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

Ruosi Shao[1], Md Shamim Seraj[1*], Kangyi Zhao[2*], Yingtao Luo[3*], Lincan Li[1], Bolin Shen[1], Averi J. Bates[4], Yue Zhao[5], Chongle Pan[4], Lisa Hightow-Weidman[1], Shayok Chakraborty[1], Yushun Dong[1]

[1]Florida State University [2]University of Pittsburgh [3]Carnegie Mellon University
[4]University of Oklahoma [5]University of Southern California

{rshao,ms19bt,ll24bb,bs24bc,lhightowweidman,schakraborty2,yd24f}@fsu.edu
KAZ78@pitt.edu yingtaoluo@cmu.edu
{Averi.J.Bates-1,cpan}@ou.edu yzhao010@usc.edu

Abstract

Large language models (LLMs) hold promise for advancing patient–provider communication, yet a persistent gap remains between benchmark-driven model development and the realities of clinical practice. This work presents a systematic, clinically grounded review of text-based medical datasets for LLM training and evaluation. We propose a scenario-based taxonomy derived from established clinical frameworks to map major knowledge-based and conversation-based corpora against core communication scenarios. We further synthesize core communication skills from gold-standard clinical assessment instruments and meta-analyze state-of-the-art medical LLM performance, highlighting how dataset properties, fine-tuning strategies, and evaluation metrics shape both knowledge acquisition and communicative competence. To empirically validate these findings, we conducted controlled fine-tuning experiments across representative LLMs, demonstrating that data composition and scenario alignment critically affect model performance. Our findings highlight the urgent need for scenario-rich datasets and standardized, human-centered evaluation protocol to advance clinically relevant medical LLMs.

1 Introduction

Large language models (LLMs) are rapidly transforming the landscape of patient–provider communication by offering scalable, accessible support across a spectrum of clinical tasks, ranging from symptom triage and patient education to behavioral counseling and chronic care management (Busch et al., 2025; Huo et al., 2025; Omar et al., 2024). Recent medical LLMs such as MedPaLM-2 (Singhal et al., 2025), Meditron (Chen et al., 2023b), Med42 (Christophe et al., 2024a,b), and GPT-4 (Nori et al., 2023) have achieved near-expert performance on knowledge-based benchmarks, such as MedMCQA (Pal et al., 2022) and PubMedQA (Jin et al., 2019). These models now match or exceed human clinicians in factual accuracy, response relevance, and socio-communication dimensions such as empathy (Singhal et al., 2025; Paiola et al., 2024; Calle et al., 2024).

Despite these advances, real-world clinical deployments reveal persistent gaps between benchmark-driven LLM development and the multifaceted demands of clinical communication (Busch et al., 2025; Liu et al., 2024b; Shi et al., 2024b,a). Empirical studies highlight that clinical feasibility, acceptability, and effectiveness depend not only on model performance but also on the clinical properties and scenario diversity of the training data (Wu et al., 2024). Current training corpora are highly heterogeneous, spanning knowledge-based datasets (e.g., PubMedQA, MedQA) focused on factual recall and clinical reasoning, and conversation-based corpus (e.g., NoteChat, Psych8K, CMtMedQA) designed to reflect the nuance and contextual richness of patient–provider interaction. Yet, the field lacks a systematic understanding of how these dataset properties align with the spectrum of communicative competencies required in practice, leading to persistent mismatches between model capabilities and the realities of clinical implementation.

To address this critical gap, our paper combines a systematic, clinically grounded review of text-based medical datasets with experimental evidence that directly tests key claims of the review. Specifically, our contributions are:

- **Scenario-Based Dataset Taxonomy:** We propose a taxonomy of clinical scenarios in patient–provider communication, derived from gold-standard clinical frameworks (e.g., OSCE, SEGUE, Calgary–Cambridge), map the coverage of widely used medical datasets to these scenarios, and systematically identify persistent

[*]Equal contribution.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 684–705

gaps in the representation of clinical scenarios within existing corpora.

- **Framework for Clinical Communication:** We provide the first systematic synthesis of core patient–provider communication skills, extracted from validated clinical instruments, to guide dataset development and model evaluation aligned with clinical practices.
- **Meta Review and Empirical Validation:** We present a meta-analytical review of state-of-the-art medical LLM performance on major benchmarks, and conduct controlled experiments to empirically validate how dataset properties and evaluation metrics influence both knowledge retention and communication performance.

These contributions provide a unified clinical framework for the design, evaluation, and empirical benchmarking of medical LLMs, bridging the gap between model development and the practical requirements of patient-centered communication in real-world clinical implementation.

Paper Structure: Section 2 introduces our scenario-based taxonomy and maps dataset coverage. Section 3 critically reviews evaluation metrics and their alignment with clinical frameworks. Section 4 synthesizes current LLM performance through meta-analytical review. Section 5 details our experimental approach and results. Section 6 reviews related work, and Sections 7 and 8 discuss limitations and future directions for developing clinically aligned medical LLMs.

2 Scenario-Based Taxonomy of Medical Datasets

Text-based medical datasets are foundational resources for the development of LLMs in healthcare. These corpora encompass a broad spectrum of content types, including medical question-answering (QA) sets, clinical examination questions, multi-turn dialogues, and counseling transcripts, designed to support both the acquisition of medical knowledge and the development of communicative competence (Ha and Longnecker, 2010). The clinical properties, data structure, and annotation methodologies embedded in these datasets fundamentally influence the accuracy, trustworthiness, and real-world performance of medical LLMs.

2.1 Categorizing Text-Based Medical Datasets

Defining Knowledge-Based Datasets. Text-based medical datasets can be categorized into (i) *knowledge-based* and (ii) *conversation-based*. Knowledge-based datasets are constructed primarily to encode and evaluate structured medical knowledge and clinical reasoning. Derived from medical board or licensing examinations, biomedical literature, clinical guidelines, or expert-authored repositories, these datasets employ highly structured formats such as multiple-choice questions (MCQs), single-turn factoid queries, or brief open-ended questions with expert-provided annotations for clinical topics and difficulty levels. The primary goal is to evaluate an LLM's capacity for factual recall and clinical reasoning in decontextualized, non-dialogic scenarios. Examples include MedQA, MedMCQA, BioASQ, and PubMedQA, which serve as standard benchmarks for knowledge-centric evaluation.

Defining Communication-Based Datasets. Communication-based datasets are explicitly constructed from, or designed to simulate, authentic patient–provider communication. These corpora are presented as single-turn or multi-turn dialogues from real-world clinical encounters, online health platforms, counseling sessions, or simulated clinical interactions, modeling the linguistic, interpersonal, and contextual dynamics of patient-provider communication. Common annotations include speaker roles, turn types, medical entities, or specific communication skills, such as empathy and shared decision making. Examples include HealthCareMagic-100K, iCliniq10K, Huatuo-26M, BianQue Corpus, NoteChat, and Psych8K, increasingly recognized for their ability to support the training and evaluation of LLMs on communication competence, patient-centeredness, and effective dialogue management.

2.2 Mapping Clinical Scenarios Coverage

Clinical Scenario Taxonomy. To systematically evaluate the clinical utility of medical datasets for LLM training, we introduce a taxonomy of core clinical scenarios that define patient–provider communication, drawing on established frameworks in clinical communication (e.g., OSCE (Newble, 2004), Calgary–Cambridge (Kurtz and Silverman, 1996; Silverman et al., 2016), SEGUE (Makoul, 2001b)). This taxonomy delineates 12 fundamental scenarios, including history taking, routine and preventive care, patient education, informed consent, behavioral counseling, shared decision making, and care transitions (Table 5).

Systematic Mapping of Dataset Coverage. Us-

Table 1: Coverage of clinical scenarios in patient-provider communication across major medical datasets.

