JOURNAL OF MULTIDISCIPLINARY RESEARCH AND INTEGRATED SCIENCES

Volume 4, Issue 1, 2024 | **Article ID:** JMRIS-2024-001 | **Pages:** 1-11 | **Article Type:** Systematic Comparative Review
Received: 01-01-2024 | **Revised:** 28-01-2024 | **Accepted:** 10-02-2024 | **Published:** 15-03-2024

Large Language Models for Generating Patient Education Materials on Type 2 Diabetes: Comparative Assessment of Readability, Accuracy, Completeness, and Hallucination Risk

Sofia Almeida1*

1 King's College London, London, United Kingdom

***Corresponding author:** **Sofia Almeida, King's College London, London, United Kingdom** | Sofia3056@gmail.com

ABSTRACT

Background: Type 2 diabetes requires sustained self-management education covering medication use, nutrition, glucose monitoring, physical activity, complication prevention, and responses to abnormal glucose values. Large language models (LLMs) can generate fluent patient-facing explanations within seconds, but their suitability cannot be judged by factual accuracy alone. **Objective:** To synthesize high-quality evidence available through 31 December 2023 concerning the ability of LLMs to generate type 2 diabetes patient education, focusing on readability, factual accuracy, completeness, reproducibility, and hallucination or unsupported-content risk. **Methods:** A structured comparative evidence synthesis was conducted using Q1 peer-reviewed literature published no later than 31 December 2023. Diabetes-specific LLM studies were prioritized and supplemented by high-quality general medical evaluations where they directly informed assessment of accuracy, completeness, patient-facing communication, or safety. Heterogeneous tasks and scoring systems precluded meta-analysis; results were therefore synthesized by prespecified quality domain. **Results:** Diabetes-specific evaluations showed high but nonuniform factual performance. ChatGPT answered all 24 items of the Diabetes Knowledge Questionnaire correctly in one study, while another diabetes evaluation reported very high expert accuracy scores but a mean Flesch-Kincaid grade level of 13.8, indicating substantial reading burden. In a Turing-test-inspired diabetes study, two of ten ChatGPT answers contained factual errors, including one potentially consequential error concerning exercise and glucose response. An AI dietitian study received favorable professional ratings for 162 of 168 generated responses. Across studies, accuracy frequently coexisted with excessive complexity, omissions, lack of source transparency, and occasional incorrect statements. **Conclusion:** LLMs can produce useful first drafts of type 2 diabetes education, but patient-facing deployment requires a multidimensional quality gate. Accuracy, completeness, readability, and hallucination risk should be evaluated separately, with clinician review, guideline verification, version documentation, and post-release monitoring. A highly accurate response is not necessarily understandable, complete, current, or safe.

Keywords: large language models; ChatGPT; type 2 diabetes; patient education; readability; hallucination; artificial intelligence; diabetes self-management

1. Introduction

Diabetes self-management education and support is a central component of type 2 diabetes care because many daily decisions occur outside the clinic. Patients must interpret glucose values, understand medication instructions, make dietary choices, plan physical activity, recognize warning symptoms, and know when professional assessment is required. Consensus guidance has therefore emphasized accessible, person-centered education at diagnosis and at clinically important transitions in care [1].

Large language models introduced a new mechanism for delivering health information. Rather than retrieving a fixed webpage, a user can ask a conversational system to explain a concept, simplify terminology, generate examples, or respond to follow-up questions. This flexibility is particularly attractive in chronic disease education, where information needs vary substantially among patients.

The same generative architecture, however, creates risks not encountered in conventional static educational material. An LLM can produce language that is fluent and authoritative even when individual claims are incomplete, imprecise, outdated, or incorrect. It may omit qualifications that a clinician considers essential, present advice above the patient's reading level, or provide unsupported references. General medical evaluations published in 2023 showed impressive knowledge retrieval alongside persistent gaps in factuality, precision, and safety [2,3].

