

Large Language Models for Diabetes Care Planning, Patient Education and Patient Safety

Enoch Chi Ngai Lim^{a, c} , Junqi Cui^b , Nga Chong Lisa Cheng^a ,
Chi Eung Danforn Lim^{a, c, d} 

Abstract

An exploration into the application of large language models in diabetes care as a support tool for clinical decision-making is warranted due to the model's ability to integrate clinical guidelines, synthesize large volumes of patient data, construct clinical notes, assist in clinical teaching, and provide a rationale for complex management decisions. Diabetes, in particular, presents a complex clinical scenario, as the formulation of management decisions must consider multiple factors, including, but not limited to, multiple targets and risk factors, monitoring data, medication safety, and the patient's behavioral and psychosocial needs, burdens, and nutritional preferences. This narrative review describes the possibilities of clinician-facing diabetes decision-support systems utilizing large language models and proposes a patient safety architecture and evaluation framework for their clinical assessment. The emphasis is on the safe implementation of these systems, not in an autonomous care setting. Based on the current literature, the systems will most likely assist with preparing clinical visits, reviewing clinical medications, preparing patient summaries for continuous glucose monitoring, planning clinical procedures, and coordinating communication within the clinical team. Despite the above benefits, these systems remain limited, and the constraints of large language models may have serious consequences in the field of diabetes, considering the potential risks of treatment-related adverse events, including hypoglycemia, diabetic ketoacidosis, and other complications based on inequitable provision of care. The safest and most viable near-term solution for large language models in clinical practice is a controlled, evidence-based, and clinician-supervised model that supports the retrieval of clinical guidelines, indicates

appropriate levels of uncertainty, identifies gaps in the data, and assumes professional accountability. The focus of future studies should extend beyond validation studies and accuracy benchmarks to evaluating the impact on clinical workflows, equity, patient care and safety, and to monitoring these systems once deployed. Large language models have the potential to support diabetes care; however, they must act to enhance the clinical voice and judgment rather than diminish it.

Keywords: Large language models; Diabetes; Clinical decision support; Patient safety; Digital health; Carbohydrate counting; Health equity

Introduction

Diabetes care is increasingly multifaceted. It now encompasses the management of diabetes-associated obesity, renal care, cardiovascular protection, the avoidance of hypoglycemia, and the interpretation of diabetes technologies and the provision of long-term behavioral support. The use of glucagon-like peptide-1 receptor agonists (GLP-1 RAs) has helped many patients achieve more convenient and effective glycemic control, often with additional benefits in weight management. However, these agents also introduce specific clinical considerations, particularly in perioperative preparation, where delayed gastric emptying may increase the risk of aspiration, and clinicians must balance this risk against the metabolic consequences of withholding treatment, including hyperglycemia and, in insulin-deficient patients, diabetic ketoacidosis [1]. The nutrition aspect of diabetes is also expanding. Various research initiatives attempt to incorporate ethnic culinary medicine, modern nutritional science, fermented foods and the microbiome, biotechnology, and functional foods into diabetes care [2–4]. This creates a complicated picture for the clinician. It shows that diabetes care requires a careful balance between pharmacotherapy, nutrition, technology, behavior, and the social sciences.

In the American Diabetes Association (ADA) Standards of Care in Diabetes 2026, there is greater emphasis on diabetes technology, continuous glucose monitoring (CGM), and individualized pharmacotherapy, including cardiovascular and kidney protection, avoidance of therapeutic inertia, and the social determinants of health [5, 6]. This greater emphasis places an increasing information burden on clinicians. In making a

Manuscript submitted June 20, 2026, accepted August 13, 2026
Published online August 31, 2026

^aTranslational Research Department, Specialist Medical Services Group, Earlwood, NSW 2206, Australia

^bVanke School of Public Health, Tsinghua University, Beijing 100084, China

^cNICM Health Research Institute, Western Sydney University, Westmead, NSW 2145, Australia

^dData Science Institute, University of Technology Sydney, Ultimo, NSW 2007, Australia

^eCorresponding Author: Enoch Chi Ngai Lim, Translational Research Department, Specialist Medical Services Group, Earlwood, NSW 2206, Australia. Email: enoch@smsg.au

doi: <https://doi.org/10.14740/aicm30>
AI in Clinical Medicine
2819-7437 (online)

clinical decision, a clinician may need to consider and synthesize a multitude of factors, including glycated hemoglobin (HbA1c), estimated glomerular filtration rate (eGFR), albuminuria, cardiovascular disease and heart failure risk, weight and hypoglycemia history, medication costs, CGM outputs, and patient preferences.