Dataset	Clinical Scenarios											
	1	2	3	4	5	6	7	8	9	10	11	12
Knowledge-Based												
MedQA	X	X	X						X	X		
CMExam	X	X	X						X	X		
MedMCQA	X	X	X						X	X		
XMedBench	X	X	X						X	X		
MultiMedQA	X	X	X						X	X		X
BioASQ-QA		X	X							X		
PubMedQA	X		X							X		
MedQuAD		X	X							X		
BiMed1.3M	X	X	X						X	X		X
Medication_QA		X	X							X		X
emrQA	X										X	
C-Eval	X	X	X						X	X		
Communication-Based												
HealthCareMagic	X	X	X	X			X				X	
iCliniq10k	X	X	X	X			X				X	
cMedQA	X	X	X	X			X				X	
Huatuo-26M	X	X	X	X		X	X		X	X		X
BianQueCorpus	X	X	X	X		X	X		X	X		X
MedDG	X	X	X	X			X		X	X		X
MedDialog	X	X	X	X			X		X	X		X
NoteChat	X	X	X	X			X		X	X		X
CMtMedQA	X	X	X	X		X	X		X	X		X
Psych8K				X			X					X

Notes: (a) Scenario numbers: 1. History Taking and Initial Assessment, 2. Screening, Routine and Preventive Care, 3. Patient Education, 4. Counseling and Behavioral Intervention, 5. Informed Consent, 6. Counseling on Procedures, 7. Shared Decision Making, 8. Breaking Bad News, 9. Acute and Emergency Encounters, 10. Rehabilitation/Chronic Disease Management, 11. Referrals and Care Transitions, 12. Addressing Patient Concerns/Barriers. (b) X = Explicit/strong coverage; (blank) = Not covered. Only datasets with explicit, process-based, or dialogic examples were counted as covering a scenario. Partial coverage (e.g., factoid or one-way advice) was not counted.

ing this taxonomy, we systematically mapped the coverage of clinical scenarios in existing medical datasets (Table 1). Knowledge-based datasets excel in foundational areas like history taking and routine care but show limited representation of process-based or ethically complex scenarios such as informed consent or care transitions. These gaps reflect their design focus on factual recall over contextual interaction. By contrast, conversation-based datasets offer broader scenario coverage, particularly in patient education, behavioral interventions, and multi-turn consultations. However, even the most comprehensive dialogue corpora lack annotated process-driven exchanges (e.g., informed consent, procedural counseling, structured handoffs) essential for modeling high-stakes clinical tasks.

2.3 Alignment Gaps and Clinical Implications

Despite rapid growth in text-based medical datasets, a critical examination reveals persistent misalignments between available data and the full spectrum of patient–provider communication (Kurtz, 2002; Matusitz and Spear, 2014). Knowledge-based datasets support LLM development for factual recall and structured reasoning but insufficiently capture interpersonal, context-rich, or ethically complex interactions (Shi et al., 2024b). As a result, LLMs trained predominantly on these resources tend to excel at identifying medical facts, diagnostic pathways, and treatment protocols, but often underperform in tasks demanding empathy, clarification, shared decision making, or negotiation of care plans (Christophe et al., 2024b).

Conversely, communication-based datasets offer more realistic representations of clinical dialogue, enabling improvements in conversational coherence, context retention, and adaptive questioning (Li et al., 2023b; Liu et al., 2023; Chen et al., 2023a). Yet, even these corpora typically lack comprehensive coverage of process-oriented and ethically complex interactions, such as formal informed consent or structured transitions of care.

Implications for LLM Training and Deployment. Partial scenario coverage limits the ability of LLMs to address the full spectrum of clinical communication needs. While existing datasets support transactional interactions (e.g., information provision, triage), they remain inadequate for nuanced, high-stakes communication, such as counseling on uncertainty, consent processes, and negotiation of preferences. The absence of standardized, scenario-based annotation frameworks further impedes systematic benchmarking and improvement of LLM communication competence. Addressing these limitations will require targeted expansion of scenario coverage, clinically validated annotation schemas, and hybrid training pipelines integrating both factual rigor and interactional nuance.

3 Evaluation Metrics for Medical LLMs and Clinical Communication Skills

The current landscape for evaluating medical LLMs is dominated by metrics that focus on factual accuracy, linguistic fidelity, and surface-level communicative behaviors (Abbasian et al., 2024). This section reviews standard metrics, their limitations, and emerging approaches to align model evaluation with real-world clinical communication needs.

3.1 Standard Evaluation Metrics

Factual and Linguistic Metrics. Accuracy remains the primary metric in medical QA benchmarks such as MedQA and MedMCQA (Wang et al., 2024a), directly measuring factual correctness on objective queries. While accuracy serves as a direct metric of factual correctness, it is inherently constrained to objective, well-defined queries. Token-level metrics (precision, recall, F1, exact match) and string similarity scores (BLEU, ROUGE, BERTScore) have become standard in evaluating information extraction, summarization, and dialogue tasks (Zhang et al., 2024; Abbasian et al., 2024). However, these metrics predominantly reward surface-form or reference overlap rather than assessing clinical appropriateness, contextualization, or communication effectiveness. As a result, LLMs may achieve high scores while producing responses that are technically correct but poorly suited to patient needs (Abbasian et al., 2024).

Human-Rated and Composite Metrics. To address these gaps, human-rated metrics, such as fluency, coherence, completeness, adequacy, and hallucination, have been incorporated, relying on clinician or annotator judgments to determine whether outputs are clear, actionable, and contextually relevant (Liu et al., 2022; Pieri et al., 2024; Yagnik et al., 2024). For example, MedDialog and NoteChat include human assessments of logical flow and informativeness (Zeng et al., 2020; Wang et al., 2023). Nevertheless, such evaluations are inconsistently applied and often lack standardized rubrics or widely accepted benchmarks, impeding comparability and reproducibility.

Sparse Assessment of Communication Skills. Explicit evaluation of patient-centered communication skills remains rare. Only a small subset of recent datasets (e.g., BianQueCorpus, Psych8K) introduce dedicated metrics for clinical communication skills, including empathy, proactive questioning, or reflective listening (Chen et al., 2023a; Liu et al., 2023; Abbasian et al., 2024). Even in these cases, annotation schemas vary and are not systematically aligned with clinical guidelines, and there is little consensus on operationalizing these skills for automated or scalable evaluation (Abbasian et al., 2024). Recent critical surveys and systematic reviews have repeatedly highlighted these gaps, noting that "existing evaluation metrics proposed for generic LLMs demonstrate a lack of comprehension regarding medical and health concepts and their significance in promoting patients' well-being. Moreover, these metrics neglect pivotal user-centered aspects, including trust-building, ethics, personalization, empathy, user comprehension, and emotional support" (Abbasian et al., 2024).

3.2 Frameworks for Human-Centered Clinical Communication Assessment

In clinical education and practice, patient–provider communication is evaluated via multi-dimensional, human-centered frameworks, such as the Calgary–Cambridge Guide, SEGUE Framework, and OSCE checklists. These instruments serve as gold standards for high-stakes assessment, medical education, and research across global context, employed both formatively (for feedback and education) and summatively (for licensing and certification), supporting reliable, reproducible, and actionable judgments about communication competence (Kurtz and Silverman, 1996; Makoul, 2001b; Silverman et al., 2016; Newble, 2004).

Core Communication Skills. These frameworks systematically decompose clinical communication into a set of core skills (Table 8), including rapport building and introduction, information gathering and questioning, active listening, information giving and explanation, empathy and patient support, shared decision making, teach-back, uncertainty communication, motivational interviewing and behavioral counseling, dialogue structure and organization, and effective closing and outlining next steps (Kurtz and Silverman, 1996; Makoul, 2001b; Abbasian et al., 2024). Each skill is linked to distinct clinical functions: rapport facilitates trust and treatment adherence; teach-back ensures patient comprehension of care instructions; shared decision-making operationalizes patient autonomy. Assessment is conducted at both the message or utterance level (e.g., "Did the provider elicit relevant medical history? Did the provider acknowledge and validate the patient's concerns?") and the encounter or structure level (e.g., "Was the conversation logically well-organized, with clear transitions between topics and a defined summary or next steps?"), allowing for both granular and holistic evaluation of communication competence (Makoul, 2001b; Abbasian et al., 2024).

3.3 Aligning LLM Evaluation with Clinical Communication Needs

A fundamental misalignment persists between standard LLM evaluation metrics and the actual de-

mands of clinical communication. Conventional metrics such as accuracy, BLEU, and ROUGE, while effectively quantify knowledge retrieval and linguistic fluency, remain largely agnostic to the interpersonal, contextual, and ethical dimensions of patient care (Abbasian et al., 2024; Wang et al., 2024a; Zhang et al., 2024). An LLM might achieve high scores on these metrics yet fail in real clinical scenarios by overlooking emotional cues, failing to check patient understanding, or bypassing shared decision-making processes (Makoul, 2001b; Abbasian et al., 2024).

Recent efforts, such as HealthBench, have sought to address this gap by introducing 48,000 itemized, expert-crafted rubrics that evaluate not only factual completeness but also context awareness, clinical reasoning, and communication quality for specific medical scenarios (Arora et al., 2025). For instance, in "expertise tailoring," models are assessed on their ability to shift between layperson-friendly and professional communication. While this approach enables scenario-specific and granular scoring, it also exposes a critical limitation: each query is paired with a unique, handcrafted rubric, which hinders scalability, standardization, and systematic benchmarking of broader communication skills and limits comparability across tasks and models. As a result, the field still lacks unified, transferable, and scalable metrics that can capture core communication competencies across the full continuum of clinical interactions (Abbasian et al., 2024; Arora et al., 2025).

