Between Knowledge and Care: Evaluating Generative AI-Based IUI in Type 2 Diabetes Management Through Patient and Physician Perspectives

Yibo Meng[1,*,†], Ruiqi Chen[2,†], Bingyi Liu[3], Yan Guan[1] and Xiaolan Ding[4]

[1]Tsinghua University, 30 Shuangqing Rd, Haidian District, Beijing 100190, China
[2]University of Washington, 1410 NE Campus Pkwy, Seattle, WA 98195, USA
[3]University of Michigan, Ann Arbor – School of Information, 105 South State Street Ann Arbor, MI 48109-1285, USA
[1]Tsinghua University, 30 Shuangqing Rd, Haidian District, Beijing 100190, China
[4]North China University of Science and Technology Health Science Center, Tangshan, Hebei, 063210, China

Abstract

Generative AI systems are increasingly used by patients seeking everyday health guidance, yet their appropriateness in chronic care contexts remains unclear. Focusing on Type 2 Diabetes Mellitus (T2DM), this paper presents a mixed-methods investigation into how AI-generated health information is interpreted by patients and evaluated by physicians in China. Drawing on formative patient grounding and a dimension-based physician evaluation, we examine AI responses along five quality dimensions: *Accuracy*, *Safety*, *Clarity*, *Integrity*, and *Action Orientation*. Our findings reveal that while current systems perform well in factual explanation and general lifestyle guidance, they frequently break down in safety signaling, contextual judgment, and responsibility boundaries—particularly when fluent responses invite overtrust. By treating quality dimensions as an interpretive lens rather than a fixed framework, this work highlights the need for intelligent user interfaces that actively mediate AI outputs in chronic disease management, supporting calibrated trust and responsible boundary-setting in long-term care.

Keywords

Generative AI, T2DM Management, Chronic Care Support

1. INTRODUCTION

Generative AI systems are rapidly entering consumer-facing health contexts, raising renewed interest in how large language models (LLMs) might support chronic disease self-management. While recent work demonstrates that LLMs can synthesize biomedical knowledge and produce fluent explanations [1, 2], fluency and coverage alone do not guarantee that health information is safe, actionable, or interactionally appropriate for real users [3]. These concerns are particularly salient in chronic care, where patients repeatedly seek guidance under uncertainty, and where small omissions or poorly framed advice can accumulate into meaningful risk over time [4, 5, 6].

In this work, we focus on Type 2 Diabetes Mellitus (T2DM) as a representative chronic condition to examine the role boundaries and reliability of generative AI in everyday self-management. T2DM requires sustained decisions around medication, diet, physical activity, and psychosocial regulation [7]. In China, where clinical workloads are high and specialist access is uneven, patients increasingly turn to always-available digital sources—including generative AI tools—to interpret symptoms, translate medical terminology, and explore self-care options outside the clinic [2, 8, 9]. This usage creates a tension central to Human Computer Interaction (HCI) and IUI: AI can lower barriers to health information access, yet it can also invite overreliance when its outputs appear confident but fail to reflect contextual constraints and responsibility boundaries.

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
*Corresponding author.
†These authors contributed equally.
mengyb22@tsinghua.org.cn (Y. Meng); ruiqich@uw.edu (R. Chen); bingyi@umich.edu (B. Liu); guany@tsinghua.edu.cn (Y. Guan); 13838038961@163.com (X. Ding)
0009-0009-0370-5456 (Y. Meng); 0000-0001-7255-5353 (R. Chen); 0009-0000-4312-153X (B. Liu); 0009-0004-5107-5050 (Y. Guan)

A core limitation in prior evaluations is that health information quality is often operationalized as a largely one-dimensional property—typically correctness on benchmark-style question answering—despite mounting evidence that patients and clinicians attend to additional dimensions such as safety signaling, coherence, and the practical framing of next steps [2, 3]. More broadly, HCI research has shown that interface fluency can inflate perceived credibility, complicating calibrated trust in safety-critical settings [4]. What remains underexplored is how clinicians interpret these quality dimensions in practice when reading AI-generated responses to authentic patient questions, and how such dimension-level judgments can inform the design of health IUIs that support appropriate reliance rather than superficial trust.

To address this gap, we conducted a two-phase mixed-methods study. Phase 1 served as a formative grounding study to capture how people with T2DM engage with generative AI in real-world self-management contexts and to inform ecologically grounded evaluation stimuli. Phase 2 then conducted a physician-centered evaluation of four widely used generative AI systems using a dimension-based rubric that reflects clinically meaningful criteria for information adequacy and risk. Across quantitative ratings and follow-up interviews, we examine how models exhibit systematic trade-offs across *accuracy*, *safety*, *clarity*, *integrity*, and *action orientation*, and how these trade-offs shape clinicians' trust judgments.

Our contributions are threefold: (1) a dimension-based expert evaluation of contemporary generative AI systems on patient-authored T2DM questions, highlighting systematic trade-offs rather than a single "best model" narrative; (2) qualitative evidence on how physicians operationalize quality dimensions through concrete judgment moments and failure modes; and (3) design implications for dimension-aware health IUIs, including mediation strategies that calibrate trust by aligning actionability and presentation with clinically grounded risk boundaries. Together, our findings argue that chronic care contexts demand a shift from answer-centric conversational agents toward IUIs that actively mediate model outputs to support safe, context-sensitive reliance.

2. RELATED WORK

2.1. Information Needs and Quality Judgments in AI-Mediated Chronic Care

Type 2 diabetes mellitus is a chronic condition that requires patients to engage in sustained information seeking and self-management across medication, diet, physical activity, and emotional regulation [7, 10]. Prior work in HCI and health informatics has shown that patients often struggle to interpret clinical guidelines, contextualize numerical indicators such as HbA1c, and translate abstract medical knowledge into everyday action [2, 11].

In practice, patients frequently supplement limited clinical consultations with online sources and AI-powered tools, particularly for routine questions and early-stage sensemaking [12, 3]. Importantly, studies consistently show that patient needs extend beyond factual explanation to include lifestyle adaptation, behavioral guidance, and emotional coping [13, 14]. These findings suggest that the perceived usefulness of health information is shaped not only by correctness, but also by how well information is contextualized, interpreted, and translated into action.

Within HCI, T2DM has served as a recurring design context for intelligent health technologies such as education tools, self-tracking systems, and coaching interfaces [15, 16]. While these systems demonstrate the value of tailored feedback, they typically rely on rule-based logic or pre-scripted content. The emergence of generative AI introduces new flexibility in health information access, but also raises questions about how patients assess the appropriateness, safety, and actionability of AI-generated content in chronic care contexts.

2.2. Evaluating and Mediating AI-Generated Health Information

Recent advances in large language models (LLMs) have enabled conversational access to health information on topics such as diet, exercise, and medication management [17, 18]. At the same time, growing

evidence suggests that AI-generated health content may be inaccurate, incomplete, or misleading when applied to domain-specific scenarios such as diabetes care [9, 19].

Most existing evaluations rely on benchmark-style medical question–answering datasets that prioritize factual correctness at the answer level [20, 21]. While valuable for model comparison, such approaches offer limited insight into how information is interpreted and acted upon by patients. In response, researchers in HCI and intelligent user interfaces have argued for multidimensional evaluation frameworks that account for safety, clarity, and actionability in addition to accuracy [4, 22].

Despite this shift, chronic disease contexts remain underexplored from a clinical evaluation perspective. Existing studies on AI use in T2DM primarily focus on usability or patient satisfaction, with limited involvement of clinicians in assessing risk, appropriateness, and responsibility boundaries [23, 24]. Prior work in health-oriented IUI design highlights the importance of risk-sensitive interaction strategies such as uncertainty communication and deferral to clinicians [5, 25], yet these mechanisms are rarely grounded in systematic clinical judgment.

Together, these gaps motivate the need for clinician-informed, dimension-based evaluation of AI-generated health information, as well as interface-level mediation strategies that support calibrated trust and appropriate use in chronic care settings.