There are now diabetes-specific large language models (LLMs), as LLMs can develop, generate, and restructure clinical messages. Sheng et al proposed some applications of LLMs in the field of diabetes for education, research, and clinical [7]. Pavon et al suggested that diabetes management will require collaboration between humans and artificial intelligence (AI), rather than the replacement of clinicians [8]. Wang et al used LLMs to improve the management of complex type 2 diabetes and to develop personalized treatment plans for diabetologists [9]. Mondal et al compared LLM-generated type 2 diabetes mellitus management plans with physician-generated plans using real patient records, with attention to both performance and safety [10]. Li et al initiated work on LLMs, diabetes, and professional training. This work captured interest in the area of education and continuing professional development and contributed to ongoing research [11].

There are both promising and concerning aspects in the emerging area of medical LLMs. Med-PaLM and Med-PaLM 2 demonstrated that LLMs could incorporate substantial medical knowledge and perform very well in a medical-centric question-and-answer (Q&A) interface [12, 13]. A rapid review of LLMs focused on clinical medicine showed that most LLMs were highly heterogeneous in methods, outcomes, and, most importantly, safety evaluations [14]. Clinical reasoning has emerged as a very pertinent and valid concern: there is good evidence that LLMs perform very well when the case is fully described; however, reasoning in the early stages (where the clinician decides what is unknown, what is to be questioned, and what is safe to leave out) is very variable [15, 16]. Therefore, regulation and governance become key. The US Food and Drug Administration (FDA) guidance on clinical decision support (CDS) software emphasizes intended use, transparency, and whether the clinician can independently review the basis of a recommendation [17]. World Health Organization (WHO) guidance on large multimodal models emphasizes safety, transparency, accountability, and inclusiveness [18], while its broader AI ethics guidance addresses autonomy and protection from harm [19]. Reporting frameworks, such as the Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI) extension, and the Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence (DECIDE-AI) guideline, stress that AI systems should be evaluated in a prospective, and workflow-aware manner [20, 21]. This review considers the potential of LLMs in diabetes CDS, their clinical risks, and the safeguards that need to be in place.

Methods

This article was designed as a structured narrative review rather than a systematic review or meta-analysis. The purpose

was not to provide pooled estimates of model accuracy, safety, or clinical effectiveness, but to develop a reproducible synthesis of the emerging literature on LLMs in diabetes care, with particular attention to CDS, patient education, patient safety, workflow integration, equity, and governance. The review did not follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology and was not prospectively registered.

A structured search strategy was used to improve transparency and reproducibility. Literature was identified through PubMed/MEDLINE, Google Scholar, prominent biomedical journals, and regulatory, policy, and guideline sources. Searches combined terms relating to LLMs and AI with terms relating to diabetes care and clinical implementation. Search terms included “large language model,” “LLM,” “diabetes,” “type 2 diabetes,” “diabetes care,” “clinical decision support,” “patient education,” “continuous glucose monitoring,” “glucose forecasting,” “medical artificial intelligence,” “clinical reasoning,” “AI governance,” “software as a medical device,” “clinical decision support software,” “patient safety,” “AI regulation,” “AI reporting guidelines,” “CONSORT-AI,” “DECIDE-AI,” “carbohydrate counting,” “meal estimation,” “workflow integration,” “clinician training,” “algorithmic bias,” and “fairness.” Searches were supplemented by targeted review of reference lists from relevant articles and by inclusion of authoritative clinical, regulatory, and governance documents that were directly relevant to the implementation of LLMs in healthcare. Searches were iterative rather than generated from a single deduplicated database export; exact counts of initially identified and screened records were therefore not prospectively retained. The final synthesis included 27 sources, comprising peer-reviewed empirical studies and reviews, together with clinical, regulatory, governance, and reporting guidance.

The review focused primarily on literature published from 2020 to May 2026, reflecting the rapid development of medical LLMs, generative AI, diabetes technology, and AI governance frameworks during this period. Earlier sources were included where they provided important background or context for diabetes management, digital health, CDS, or AI evaluation. Particular emphasis was placed on peer-reviewed empirical studies, reviews, diabetes-specific LLM publications, clinical guidelines, regulatory guidance, and reporting frameworks relevant to the safe clinical evaluation of AI-enabled decision support.

Sources were considered eligible for inclusion if they addressed at least one of the following domains: LLMs in diabetes care; LLM-supported clinical decision-making; diabetes patient education; carbohydrate counting or meal estimation; CGM interpretation or diabetes data synthesis; medication safety; clinical reasoning; AI-enabled CDS; software as a medical device; AI governance; regulatory oversight; patient safety; workflow evaluation; or equity in AI-supported healthcare. Sources were excluded if they focused solely on technical model development without clinical relevance, were unrelated to diabetes or healthcare decision support, did not address implementation or safety implications, or provided only speculative commentary without a clear link to clinical practice, education, governance, or patient safety.