4 Performance of Medical LLMs Across Datasets and Evaluation Metrics

The empirical trajectory of medical LLMs echoes dataset taxonomy and evaluation frameworks in Sections 2 and 3. Recent development draws upon both knowledge-based corpora (MedQA, MedMCQA, PubMedQA, CMExam, MultiMedQA) and conversation-based corpora (BianQue Corpus, MedDialog, NoteChat, Psych8K), yielding two distinct but complementary competence axes: knowledge recall and clinical communication.

4.1 Trends in LLM Fine-Tuning

Fine-tuning methodology has undergone rapid innovation. While full-parameter fine-tuning remains effective for moderate-scale models (Christophe et al., 2024a), parameter-efficient strategies such as LoRA and QLoRA now predominate, enabling scalable adaptation across languages and datasets (Wang et al., 2024c; Pieri et al., 2024; Li et al., 2023a; Ye et al., 2023). Instruction tuning, retrieval augmentation, and alignment techniques, including RLHF, DPO, and domain-specific preference optimization, further calibrate model outputs to clinical contexts (Arora et al., 2025; Singhal et al., 2025; Dou et al., 2023).

4.2 Knowledge-Centric Performance

Knowledge-centric evaluation demonstrates that leading models, such as MedPaLM-2 and Med42, consistently achieve 70–80% accuracy on MedQA, 65–75% on MedMCQA, 75–79% on PubMedQA, and 59–62% on CMExam, with MedPaLM-2 rivaling GPT-4 on several closed benchmarks (Singhal et al., 2025; Christophe et al., 2024a; Wang et al., 2024c; Liu et al., 2024a). Retrieval-augmented and domain-specialized LLMs (JMLR-13B, BiMediX) produce 2–10% absolute gains over base models and, in select biomedical subdomains, can outperform GPT-4 (Wang et al., 2024c; Pieri et al., 2024; Krithara et al., 2023). Domain-adapted small models (e.g., PubMedBERT at 47% on MedMCQA (Pal et al., 2022); T5 with 6–10 BLEU/ROUGE on MedQuAD (Lamichhane and Kahanda, 2023)) improve over non-adapted baselines but remain below top-tier performance. Recent multilingual and Chinese-specific models, including ChatGLM, Huatuo, and Qilin-Med, regularly achieve or exceed 60% accuracy on national licensing-style exams after LoRA-based tuning and chain-of-thought prompting, narrowing gaps with English-centric or general-purpose LLMs (Li et al., 2023a; Liu et al., 2024a; Ye et al., 2023; Chen et al., 2023a).

4.3 Communication-Centric Performance and Human Evaluation

In communication tasks, pretraining and fine-tuning on dialogue corpora have produced measurable advances. LLMs achieve BLEU-1/BLEU-4 scores from 20–44 and ROUGE-L up to 27 (MedDialog, BianQue, Huatuo, NoteChat) (Chen et al., 2023a; Zeng et al., 2020; Wang et al., 2023; Li et al., 2023a). Human and expert evaluations reveal that models fine-tuned on communication-rich corpora—such as BianQue, NoteChat, PlugMed, and ChatCounselor—exhibit increased informativeness, proactive questioning, and scenario-based empathy, and can surpass ChatGPT/GPT-4 in pairwise preference for synthetic physician–patient dialogue (Chen et al., 2023a; Wang et al., 2023;

Dou et al., 2023; Liu et al., 2023). PlugMed and Psych8K introduce multidimensional, human-centered metrics, and ChatCounselor achieves near-ChatGPT performance on counseling skills (Dou et al., 2023; Liu et al., 2023). However, rigorous evaluation of advanced skills (e.g., shared decision making, uncertainty communication, teach-back) remains uncommon due to annotation burden. HealthBench demonstrates the feasibility of expert-annotated, scenario-grounded evaluation, yet systematic adoption, standardization, and scalability remain limited (Arora et al., 2025).

4.4 Comparative Analysis and Challenges

Parameter-efficient, domain-finetuned models (e.g., Llama2-7B/13B, ChatGLM, BiMediX) can match or exceed much larger LLMs (GPT-4, Meditron-70B) in in-domain or low-resource settings (Wang et al., 2024c; Pieri et al., 2024; Li et al., 2023a; Ye et al., 2023). LoRA-based approaches have become essential for scalable, iterative adaptation. Yet despite outperforming generic LLMs in multi-turn, patient-centered dialogue, communication-optimized models still fall short of human clinicians, especially in process-oriented or ethically complex encounters (Arora et al., 2025; Makoul, 2001b). This persistent gap underscores the need for high-fidelity, process-annotated dialogue corpora and robust, scenario-based clinical-communication metrics.

4.5 Meta-Analytical Performance Review of Prior Work

Building on the scenario taxonomy and evaluation framework in Sections 2–3, our meta-analytic synthesis of prior work (Table 2) reveals a consistent composition effect: dataset properties systematically shape distinct competence axes in medical LLMs. Knowledge-oriented corpora (e.g., MedQA, PubMedQA), combined with instruction tuning, CoT prompting, and LoRA/QLoRA, reliably improve factual performance on licensing/board-style items (e.g., Med-PaLM2 at 86.5% on MedQA), whereas conversation-oriented corpora (e.g., NoteChat, Psych8K, Zhongjing) drive gains in dialogue quality (BLEU, ROUGE, BERTScore), engagement, and proactive questioning. However, both streams underrepresent process-oriented, guideline-concordant behaviors, such as informed consent, shared decision making, or uncertainty communication, mirroring the evaluation gaps identified in Section 3.

Implications for Experimental Hypotheses. Taken together, these patterns operationalize our central claim that *data composition and scenario coverage* determine where models improve (knowledge recall vs. communicative competence) and where they fall short (process-aligned clinical communication). To convert these correlational trends into causal evidence, Section 5 presents a controlled study that varies dataset composition (knowledge vs. conversation vs. mixed) and jointly evaluates knowledge and communication outcomes. We test three hypotheses derived directly from the synthesis above:

- **H1 (Knowledge-only → Factual Accuracy).** Relative to conversation-only fine-tuning, knowledge-only fine-tuning yields higher factual accuracy with little material gains on communication metrics.
- **H2 (Communication-only → Conversation Quality).** Relative to knowledge-only fine-tuning, conversation-only fine-tuning yields higher scores on conversation quality and better readability, with little material gains in factual accuracy.
- **H3 (Mixed → Balanced Performance).** A balanced mixed regimen (a) outperforms conversation-only on factual accuracy, (b) outperforms knowledge-only on communication metrics/readability, and (c) achieves the highest combined score across knowledge and communication (defined in Section 5).

5 Empirical Evaluation of Data Composition and Scenario Alignment

To empirically validate the core claims from Section 4, we conducted a controlled study testing how *dataset composition* shapes medical LLM performance across knowledge and clinical-communication axes. We test the three hypotheses **H1–H3** as defined in Section 4.5.

5.1 Experimental Setup

We selected two open-source instruction-tuned LLMs (LLaMA3-8B and Qwen2.5-7B). Datasets are grouped into (i) *knowledge-based* (MedQA, MedMCQA, PubMedQA, MedQuAD) and (ii) *conversation-based* (HealthCareMagic-100K, iCliniq10K, MedDialog, NoteChat). To isolate composition effects under a comparable budget,

Table 2: Meta-Analytic Overview of LLM Performance on Knowledge- and Conversation-Based Medical Datasets