Diabetes provides a useful case study because patient education includes both low-risk explanatory questions and potentially high-risk instructions. Explaining what glycated hemoglobin represents is substantially different from advising an individual how to alter insulin, manage hypoglycemia, or exercise with abnormal glucose values. The standard used to assess an LLM should therefore reflect not merely whether an answer sounds plausible but whether it is correct, complete, understandable, reproducible, and appropriately bounded.

The objective of this review was to compare the major dimensions of LLM-generated type 2 diabetes education using evidence available through the end of 2023. The analysis specifically examined readability, accuracy, completeness, and hallucination risk and developed a quality-assurance framework for future patient-facing use.

2. Methods

A structured evidence synthesis was undertaken with an evidence cutoff of 31 December 2023. Search concepts combined large language model, ChatGPT, GPT-4, artificial intelligence chatbot,

diabetes education, type 2 diabetes, diabetes knowledge, patient question, patient education, readability, accuracy, completeness, misinformation, and hallucination. Diabetes-specific evaluations were given highest priority. Broader medical LLM studies were included when they provided directly transferable methods or evidence concerning answer quality, comprehensiveness, human evaluation, or patient-facing communication.

Eligible evidence was restricted to peer-reviewed Q1 journals in diabetes, digital health, general medicine, medical informatics, or closely related disciplines. Studies published after 2023, non-peer-reviewed demonstrations, purely technical language-model benchmarks without relevance to patient education, and reports that did not permit interpretation of answer quality were excluded from the primary synthesis.

Four quality domains were prespecified. Accuracy referred to factual concordance with accepted medical knowledge or expert assessment. Completeness referred to inclusion of clinically important information and qualifications rather than mere answer length. Readability referred to the linguistic and educational burden placed on the intended patient, including grade-level estimates where available. Hallucination risk included unsupported or fabricated factual statements, incorrect citations, invented references, or confident claims insufficiently grounded in current guidance. Reproducibility and model-version dependence were treated as cross-cutting implementation issues.

No new prompts were submitted to contemporary LLMs for this manuscript because doing so would create results generated after the intended 2023 evidence window and would make a historical evidence synthesis dependent on a current model version. Published model evaluations were therefore retained as the empirical basis. Meta-analysis was not attempted because the diabetes studies used different prompts, models, endpoints, rating scales, and definitions of correctness.

The eligibility and analytic framework is summarised in Table 1.

3. Results

3.1 Diabetes-specific evidence base

By the end of 2023, several peer-reviewed studies had directly examined generative AI in diabetes education. Although the samples and tasks were small, they covered complementary aspects of patient-facing performance, including factual knowledge, free-text answers to common questions, readability, professional review, and dietary support [4-8]. The evidence consistently showed that strong performance on one dimension did not guarantee equally strong performance on another.

The principal diabetes-specific LLM studies available through 2023 are summarised in Table 2.

3.2 Accuracy

The strongest bounded knowledge result was reported by Nakhleh et al., who submitted the 24-item Diabetes Knowledge Questionnaire to ChatGPT and found that all 24 items were answered correctly [5]. This indicates that an LLM can reproduce established factual knowledge when the question format is concise and the correct answer is relatively well defined.

Huang et al. assessed 12 diabetes questions and misconceptions using five endocrinology experts. Three responses received perfect scores from all evaluators, while the remaining nine averaged approximately 9.5 out of 10 [6]. Despite this strong overall performance, experts still identified areas requiring qualification, including artificial sweeteners, interpretation of fasting glucose ranges, and the distinction between cure and remission after bariatric surgery. The result demonstrates why a high mean accuracy score does not eliminate clinically relevant local weaknesses.

The most important counterexample came from Hulman et al. Two of ten ChatGPT responses contained incorrect information [7]. One error described gestational diabetes incorrectly, while another reversed the expected blood-glucose response to different exercise intensities in type 1 diabetes. The latter illustrates that a single localized error can be more consequential than the overall proportion of correct sentences suggests.