The synthesis proceeded in four stages. First, the literature

Table 1. Potential Clinician-Facing LLM-CDS Use Cases In Diabetes Care

Clinical domain	Potential LLM contribution	Required clinical inputs	Safeguard	Supporting references
Medication optimization	Draft options for treatment intensification, de-intensification, renal-dose review, hypoglycemia-risk review, and cardiorenal risk reduction	Diagnosis, HbA1c, eGFR, albuminuria, atherosclerotic cardiovascular disease status, heart failure status, hypoglycemia history, current therapy, treatment goals	Link each suggestion to current guideline or formulary source; require clinician approval	[5, 6, 9, 10, 17]
GLP-1 RA perioperative planning	Summarize aspiration-risk considerations, glycemic consequences of withholding therapy, and perioperative trade-offs	Procedure type, fasting plan, gastrointestinal symptoms, diabetes control, anesthetic plan, indication for GLP-1 RA	Require anesthetist or endocrinologist review; avoid automatic continuation or withholding advice	[1, 6, 17]
CGM narrative interpretation	Convert glucose metrics and patterns into clinician-readable summaries, including time in range, nocturnal hypoglycemia, variability, and possible behavioral triggers	CGM metrics, insulin timing, meals, activity, symptoms, medication changes	Verify LLM-generated summaries against raw CGM data and device reports	[5, 7, 8]
Visit preparation	Generate structured pre-consult summaries from electronic health record data, recent laboratory results, medication lists, and complication screening status	Diagnoses, medications, laboratory trends, screening results, prior notes, admissions	Clinician confirms factual accuracy before use in documentation or decision-making	[7–14, 17, 18]
Patient education	Draft tailored explanations on medicines, monitoring, diet, weight, physical activity, and culturally relevant nutritional approaches	Literacy level, language, dietary pattern, cultural context, treatment plan, comorbidities	Use approved patient-education material; clinician review before release	[2–4, 7, 8, 18, 19]
Multidisciplinary care coordination	Draft referral letters, shared care plans, follow-up checklists, and team communication summaries	Complication status, care goals, pending investigations, team roles, patient preferences	Maintain audit trail; clarify responsible clinician and intended recipient	[17, 20, 21]

CDS: clinical decision support; CGM: continuous glucose monitoring; eGFR: estimated glomerular filtration rate; GLP-1 RA: glucagon-like peptide-1 receptor agonist; HbA1c: glycated hemoglobin; LLM: large language model.

was grouped into thematic domains: diabetes care complexity, clinician-facing LLM support, patient education, carbohydrate counting and meal estimation, medication and perioperative safety, CGM and longitudinal data interpretation, clinical reasoning, AI governance, regulatory requirements, reporting standards, workflow evaluation, and equity. Second, diabetes-specific use cases were identified and mapped to potential clinical inputs, expected LLM contributions, and required safeguards. Third, safety architecture principles were derived by integrating CDS guidance, AI governance documents, and healthcare AI reporting frameworks. Fourth, evaluation domains were developed to reflect issues particularly relevant to diabetes care, including factual accuracy, guideline concordance, missing-data awareness, safety, calibration, equity, workflow impact, and patient outcomes.

Diabetes-specific empirical studies were qualitatively appraised for design, sample size and setting, model and prompt transparency, reference standard, external validation, subgroup reporting, workflow assessment, and patient outcomes; no numerical quality score was assigned. Because this was a narrative review, formal risk-of-bias assessment, quality scoring, and meta-analysis were not undertaken. The review does not claim to identify every available publication on LLMs in dia-

betes care, nor does it establish the effectiveness or safety of any specific LLM-CDS system. Instead, it provides a transparent and reproducible method for selecting, organizing, and interpreting literature relevant to the near-term use of LLMs in diabetes care. The findings should therefore be interpreted as a clinically oriented synthesis and implementation framework rather than as definitive evidence of clinical benefit.

Results

The literature indicates that LLMs may be particularly useful for diabetes care when treated as structured support systems rather than fully automated ones. The most promising near-term applications may not revolve around autonomous diagnosis or treatment decisions. Detailed potential use cases and safeguards are summarized in Table 1 [1–14, 17–21], while the main text focuses on their clinical implications and limitations. Diabetes care is a strong use case for clinician-supervised LLM support because it requires rapid integration of glycemic control, renal and cardiovascular risk, medication safety, nutrition, technology data, and patient preferences [5–9].