Fine-Tuning Methods	Evaluation Metrics	Performance Range (reported)	Examples
Knowledge-Based			
Instruction Fine-Tuning (IFT)	Accuracy, F1, Exact Match (EM)	70–86% accuracy (e.g., Med-PaLM2: 86.5% on MedQA)	MedQA, PubMedQA, MultiMedQA
LoRA / QLoRA	Accuracy, Training Efficiency	65–75% accuracy (e.g., Llama2-7B: 72.4% on PubMedQA)	BioASQ-QA, CMExam, XMedBench
Domain-Specific Vocabulary	Accuracy, Semantic Answer Similarity	68–79% SAS (e.g., BioMed-RoBERTa: 90% RQE accuracy)	MedQuAD, emrQA
Chain-of-Thought (CoT)	CoT Accuracy, Stepwise Reasoning	75–80% CoT accuracy (e.g., Med-PaLM2 on multi-step tasks)	MedQA, C-Eval
Multi-Stage Training	Accuracy, Generalization	60–75% accuracy (e.g., Qilin-Med: 40% on CMExam)	Huatuo-26M, PubMedQA
Data Augmentation	Accuracy, Robustness	+10–15% improvement (e.g., BioGPT-Large: 75.4% vs. 63% baseline)	PubMedQA, MedQuAD
Communication-Based			
RLHF	BLEU, ROUGE, Human Preference	BLEU-4: 25–35; ROUGE-L: 40–50 (e.g., Zhongjing: ROUGE-L 45.3)	CMtMedQA, BianQueCorpus
Supervised Fine-Tuning (SFT)	Dialogue Coherence, Fluency	BLEU-4: 20–28; ROUGE-L: 30–40 (e.g., ChatDoctor: BERTScore F1 = 0.8446)	HealthCareMagic-100k, iCliniq10k
Conversational Preference Training	Proactive Questioning (PQA), Safety	PQA: 7–9 (e.g., Zhongjing: PQA 8.7)	CMtMedQA, NoteChat
LoRA / QLoRA	Token Efficiency, Diversity	Self-BLEU: 0.12–0.18 (e.g., NoteChat: 0.12 vs. GPT-4's 0.18)	NoteChat, BiMed1.3M
Multi-Turn Dialogue Training	Context Retention, Engagement	Human Preference: 70–85% (e.g., NoteChat preferred over GPT-4)	Psych8K, BianQueCorpus
Synthetic Data Augmentation	Factuality, Diversity	+15–20% ROUGE-L improvement (e.g., MEDSAGE: 14.8% F1 gain)	NoteChat, MedDG

Note: Ranges are from reported results in prior work; metrics are not directly comparable across datasets or tasks. Examples show typical, not universal, baselines.

Table 3: Performance results for LLaMA3-8B and Qwen2.5-7B after fine-tuning on knowledge-based, conversation-based, and mixed strategies. MedCommEval is our proposed rubric. "Difference" is the change vs. the original model without fine-tuning. Mixed strategies combine knowledge and conversation datasets at different ratios (knowledge:conversation). "Combined" merges datasets within each strategy; "All Combined" includes all eight datasets from both streams.

Model	FT Strategy	Dataset	Accuracy (↑)		BERTScore (↑)		HealthBench (↑)		MedCommEval (↑)		ΔFKGL (↓)	
			Value	Difference	Value	Difference	Value	Difference	Value	Difference	Value	Difference
LLaMA3-8B	w/o	w/o	52.6	N/A	82.7	N/A	0.21	N/A	**0.47**	N/A	2.84	N/A
	Knowledge	MedQA	54.7	+2.1	82.7	+0.0	0.18	-0.03	0.45	-0.02	2.77	-0.07
	Knowledge	MedMCQA	54.5	+1.9	82.8	+0.1	0.18	-0.03	0.46	-0.01	2.79	-0.05
	Knowledge	PubMedQA	52.9	+0.3	82.7	+0.0	0.18	-0.03	0.45	-0.02	2.86	+0.02
	Knowledge	MedQuAD	52.8	+0.2	82.7	+0.0	0.19	-0.02	0.45	-0.02	2.74	-0.10
	Knowledge	Combined	54.8	+2.2	82.8	+0.1	0.18	-0.03	0.45	-0.02	2.76	-0.08
	Conversation	HealthCareMagic-100K	52.3	-0.3	86.8	+4.1	0.23	+0.02	0.42	-0.05	2.25	-0.59
	Conversation	iCliniq10K	52.4	-0.2	85.6	+2.9	0.20	-0.01	0.38	-0.09	2.18	-0.66
	Conversation	MedDialog	52.3	-0.3	83.9	+1.2	0.20	-0.01	0.44	-0.03	2.26	-0.58
	Conversation	NoteChat	52.3	-0.3	85.6	+2.9	0.22	+0.01	0.42	-0.05	**2.04**	**-0.80**
	Conversation	Combined	52.4	-0.2	86.1	+3.4	**0.23**	**+0.02**	0.43	-0.04	2.16	-0.68
	Mixed (20:80)	All Combined	54.2	+1.6	**86.6**	**+3.9**	0.22	+0.01	0.43	-0.04	2.32	-0.52
	Mixed (50:50)	All Combined	**54.8**	**+2.2**	86.4	+3.7	0.22	+0.01	0.41	-0.06	2.48	-0.36
	Mixed (80:20)	All Combined	54.7	+2.1	85.9	+3.2	0.22	+0.01	0.42	-0.05	2.69	-0.15
Qwen2.5-7B	w/o	w/o	51.3	N/A	82.4	N/A	0.21	N/A	**0.45**	N/A	3.01	N/A
	Knowledge	MedQA	**52.8**	**+1.5**	82.4	+0.0	0.19	-0.02	0.43	-0.02	2.95	-0.06
	Knowledge	MedMCQA	52.4	+1.1	82.4	+0.0	0.19	-0.02	0.44	-0.01	2.93	-0.08
	Knowledge	PubMedQA	51.3	+0.0	82.4	+0.0	0.18	-0.03	0.41	-0.04	2.99	-0.02
	Knowledge	MedQuAD	51.4	+0.1	82.3	-0.1	0.21	+0.00	0.43	-0.02	2.93	-0.08
	Knowledge	Combined	54.0	+1.4	82.6	+0.2	0.19	-0.02	0.42	-0.03	2.95	-0.06
	Conversation	HealthCareMagic-100K	51.3	+0.0	84.6	+2.2	0.21	+0.00	0.38	-0.07	2.56	-0.45
	Conversation	iCliniq10K	51.3	-0.0	85.2	+2.8	0.21	+0.00	0.41	-0.04	2.55	-0.46
	Conversation	MedDialog	51.3	+0.0	84.1	+1.7	0.21	+0.00	0.42	-0.03	2.59	-0.42
	Conversation	NoteChat	51.3	-0.0	84.4	+2.0	0.22	+0.01	0.41	-0.04	2.44	-0.57
	Conversation	Combined	51.3	+0.0	85.1	+2.7	0.21	+0.00	0.40	-0.05	**2.39**	**-0.62**
	Mixed (20:80)	All Combined	52.0	+0.7	**86.4**	**+4.0**	**0.22**	**+0.01**	0.42	-0.03	2.59	-0.42
	Mixed (50:50)	All Combined	52.7	+1.4	86.2	+3.8	0.22	+0.01	0.40	-0.05	2.76	-0.25
	Mixed (80:20)	All Combined	52.5	+1.2	85.7	+3.3	0.22	+0.01	0.41	-0.04	2.91	-0.10

Note: ΔFKGL measures the readability gap between question and model response. Lower ΔFKGL indicates better alignment with patient comprehension levels.

we subsample 10,000 instances per dataset, train for up to three epochs with early stopping, and hold model family, optimizer, and schedules constant across all conditions. We compare three fine-tuning strategies: knowledge-only, conversation-only, and mixed (20:80, 50:50, 80:20 knowledge:conversation).

5.2 Evaluation Protocol and Metrics

Models were evaluated on held-out test sets disjoint from the fine-tuning data using five complementary metrics: (1) **Accuracy** on 1,000 QA pairs from knowledge-based datasets (knowledge retention); (2) **BERTScore** on 1,000 randomly sampled conversation-based instances (semantic similarity and dialogue coherence); (3) **HealthBench**, a scenario-based expert rubric of clinical communication across eight domains; (4) **MedCommEval**, our guideline-aligned rubric of professional appropriateness and context sensitivity (scored 0–1; see Section A.7), computed on the same instances as HealthBench; and (5) **ΔFKGL**, the Flesch–Kincaid Grade Level (FKGL) difference between the patient question and the model response (lower indicates better alignment with patient health literacy) (Kincaid et al., 1975), also computed on 1,000 conversation-based samples. This multi-metric protocol enables rigorous comparison of how dataset composition and scenario alignment affect both factual and communicative performance. For overall comparisons (used in H3),

we report a composite score defined as the mean of within-model min–max normalized {Accuracy, BERTScore, HealthBench, $-\Delta$FKGL}; MedCommEval score is analyzed and reported separately given rubric specificity.