3. METHOD

Our study followed a two-phase mixed-methods design. Phase 1 served as a formative grounding study to surface how patients conceptualize and evaluate AI-generated health information in everyday self-management contexts. Phase 2 then built on this grounding to conduct a dimension-based expert evaluation with physicians. Importantly, Phase 1 was not intended to produce standalone empirical findings; instead, it established the evaluative dimensions, stimuli, and scope for the physician study that follows.

3.1. Phase 1: Formative Grounding of Evaluation Dimensions

Phase 1 explored how patients with T2DM interact with generative AI tools in real-world self-management scenarios. The goal of this phase was epistemic rather than confirmatory: to surface the implicit criteria patients use when judging AI-generated health information, and to ground the construction of evaluation dimensions used in Phase 2.

We focused on individuals with type 2 diabetes mellitus, a chronic condition that requires sustained information seeking, behavioral regulation, and interpretive judgment. This context allowed us to observe how patients engage with AI-generated health information across routine, ambiguous, and emotionally laden situations, thereby revealing tensions around correctness, safety, interpretability, coherence, and actionability that are difficult to capture through expert-authored prompts alone.

3.1.1. Participants

We recruited participants through a combination of online and offline channels. Online recruitment notices were posted on widely used Chinese social platforms, including Xiaohongshu, Bilibili, and Baidu Tieba. Offline recruitment was conducted in community centers and public spaces across four provinces in Northern China (Henan, Hebei, Shanxi, and Shandong). Eligibility criteria required participants to be at least 18 years old, have a clinical diagnosis of T2DM, and possess prior experience using generative AI tools (e.g., ChatGPT, Wenxin Yiyan, or similar systems) for diabetes-related information seeking or self-management.

Participants were excluded if they had not been diagnosed with T2DM, were experiencing severe complications or other serious illnesses that could interfere with participation, or had recently taken part in other clinical studies with overlapping objectives. All participants provided informed consent prior to participation.

A total of 21 participants were enrolled, including 11 males and 10 females, ranging in age from 33 to 65 years ($M = 47.76$, $SD = 8.14$). Duration of T2DM ranged from 1 to 23 years ($M = 8.10$, $SD = 5.52$). Twelve participants resided in rural areas and nine in urban settings. Educational backgrounds varied substantially, ranging from no formal education to bachelor's degrees. Table 1 summarizes participant demographics.

The study protocol was reviewed and approved by the Institutional Review Board (IRB) of the Author's Institute. All collected data were anonymized prior to analysis, and participants retained the right to withdraw or request data deletion at any time. Each participant received a compensation of 30 RMB.

Table 1
Demographic characteristics of participants in the observational study (N=21).

ID	Age	Years with T2DM	Gender	Residence	Education
1	44	5	F	Urban	High school
2	45	5	M	Urban	High school
3	54	4	M	Rural	Middle school
4	55	12	M	Urban	Middle school
5	56	13	F	Rural	Primary school
6	65	23	F	Urban	Primary school
7	57	17	M	Rural	None
8	56	11	F	Urban	High school
9	45	11	M	Rural	High school
10	38	3	F	Urban	Bachelor
11	36	3	M	Urban	Bachelor
12	39	2	F	Urban	High school
13	33	4	M	Rural	Bachelor
14	45	11	M	Rural	High school
15	56	12	F	Urban	None
16	52	10	F	Rural	None
17	49	6	M	Rural	High school
18	41	1	M	Rural	Bachelor
19	46	3	F	Urban	Middle school
20	44	7	M	Rural	None
21	47	7	F	Rural	None

3.1.2. Procedure

Procedures in Phase 1 were intentionally designed to elicit implicit evaluative concerns that patients bring to AI-mediated health interactions, rather than to benchmark usability or system performance. The study followed a multi-stage protocol combining structured questionnaires and semi-structured interviews.

Participants first completed two structured instruments: an AI Attitude Questionnaire (Appendix A) and an AI Usage Questionnaire (Appendix B). The attitude questionnaire comprised ten items rated on a 10-point Likert scale (1 = strongly disagree, 10 = strongly agree), probing perceptions related to convenience, perceived usefulness, personalization, psychological support, risk awareness, and overall satisfaction. The usage questionnaire captured participants' prior experiences with generative AI systems, including platforms used, interaction modalities (e.g., mobile applications or embedded assistants), and examples of typical information-seeking scenarios.

To contextualize questionnaire responses and surface evaluative reasoning not easily expressed through fixed-response items, we conducted 35–55 minute semi-structured interviews with each participant. Interview questions explored participants' medical background, daily self-management practices, experiences with AI-generated health information, perceived benefits and shortcomings, and expectations for future AI support. Interviews were conducted either in person or via phone, depending

on participant availability. With participant consent, all interviews were audio-recorded and transcribed verbatim within 24 hours. Each transcript was cross-checked to ensure accuracy and completeness.

All research staff received standardized training in qualitative interviewing and completed preparatory reviews on T2DM management to ensure ethically and contextually appropriate data collection. Data were stored securely on the institutional cloud infrastructure of [Anonymous University], with access restricted to authorized researchers.

3.1.3. Data Analysis

We adopted a mixed-methods analytic approach that combined descriptive analysis of structured questionnaire data with thematic analysis of qualitative materials. Questionnaire responses were summarized using descriptive statistics to characterize participant backgrounds and general AI usage patterns.

Qualitative data—including interview transcripts and patient-authored AI queries—were analyzed using thematic analysis [26]. Two researchers independently coded an initial subset of transcripts to construct a shared codebook, which was iteratively refined through discussion. Inter-coder agreement exceeded 85%, with disagreements resolved by consensus. Coding combined inductive strategies, which captured emergent patterns in patients' information-seeking behaviors, and deductive strategies, which were informed by preliminary evaluative concerns relevant to subsequent expert assessment.

Importantly, analytic codes and themes identified in Phase 1 were not treated as empirical findings to be interpreted or generalized. Instead, they functioned as sensitizing constructs that informed: (1) the definition of evaluation dimensions, (2) the scope of physician judgment criteria, and (3) the selection of patient-authored questions used as evaluation stimuli in Phase 2. By design, insights from Phase 1 are reflected in the structure of the expert evaluation rather than reported as standalone results.

3.2. Phase 2: Dimension-based Physician Evaluation

Building on insights from the formative study in Phase 1, we conducted a physician-centered evaluation to assess the quality of AI-generated health information for T2DM management. This phase shifts the analytic perspective from patient experiences to expert clinical judgment, examining how generative AI responses perform along core quality dimensions, including *accuracy*, *clarity*, *safety*, *integrity*, and *action orientation*. The goal of this phase is to identify system-level strengths and limitations as perceived by clinicians, and to inform the design of safer and more trustworthy AI-assisted health interfaces.

3.2.1. Participants

We recruited participating physicians through online outreach on major Chinese platforms, including Xiaohongshu, Bilibili, and Baidu Tieba. To be eligible, physicians were required to meet the following criteria: (1) hold a valid medical license; (2) be currently practicing in endocrinology or have long-term experience treating T2DM; (3) have a minimum of five years of clinical experience in T2DM care; and (4) hold the professional rank of attending physician or higher. All participants were required to provide written informed consent. Physicians were excluded if they were directly involved in the development of AI tools for diabetes care or held financial interests in related companies, to avoid conflicts of interest. Physicians currently suspended or not actively practicing were also excluded.

A total of seven physicians participated in this evaluation study. Their ages ranged from 34 to 62 years ($M = 48.86$, $SD = 11.57$), with years of clinical practice ranging from 5 to 33 years ($M = 19.86$, $SD = 11.60$). All participants held a PhD degree. Five physicians specialized in endocrinology or metabolic disorders, one in rehabilitation medicine, and one in cardiology. Four practiced in urban settings and three in rural areas. Table 2 summarizes the demographic characteristics of the participating physicians.

This study was approved by the IRB of the Author's Institute. All participants were fully informed of the study's goals, procedures, and potential implications. They retained the right to withdraw from the

study or request data deletion at any time. All data were anonymized during collection and analysis. Each physician received a compensation of 100 RMB for their participation.