Recent diabetes care research has examined LLMs for

patient education, clinician support, professional training, and care plan development. Sheng et al identified several domains in which LLMs may be valuable to diabetes care, including patient education, clinician support, and research [7]. Pavon et al underscored the need for collaboration between humans and LLMs for diabetes care but cited clinician replacement as a secondary concern [8]. Wang et al analyzed the use of LLMs to refine the development of personalized diabetes care plans for complex type 2 diabetes and suggested the technology may be of value in the context of structured, clinician-supported care plans [9]. Mondal et al [10] further extended this line of inquiry by comparing LLM-generated type 2 diabetes mellitus management plans with physician-generated plans using real patient records. Their study is particularly relevant because it goes beyond theoretical use cases to examine both performance and safety in a clinically grounded context. These findings support the view that LLMs may assist with structured diabetes care planning, but their outputs require clinician review, especially when patient-specific risk factors, medication safety, comorbidities, and treatment individualization are involved [10]. Li et al showed LLMs may be used in diabetes care for education and training of professionals [11]. A summary of the use cases for clinician-facing LLMs support in diabetes care is provided in Table 1 [1–14, 17–21].

The diabetes-specific evidence remains preliminary and methodologically heterogeneous. The studies by Sheng et al and Pavon et al are perspective articles rather than clinical validation studies [7, 8]. Wang et al used records from a single tertiary academic setting; Mondal et al compared plans from 50 records in one setting; and Li et al assessed examination performance rather than clinical outcomes [9–11]. Model versions, prompts, reference standards, and outcomes differed, while external validation, subgroup analysis, workflow testing, and patient outcomes were limited. Current evidence therefore supports feasibility more strongly than effectiveness or safety across diverse diabetes populations.

LLMs should not be treated as a single category. General-purpose models such as GPT-4 differ from medically adapted models such as Med-PaLM and from diabetes-specific or locally configured systems that use fine-tuning, structured prompts, approved-source retrieval, or clinical-data connections [7–14]. Reliability depends on model version, data source, intended use, and clinical validation.

Carbohydrate counting is a safety-sensitive use case because errors may influence prandial insulin dosing. Studies of general-purpose multimodal LLMs found variable accuracy: Goncalves et al reported occasional clinically important overestimation [22], Johansen et al found better performance for individual foods than composite meals [23], and Aslan et al found that mean carbohydrate targets could be approximated despite variability and inaccurate total energy [24]. These findings support preliminary meal estimation or education, not autonomous bolus calculation. Portion size, local food-composition data, and uncertainty should be shown, with confirmation against a food database and clinician or dietitian review before the estimate informs insulin dosing.

Another persistent idea was that LLMs could aid in longitudinal synthesis. Diabetes care seldom hinges on a single assessment. Values such as HbA1c, CGM, eGFR, albuminu-

ria, and medication lists acquire clinical significance when assessed in context. LLMs might help synthesize fragmented electronic health records into a preliminary clinical summary, flagging overdue screenings, potential medication-safety issues, and renal-, cardiovascular-, or hypoglycemia-related risk factors for clinician review before treatment intensification [5, 6, 9]. This is also not the same as an autonomous recommendation. The clinician's view of the patient is beneficial.

The literature also frames patient education and communication as lower-risk but still safety-sensitive applications of LLMs. In this framework, LLMs could craft health literacy-sensitive explanations across a spectrum of languages, cultural food preferences, medications, and behavior change. This is important in diabetes, since culturally relevant food practices and other behavior changes dominate peripheral gut health and nutrition, family routines, and functional food lifestyle interventions within real-world management choices [2–4]. There is much risk in breakdowns in safe and unsafe trade spaces in patient-centered, LLM-produced outputs. The guidance reviewed recommended that patient-centered LLM outputs be based on educational content that is approved, grounded in safety, and retained for clinician review prior to patient release [17–19].

The overall medical LLM literature discusses several key drawbacks. Systems like Med-PaLM show LLMs can incorporate valuable clinical knowledge and can succeed at medical question answering [12, 13]. However, evidence from systematic reviews shows that evaluation methods, outcome measures, and even safety tests were often quite different from one another [14]. Recent studies of clinical reasoning show that LLMs may handle clinical scenarios better when they are provided with all the case information and perform less reliably when they must determine what information is missing, what uncertainties are present, and what the next clinical question should be [15, 16]. This shortcoming especially impacts diabetes, since missing information about kidney function, pregnancy status, history of hypoglycemia, albuminuria, perioperative status and medication availability may substantially alter the safest recommendation.

The literature reviewed also creates justification for more controlled technical and clinical designs. General-purpose chatbots are inadequate for performing high-level diabetes CDS. This is the case unless such chatbots are part of a system that retrieves sanctioned evidence, bounds responses to clinical guidelines, articulates uncertainty, and preserves the clinician in the decision-making process. Retrieval-augmented generation, local clinical guidelines and design frameworks, formulary connections, trace logs, and structured prompts may mitigate some risks; however, they will not eliminate all risks. Table 2 [1, 5, 6, 14, 17–21, 25, 26] outlines proposed safety measures and designs for integrating LLMs into diabetes CDS.