Proposed Clinical Communication Metric. We introduce **MedCommEval**, an eight-domain rubric aligned with established clinical guidance (rapport/intro, information gathering, active listening, plain-language explanation, empathy/support, structure/organization, closing/next steps, shared decision making). Domain definitions and indicators are provided in *Appendix* Table 8; scoring anchors (0=*absent*, 1=*partial/unclear*, 2=*clear/present*) are in *Appendix* Table 9 (Section A.7). Each response is double-rated by trained annotators blinded to model identity; disagreements are adjudicated using rubric anchors. Domain scores {0,1,2,NA} are averaged across raters, domains marked NA are excluded, and item scores are rescaled to $[0, 1]$ by division by 2. The corpus-level MEDCOMMEVAL is the mean across items. This rubric complements HealthBench by sensitively capturing guideline-concordant behaviors beyond text similarity.

5.3 Key Findings

Results in Table 3 reveal clear and consistent trends across both LLaMA3-8B and Qwen2.5-7B models regarding the influence of dataset composition and fine-tuning strategy on medical LLM performance.

Finding 1: Knowledge-only improves factual accuracy (supports H1). Knowledge-only fine-tuning yields the largest accuracy gains with negligible lift on communication metrics. For LLaMA3-8B, *Knowledge:Combined* reaches **54.8** Accuracy (+2.2 over baseline) with flat BERTScore/HealthBench and a slight MedCommEval decline (0.47→0.45). Qwen2.5-7B shows the same pattern (e.g., MedQA **52.8**, +1.5).

Finding 2: Conversation-only improves communication and readability (supports H2). Conversation-only fine-tuning delivers the largest gains in BERTScore and modest gains in HealthBench, while substantially narrowing ΔFKGL, with no accuracy improvement. For LLaMA3-8B, *Conversation:Combined* increases BERTScore by **+3.4** (82.7→86.1) and HealthBench by **+0.02**; ΔFKGL improves by 0.68. Qwen2.5-7B mirrors these effects (BERTScore +2.7; ΔFKGL 0.62).

Finding 3: Mixed strategies balance competencies (supports H3). Mixed fine-tuning attains a more integrated profile. For LLaMA3-8B, *Mixed 50:50 (All Combined)* ties the top Accuracy (**54.8**, +2.2) while maintaining high BERTScore (86.4, +3.7) and competitive HealthBench; Qwen2.5-7B shows a parallel trend (52.7, +1.4; 86.2, +3.8). The Composite score is highest for 50:50 mixes in both models. A caveat is that MedCommEval declines modestly across strategies, indicating persistent shortfalls in guideline-concordant behaviors.

Takeaway. The controlled study corroborates the composition effect identified in Section 4: knowledge- and conversation-centric data confer complementary competencies, and mixed fine-tuning best reconciles factual and communicative performance. The residual deficits on MedCommEval motivate process-annotated, guideline-grounded supervision layered atop mixed fine-tuning.

6 Related Work

Recent surveys and systematic reviews have mapped the expanding landscape of text-based datasets and evaluation benchmarks central to medical LLM development. Yan et al. (Yan et al., 2024) provide a systematic taxonomy spanning medical QA, clinical dialogues, and multimodal benchmarks. Zhang et al. (Zhang et al., 2024) categorize datasets by clinical source (e.g., EHRs, biomedical literature, online forums) and data structure (QA pairs, dialogues, clinical notes), highlighting the heterogeneity and evolving complexity of LLM training resources. Wang et al. (Wang et al., 2024b) further analyze dataset properties and training paradigms, including instruction tuning, parameter-efficient fine-tuning, and RLHF, documenting the transition from factoid QA toward open-ended, multi-turn dialogue tasks. Reviews focused on clinical text corpora, such as Spasic and Nenadic (Spasic and Nenadic, 2020) and Wu et al. (Wu et al., 2024), describe the diversity of narrative clinical documents and the challenges of data scarcity, annotation inconsistency, and limited coverage of process-based communication.

Recent surveys more directly address the evaluation of medical LLMs and their supporting resources. Liu et al. (Liu et al., 2024b) provide a critical review of medical LLM datasets, evaluation protocols, and alignment methods, emphasizing the need for domain-specific benchmarks and model adaptation. Wang et al. (Wang et al.,

2024b) and Shi et al. (Shi et al., 2024b) detail the current spectrum of evaluation methods for medical conversational agents, ranging from automatic metrics (e.g., BLEU, ROUGE, BERTScore) to human-centered and scenario-based assessments. Abbasian et al. (Abbasian et al., 2024) propose a comprehensive framework for evaluating generative AI in healthcare, integrating user-centered criteria such as accuracy, trustworthiness, empathy, and computing performance, and systematically critique the inability of surface-level, machine-centric metrics to capture the breadth of clinical and interpersonal requirements. These reviews collectively highlight limitations in prevailing benchmarks, including the neglect of contextual relevance, emotional support, personalization, and ethical dimensions in chatbot deployment.

Although these contributions have advanced the field, prior reviews primarily enumerate datasets or summarize evaluation metrics, often overlooking systematic mapping to real-world clinical scenarios and rigorous critique of metric alignment with clinical communication practice. Building on these foundations, our work introduces a clinical-communication-grounded taxonomy (Section 2) that explicitly maps major text-based medical datasets to core patient–provider scenarios, drawing on established frameworks such as OSCE, SEGUE, and Calgary–Cambridge. This mapping enables systematic identification of coverage gaps, particularly for process-based and ethically complex interactions that are underrepresented in existing corpora. In Section 3, we provide a comprehensive review of evaluation metrics for medical LLMs, synthesizing both standard automatic metrics and human-centered criteria derived from clinical communication frameworks, and critically analyze the extent to which current LLM evaluation practices align or misalign with the demands of real clinical encounters. Section 4 presents a meta-analytical synthesis of the current performance landscape for state-of-the-art medical LLMs, drawing on results from major datasets and a range of evaluation metrics. This integrative perspective advances a unified clinical frame for dataset selection, model evaluation, and performance benchmarking, and provides a roadmap for the patient-centered development of medical LLMs.

7 Conclusion

This study bridges the persistent gap between benchmark-driven development of medical LLMs and the multifaceted demands of patient–provider communication in clinical care. First, we introduce a scenario-based taxonomy derived from established clinical frameworks and systematically map the coverage of widely used medical corpora, revealing both strengths and significant gaps, particularly in process-oriented and ethically complex scenarios, across current knowledge-based and conversation-based datasets. Second, we propose a synthesized framework of core clinical communication skills, rigorously derived from gold-standard assessment instruments, to guide dataset construction, annotation, and evaluation in alignment with real-world clinical competence. Third, our meta-analytical review of state-of-the-art medical LLMs highlights substantial progress in knowledge-centric tasks, while persistent challenges remain in modeling the contextual and interpersonal dimensions of patient-centered communication. Finally, our empirical experiment demonstrates that integrating both knowledge-based and conversation-based data yields more balanced improvements in factual accuracy and communicative competence, but also underscores ongoing limitations of current datasets for supporting guideline-aligned, patient-centered interaction. Collectively, these contributions provide a unified, clinically grounded framework for the systematic development, evaluation, and empirical benchmarking of medical LLMs. Our findings underscore the urgent need to expand scenario coverage, adopt standardized, human-centered evaluation metrics, and integrate clinically validated frameworks throughout the LLM development lifecycle to fully realize the promise of these models for real-world healthcare.

8 Limitations

Despite its comprehensive scope, this review has several limitations. *First*, we focused exclusively on publicly available, text-based datasets to prioritize the review of LLM applications in patient-provider communication, and did not include multimodal corpora (e.g., audio or video).

Additionally, while our systematic review of communication skills draws from gold-standard clinical education frameworks, the operationalization and annotation of these competencies remain inconsistent across current datasets. As a result, the applicability of our evaluation metrics may depend on dataset structure and interaction type; for instance, competencies such as information gathering and active listening are central to multi-turn dialogues but less relevant for single-turn QA.

Lastly, although we critically examine existing evaluation metrics and highlight the need for human-centered, scenario-based assessment, there remains no consensus standard for large-scale benchmarking of communicative competencies. Addressing these gaps will require collaborative development of richer, process-annotated datasets and unified evaluation rubrics to enable robust, clinically meaningful assessment of medical LLMs.

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.

Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, et al. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. *NPJ Digital Medicine*, 7(1):82.

ACGME. 2023. Common program requirements (residency), accreditation council for graduate medical education. Accessed July 23, 2025.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

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*.

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*.

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

B. Hodges, G. Regehr, M. Hanson, and N. McNaughton. 1996. Osce checklists: Do they reflect communication skills? *Academic Medicine*, 71(2):154–158.

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*.

A. Kalet, C. Chou, and R. Ellaway. 2012. Teaching history taking: clinical skills and communication. *Medical Education*, 46(12):1162–1172.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

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.

