doctor-patient conversations from online consultation website. Each consultation includes: description of medical conditions and patient history + doctor-patient conversation + diagnosis and treatment suggestions	N/A	~3.4M conversations in Chinese, 0.26M conversations in Engligh	(Zeng et al., 2020) (Dou et al., 2023) (Tian et al., 2023)
NoteChat (Wang et al., 2023a)		Synthetic doctor-patient conversations generated via LLMs based on 167K case reports in the PMC-Patients dataset and 1.7K structured short doctor-patient conversations in the MTS-Dialog dataset	N/A	~10K	(Wang et al., 2023a) (Binici et al., 2024)
CMtMedQA (Yang et al., 2024b)		Real-world multi-turn doctor-patient communication standardized with self-instruction method	N/A	70K multi-turn dialogues and 400K single-turn conversations	(Yang et al., 2024b) (Zhao et al., 2024) (Zhang et al., 2025)
Psych8K (Liu et al., 2023)	Mental Health Counseling Transcripts	Real-world in-depth counseling transcripts, de-identified and segmented into 10-round short conversations via GPT-4	Annotated on counseling metrics via GPT-4 (e.g., direct guidance, approval & reassurance, interpretation, self-disclosure, etc.)	~8K conversation fragments	(Liu et al., 2023)