Table 2
Demographic characteristics of participating physicians (N=7).

ID	Age	Years of Practice	Department	Gender	Residence	Degree
1	57	28	Endocrinology	M	Urban	PhD
2	62	33	Endocrinology	M	Urban	PhD
3	57	29	Endocrinology	M	Urban	PhD
4	55	25	Rehabilitation	F	Rural	PhD
5	34	5	Endocrinology and Metabolism	M	Urban	PhD
6	42	13	Cardiology	F	Urban	PhD
7	35	6	Endocrinology	F	Rural	PhD

3.2.2. Procedures

The evaluation procedure consisted of three tightly coupled steps: clinician-guided question curation, independent AI output evaluation, and post-evaluation expert reflection. Together, these steps enabled a systematic examination of AI-generated health information grounded in real patient queries and expert clinical judgment.

Clinician-guided Question Curation. To ensure realism and clinical relevance, the physician team first curated a fixed set of patient-authored questions drawn from the corpus collected in Phase 1. Rather than constructing a taxonomy or performing category-based sampling, this step served solely as stimulus preparation for the evaluation task.

Questions were selected through iterative discussion based on three criteria: (1) frequency of occurrence in patient submissions; (2) clinical significance for diabetes self-management; and (3) potential risk if answered incorrectly or acted upon without professional supervision. This process resulted in a representative set of real-world patient questions that balanced common informational needs with clinically sensitive scenarios. The curated question set was used consistently across all evaluated AI systems. [1]

AI Output Evaluation. Each curated question was independently entered into four widely used generative AI systems: GPT-4-turbo [27], DeepSeek-R1 [28], Kimi K2 [29], and ERNIE Bot 4.5 [30]. To ensure comparability, each query was submitted in a fresh session, with memory or context-retention features disabled where possible. All AI-generated responses were saved verbatim and anonymized by model and question ID.

Physicians independently evaluated each AI response using a standardized five-dimensional rubric developed collaboratively by the clinical team. The evaluation dimensions were:

- **Accuracy**: alignment with current clinical knowledge and evidence-based practice;
- **Safety**: appropriate signaling of risks, contraindications, and the need for professional consultation;
- **Clarity**: accessibility of language and transparency of explanation;
- **Integrity**: internal coherence, completeness, and prioritization of information;
- **Action Orientation**: provision of concrete and practically useful guidance without overstepping clinical authority.

[1] To support transparency and future reuse, the curated set of patient-authored questions, together with metadata on occurrence frequency, clinical relevance, and potential risk considerations, will be made publicly available as part of a future release accompanying this line of work.

Each dimension was scored on a fixed numerical scale. Physicians were instructed to base their ratings on clinical judgment rather than personal preferences or model familiarity. In addition to numerical scores, physicians could optionally record brief qualitative comments to highlight perceived strengths, concerns, or safety issues in specific responses.

Post-evaluation Expert Reflection. Following the scoring task, we conducted semi-structured interviews with each physician to probe their evaluative reasoning and interpretation of dimension-level trade-offs. The interview protocol focused on how physicians made judgments about safety, actionability, coherence, and trustworthiness, and why certain responses were perceived as appropriate, risky, or misleading.

Rather than prompting discussion around question categories, interviews centered on dimension-level failure modes, boundary conditions for acceptable action-oriented guidance, and the interaction between fluency and clinical reliability. Interviews also explored physicians' broader expectations for how generative AI should be integrated into clinical workflows and patient-facing systems. Each interview lasted 30–60 minutes, was audio-recorded with consent, transcribed verbatim, and verified for accuracy.

3.2.3. Data Analysis

Our analysis focused on dimension-based evaluation of model quality, aligning with the paper's scope of characterizing trade-offs across *Accuracy*, *Safety*, *Clarity*, *Integrity*, and *Action Orientation*. Although the patient-authored question set covered diverse real-world concerns, we did not treat question type as an analytic factor in this short paper; instead, questions served as ecologically grounded stimuli to probe dimension-level reliability.

We first screened physician rating data for completeness and consistency, and aggregated scores at the model $\times$ dimension level. For each AI model, we computed descriptive statistics (mean, standard deviation, and interquartile range) for each of the five dimensions, as well as an overall score obtained by summing dimension scores (when reported). To examine whether models differed in quality, we conducted a two-way repeated-measures ANOVA [31] with *Model* (four levels) and *Dimension* (five levels) as within-subject factors. When sphericity assumptions were violated, Greenhouse–Geisser corrections were applied. We performed post-hoc pairwise comparisons with multiplicity correction and reported effect sizes to characterize practical significance. Visualizations (e.g., boxplots, radar charts, and heatmaps) were used to summarize model performance profiles across dimensions and to highlight stability versus variability in physician judgments.

Interview transcripts were analyzed using reflexive thematic analysis [26] to capture physicians' reasoning about dimension-level judgments, trade-offs, and failure modes. Two researchers conducted iterative coding across all transcripts, beginning with open coding to identify recurrent evaluative heuristics (e.g., how clinicians detect weak safety signaling or insufficient prioritization), followed by axial organization around the five evaluation dimensions and cross-cutting themes (e.g., fluency-driven overtrust, responsibility boundaries, and conditions for acceptable action-oriented guidance). Disagreements were resolved through discussion, and representative quotes were selected to illustrate how clinicians operationalized quality dimensions in practice and how these interpretations informed design expectations for dimension-aware health IUIs.

4. RESULTS

4.1. Quantitative Findings

4.1.1. Overall Performance Variability Across Models

To assess the overall quality of health information generated by different AI models, we aggregated physician ratings across all questions and evaluation dimensions. As shown in Figure 1, ChatGPT

achieved the highest mean score ($M = 103.07$, $SD = 7.61$), followed by DeepSeek ($M = 97.51$, $SD = 7.18$). ERNIE Bot ($M = 88.36$, $SD = 5.95$) and Kimi ($M = 86.46$, $SD = 6.35$) received considerably lower ratings. In addition to higher means, ChatGPT and DeepSeek displayed narrower interquartile ranges and fewer outliers, indicating more consistent performance. In contrast, Kimi and ERNIE Bot exhibited larger dispersion and more frequent outliers, suggesting unstable behavior across diverse evaluation contexts.

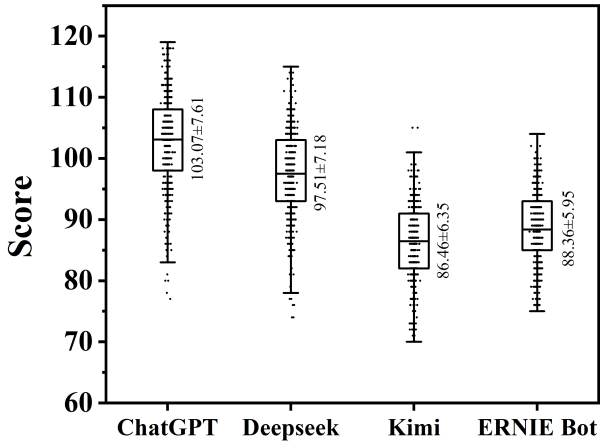

Figure 1: Overall quality evaluation of four AI models. Each box represents the distribution of aggregated scores across all evaluation dimensions and questions for one AI system.

A repeated-measures ANOVA confirmed a significant effect of model type on overall quality, $F(3, 18) = 2315.46$, $p < .001$, $\eta^2_G = 0.72$. Post-hoc tests showed that ChatGPT significantly outperformed all three alternatives ($p < .001$ for each), and DeepSeek also scored higher than both Kimi and ERNIE Bot. No reliable difference was observed between the latter two. Together, these results reveal a pronounced performance hierarchy, with ChatGPT providing the most consistent overall output, while the lower-ranked systems showed greater volatility in quality and presentation.

4.1.2. Dimension-wise Performance Profiles and Trade-offs

To understand why overall model performance differed, we further examined physician ratings across five evaluation dimensions: *Accuracy*, *Safety*, *Clarity*, *Integrity*, and *Action Orientation* (Figure 2, Figure 3). Clear dimension-level disparities emerged across models, revealing distinct strength–weakness profiles.