Suzanne M Kurtz and Jonathan D Silverman. 1996. The calgary—cambridge referenced observation guides: an aid to defining the curriculum and organizing the teaching in communication training programmes. *Medical education*, 30(2):83–89.

Prajwol Lamichhane and Indika Kahanda. 2023. Enhancing health information retrieval with large language models: A study on medquad dataset. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 2147–2152. IEEE.

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*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. 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*.

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.

Gregory Makoul. 2001a. Essential elements of communication in medical encounters: the kalamazoo consensus statement. *Academic medicine*, 76(4):390–393.

Gregory Makoul. 2001b. The segue framework for teaching and assessing communication skills. *Patient education and counseling*, 45(1):23–34.

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

S.W. Mercer, M. Maxwell, D. Heaney, and G.C. Watt. 2004. The consultation and relational empathy (care) measure: Development and preliminary validation and reliability of an empathy-based consultation process measure. *Family Practice*, 21(6):699–705.

W.R. Miller and S. Rollnick. 2012. *Motivational Interviewing: Helping People Change*. Guilford Press.

David Newble. 2004. Techniques for measuring clinical competence: objective structured clinical examinations. *Medical education*, 38(2):199–203.

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*.

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*.

C. Ryan, D. Safran, and A. Brown. 2001. The roter interaction analysis system (rias): Utility and limitations for health communication research. *Patient Education and Counseling*, 45(2):93–99.

D. Schillinger, J. Piette, and K. et al. Grumbach. 2003. Closing the loop: physician communication with diabetic patients who have low health literacy. *Archives of Internal Medicine*, 163(1):83–90.

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.

Jonathan Silverman, Suzanne Kurtz, and Juliet Draper. 2016. *Skills for communicating with patients*. crc press.

M. Simpson and R. Buckman. 2007. Uncertainty in medical communication: A review and exploratory model. *Patient Education and Counseling*, 67(3):267–277.

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.

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

Haochun Wang, Sendong Zhao, Zewen Qiang, Bing Qin, and Ting Liu. 2024a. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. *CoRR*.

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. 2023. Notechat: a dataset of synthetic doctor-patient conversations conditioned on clinical notes. *arXiv preprint arXiv:2310.15959*.

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. 2024. Clinical text datasets for medical artificial intelligence and large language models — a systematic review. *NEJM AI*.

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*.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024. 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*.

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.

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.

A Overview of Text-based medical datasets

In this appendix, we present Table 4, 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 4 and Objectives

Table 4 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 4 are:

- Publicly available or well-documented.
- Explicit focus on medical patient communication, including diagnostic QA, 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 4 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 QA, 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 QA 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 QA**: Datasets such as *PubMedQA, BioASQ, MedQuAD*, extracting knowledge from academic sources.
- **Electronic Health Record (EHR)-Based QA**: 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.

A.4 Construction and Purpose of Table 5

Table 5 presents a structured taxonomy of core clinical scenarios that define patient–provider communication. This taxonomy is systematically synthesized from gold-standard frameworks widely adopted in medical education and assessment, including the Objective Structured Clinical Examination (OSCE), Accreditation Council for Graduate Medical Education (ACGME) competencies, the SEGUE framework, the Calgary–Cambridge Guide, and the Kalamazoo Consensus Statement.

The table enumerates 12 distinct clinical scenarios, each corresponding to a critical communicative context in patient care. For each scenario, the table provides:

A concise scenario name (e.g., "History Taking and Initial Assessment"), A definition clarifying its scope and communicative purpose, The frameworks or guidelines from which the scenario is derived. This taxonomy underpins our scenario-based mapping of dataset coverage (see Section 2), enabling fine-grained analysis of whether existing datasets adequately reflect the full range of clinical communication requirements. It also supports downstream evaluation and model development by explicitly connecting scenario types to the competencies expected in real-world practice.

A.5 Overview and Organization of Table 6

Table 6 offers a comprehensive summary of the evaluation metrics most commonly used in the assessment of medical LLMs. Metrics are grouped by their methodological approach (automated vs. human-rated), the type of task they are designed to assess, and their relevance to either knowledge-based or communication-based evaluation.

Columns in this table include:

Metric: The name of the metric (e.g., Accuracy, F1 Score, BLEU, Empathy Score). Definition: A brief description of what the metric measures and how it is calculated. Rater: Whether the metric is computed automatically or via human annotation. Application: The context or task where the metric is typically applied (e.g., MCQA, dialogue, summarization, counseling). Reference: Key literature sources or benchmarks that have used the metric. The table distinguishes between classical metrics for factual correctness (e.g., Accuracy, F1, Exact Match), generative metrics for open-ended or dialogue tasks (e.g., BLEU, ROUGE, BERTScore), and human-centered metrics for nuanced attributes like empathy, fluency, or counseling effectiveness. Composite or task-specific metrics, such as Hallucination Score or Proactive Questioning, are also included to highlight recent methodological advances in model assessment.

This overview enables practitioners and researchers to select appropriate evaluation criteria for different clinical communication scenarios and model objectives, while highlighting the limitations of relying solely on surface-level or automated metrics.

A.6 Structure and Rationale of Table 7

Table 7 catalogs the principal frameworks and checklists used to evaluate patient–provider communication, especially in the context of human-centered or scenario-based model evaluation. These instruments are drawn from validated tools in clinical education, such as the Calgary–Cambridge Guide, SEGUE Framework, OSCE communication subscores, and specialty checklists for empathy, teach-back, and uncertainty communication.

Key columns include:

Metrics: The name of the evaluation framework or checklist. Definition: A short description of the framework's overall focus and intended use. Core Elements: The principal domains or items assessed (e.g., rapport, information gathering, explanation, empathy, shared decision making). Scene: The clinical scenarios (by scenario number from Table 5) where the framework is most applicable. The table clarifies how each instrument decomposes clinical communication into observable skills, offering granularity for both formative (feedback, education) and summative (high-stakes assessment, benchmarking) evaluation. For instance, the Calgary–Cambridge Guide and SEGUE Framework assess communication across all phases of the clinical encounter, while empathy scales and teach-back checklists provide focused measurement for specific communicative functions.

This structured inventory is intended to guide both model developers and evaluators in aligning dataset annotation, training objectives, and LLM evaluation protocols with authentic clinical standards and gold-standard assessment practices.

A.7 Evaluation Criteria and Scoring Rubric for Medical LLM Patient-Provider Communication

To systematically evaluate the patient-provider communication skills of medical LLMs, we

adopted a comprehensive set of communication metrics and a structured scoring rubric. This framework was designed to assess the quality of single-turn, open-ended responses generated by medical LLMs in simulated clinical consultations.

Domains of Evaluation in Table Table 8: The evaluation rubric covers eight key domains of patient-provider communication, each reflecting a critical aspect of effective clinical interaction:

- *Rapport/Introduction*: Initiating the conversation warmly, establishing trust, and clarifying the purpose.
- *Information Gathering/Questioning*: Actively eliciting patient details, clarifying symptoms, and seeking relevant background.
- *Active Listening/Clarification*: Demonstrating understanding through paraphrasing, summarization, or clarification requests.
- *Information Giving/Explanation*: Providing clear, relevant, and accessible information in response to patient queries.
- *Empathy/Support*: Recognizing and validating patient emotions, and offering support or encouragement.
- *Structure/Organization*: Presenting information logically, with coherent transitions and clear progression.
- *Closing/Next Step*: Summarizing key points, outlining follow-up actions, and providing closure.
- *Shared Decision Making*: Presenting care options, discussing pros and cons, and inviting patient participation in decision-making.

Scoring Guidelines in Table 9: Each domain is scored based on the presence and quality of relevant communicative behaviors, according to the following scale:

Communication Scoring Guidelines

0 = Absent: The domain is not demonstrated in the LLM's response.
1 = Partial/Unclear: The domain is present but weak, unclear, generic, or incomplete.
2 = Clear/Present: The domain is clearly and effectively demonstrated.
NA = Not Applicable: Used only if the domain cannot logically be expressed in a single-turn response.

For each response, raters assign scores and provide brief explanations or direct quotes to justify their ratings for each domain.

Evaluation Procedure: Raters are instructed to:

- Read the patient query and the LLM's response in full.
- Assess each communication domain using the definitions, behavioral indicators, and anchor examples provided in the rubric.
- Assign a score (0,1,2,or NA) for each domain and record a justification.
- Enter scores and explanations into a structured output table.

This rubric is the primary instrument employed in our experiments to evaluate the patient-provider communication quality of various medical LLMs, including Qwen and Llama, across different test sets and interaction scenarios in our experiments.