For diabetes CDS, the importance of evaluation extends beyond accuracy. A proposed model answer may be factually correct but still unsafe if renal dosing, recurrent hypoglycemia, the risk of unsafe perioperative medication, affordability, and the treatment burden have been ignored. The guidance evaluated provides support for evaluation across many areas, including factual accuracy, adherence to guidelines, recognition of missing data, calibration, safety, workflow, equity, and

Table 2. Safety Architecture for LLM-Enabled Diabetes CDS

Layer	Function	Practical design feature	Example in diabetes care	Supporting references
Data governance	Protect privacy, confidentiality, and data integrity	Role-based access, audit logs, data minimization, de-identification where appropriate	Restrict CGM and electronic health record access to authorized members of the care team	[18, 19]
Retrieval layer	Ground outputs in approved clinical evidence	Retrieval-augmented generation using local guidelines, formularies, institutional protocols, and current standards	Retrieve current diabetes pharmacotherapy, CGM, renal, and cardiovascular guidance before drafting advice	[5, 6, 14, 17, 18]
Clinical logic layer	Constrain model outputs using safety rules	Rule checks for eGFR, pregnancy, hypoglycemia history, allergies, sodium-glucose cotransporter-2 inhibitor sick-day rules, and perioperative status	Flag renal dosing concerns, hypoglycemia risk, or perioperative medication issues	[1, 5, 6, 17]
Explanation layer	Make the basis of outputs reviewable	Source-linked rationale, missing-data list, uncertainty statement, and distinction between retrieved facts and generated suggestions	State: “albuminuria unavailable; renal-risk recommendation incomplete”	[14, 17, 18]
Human review layer	Preserve professional judgment and accountability	Mandatory clinician approval before patient-facing advice or treatment changes	Clinician edits and approves medication or education plan before release	[17–19, 21]
Monitoring layer	Detect failures after deployment	Error reporting, performance audit, subgroup analysis, model-version tracking, and drift assessment	Monitor advice quality across age, language, renal function, ethnicity, disability, and technology access	[18–21, 25, 26]

CDS: clinical decision support; CGM: continuous glucose monitoring; eGFR: estimated glomerular filtration rate; LLM: large language model.

patient outcomes [17–21]. These areas are greatly emphasized because diabetes presents in unique populations, and outcomes are influenced by access to technology, social determinants of health, language, comorbidities, and health literacy. Table 3 [5,

6, 14–21, 25–27] presents the proposed evaluation areas for the LLM-CDS for diabetes care.

In general, the findings suggest a careful but optimistic approach. LLMs could ease both cognitive and administrative

Table 3. Suggested Evaluation Domains for LLM-CDS in Diabetes Care

Evaluation domain	Example metric	Why it matters	Supporting references
Factual accuracy	Percentage of statements supported by cited guideline, formulary, or patient record	Reduces hallucinated, unsupported, or outdated advice	[14, 17, 18]
Guideline concordance	Agreement with ADA or local diabetes standards across predefined scenarios	Tests whether outputs align with accepted diabetes care	[5, 6, 17]
Missing-data awareness	Rate of appropriate requests for absent HbA1c, eGFR, albuminuria, pregnancy status, hypoglycemia history, or CGM data	Prevents premature recommendations when critical information is absent	[5, 6, 15, 16]
Safety	Serious error rate, near-miss rate, and rate of unsafe treatment suggestions in simulation or live use	Captures clinically meaningful harm beyond answer accuracy	[17–21]
Calibration	Agreement between expressed certainty and correctness	Reduces overconfident incorrect outputs	[14–16]
Equity	Performance across language, age, sex, ethnicity, socioeconomic status, disability, rurality, renal disease, and technology access	Prevents amplification of diabetes-related health disparities	[18, 19, 21, 26, 27]
Workflow impact	Consultation time, documentation burden, alert fatigue, clinician acceptance, and override rate	Determines whether the tool improves or disrupts clinical work	[20, 21, 25, 26]
Patient outcomes	HbA1c, time in range, severe hypoglycemia, admissions, medication adherence, patient understanding, and treatment satisfaction	Links LLM-CDS to clinically meaningful benefit	[5, 6, 20, 21]

ADA: American Diabetes Association; CDS: clinical decision support; CGM: continuous glucose monitoring; eGFR: estimated glomerular filtration rate; HbA1c: glycated hemoglobin; LLM: large language model.

strains in diabetes care via record summarization, evidence retrieval, structured plan drafting, and gap identification. LLMs should never be viewed as autonomous clinical authorities. They should assist clinical reasoning, highlight and articulate uncertainty, and enhance communication and professional accountability.