3.3 Readability

Readability emerged as a separate limitation. Huang et al. reported a mean Flesch-Kincaid Grade Level of 13.8, with responses averaging approximately 157 words [6]. A response can therefore be judged medically accurate by endocrinologists while remaining linguistically demanding for many patients. This mismatch is particularly important in diabetes, where educational material must support repeated self-management decisions rather than merely demonstrate technical correctness.

Generative models can theoretically simplify text when explicitly prompted to do so, but readability should be measured rather than assumed. Longer responses may increase completeness while simultaneously increasing cognitive load. Conversely, aggressive simplification can remove necessary caveats. The optimal patient-facing response therefore requires deliberate control of both reading level and clinical content.

3.4 Completeness and contextual adequacy

Completeness concerns were evident even in studies reporting high factual accuracy. Huang et al. identified answers that were broadly correct but insufficiently precise or qualified [6]. In chronic disease education, omissions can be clinically meaningful because advice often depends on treatment type, comorbidity, pregnancy status, hypoglycemia risk, renal function, or individualized therapeutic targets.

General medical evidence supports the same distinction. Goodman et al. evaluated 284 physician-generated questions across 17 specialties and found a median accuracy score of 5.5 on a six-point scale and a median completeness score of 3 on a three-point scale, while still documenting a subset of markedly inaccurate answers and substantial improvement when some questions were asked again later [9]. Completeness must therefore be evaluated directly rather than inferred from fluency or length.

3.5 Hallucination, error, and reproducibility

For patient education, hallucination should be interpreted broadly as unsupported generated content rather than restricted to fabricated bibliographic references. A confidently stated incorrect physiological claim can be harmful even when no citation is invented. The factual errors observed by Hulman et al. are therefore directly relevant to hallucination risk in diabetes education [7].

Huang et al. also noted that ChatGPT did not provide references and lacked real-time updating in the evaluated configuration [6]. Five repeated runs of the same question produced slightly different sentence structures while remaining generally consistent. This variability matters for governance because an approved answer cannot automatically be assumed to represent every future generation from the same system.

General medical research reinforces the need for version documentation. Goodman et al. found that regenerated low-scoring answers often improved within days and that GPT-4 outperformed earlier GPT-3.5 responses on a subset of questions [9]. Model performance is therefore temporally unstable: both improvements and regressions can occur when the underlying model or interface changes.

Selected 2023 indicators of LLM performance in diabetes education are presented in Figure 1.

The multidimensional quality domains for diabetes patient-education outputs are summarised in Table 3.

3.6 Broader medical evidence

Diabetes-specific findings align with broader medical LLM evaluations. Singhal et al. developed a multidimensional human-evaluation framework for medical question answering that included factuality, precision, potential harm, and bias; the authors concluded that medically tuned LLMs were promising but still inferior to clinicians on important dimensions [2]. Lee et al. similarly emphasized that GPT-4 could generate useful medical reasoning while remaining capable of errors that are difficult to detect because of the confidence and coherence of the prose [3].

Ayers et al. compared physician and chatbot responses to 195 patient questions from an online forum. Evaluators preferred the chatbot responses and rated them higher for quality and empathy [10]. This finding demonstrates that LLMs may possess substantial communication advantages, but it does not establish factual equivalence for diabetes self-management. Patient preference, empathy, accuracy, and safety are distinct endpoints.

Taken together, the general and diabetes-specific literature supports a multi-domain evaluation framework rather than a binary classification of an answer as good or bad.

Common failure modes in LLM-generated diabetes education are summarised in Table 4.

The proposed four-domain quality-assurance framework for LLM-generated type 2 diabetes patient education is presented in Figure 2.