Table 4: Summary of Medical QA and Communication Datasets

Dataset	Source	Data Type	Annotation	Scale
Knowledge-Based Datasets				
MedQA (Jin et al., 2021)	Medical Licensing and Board Exam Question Bank	MCQA medical licensing exam		∼60K
MedMCQA (Pal et al., 2022)		MCQA medical exam and mocked tests created by human experts	Explanations provided	∼193K
MultiMedQA (Singhal et al., 2022)		MCQA and Open QA synthesized from 7 medical QA datasets (MedQA, MedMCQA, PubMedQA, MMLU, LiveQA, MedicationQA, HealthSearchQA)		∼474K development set and 9K test set
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
XMedBench (Wang et al., 2024d)		MCQA synthesized from multilingual medical QA datasets		
BioASQ (Krithara et al., 2023)	Scientific Literature-Based Medical QA	Biomedical QA (including both exact answer and ideal answer) from scientific literature with references and supporting material	Structured QA labels (e.g., question type, concept, answer, reference, supporting material)	∼5K
PubMedQA (Jin et al., 2019)		Biomedical QA collected from PubMed abstracts	Each QA instance labeled: Question + Context + Long Answer + Final Answer (yes/no/maybe)	∼1k expert-annotated, 211.3k generated QA, 61.2k unlabeled
MedQuAD (Abacha et al., 2019)		Medical QA sourced from NIH websites	Each QA instance labeled: Question + Answer + Source + Focus Area	∼47K
BiMed1.3M (Pieri et al., 2024)		MCQA, medical Q&A, and multi-turn patient communication simulated with ChatGPT		∼1.3M samples (423.8K Q&A, 638.1K MCQA, 249.7K chat)
Medication QA (Abacha et al., 2019)		Medication Q&A	Each QA instance labeled: Focus (Drug) + Question Type + Answer + Section Title + URL	674
emrQA (Pampari et al., 2018)	Electronic Health Record-Based QA	EHR-based Q&A, including both question-logical form pairs and QA pairs	EHR documents annotated with QA (QA and Question-Logical Form-Answer Evidence)	∼1M question-logical form pairs, 400K Q&A
Communication-Based Datasets				
HealthCareMagic-100K (Li et al., 2023b)	Single-Turn Online Symptom Inquiry	Real-world user queries with doctor responses on an online health platform		∼100K
iCliniq10K (Li et al., 2023b)		Real-world user queries with doctor responses on an online health platform		∼10K
cMedQA (Zhang et al., 2017)		Real-world patient queries answered by doctors from online medical QA forum	Question with a pair of ground truth answer and an incorrect answer; Total Questions: Q (54K) & A (102K); Training Set: Q (50K) & A (94K); Development Set: Q (2K) & A (4K); Test Set: Q (2K) & A (4K)	
Huatuo-26M (Li et al., 2023a)		Real-world patient queries answered by doctors from online medical QA forum; Medical QA collected from medical encyclopedia; Medical QA collected from knowledge graph		∼26M
BianQue Corpus (Chen et al., 2023a)	Multi-Turn Patient-Provider Conversation	Real-world multi-turn doctor-patient conversations		∼2.4M conversation samples
MedDG (Liu et al., 2022)		Real-world multi-turn doctor-patient conversations	Each sentence labeled: Role (Doctor/Patient) + Symptom + Medicine + Examination + Attribute + Disease	18K
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		∼3.4M conversations in Chinese, 0.26M conversations in English
NoteChat (Wang et al., 2023)		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		∼10K
CMtMedQA (Yang et al., 2024)		Real-world multi-turn doctor-patient conversations standardized with self-instruction method		70K multi-turn dialogues and 400K single-turn conversations
Psych8K (Liu et al., 2023)	Mental Health Counseling Transcript	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

Table 5: Taxonomy of Clinical Scenarios of Patient-Provider Communication

No.	Scenario	Definition	Frameworks
1	History Taking and Initial Assessment	*Systematic gathering of information about a patien's symptoms, concerns, and relevant background to inform diagnosis and initial care.*	OSCE, ACGME, SEGUE, Calgary–Cambridge
2	Screening, Routine and Preventive Care	*Conducting proactive health maintenance activities, including screening, immunizations, and preventive advice, typically during well or follow-up visits.*	OSCE, ACGME, Calgary–Cambridge
3	Patient Education	*Provision of clear, relevant, and accessible health information about conditions, tests, treatments, or prevention, tailored to patient understanding.*	OSCE, ACGME, SEGUE, Calgary–Cambridge, Kalamazoo
4	Counseling and Behavioral Intervention	*Providing guidance, motivation, and support for health-related behavior change or psychosocial issues (e.g., lifestyle, risk reduction, coping).*	OSCE, ACGME, SEGUE, Calgary–Cambridge
5	Informed Consent	*Ensuring the patient understands and voluntarily agrees to a proposed intervention, including its risks, benefits, and alternatives.*	OSCE, ACGME, SEGUE, Calgary–Cambridge
6	Counseling on Procedures	*Explaining the purpose, steps, risks, benefits, and aftercare of medical or surgical procedures, and addressing any patient questions or concerns.*	OSCE, ACGME, SEGUE, Calgary–Cambridge
7	Shared Decision Making	*Engaging the patient as a partner in choices about their care, including eliciting values, presenting options, and supporting deliberation.*	ACGME, SEGUE, Calgary–Cambridge, Kalamazoo
8	Breaking Bad News	*Delivering difficult, unexpected, or life-altering information with empathy and clarity, while supporting the patient's emotional response.*	OSCE, SEGUE, Calgary–Cambridge, Kalamazoo
9	Acute and Emergency Encounters	*Assessing and managing new, urgent, or life-threatening symptoms or events, requiring rapid clinical action and clear, focused communication.*	OSCE, ACGME, Calgary–Cambridge
10	Rehabilitation / Chronic Disease Management	*Supporting ongoing care, self-management, and functional recovery for patients with chronic illness or disability, emphasizing partnership and follow-up.*	OSCE, ACGME, Calgary–Cambridge
11	Referrals and Care Transitions	*Coordinating and communicating the handover of care to another provider or service to ensure continuity and patient understanding.*	OSCE, ACGME, SEGUE, Kalamazoo
12	Addressing Patient Concerns / Barriers	*Identifying and helping patients overcome social, cultural, financial, or psychological barriers to care and responding to specific worries or needs.*	OSCE, ACGME, SEGUE, Calgary–Cambridge, Kalamazoo

Note: OSCE = Objective Structured Clinical Examination (Newble, 2004); ACGME = Accreditation Council for Graduate Medical Education Common Program Requirements (ACGME, 2023); SEGUE = SEGUE Framework for communication (Makoul, 2001b); Calgary–Cambridge = Calgary–Cambridge Guide to the Medical Interview (Kurtz and Silverman, 1996; Silverman et al., 2016); Kalamazoo = Kalamazoo Consensus Statement (Makoul, 2001a).

Table 6: Common Evaluation Metrics for Medical LLM Performance

Metric	Definition	Rater	Application	Reference
Accuracy	*Proportion of correct answers out of total questions/instances*	Automated	MCQ, open-ended QA	(Jin et al., 2021; Pal et al., 2022)
F1 Score	*Harmonic mean of precision and recall*	Automated	QA, entity prediction, named entity recognition (NER)	(Jin et al., 2019; Abacha et al., 2019)
Precision	*Proportion of true positives out of predicted positives*	Automated	Entity extraction, QA	(Liu et al., 2022; Abacha et al., 2019)
Recall	*Proportion of true positives out of actual positives*	Automated	Entity extraction, QA	(Liu et al., 2022; Abacha et al., 2019)
Exact Match	*Percentage of answers that exactly match the reference answer*	Automated	Short answer QA, open-ended QA	(Abacha et al., 2019)
BLEU	*N-gram precision score between generated and reference text*	Automated	Dialogue generation, summarization	(Liu et al., 2022; Zeng et al., 2020)
ROUGE	*Recall-oriented measure of overlapping n-grams/sequences*	Automated	Dialogue, summarization, open-ended QA	(Liu et al., 2022; Wang et al., 2023)
BERTScore	*Embedding-based semantic similarity between generated and reference text*	Automated	Dialogue, summarization	(Liu et al., 2022; Wang et al., 2023)
Perplexity	*Measures how well the model predicts the next token*	Automated	Language modeling, dialogue, QA	(Zeng et al., 2020)
Proactive Questioning Assessment	*Measures LLM's ability to proactively ask relevant follow-up questions*	Automated / Human	Medical chatbots, proactive dialogue, triage	(Chen et al., 2023a; Yang et al., 2024)
Fluency	*Human annotator assessment of grammatical, readable, natural text*	Human	Dialogue, counseling, summarization	(Liu et al., 2023, 2022; Wang et al., 2023)
Coherence	*Logical flow and context retention in multi-turn dialogue*	Human	Dialogue, counseling	(Liu et al., 2023, 2022; Wang et al., 2023)
Completeness	*Degree to which the response covers all required information*	Human	QA, clinical case assessment, summarization	(Jin et al., 2021; Zeng et al., 2020)
Adequacy	*Appropriateness / informativeness*	Human	QA, clinical dialogue	(Zeng et al., 2020)
Hallucination Score	*Rate of unsupported or incorrect content in output*	Automated / Human	Summarization, dialogue, QA	(Liu et al., 2022; Dou et al., 2023)
Consistency Score	*Degree of alignment with ground truth and logical flow across dialogue turns*	Automated / Human	Dialogue, QA	(Liu et al., 2022; Dou et al., 2023)
Intent Accuracy	*Correct prediction of dialogue intent or action*	Automated	Dialogue act prediction, task-oriented dialogue	(Zeng et al., 2020; Dou et al., 2023)
Counseling Metrics	*Multi-dimensional scores for empathy, listening, guidance, etc.*	Human	Mental health counseling, communication evaluation	(Liu et al., 2023; Wang et al., 2023)