Discussion

Framing LLM-CDS as clinical augmentation

At present, augmentation is likely the most appropriate application of LLMs in diabetes care. An LLM may be able to summarize a given clinical history, pinpoint missing renal data, and/or suggest clinician review points, and as such, may assist in the review of clinical notes and/or the preparation of administrative documents. Augmentation is an appropriate term, as the inclusion of LLMs in the clinical decision-making process is certainly not warranted. This distinction should be maintained. Because LLM outputs are often fluent, they may be perceived as more authoritative than their underlying clinical reliability warrants. In diabetes care, answers, even with a veneer of plausibility, may be unsafe because they are likely to exclude many factors. Human oversight, therefore, should absolutely be required during the design process of a review system, and certainly should not be employed in a generic way at the close of a specific recommendation.

Value of diabetes data lies in its longitudinal capture

The review of diabetes data is heavily reliant on its longitudinal capture. Serial HbA1c results alone may still provide an incomplete clinical picture. Relevant data may include weight change, CGM metrics, renal function, albuminuria, hypoglycemic events, medication history, comorbidities, and clinical or social factors affecting individual care. LLMs may assist in consolidating longitudinal data into a clinical summary ahead of a proposed appointment. This capability will likely be useful in primary care, endocrinology, perioperative medicine, and in multidisciplinary diabetes clinics. A clinician may use the system to explain how glycemic control may be declining, whether renal-protective therapy has been implemented, which screening tests are pending, and what is lacking to justify escalation of treatment. The most appropriate systems should not provide final recommendations; rather, they should surface relevant context that broadens the clinician's understanding of the patient.

Clinical reasoning within the context of uncertainty

The most significant limitation is not that LLMs lack a clinical lexicon. It is that they cannot be relied upon to reason under uncertainty. Real-world patients do not present with complete case vignettes. A patient may refer to "hypos" with no confirmation of glucose levels. A CGM may present with a compres-

sion artefact and may show a false low. An elevated HbA1c may be due to an insulin gap, steroid exposure, an infection, depression, diminished food security, or disease progression. A GLP-1 RA may be of concern with regard to the risk of pulmonary aspiration during anesthesia. In all of these cases, the safe course of action will rely on identifying the appropriate missing question. For these reasons, the diabetes LLM-CDS must be assessed with respect to its behavior when data are missing. A safe system will identify knowledge gaps. It will also request the patient's renal function before suggesting a change in renal-protective therapy, the urine albumin-to-creatinine ratio before renal risk stratification, the patient's level of hypoglycemia before safely recommending insulin or sulfonylurea therapy, and the patient's pregnancy status before reviewing potentially restricted medications. Knowing when not to provide an answer is a top safety concern.

Patient education requires cultural and clinical grounding

LLMs have the potential to personalize patient education materials based on literacy levels and language, and to provide education in the context of the patient's limitations. Education on diabetes encompasses more than just the transfer of information. It incorporates learning about medications, eating habits, weight, family and cultural food choices, costs, fear of complications, and self-management confidence. There is potential for constrained LLMs to assist clinicians in formulating language-centered evidence-based recommendations.

Regarding patient safety, LLMs may pose a danger. An example of this would be an LLM that recommends consuming incorrect food, suggests interventions that reinforce false outcomes, and fails to recognize the imminent need for medical intervention. Education given to patients by an LLM should be based on culturally grounded educational resources and reviewed if it alters medical treatment.

Governance must be local and auditable

LLM-CDS systems should not be viewed as a generalized productivity solution. Each new LLM-CDS system or intervention should be viewed as a purpose-built product, as drafting clinic letters differs from systems that suggest pharmacotherapy, summarize CGM data, or offer insulin recommendations. Each of these products has different safety, regulatory, evidentiary and risk concerns. Healthcare systems must be auditable. Clinicians must be able to evaluate the patient data used by the system, the evidence utilized, and the rationale. Version control, error reporting, and post-deployment surveillance are necessary. Diabetes care is especially sensitive to subgroup inequities identified in the literature, as technology access and localization, language, age, impaired renal function, disability, and socioeconomic status may influence model performance and access to care.

Practical implementation should begin with a defined use case, workflow, intended user, and accountable clinician. Secure electronic health record or device interfaces, role-based

access, approved evidence sources, interoperability, audit logs, cybersecurity, version control, and downtime or update procedures may be required [25, 26]. Clinicians need training in intended use, limitations, source checking, missing-data management, escalation, and error reporting. Staged simulation, limited piloting, and prospective monitoring should assess workload, alert fatigue, overrides, subgroup performance, and unintended effects before scale-up [20, 21, 25]. Local clinical, information technology, privacy, legal, and consumer governance should review predefined safety and equity thresholds.