ChatGPT consistently achieved the highest scores across all five dimensions, including *Accuracy* ($M = 25.66$, $SD = 2.46$) and *Safety* ($M = 25.62$, $SD = 2.51$), indicating reliable factual correctness and low-risk phrasing. DeepSeek followed similar patterns but showed noticeable drops in *Clarity* ($M = 16.53$) and *Integrity* ($M = 16.52$), often producing responses that were accurate yet linguistically fragmented or lacking logical cohesion.

Kimi scored lowest overall, with pronounced weaknesses in *Clarity* ($M = 15.53$) and *Integrity* ($M = 15.37$), where responses frequently included incomplete reasoning chains or oversimplified advice. ERNIE Bot demonstrated a distinct yet uneven profile: although weaker in factual precision, it performed relatively better in *Action Orientation* ($M = 17.14$), tending to generate more procedural and directive recommendations. However, physicians cautioned that such actionable content occasionally lacked appropriate safety framing, highlighting a tension between directness and risk management in patient-facing design.

A two-way repeated-measures ANOVA revealed significant main effects of both Model ($F(3, 18) = 2315.46$, $p < .001$, $\eta^2 = 0.713$) and Dimension ($F(4, 24) = 4823.12$, $p < .001$, $\eta^2 = 0.922$), as well as a significant interaction ($F(12, 72) = 351.92$, $p < .001$, $\eta^2 = 0.606$). These findings confirm that performance differences depend not only on the model but also on the dimension of evaluation, forming distinctive strength–weakness profiles across systems.

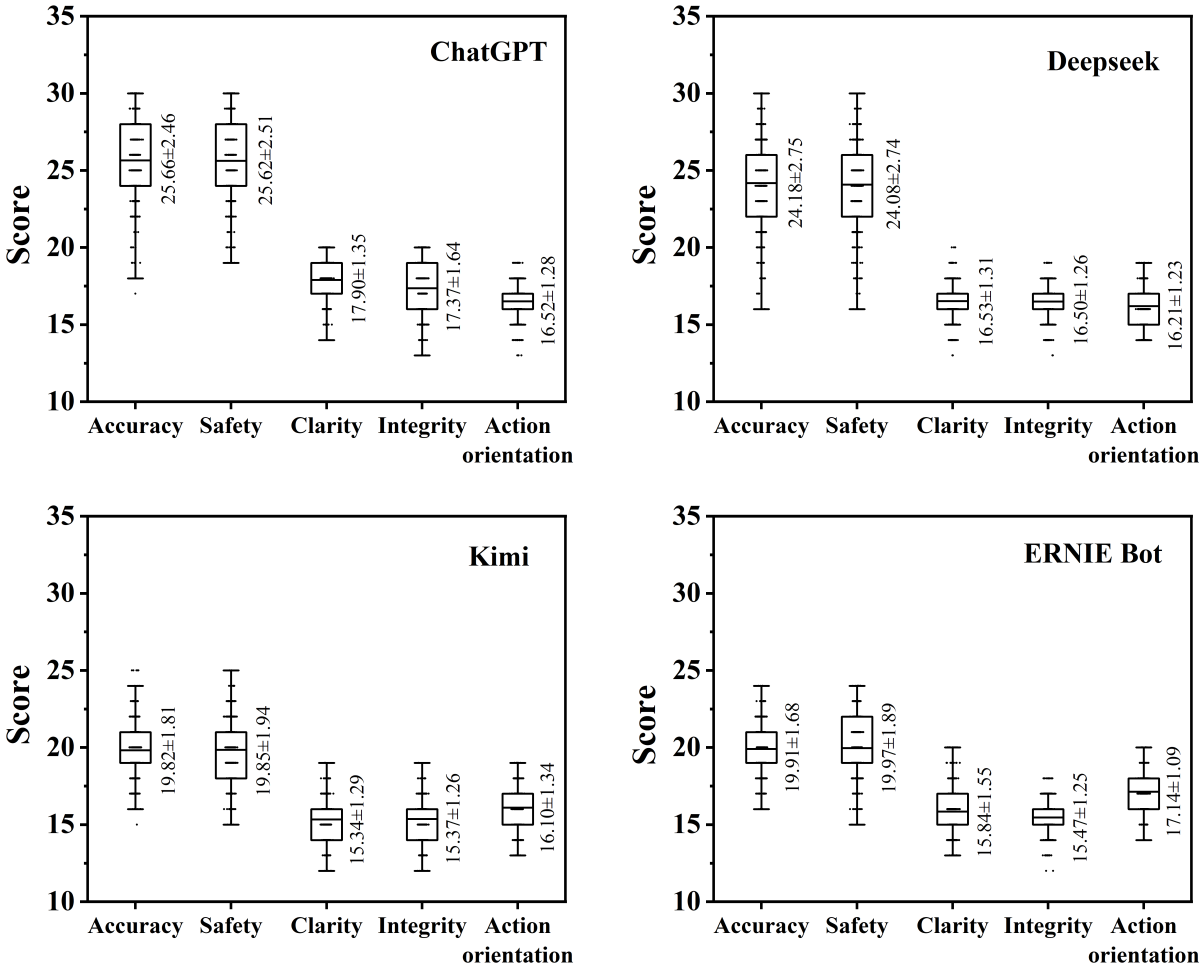

Figure 2: Dimension-wise quality evaluation of four AI models across five key criteria: Accuracy, Safety, Clarity, Integrity, and Action Orientation. Each subplot represents one AI model. Boxplots depict the distribution of scores across all evaluated questions under each criterion.

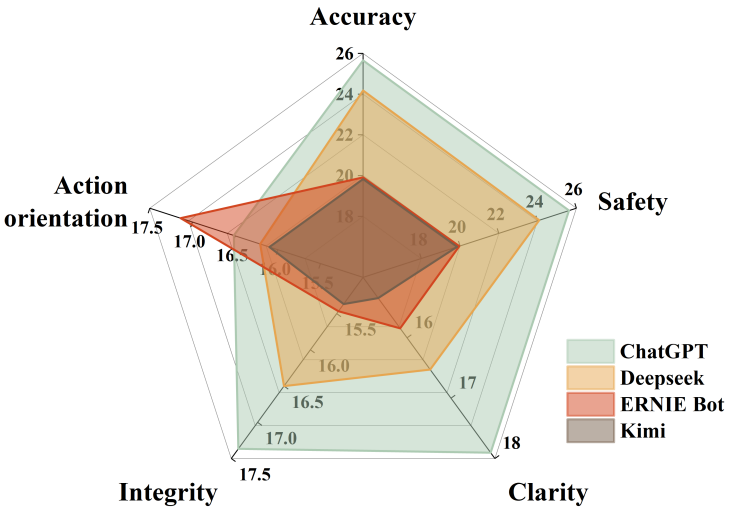

Figure 3: Dimension-wise comparison of AI models across five key quality dimensions: Accuracy, Safety, Clarity, Integrity, and Action Orientation. Each polygon represents one AI model's average score on the five dimensions. Larger enclosed areas indicate stronger overall performance across dimensions.

Overall, these findings indicate that no single model performs optimally across all quality dimensions. Instead, models exhibit complementary strengths and weaknesses, suggesting the need for adaptive interface strategies that allocate models based on dimension-level reliability. For example, ChatGPT may serve as a default for patient-facing educational content, whereas ERNIE Bot's procedural strength could be selectively leveraged under additional safety constraints. Such dimension-aware orchestration would allow interactive systems to balance usability and clinical risk more effectively.

4.2. Qualitative Findings

4.2.1. How Physicians Interpret Quality Dimensions in Practice

Physicians' qualitative accounts revealed that evaluations of AI-generated health information were grounded in concrete judgment moments rather than abstract notions of quality. In practice, clinicians assessed responses by closely examining how information was framed, prioritized, and translated into guidance, often identifying subtle but consequential breakdowns across multiple quality dimensions. These judgments reflected not only whether content was factually correct, but whether it aligned with clinical expectations for safety, coherence, responsibility boundaries, and appropriate action.