Table 7: Performance Evaluation Metrics for Patient-Provider Communication

Metrics	Definition	Core Elements	Scene
Calgary-Cambridge Guide (Kurtz and Silverman, 1996)	*Comprehensive assessment of all phases of clinical communication; used for education, feedback, research.*	5 domains: initiating session, gathering info, building relationship, explanation/planning, closing session. Items: rapport, agenda-setting, open/closed Qs, active listening, summarizing, patient cues, explanations, shared decisions, closure.	1–12
SEGUE Framework (Makoul, 2001b)	*Structured observer rating for patient-centered provider communication.*	5 domains: set stage, elicit info, give info, understand perspective, end encounter. Items: greeting, respect, open Qs, info delivery, clarification, empathy, follow-up.	1–12
OSCE Communication Subscores (Hodges et al., 1996)	*Standardized patient/examiner checklist for core communication in clinical scenarios and high-stakes exams.*	Items: rapport, open/closed Qs, listening, empathy, info giving, structure, summary, shared decision making, professionalism, closing.	1–12
Empathy Scales (CARE (Mercer et al., 2004), RIAS (Ryan et al., 2001))	*Evaluate relational and affective communication.*	CARE: at ease, listening, understanding, compassion, clear explanations, empowerment. RIAS: frequency coding for empathy, partnership, social talk, open/closed Qs, info-giving.	3, 4, 5, 6, 8, 10, 12
Teach-Back Checklist (Schillinger et al., 2003)	*Assess whether provider checks patient understanding via teach-back.*	Items: use plain language, ask patient to restate info, clarify misunderstanding, check understanding at key points, encourage questions.	2, 3, 4, 10, 12
History-Taking Competency Checklist (Kalet et al., 2012)	*Evaluate comprehensiveness and structure of clinical history-taking.*	Items: introduction, agenda, open/closed Qs, listening, summarize/clarify, HPI, PMH, meds, allergies, family/social, risk factors, closure with plan.	1, 9, 10, 12
Uncertainty Communication Checklist (Simpson and Buckman, 2007)	*Assess provider skill in discussing diagnostic/prognostic uncertainty.*	Items: state uncertainty, explain reasons, share diagnostic reasoning, next steps, discuss risks/benefits, invite questions, check understanding.	3, 5, 6, 8, 10, 12
Counseling & Intervention Skills Checklist (Miller and Rollnick, 2012)	*Assess skills in motivational interviewing and behavioral counseling.*	Items: elicit readiness to change, reflective listening, affirm strengths, avoid argument, offer advice w/ permission, summarize, collaborate on plan, follow-up.	4, 7, 10, 12

Table 8: Comprehensive Evaluation Metrics for Patient-Provider Communication

Domain	Definition	Example Behavioral Indicators	Sample Rating
1. Rapport/ Introduction	*Initiates warmly, establishes trust, clarifies purpose or agenda.*	Greets or welcomes patient; addresses respectfully; introduces self or role; thanks patient; clarifies purpose; acknowledges receipt of query.	0–2 scale (absent/-partial/clear)
2. Information Gathering/ Questioning	*Attempts to elicit more information, clarify symptoms, or check details.*	Asks open-ended questions ("Can you tell me more about..."); seeks details; clarifies ambiguous symptoms; checks for associated symptoms ("Do you also experience...?").	0–2 scale; checklist for open/focused Qs
3. Active Listening/ Clarification	*Demonstrates understanding by reflecting, paraphrasing, summarizing, or checking comprehension.*	Restates or paraphrases patient's main complaint; summarizes patient's symptoms; reflects emotion or concern ("You mentioned you feel dizzy and nauseous..."); seeks clarification.	0–2 scale; number of clarifications
4. Information Giving/ Explanation	*Provides relevant information or advice in clear, accessible language.*	Clear explanation of diagnosis/condition; uses plain language; avoids jargon; organizes explanation logically; checks for understanding ("Does that make sense?"); uses teach-back techniques.	0–2 scale; checklist for plain language
5. Empathy/ Support	*Expresses understanding, validation, support, or reassurance.*	Recognizes and responds to patient emotions ("I can understand this is upsetting..."); affirms patient's concerns; offers encouragement; provides supportive statements.	0–2 scale; empathy/support phrase count
6. Structure/ Organization	*Arranges information logically, uses transitions and summaries, avoids confusion.*	Clear order of ideas; logical flow; uses transition phrases ("Now, let's discuss..."); summarizes before moving to next topic; no abrupt topic changes or disorganization.	0–2 scale; checklist for transitions
7. Closing/ Next Step	*Concludes with summary, next steps, or explicit invitation for further questions.*	Summarizes advice or diagnosis; outlines follow-up steps; invites further questions ("Is there anything else?"); thanks patient; gives clear or explicit action.	0–2 scale; presence of summary/plan
8. Shared Decision Making	*Engages the patient in choices about their care, values/preferences, or options.*	Presents more than one option; discusses pros/cons; asks for patient's preferences ("Would you prefer...?"); supports patient autonomy; encourages participation in decisions.	0–2 scale; checklist for SDM behaviors

Table 9: Scoring Rubric for Patient–Provider Communication. Each domain is scored on a 3-point scale: 0 = Absent (behavior not demonstrated), 1 = Partial/Unclear (behavior partially demonstrated or unclear), 2 = Clear/Present (behavior clearly demonstrated).

Domain	**Score = 0**	**Score = 1**	**Score = 2**	**Example Anchors / Behaviors**
Rapport / Introduction	No greeting or acknowledgment	Generic greeting; lacks warmth or purpose	Warm greeting, thanks, and clear purpose/role introduction	"Hello, thank you for your question. I'm Dr. X. How can I help you today?"
Information Gathering	No questions; no attempt to elicit information	Vague or closed questions; limited clarification	Open-ended focused questions; adapts to patient context	"Can you tell me more about your symptoms? When did they start?"
Active Listening	Ignores or misinterprets concerns	Some acknowledgment, but incomplete or vague	Restates/paraphrases key concerns; summarizes points	"You mentioned dizziness when standing up and nausea—that sounds uncomfortable."
Information Giving	No explanation or unclear jargon-heavy info	Partial or unclear information	Clear, relevant, organized explanation; checks understanding	"BPPV is a common cause of dizziness and usually resolves in days."
Empathy / Support	No emotional recognition or support	Generic sympathy ("I understand"), nonspecific	Explicit recognition and validation; offers encouragement	"I can understand this is worrying for you. You're not alone—many feel this way."
Structure / Organization	Disorganized; abrupt topic changes	Some structure but weak sequencing	Logical flow; clear transitions and signposting	"First, let's review your symptoms. Then we will discuss treatment options."
Closing / Next Steps	Abrupt ending; no summary or plan	Mentions next steps but unclear or brief	Summarizes points; outlines plan; invites questions; expresses thanks	"To sum up, I recommend seeing an ENT. Let me know if you have more questions."
Shared Decision Making	Presents only one option; ignores preferences	Mentions options but no meaningful engagement	Discusses options, pros/cons; elicits preferences; supports autonomy	"There are several approaches. Would you prefer exercises or seeing a specialist?"