Integrating equity into evaluation

If not appropriately implemented, LLM-CDS may exacerbate inequity. Diabetes-specific evidence remains limited, so poorer performance in patients with fragmented records, limited English or health literacy, rural residence, disability, or restricted technology access should be treated as a plausible risk requiring direct evaluation, not an established effect [26]. Training and evaluation data may underrepresent demographic or clinical groups, producing performance variation across language, age, sex, ethnicity, socioeconomic status, disability, rurality, renal function, and technology access [26, 27]. Equity evaluation should report subgroup sample sizes, missingness, and performance, examine clinically relevant intersections, and assess whether errors concentrate in particular groups [26, 27]. It should also consider digital access, literacy, and affordability. A pharmacologically sound recommendation that is financially inaccessible or depends on unavailable technology remains poor CDS.

Limitations and future directions

As this is a narrative review, it cannot provide consolidated measures of accuracy, safety, and outcome benefit. The rapid pace of literature may lead to drastic changes to the performance of these models. Furthermore, many published studies fail to address the actual clinical environment, opting instead for simulated cases, exam questions, or prompted retrospectives. Diabetes-related studies conducted in clinical settings remain largely unexplored. The search was not exhaustive and may have missed studies indexed elsewhere or published after May 2026. The narrative design permits subjective selection and interpretation; duplicate screening and formal risk-of-bias assessment were not undertaken, and publication bias may favor positive results. Model version, prompt, language, and local data context may also limit reproducibility and generalizability.

Research must focus on real-world evaluations of LLM-CDS in the context of diabetes and be designed to reflect the burden of multiple coexisting conditions, the burden on treated patients, the cost of treatment, and patients' preferences. Safety must be evaluated, and the impacts of clinical support decision systems on clinician behavior and workflow, equity, and patient outcomes must be assessed. LLM-CDS must be compared to standard practice, rule-based clinical support decision

systems, and retrieval-based systems. Clear, prospective studies are needed for determining the processes and procedures for addressing output errors within the system and the actions to be taken in case of a disagreement on recommendations between the clinical support decision system and the clinician, as well as clarifying legal responsibilities in the case of model support that leads to adverse outcomes.

Conclusions

LLMs can be integrated into diabetes care, provided they are carefully implemented. Finding evidence, summarizing patient records, identifying missing data, assisting with documentation, creating educational content, and drafting care plans for clinician review are tasks LLMs can support. However, LLMs should not be used for self-directed care. Diabetes treatment relies on a multitude of longitudinal patient data as well as behavioral and pharmacological interventions, making LLM-CDS difficult to deploy in most care contexts. LLM-CDS will help provide clinical care only if it is evidence-based, transparent about evidence gaps or uncertainties, fully traceable, equity-focused, designed for *ex-ante* evaluation, and aimed at strengthening the clinician-patient relationship.

Acknowledgments

None to declare.

Financial Disclosure

None to declare.

Conflict of Interest

None of the authors declared conflicts of interest regarding this manuscript.

Author Contributions

Conceptualization: ECNL, NCLC, CEDL. Data curation: ECNL, JC. Formal analysis: ECNL, JC, NCLC, CEDL. Funding acquisition: CEDL. Investigation: ECNL, JC. Methodology: ECNL, JC. Project administration: ECNL. Supervision: NCLC, CEDL. Visualization: ECNL, CEDL. Writing – original draft: ECNL, NCLC, CEDL. Writing – review and editing: ECNL, JC, NCLC, CEDL.

Data Availability

The authors declare that data supporting the findings of this study are available within the article.

AI Use Declaration

The authors acknowledge the use of Grammarly (writing assistance software) solely for language editing, grammar checking, spelling correction, and stylistic refinement of the manuscript. The authors take full responsibility for the content, interpretation, and conclusions presented in this work.