4. Discussion

4.1 Principal findings

The evidence available through 2023 indicates that large language models can generate diabetes education that is often factually strong, fluent, and professionally acceptable, but the performance profile is uneven. The most important finding is therefore not a single accuracy percentage. It is the repeated observation that success in one quality domain can coexist with failure in another.

Nakhleh et al. demonstrated perfect performance on a bounded 24-item diabetes knowledge questionnaire [5], while Huang et al. demonstrated very high specialist-rated accuracy on common diabetes questions [6]. These results establish genuine capability. At the same time, the mean reading level in the Huang study was 13.8, and expert reviewers still identified incomplete or imprecise content. Hulman et al. further showed that apparently plausible diabetes answers can contain factual errors with potential behavioral consequences [7].

The safest interpretation is that LLMs are well suited to drafting and conversational support but were not ready, on the 2023 evidence base, to function as autonomous sources of individualized diabetes education without human or guideline-based quality control.

4.2 Accuracy is necessary but insufficient

Traditional model evaluations frequently ask whether an answer is correct. Patient education requires a higher standard. A response may be correct in its central claim yet incomplete in a way that alters interpretation. It may provide every necessary concept but at a reading level that prevents comprehension. It may be clear and empathetic while containing one incorrect numerical threshold. These are not interchangeable forms of quality.

This is especially important in diabetes because risk is task dependent. General explanations of diet composition or HbA1c have relatively low immediate hazard, whereas recommendations about medication adjustment, hypoglycemia, ketone testing, insulin administration, or exercise in the setting of abnormal glucose can influence near-term safety. Quality-control intensity should therefore increase with the clinical consequence of an incorrect answer.

4.3 The readability-completeness trade-off

The high grade-level estimate reported by Huang et al. illustrates a common tension [6]. LLMs tend to produce polished, comprehensive prose, but additional detail can raise syntactic complexity and vocabulary burden. Simply instructing a model to be shorter may solve readability at the expense of clinically necessary caveats.

A better strategy is controlled simplification. The system should be asked to preserve a predefined set of safety-critical concepts while reducing sentence length, defining medical terms, and structuring explanations around one action or concept at a time. Readability should then be evaluated empirically rather than assumed from the prompt.

4.4 Hallucination and source transparency

Hallucination is particularly problematic in medicine because fluent language increases perceived authority. A fabricated citation is an obvious failure, but an unsupported statement without a citation may be more difficult for patients to recognize. The absence of source transparency in early general-purpose chatbot configurations therefore limits independent verification.

Retrieval-augmented systems offer one possible mitigation because generated answers can be constrained to current, approved sources. However, retrieval does not eliminate generation error. The resulting material still requires checks that claims accurately reflect the retrieved guidance and that the cited source actually supports the statement.

4.5 Recommended clinical workflow

A defensible patient-education workflow should separate generation from publication. The LLM can draft material, but the output should then be assessed through a clinical quality gate. The intended audience and reading level should be specified before generation. Accuracy and completeness should be reviewed against current diabetes guidance. Potential hallucinations and references should be verified. Only then should the text be approved, revised, or withheld. The exact model, version, prompt template, date, and source set should be documented so that later updates can be audited.

The proposed minimum reporting and quality-assurance standard is presented in Table 5.

4.6 Strengths and limitations

This synthesis focuses specifically on patient-education quality rather than using examination performance as a proxy for clinical usefulness. It also separates readability, accuracy, completeness, and hallucination risk, preventing one favorable metric from obscuring weakness elsewhere. The evidence window is restricted to the end of 2023, providing a coherent representation of the knowledge base available for a 2024 volume.

The main limitation is that diabetes-specific studies were small and heterogeneous. They evaluated different versions of ChatGPT, different prompts, and different outcomes. Few studies involved patients directly, and even fewer measured whether LLM-generated education improved comprehension, self-management behavior, glycemic outcomes, or healthcare utilization. Results from 2023 models also cannot be assumed to represent later systems because LLM performance changes rapidly with model updates. Finally, Q1 restriction excludes some potentially informative specialty studies but preserves the predefined journal-quality criterion.