Across interviews, clinicians consistently described generative AI as a useful informational resource for supporting patients' early-stage understanding prior to clinical encounters. Several participants characterized AI as a "pre-visit primer" that could help patients grasp foundational concepts and formulate more focused questions. As P3 noted, *"As an information aggregator, it performs remarkably well. A newly diagnosed patient might come in already knowing what insulin resistance is, and then ask more specific questions."* Models such as ChatGPT were frequently cited as effective in this preparatory role due to their coherent narratives and pedagogical clarity. At the same time, several clinicians cautioned that clarity was not a neutral attribute. They anticipated that fluent and well-organized explanations could amplify the perceived authority of AI outputs, leading patients to misinterpret educational descriptions as implicit clinical endorsement when responsibility boundaries were not made explicit. In this sense, clarity functioned as a double-edged quality that supported comprehension while potentially obscuring limits of AI authority.

Concerns about *Safety* centered on responses that adopted a neutral or balanced tone in situations where explicit prioritization or prohibition was clinically expected. Physicians described instances in which AI systems presented options symmetrically without clearly signaling when an option should be avoided or when professional consultation was non-negotiable. As P6 remarked, *"this kind of neutrality can be dangerous for high-risk patients."* Such failures were rarely attributed to incorrect facts, but instead to omissions or de-emphasis of critical warnings that reduced the perceived urgency of risk. ChatGPT was generally regarded as more conservative in risk signaling, whereas other models were more likely to understate safety constraints.

The dimension of *Action Orientation* further highlighted tensions between safety and usability. Physicians differed in their tolerance for directive language, but shared a clear boundary regarding clinical authority. As P2 stated, *"I'm fine with AI explaining options, but once it tells a patient what dosage to change, that crosses a line."* While clinicians recognized that deferring responsibility through statements such as "consult your doctor" often reflects safety-by-design constraints in current AI systems, they noted that repeated reliance on generic disclaimers could frustrate patients seeking practical guidance. Conversely, models such as ERNIE Bot were perceived as overly action-oriented, providing step-by-step instructions without sufficient contextual qualification. Physicians emphasized that procedural guidance was only acceptable when tightly coupled with explicit conditions and escalation criteria.

Finally, clinicians evaluated responses in terms of *Integrity*, focusing on whether information was synthesized and prioritized to support interpretation. Outputs that appeared comprehensive yet failed to guide patients toward next steps were described as "data points, not strategies" (P3). DeepSeek was frequently cited as accurate at the sentence level but fragmented at the response level, requiring clinicians to reconstruct a coherent narrative.

In sum, these findings illustrate that physicians do not assess quality dimensions in isolation, but

reason about how they interact within specific judgment contexts. Breakdowns were rarely attributed to incorrect information alone, but instead to failures of contextualization, prioritization, or alignment with clinical responsibility, helping to explain the dimension-level trade-offs observed in our quantitative analysis.

4.2.2. Implications for Designing Dimension-aware Health AI Systems

Physicians' qualitative reflections suggest that many limitations of current generative AI systems arise not from insufficient medical knowledge, but from misalignment with how quality is evaluated in clinical practice. Rather than treating quality dimensions as independent attributes, clinicians emphasized the need to coordinate system behavior across dimensions to ensure that AI-generated guidance remains interpretable, safe, and clinically appropriate.

First, physicians stressed that actionable guidance becomes meaningful only when grounded in patient-specific context. In our study, breakdowns in actionability frequently co-occurred with failures of coherence, particularly when systems issued broadly applicable recommendations without accounting for comorbidities, treatment history, or lived constraints. As P6 explained, *"I want AI to stop answering like a textbook. If it could say, 'Because this patient has nephropathy, avoid high-protein diets,' that's when it becomes truly useful."* Clinicians consistently framed personalization not as an optional enhancement, but as a prerequisite for safe action, suggesting that the specificity and assertiveness of recommendations should depend on available patient context and clinical risk.

Second, concerns about safety were closely tied to the interpretability of AI-generated outputs. Physicians noted that even factually correct responses became difficult to trust when information sources, uncertainty, or guideline alignment were not made explicit. As P2 remarked, *"I need to know where the information is coming from. Was it a national guideline? A journal article? A user forum?"* Making provenance and confidence visible enabled clinicians to interpret AI outputs as informational inputs rather than implicit clinical judgments.

Third, physicians highlighted the role of contextual grounding in maintaining coherence. Breakdowns occurred when AI systems presented comprehensive but decontextualized information, leaving patients uncertain about relevance or next steps. Several participants described such responses as technically informative yet practically disorienting. To address this, physicians envisioned systems that anchor explanations in concrete patient data or recent behaviors. As P5 suggested, *"If AI could pull my patient's continuous glucose monitor readings or even step counts, it could explain why their blood sugar spiked last night—and what to do next."* Here, contextual integration supported narrative coherence by connecting medical information to patients' lived experience.

Finally, clinicians emphasized the need for explicit governance mechanisms to regulate AI autonomy in high-stakes scenarios. Overly directive responses were perceived as particularly risky when they encroached on clinical decision-making without safeguards. As P7 stated, *"It should be a draft assistant. Let it write, but require a doctor to review before the patient sees anything involving medication changes."* Such perspectives point to the importance of constraining action-oriented outputs based on risk and clearly delineating AI-generated drafts from clinician-endorsed guidance.

5. DISCUSSION

5.1. Generative AI as a Scalable Health Educator

Our findings highlight the potential of generative AI as an accessible and scalable tool for patient education, particularly in chronic disease management such as T2DM. Across both quantitative and qualitative analyses, physicians consistently recognized that LLMs can deliver structured and digestible explanations for factual knowledge and general lifestyle guidance. This observation aligns with prior work showing that LLMs can support patient understanding of medical terminology, risk factors, and self-care concepts in controlled settings [3, 2].

A key contribution of our study lies in the ecological grounding of these findings. Unlike prior evaluations based on static prompts or expert-authored question sets [32, 33], our questions emerged directly from patient-initiated information-seeking behavior under real-world constraints. Physicians described AI as a "pre-visit primer" that helps patients arrive at clinical encounters with a baseline understanding, enabling more focused questions and reducing communication overhead during consultations.

At the same time, our results clarify the limits of this educational role. Physicians emphasized that clear and accurate explanations do not substitute for clinical reasoning or judgment. Even high-quality informational responses were seen as potentially problematic when fluency blurred the boundary between education and decision-making. This concern echoes prior work warning that overly confident or polished AI outputs may be misinterpreted as authoritative guidance [4, 34, 35].

Together, these findings suggest that generative AI is best positioned as an educational scaffold rather than a diagnostic or decision-making agent. For IUI systems, this implies the importance of explicitly framing AI as a support for health literacy, for example through role-based messaging, exploratory scaffolds, and cues that encourage follow-up with human professionals. In this way, AI can function as a "health literacy amplifier" embedded in the patient journey, improving not only access to information but also the quality of subsequent clinician–patient communication.

5.2. Dimension-wise Trade-offs Explain Why Model Quality Is Not One-dimensional

Our findings indicate that the quality of AI-generated health information cannot be adequately characterized by a single aggregate score or simple model ranking. Instead, model behavior is shaped by systematic trade-offs across multiple quality dimensions, which jointly determine whether responses are interpretable and clinically appropriate in practice. This perspective complements prior evaluations that focus on variability in factual correctness or linguistic fluency across LLMs [2, 1], and aligns with human-centered AI work showing that surface-level usability signals can obscure deeper issues of reliability and appropriateness in real-world use [36, 37].

A central trade-off observed in our study concerns the relationship between accuracy and integrity. Even when responses contained largely correct medical statements, insufficient prioritization, weak internal coherence, or overly exhaustive enumerations often reduced their practical value. This echoes broader HCI and health-informatics critiques that "more information" does not necessarily translate into better support, particularly for lay users who rely on systems to structure and contextualize guidance [38, 39]. In our data, integrity mediated the translation of correctness into usefulness: responses that failed to impose hierarchy or synthesize content into a coherent explanatory narrative were frequently perceived as fragmented despite being factually accurate.