References

1. Lim ECN, Lim CED. Perioperative management of GLP-1 receptor agonists: Balancing aspiration risk with therapeutic benefit. *SN Compr Clin Med*. 2025;7:313. [doi](#)
2. Lim ECN, Yu XF, Lim CED. Beyond pharmaceuticals: Integrating Chinese culinary medicine with modern nutritional science for holistic diabetes management. *J Tradit Chin Med Sci*. 2026;13(2):203-216. [doi](#)
3. Lim ECN, Yu WTS, Lim CED. Integrative gut health: How fermented foods bridge ancient Eastern wisdom and modern microbiome science. *J Tradit Chin Med Sci*. 2025;12(4):499-508. [doi](#)
4. Lim ECN, Lim CED. Review of integrating biotechnology and functional foods to enhance health and sustainability. *Yangtze Med*. 2025;9(4):175-187. [doi](#)
5. American Diabetes Association Professional Practice Committee for Diabetes. 7. Diabetes technology: standards of care in diabetes-2026. *Diabetes Care*. 2026;49(Suppl 1):S150-S165. [doi](#) [pubmed](#)
6. American Diabetes Association Professional Practice Committee for Diabetes. 9. Pharmacologic approaches to glycemic treatment: standards of care in diabetes-2026. *Diabetes Care*. 2026;49(Suppl 1):S183-S215. [doi](#) [pubmed](#)
7. Sheng B, Guan Z, Lim LL, Jiang Z, Mathioudakis N, Li J, Liu R, et al. Large language models for diabetes care: potentials and prospects. *Sci Bull (Beijing)*. 2024;69(5):583-588. [doi](#) [pubmed](#)
8. Pavon JM, Schlientz D, Maciejewski ML, Economou-Zavlanos N, Lee RH. Large language models in diabetes management: the need for human and artificial intelligence collaboration. *Diabetes Care*. 2025;48(2):182-184. [doi](#) [pubmed](#)
9. Wang M, Sushil M, Williams CYK, Miao BY, Kim S, Masharani U, Ku G, et al. Assessment of large language models for enhancing diabetologist-developed personalized treatment plans in complex type 2 diabetes. *Clin Diabetes*. 2025;43(4):545-553. [doi](#) [pubmed](#)
10. Mondal A, Naskar A, Roy Choudhury B, Chakraborty S, Biswas T, Sinha S, Roy S. Evaluating the performance and safety of large language models in generating type 2 diabetes mellitus management plans: a comparative study with physicians using real patient records. *Cureus*. 2025;17(3):e80737. [doi](#) [pubmed](#)
11. Li H, Jiang Z, Guan Z, Bao Y, Liu Y, Hu T, Li J, et al. Large language models for diabetes training: a prospective study. *Sci Bull (Beijing)*. 2025;70(6):934-942. [doi](#) [pubmed](#)
12. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. [doi](#) [pubmed](#)
13. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, Hou L, et al. Toward expert-level medical question answering with large language models. *Nat Med*. 2025;31(3):943-950. [doi](#) [pubmed](#)
14. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak*. 2025;25(1):117. [doi](#) [pubmed](#)
15. Rao AS, Esmail KP, Lee RS, Jiang S, Arraiza Carlo B, Gill J, Khanna P, et al. Large language model performance and clinical reasoning tasks. *JAMA Netw Open*. 2026;9(4):e264003. [doi](#) [pubmed](#)
16. Tordjman M, Mei X. Limitations of large language models in clinical diagnostic reasoning. *JAMA Netw Open*. 2026;9(4):e264014. [doi](#) [pubmed](#)
17. U.S. Food and Drug Administration. Clinical decision support software: guidance for industry and food and drug administration staff. Silver Spring, MD: U.S. Food and Drug Administration; 2026. Available from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software>. Accessed June 2, 2026.
18. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multimodal models. Geneva: World Health Organization; 2025. Available from: <https://www.who.int/publications/i/item/9789240084759>. Accessed June 2, 2026.
19. World Health Organization. Ethics and governance of artificial intelligence for health: WHO Guidance. Geneva: World Health Organization; 2021. Available from: <https://www.who.int/publications/i/item/9789240029200>. Accessed June 2, 2026.
20. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Spirit AI, The SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. [doi](#) [pubmed](#)
21. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, Denniston AK, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med*. 2022;28(5):924-933. [doi](#) [pubmed](#)
22. Goncalves S, Coelho C, Pretre L, Roussillon C, Jarlot M, Ducloux C, Penfornis A, et al. Chat, Gemini and Claude at the dinner table: assessing general-purpose AI tools for carbohydrate counting in the context of type 1 diabetes. *Diabetes Res Clin Pract*. 2026;231:113031. [doi](#) [pubmed](#)
23. Johansen AR, Laursen IK, Jacobsen V, Moller ZH, Cichosz SL. Evaluating accuracy of ChatGPT-4o in automated carbohydrate estimation from images as a self-management tool for adolescents with type 1 diabetes. *J Diabetes Sci Technol*. 2026. [doi](#) [pubmed](#)
24. Aslan S, Sozlu S. Evaluating ChatGPT for carbohydrate counting accuracy in diabetes management: a precision

- health approach. *J Nutr.* 2026;156(8):101512. [doi pubmed](#)
25. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, Weicken E, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ.* 2025;388:e081554. [doi pubmed](#)
26. Bajramagic M, Battelino T, Cos X, Cote M, Cui N, Forbes A, Galderisi A, et al. Artificial intelligence-driven clinical decision support systems to assist healthcare professionals and people with diabetes in Europe at the point of care: a Delphi-based consensus roadmap. *Diabetologia.* 2026;69(2):259-273. [doi pubmed](#)
27. Alderman JE, Palmer J, Laws E, McCradden MD, Ordish J, Ghassemi M, Pfohl SR, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. *Lancet Digit Health.* 2025;7(1):e64-e88. [doi pubmed](#)