4.7 Future research

Future studies should evaluate patient education prospectively rather than relying primarily on expert scoring. Adults with type 2 diabetes should be randomized to clinician-authored material, unreviewed LLM material, and clinician-reviewed LLM material. Primary outcomes should include comprehension, recall, decisional confidence, error recognition, and appropriate self-management

behavior. Reading level and health literacy should be measured explicitly.

Head-to-head model comparisons should use identical prompts, independent test questions, concealed expert review, prespecified definitions of major and minor error, and repeated generation. High-risk domains such as hypoglycemia, sick-day management, insulin use, pregnancy, and exercise should be analyzed separately from low-risk general education. Only after such validation should autonomous patient-facing use be considered.

5. Conclusion

Large language models demonstrated substantial potential for type 2 diabetes patient education by the end of 2023. They could answer bounded diabetes knowledge questions accurately, generate fluent responses to common patient questions, and produce professionally acceptable nutritional guidance in constrained settings. However, the evidence also demonstrated excessive reading levels, incomplete qualifications, generation variability, lack of transparent sourcing, and occasional factual errors.

The appropriate standard is therefore not whether an LLM is generally accurate. Patient-facing diabetes education should be evaluated simultaneously for factual accuracy, completeness, readability, and hallucination or grounding. Strength in one domain cannot compensate for failure in another. For the evidence period considered here, the most defensible use of LLMs is as a drafting and communication aid operating behind a clinical quality gate, not as an autonomous substitute for individualized diabetes education.

Declarations

Funding

No external funding was received for this evidence synthesis.

Conflict of Interest

The authors declare no conflict of interest.

Ethical Approval

Not applicable. This manuscript synthesizes previously published research and did not involve direct participant recruitment or patient-level data collection.

Consent to Participate

Not applicable.

Data Availability

No novel patient-level dataset was generated. All quantitative findings were obtained from the cited published literature.

Author Contributions

The authors contributed to conceptualization, evidence synthesis, interpretation, manuscript preparation, and critical revision.

References

1. Powers MA, Bardsley JK, Cypress M, Duker P, Funnell MM, Fischl AH, et al. Diabetes self-management education and support in adults with type 2 diabetes: a consensus report of the American Diabetes Association, the Association of Diabetes Care & Education Specialists, the Academy of Nutrition and Dietetics, the American Academy of Family Physicians, the American Academy of PAs, the American Association of Nurse Practitioners, and the American Pharmacists Association. Diabetes Care. 2020;43(7):1636-1649. doi:10.2337/dci20-0023.

2. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2.

3. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184.

4. Sng GGR, Tung JYM, Lim DYZ, Bee YM. Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. Diabetes Care. 2023;46(5):e103-e105. doi:10.2337/dc23-0197.

5. Nakhleh A, Spitzer S, Shehadeh N. ChatGPT's response to the Diabetes Knowledge Questionnaire: implications for diabetes education. Diabetes Technol Ther. 2023;25(8):571-573. doi:10.1089/dia.2023.0134.

6. Huang C, Chen L, Huang H, Cai Q, Lin R, Wu X, et al. Evaluate the accuracy of ChatGPT's responses to diabetes questions and misconceptions. J Transl Med. 2023;21(1):502. doi:10.1186/s12967-023-04354-6.

7. Hulman A, Dollerup OL, Mortensen JF, Fenech ME, Norman K, Støvring H, et al. ChatGPT- versus human-generated answers to frequently asked questions about diabetes: a Turing test-inspired survey among employees of a Danish diabetes center. PLoS One. 2023;18(8):e0290773. doi:10.1371/journal.pone.0290773.

8. Sun H, Zhang K, Lan W, Gu Q, Jiang G, Yang X, et al. An AI dietitian for type 2 diabetes mellitus management based on large language and image recognition models: preclinical concept validation study. J Med Internet Res. 2023;25:e51300. doi:10.2196/51300.