A second trade-off emerged between accuracy and safety. Safety failures in our dataset were rarely driven by overtly incorrect statements, but instead by subtle breakdowns in risk signaling, such as missing, softened, or de-emphasized warnings. These patterns extend prior findings in uncertainty communication and reliability signaling, which show that users often treat fluency and completeness as proxies for reliability [40, 41]. In consumer-facing health contexts, such misalignment is especially consequential because users may act on advice without professional oversight. Our results therefore suggest that safety cannot be treated as a passive byproduct of correctness alone, but depends on how risks and boundaries are explicitly foregrounded [42].

Action orientation introduced a further tension between practical usefulness and clinical risk. While actionable recommendations can reduce ambiguity and support self-care routines, the same directive tone can amplify harm when applied to situations that require individualized clinical judgment. Prior work has shown that the persuasive form of AI outputs can shape user reliance even when underlying validity is uncertain [41, 36]. Consistent with these observations, our findings indicate that actionability functions as a conditional dimension: its appropriateness depends on whether responses clearly specify contextual boundaries, escalation triggers, and non-negotiable consultation points.

Finally, comparisons across models revealed that these trade-offs are not distributed uniformly. Systems with similar aggregate performance often exhibited qualitatively distinct profiles, such as stronger fluency but weaker coherence, or higher directive actionability paired with weaker safety

signaling. This aligns with recent arguments that evaluation should move beyond identifying a single "best" model toward more granular, task- and risk-sensitive assessments of model behavior [35, 43, 44].

Taken together, these findings reframe model quality as an inherently multi-dimensional construct shaped by interacting trade-offs rather than a single performance axis. This framing highlights why aggregate scores are insufficient and underscores the need to reason about how specific dimension-level strengths and weaknesses align with the demands and risks of particular health information contexts.

5.3. Calibrated Trust through Dimension-aware Mediation

A central theoretical implication of our findings is that trust in AI-generated health information cannot be treated as a direct function of output fluency or aggregate performance. Instead, clinicians described trust as emerging from how multiple quality dimensions are interpreted in relation to clinical risk, responsibility boundaries, and anticipated patient behavior. This perspective challenges a common assumption in conversational AI deployment, namely that improvements in linguistic naturalness or factual coverage will naturally lead to appropriate trust.

Prior work in human–AI interaction has shown that fluent and well-structured language can inflate perceived credibility, even when underlying reasoning is incomplete or uncertain [37, 41, 36]. In health contexts, this phenomenon is particularly consequential because confident phrasing may be interpreted by patients as implicit clinical endorsement. Our study extends this literature by showing that clinicians do not assess trustworthiness holistically. Rather, they articulated specific concerns about how patients might over-rely on AI outputs when particular quality dimensions are misaligned, even if responses appear clear and correct on the surface.

Across interviews, physicians repeatedly emphasized that trust-related risks arise not from isolated errors, but from imbalances across quality dimensions that jointly signal responsibility and risk. Quantitatively, models with strong aggregate scores still exhibited uneven dimension-level performance. Qualitatively, clinicians expressed concern that such imbalances could mislead patients into treating AI-generated information as equivalent to clinical judgment. Importantly, these observations reflect clinicians' anticipations of patient interpretation rather than measured patient trust, a distinction that helps bound the scope of our claims.

This view aligns with prior work on calibrated trust, which argues that effective human–AI systems should support context-sensitive reliance rather than blind acceptance or blanket skepticism [40, 42]. However, existing approaches often operationalize calibration through uncertainty visualization or confidence scores alone. Our findings suggest that such mechanisms are insufficient when they fail to reflect how different quality dimensions jointly signal responsibility and risk. A response may communicate uncertainty accurately while still appearing overly authoritative if action orientation or safety boundaries are not clearly constrained.

In all, these results suggest that trust in health-related AI outputs is best understood as an interpretive outcome shaped by dimension-level cues rather than a direct reflection of model competence. From an IUI perspective, this framing shifts attention away from treating trust as a static property of AI systems and toward understanding how trust may be inadvertently amplified or attenuated through the interaction between content quality dimensions and perceived clinical stakes.

5.4. Design Implications for Dimension-aware Health IUIs

Our findings suggest that responsible deployment of generative AI in chronic care requires IUIs to actively mediate model outputs, rather than presenting responses as uniformly reliable answers. The following design implications translate the dimension-level trade-offs identified in Section 5.2 and the trust-related concerns discussed in Section 5.3 into concrete interface-level strategies grounded in prior IUI and health-informatics work.

First, IUIs should *surface dimension-level signals* to support calibrated reliance. Rather than relying on aggregate confidence cues, interfaces should help users distinguish between responses that are clear yet non-actionable, actionable yet weakly safety-framed, or comprehensive yet insufficiently prioritized.

Prior work shows that interface cues shape reliance decisions [42, 40], and our results indicate that such cues are most effective when they map onto the specific quality dimensions clinicians attend to, such as integrity and action orientation.

Second, IUIs should *mediate actionability through explicit responsibility boundaries*. Because directive language can both reduce ambiguity and increase clinical risk, interfaces should modulate how actionable a response appears and when escalation is triggered. This aligns with safety-oriented interaction patterns emphasizing guarded autonomy and graceful fallback in high-stakes systems [45]. Concretely, systems can frame outputs as educational drafts rather than recommendations, couple procedural suggestions with clear clinician-consultation gates, and suppress prescriptive phrasing when safety or integrity signals are weak.

Third, our findings motivate *dimension-aware model orchestration*. Since no single model consistently performs well across all dimensions, IUIs can act as mediators that dynamically route, constrain, or combine models based on the dimension profile required for a given interaction, rather than selecting a single "best" model globally. This implication builds on emerging work in multi-model system design [46, 47], while our study provides clinician-grounded evidence for why orchestration should be driven by dimension-level strengths and weaknesses rather than aggregate scores or surface fluency.

Fourth, IUIs should support *dynamic personalization* to improve integrity and action orientation. Physicians emphasized that generic recommendations often fail not because they are incorrect, but because they are insufficiently contextualized to comorbidities, routines, and lived constraints. Consistent with prior work on personalized health interfaces [15, 48, 49], future systems should incorporate patient context when available and ethically appropriate, allowing explanations and guidance to be anchored in relevant personal circumstances rather than abstract generalities.

Finally, IUIs should treat *emotional responsiveness* as part of interactional mediation rather than surface-level wording. Our qualitative findings suggest that linguistic fluency can create an illusion of empathy while still feeling generic or misattuned to users' situations. This resonates with critiques of decontextualized empathetic language in AI systems [50, 51]. In chronic care contexts, emotion-aware mediation may involve pacing, validation, and resource linking, while maintaining explicit boundaries around clinical authority.

In sum, these implications point to a shift from answer-centric AI toward dimension-aware mediation in health IUIs. By aligning presentation, actionability, orchestration, personalization, and emotional framing with dimension-level reliability and clinical risk, interface design becomes a primary lever for supporting safe, interpretable, and context-sensitive use of generative AI in health settings.

5.5. Limitations and Future Work

This study has several limitations that contextualize the scope of our findings and suggest directions for future research. First, although we evaluated four widely used generative AI models, our results capture model behavior at a specific moment in time. Given the rapid pace of model iteration and safety alignment, relative performance and observed dimension-level trade-offs may evolve. Longitudinal evaluations will be necessary to assess the stability of these patterns over time. Second, our evaluation focused on single-turn responses, limiting insight into how quality breakdowns and corrective mechanisms unfold across multi-turn interactions. Future work should examine interactive scenarios involving clarification, follow-up questions, and escalation, in order to assess how dimension-level signals change over the course of an exchange. Third, our qualitative findings were based on interviews with seven endocrinologists from urban hospitals in China. While this sample provided clinically grounded perspectives on diabetes care, future studies should include a broader range of clinical roles, care settings, and cultural contexts to assess the generalizability of these interpretations. Finally, while prompts were standardized across models, we did not systematically vary prompt framing. Prior work suggests that prompt design can influence both model behavior and user interpretation. Future research should explore how different prompt framings interact with dimension-level reliability and boundary signaling in health-related AI use.