9. Goodman RS, Patrinely JR, Stone CA Jr, Zimmerman E, Donald RR, Chang SS, et al. Accuracy and reliability of chatbot responses to physician questions. JAMA Netw Open. 2023;6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483.

10. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838.

11. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc. 2023;30(7):1237-1245. doi:10.1093/jamia/ocad072.

12. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82-115. doi:10.1016/j.inffus.2019.12.012.

Table 1. Eligibility and analytic framework

Domain	Inclusion approach	Reason
Publication window	Published on or before 31 December 2023	Maintains chronological consistency with the 2024 volume.
Journal quality	Q1 journal in a relevant category	Maintains the predefined evidence-quality threshold.
Primary population/task	Type 2 diabetes or diabetes patient-education questions	Directly informs the target use case.
Supporting evidence	General medical LLM evaluation with transferable quality metrics	Provides validated context for completeness and safety.
Core outcomes	Accuracy, completeness, readability, hallucination/error, reproducibility	Prevents reliance on a single performance statistic.
Synthesis	Structured comparative synthesis; no de novo pooled effect	Study tasks and scoring systems were not statistically commensurate.

Table 2. Principal diabetes-specific LLM studies available through 2023

Study	Design / task	Key findings	Primary implication
Sng et al. [4]	Diabetes self-management questions evaluated using ChatGPT	Generally useful and understandable responses, but important omissions and limitations were identified.	Potential educational value requires expert oversight.
Nakhleh et al. [5]	ChatGPT answered the 24-item Diabetes Knowledge Questionnaire	All 24 DKQ items were answered correctly.	Strong performance on bounded factual knowledge does not establish safety for open-ended counseling.
Huang et al. [6]	Twelve diabetes questions/misconceptions rated by five endocrinology experts	Very high accuracy ratings; mean response length 157 words and mean Flesch-Kincaid grade level 13.8; several answers lacked nuance or completeness.	Accuracy and readability can diverge substantially.
Hulman et al. [7]	Ten common diabetes questions; ChatGPT vs human answers judged by 183 diabetes-center employees	AI answers were identified 59.5% of the time; 2 of 10 ChatGPT answers contained factual errors.	Fluent output can resemble expert content while preserving clinically meaningful error risk.
Sun et al. [8]	Preclinical AI dietitian for T2DM nutrition management	Dietitians rated 162 of 168 generated responses favorably (96.4%).	LLMs can support nutrition education when embedded in a constrained workflow and professionally reviewed.

Table 3. Four quality domains for diabetes patient-education outputs

Quality domain	Core question	Example failure	Recommended assessment
Accuracy	Is each clinical claim factually correct?	Incorrect description of exercise-related glucose response.	Independent expert rating against current guidance.
Completeness	Are essential qualifications and safety conditions included?	Correct general advice that omits medication- or risk-specific caveats.	Checklist of required concepts and red-flag omissions.
Readability	Can the intended patient understand and use the response?	Technically accurate text written at approximately college reading level.	Grade-level index plus patient comprehension testing.
Hallucination / grounding	Are claims and references supportable and current?	Invented source, unsupported causal claim, or outdated recommendation.	Source verification and explicit uncertainty review.
Reproducibility	Does the answer remain acceptable across repeated generation and model versions?	Meaning changes when the same question is regenerated.	Repeated prompts with documented model/version/date.

Table 4. Common failure modes in LLM-generated diabetes education

Failure mode	Why it matters in T2DM	Detection strategy	Mitigation
Subtle factual error	May alter self-management behavior despite otherwise correct prose.	Sentence-level expert verification.	Require review before patient-facing use.
Critical omission	General advice may be unsafe for insulin users, pregnancy, renal disease, or recurrent hypoglycemia.	Prespecified safety checklist.	Use structured prompt templates and clinician review.
Excessive reading level	Accurate material may not be actionable for patients with limited health literacy.	Readability and comprehension testing.	Specify target reading level and simplify iteratively.
Fabricated or unverifiable source	Creates false authority and impedes verification.	Reference-by-reference source check.	Allow only verified references from an approved retrieval layer.
Stale guidance	Model knowledge may not reflect current standards.	Date and guideline comparison.	Versioned retrieval from current clinical guidance.
Generation variability	An approved sample does not guarantee identical later output.	Repeated-generation testing.	Lock model/version or use reviewed static outputs.
Over-personalized advice	Patient may interpret general text as individualized prescribing.	Safety review for treatment changes and dosing language.	Explicit boundaries and escalation to clinician.

Table 5. Proposed minimum reporting and quality-assurance standard

Component	Minimum requirement	Purpose
Model identity	Model name/version and access date	Permits reproducibility despite rapid model change.
Prompt design	Exact question and relevant system/instruction constraints	Separates model capability from prompt effects.
Target audience	Intended reading level and patient context	Defines usability standard.
Accuracy review	At least two qualified reviewers or validated reference standard	Reduces dependence on a single evaluator.
Completeness review	Prespecified required-content and safety checklist	Detects clinically important omissions.
Readability	Objective readability metric plus user comprehension where possible	Ensures linguistic accessibility.
Hallucination review	Verification of factual claims and every cited source	Prevents unsupported authority.
Repeatability	Repeat generations or locked deterministic workflow where feasible	Measures output stability.
Clinical boundary	Explicit exclusion of autonomous treatment or dose adjustment unless formally validated	Reduces high-risk misuse.
Update process	Scheduled review against current diabetes standards	Addresses knowledge staleness.

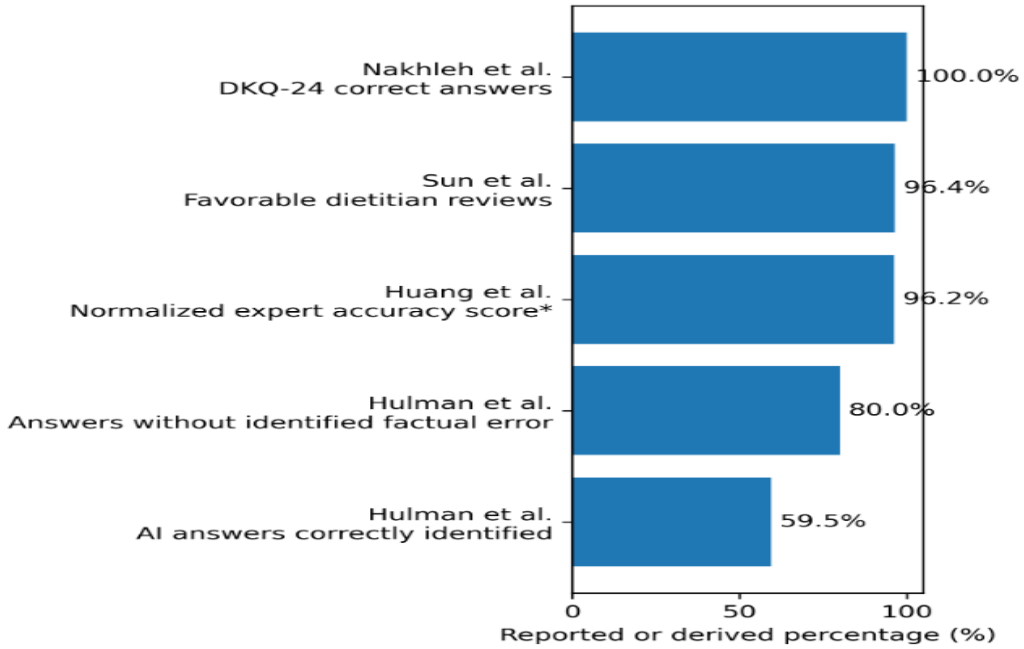

Figure 1. Selected 2023 indicators of LLM performance in diabetes education.