6. CONCLUSION

This paper examined how contemporary generative AI systems support T2DM self-management, drawing on formative patient grounding and a dimension-based evaluation by physicians. Our findings reveal persistent tensions at the dimension level—between accessibility and appropriateness, fluency and clinical reliability, and standardized responses and context-sensitive clinical judgment—highlighting how seemingly high-quality AI outputs can still break down in chronic care settings.

By treating evaluation dimensions as an interpretive lens rather than a fixed framework, this work foregrounds the role of intelligent user interfaces in mediating trust, responsibility, and action in AI-assisted health communication. We argue that long-term conditions such as T2DM call for IUI designs that explicitly support calibrated trust, contextual awareness, and responsible boundary-setting, rather than relying on fluent generation alone.

Declaration on Generative AI

Generative AI was used only for language-level support, including grammar, readability, and LaTeX formatting assistance, especially on tables and appendix formatting. It was not used to collect data, code transcripts, generate themes, conduct analysis, create interview materials, or produce participant data. The authors take responsibility for the integrity of the data, analysis, and claims.

References

[1] R. Samimi, A. Bhattacharya, L. Gosak, G. Stiglic, K. Verbert, Visual-conversational interface for evidence-based explanation of diabetes risk prediction, in: Proceedings of the 7th ACM Conference on Conversational User Interfaces, 2025, pp. 1–18.

[2] C. A. Hernandez, A. E. V. Gonzalez, A. Polianovskaia, R. A. Sanchez, V. M. Arce, A. Mustafa, E. Vypritskaya, O. P. Gutierrez, M. Bashir, A. E. Sedeh, The future of patient education: Ai-driven guide for type 2 diabetes, Cureus 15 (2023).

[3] S. Akrimi, L. Schwensfeier, P. Düking, T. Kreutz, C. Brinkmann, Chatgpt-4o-generated exercise plans for patients with type 2 diabetes mellitus—assessment of their safety and other quality criteria by coaching experts, Sports 13 (2025) 92.

[4] E. R. Burgess, I. Jankovic, M. Austin, N. Cai, A. Kapuścińska, S. Currie, J. M. Overhage, E. S. Poole, J. Kaye, Healthcare ai treatment decision support: Design principles to enhance clinician adoption and trust, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3544548.3581251. doi:10.1145/3544548.3581251.

[5] M. Mirbabaie, J. Marx, N. Möllmann, S. Matt, Digital assistants for diabetes treatment: Designing a user interface to support chronic disease self-management, ACM SIGMIS Database: the DATABASE for Advances in Information Systems 56 (2025) 55–79.

[6] E. G. Mitchell, R. Maimone, A. Cassells, J. N. Tobin, P. Davidson, A. M. Smaldone, L. Mamykina, Automated vs. human health coaching: exploring participant and practitioner experiences, Proceedings of the ACM on human-computer interaction 5 (2021) 1–37.

[7] E. Gong, S. Baptista, A. Russell, P. Scuffham, M. Riddell, J. Speight, D. Bird, E. Williams, M. Lotfaliany, B. Oldenburg, My diabetes coach, a mobile app–based interactive conversational agent to support type 2 diabetes self-management: randomized effectiveness-implementation trial, Journal of medical Internet research 22 (2020) e20322.

[8] Z. Lanshan, J. Yang, G. Fang, Factors influencing the acceptance of medical ai chat assistants among healthcare professionals and patients: a survey-based study in china, Frontiers in Public Health 13 (2025) 1637270.

[9] S. Guo, Y. Song, G. Chen, H. Han, H. Wu, J. Ma, Promoting trust and intention to adopt health

information generated by chatgpt among healthcare customers: An empirical study, Digital Health 11 (2025) 20552076251374121.
[10] R. Gerstenberg, R. Neuhaus, M. Vitt, M. Hassenzahl, Living well with diabetes: Rethinking digital diabetes management systems, in: Proceedings of the 2025 ACM Designing Interactive Systems Conference, 2025, pp. 3153–3172.
[11] L. S. Mayberry, C. R. Lyles, B. Oldenburg, C. Y. Osborn, M. Parks, M. E. Peek, mhealth interventions for disadvantaged and vulnerable people with type 2 diabetes, Current diabetes reports 19 (2019) 148.
[12] P. M. Desai, E. G. Mitchell, M. L. Hwang, M. E. Levine, D. J. Albers, L. Mamykina, Personal health oracle: Explorations of personalized predictions in diabetes self-management, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1–13. URL: https://doi.org/10.1145/3290605.3300600. doi:10.1145/3290605.3300600.
[13] L. Biernatzki, S. Kuske, J. Genz, M. Ritschel, A. Stephan, C. Bächle, S. Droste, S. Grobosch, N. Ernstmann, N. Chernyak, et al., Information needs in people with diabetes mellitus: a systematic review, Systematic reviews 7 (2018) 27.
[14] S. M. Dsouza, J. Venne, S. Shetty, H. Brand, Identification of challenges and leveraging mhealth technology, with need-based solutions to empower self-management in type 2 diabetes: a qualitative study, Diabetology & Metabolic Syndrome 16 (2024) 182.
[15] A. Ayobi, K. Stawarz, D. Katz, P. Marshall, T. Yamagata, R. Santos-Rodriguez, P. Flach, A. A. O'Kane, Co-designing personal health? multidisciplinary benefits and challenges in informing diabetes self-care technologies, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–26.
[16] A. Bhattacharya, J. Ooge, G. Stiglic, K. Verbert, Directive explanations for monitoring the risk of diabetes onset: introducing directive data-centric explanations and combinations to support what-if explorations, in: Proceedings of the 28th international conference on intelligent user interfaces, 2023, pp. 204–219.
[17] S. Reddy, Generative ai in healthcare: an implementation science informed translational path on application, integration and governance, Implementation Science 19 (2024) 27.
[18] F. G. Rebitschek, A. Carella, S. Kohlrausch-Pazin, M. Zitzmann, A. Steckelberg, C. Wilhelm, Evaluating evidence-based health information from generative ai using a cross-sectional study with laypeople seeking screening information, npj Digital Medicine 8 (2025) 343.
[19] V. Agarwal, Y. Jin, M. Chandra, M. De Choudhury, S. Kumar, N. Sastry, Medhalu: Hallucinations in responses to healthcare queries by large language models, arXiv preprint arXiv:2409.19492 (2024).
[20] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, X. Lu, Pubmedqa: A dataset for biomedical research question answering, arXiv preprint arXiv:1909.06146 (2019).
[21] Y. Kim, J. Wu, Y. Abdulle, H. Wu, Medexqa: Medical question answering benchmark with multiple explanations, arXiv preprint arXiv:2406.06331 (2024).
[22] J. Kim, H. Maathuis, D. Sent, Human-centered evaluation of explainable ai applications: a systematic review, Frontiers in Artificial Intelligence 7 (2024) 1456486.
[23] V. Swallow, J. Horsman, E. Mazlan, F. Campbell, R. Zaidi, M. Julian, J. Branchflower, J. Martin-Kerry, H. Monks, A. Soni, et al., Digibete, a novel chatbot to support transition to adult care of young people/young adults with type 1 diabetes mellitus: outcomes from a prospective, multimethod, nonrandomized feasibility and acceptability study, JMIR diabetes 10 (2025) e74032.
[24] Y.-C. Tseng, S. Chen, K.-H. Mah, Y.-C. Chen, Designing an ai chatbot for team-based diabetes care: An iterative human-in-the-loop approach, in: International Conference on Human-Computer Interaction, Springer, 2025, pp. 261–276.
[25] S. Schömbs, Designing embodied agents for impactful human-centred risk communication, in: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–5.
[26] V. Braun, V. Clarke, Using thematic analysis in psychology, Qualitative research in psychology 3 (2006) 77–101.
[27] OpenAI, ChatGPT (june 2025 version), 2025. URL: https://openai.com/, openAI ChatGPT large

language model, accessed June 2025.
[28] DeepSeek, DeepSeek (june 2025 version), 2025. URL: https://www.deepseek.com/, deepSeek large language model, accessed June 2025.
[29] M. AI, Kimi (june 2025 version), 2025. URL: https://www.kimi.com/, kimi AI assistant developed by Moonshot AI, accessed June 2025.
[30] Baidu, ERNIE Bot (june 2025 version), 2025. URL: https://yiyan.baidu.com/, eRNIE Bot developed by Baidu, accessed June 2025.
[31] E. R. Girden, ANOVA: Repeated measures, 84, sage, 1992.
[32] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, S. Singh, Autoprompt: Eliciting knowledge from language models with automatically generated prompts, arXiv preprint arXiv:2010.15980 (2020).
[33] Y. Mou, S. Zhang, W. Ye, Sg-bench: Evaluating llm safety generalization across diverse tasks and prompt types, Advances in Neural Information Processing Systems 37 (2024) 123032–123054.
[34] S. Rajagopal, J. H. Sohn, H. Subramonyam, S. Mohan, Can generative ai support patients' & caregivers' informational needs? towards task-centric evaluation of ai systems, 2025. URL: https://arxiv.org/abs/2402.00234. arXiv:2402.00234.
[35] E. Asgari, N. Montaña-Brown, M. Dubois, S. Khalil, J. Balloch, J. A. Yeung, D. Pimenta, A framework to assess clinical safety and hallucination rates of llms for medical text summarisation, npj Digital Medicine 8 (2025) 274.
[36] E. Rosbach, J. Ganz, J. Ammeling, A. Riener, M. Aubreville, Automation bias in ai-assisted medical decision-making under time pressure in computational pathology, in: BVM Workshop, Springer, 2025, pp. 129–134.
[37] O. Wysocki, J. K. Davies, M. Vigo, A. C. Armstrong, D. Landers, R. Lee, A. Freitas, Assessing the communication gap between ai models and healthcare professionals: Explainability, utility and trust in ai-driven clinical decision-making, Artificial Intelligence 316 (2023) 103839.
[38] M. Merry, P. Riddle, J. Warren, A mental models approach for defining explainable artificial intelligence, BMC Medical Informatics and Decision Making 21 (2021) 344.
[39] T. Schrills, T. Franke, How to answer why–evaluating the explanations of ai through mental model analysis, arXiv preprint arXiv:2002.02526 (2020).
[40] J. Reyes, A. U. Batmaz, M. Kersten-Oertel, Trusting ai: does uncertainty visualization affect decision-making?, Frontiers in Computer Science 7 (2025) 1464348.
[41] M. Reis, F. Reis, W. Kunde, Influence of believed ai involvement on the perception of digital medical advice, Nature Medicine 30 (2024) 3098–3100.
[42] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, et al., Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 401–413.
[43] F. Gaber, M. Shaik, F. Allega, A. J. Bilecz, F. Busch, K. Goon, V. Franke, A. Akalin, Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis, npj Digital Medicine 8 (2025) 263.
[44] I. Jahan, M. T. R. Laskar, C. Peng, J. X. Huang, A comprehensive evaluation of large language models on benchmark biomedical text processing tasks, Computers in biology and medicine 171 (2024) 108189.
[45] M. Florins, J. Vanderdonckt, Graceful degradation of user interfaces as a design method for multiplatform systems, in: Proceedings of the 9th international conference on Intelligent user interfaces, 2004, pp. 140–147.
[46] S. Naik, A. L. Toombs, A. Snellinger, S. Saponas, A. K. Hall, Designing with multi-agent generative ai: Insights from industry early adopters, in: Proceedings of the 2025 ACM Designing Interactive Systems Conference, 2025, pp. 1961–1972.
[47] C. P. Lee, J. Choi, B. Mutlu, Map: Multi-user personalization with collaborative llm-powered agents, in: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–11.
[48] H. Torkamaan, J. Ziegler, Recommendations as challenges: estimating required effort and user

ability for health behavior change recommendations, in: Proceedings of the 27th International Conference on Intelligent User Interfaces, 2022, pp. 106–119.
[49] M. Jacobs, J. Johnson, E. D. Mynatt, Mypath: Investigating breast cancer patients' use of personalized health information, Proceedings of the ACM on Human-Computer Interaction 2 (2018) 1–21.
[50] U. Genç, H. Verma, Situating empathy in hci/cscw: A scoping review, Proceedings of the ACM on Human-Computer Interaction 8 (2024) 1–37.
[51] A. Cuadra, M. Wang, L. A. Stein, M. F. Jung, N. Dell, D. Estrin, J. A. Landay, The illusion of empathy? notes on displays of emotion in human-computer interaction, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–18.

A. Attitudes Toward AI Questionnaire

Participants rated the following items on a 10-point Likert scale, where 1 indicates *Strongly Disagree* and 10 indicates *Strongly Agree*.

Table 3
Attitudes Toward AI Questionnaire

ID	Statement
Q1	AI provides information in a very convenient and timely manner, allowing me to obtain answers anytime and anywhere.
Q2	I believe the information provided by AI is accurate and reliable.
Q3	AI's responses are expressed in plain language, with clear structure, making them easy for me to understand and follow.
Q4	AI can provide personalized advice tailored to my specific condition (e.g., disease course, symptoms), rather than only general answers.
Q5	After consulting AI, I feel more confident in managing my condition.
Q6	AI's responses help me organize my thoughts, enabling more effective communication with my doctor.
Q7	I worry that over-reliance on AI may cause me to ignore important bodily signals or delay seeking medical care.
Q8	Interacting with AI provides me with a certain degree of psychological comfort and support when I feel anxious or confused.
Q9	AI can objectively compare different treatment options (e.g., medication, lifestyle adjustments) and highlight their advantages and disadvantages.
Q10	Overall, I am satisfied with using AI as an auxiliary tool for diabetes management.

B. AI Usage Questionnaire

This questionnaire was designed to systematically capture the actual use of AI platforms by patients with T2DM. Platforms are categorized by functionality (general-purpose vs. healthcare-specific) and further distinguished by access modality (standalone apps, system-integrated AI, wearable/embodied devices, etc.). Participants were asked to indicate whether they had used each platform.

Table 4
AI Usage Questionnaire

Function	Type	Platform	Access Modality
General-purpose AI			
	Large Language Models	ChatGPT	Website
		DeepSeek	App, Website
		Wenxin Yiyan	App, Website
		Doubao	App, Website
		Tongyi Qianwen	App, DingTalk-integrated, Website
		Zhipu Qingyan	App, Website
		Kimi Chat	App, Website
		iFlytek Spark	App, Website
System-integrated AI			
		Tencent Yuanbao	App (Mini Program)
		SenseChat	App, Website
		Xiao Ai Assistant	Xiaomi phone / smart speaker
		Vivo Blue Heart Assistant	vivo / iQOO phones
		OPPO Breeno AI Assistant	OPPO / OnePlus / Realme phones
		Huawei Xiaoyi	Huawei phones
Other General-purpose			
		Tiangong AI	App, Website
		360 Zhinao	App, Website
Healthcare AI			
	General-purpose Medical AI	DingXiang Doctor	App, Mini Program
		Chunyu Doctor	App, Mini Program
		HaoDF Online	App, Mini Program
		Ping An Health	App, Mini Program
		WeDoctor	App, Mini Program
	Diabetes-specific AI	Sugar Nurse	App
		Weitang	App
		Glucose Manager	App
Embodied Interfaces			
		Tmall Genie	Smart speaker
		Apple Watch	Smartwatch / wristband
		Huawei Watch	Smartwatch / wristband
		Xiaomi Band	Smartwatch / wristband
		In-car AI	Vehicle interface
Other			
		Please specify any additional AI platform related to T2DM that you use.	