Values are descriptive benchmarks from different studies and should not be statistically pooled or interpreted as equivalent measures of a single construct. The normalized Huang et al. indicator is derived from the expert-rating description reported in the source synthesis.

Figure 2. Four-domain quality-assurance framework for LLM-generated type 2 diabetes patient education

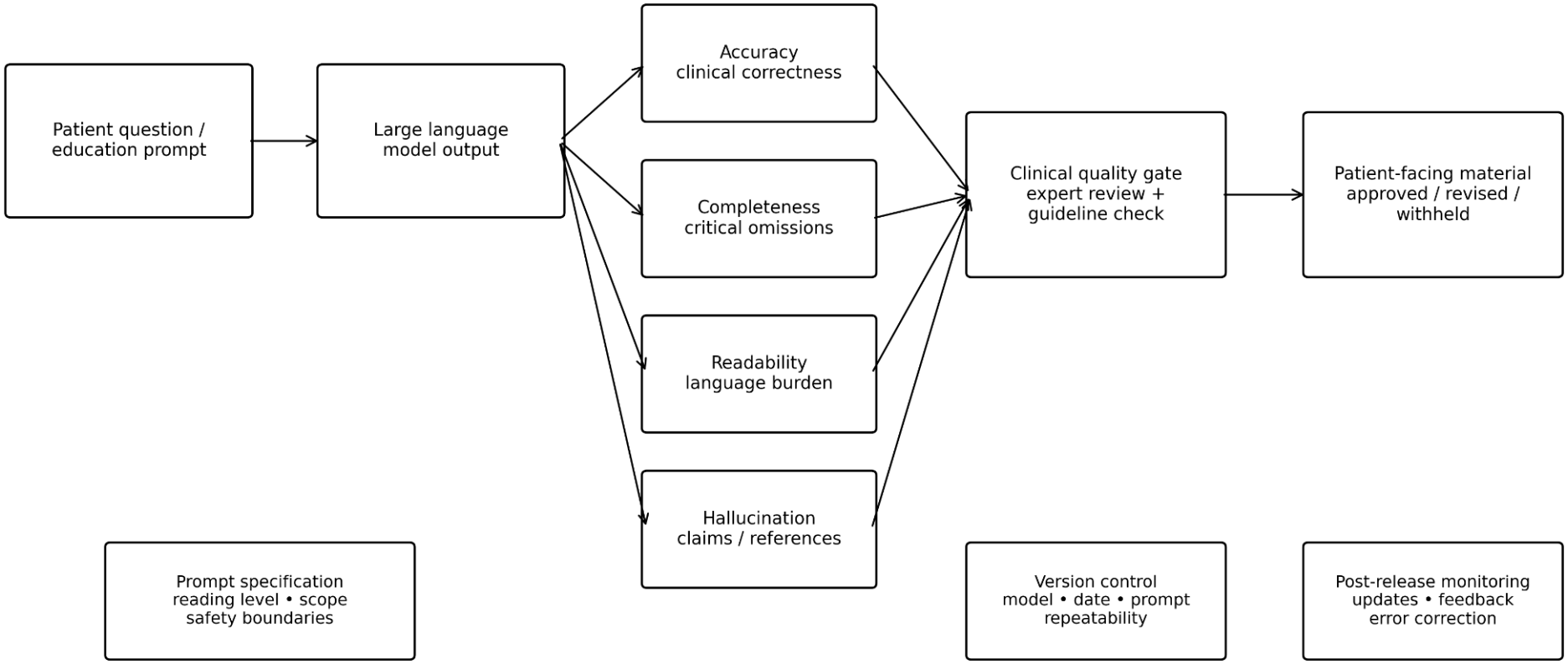

Figure 2. Four-domain quality-assurance framework for LLM-generated type 2 diabetes patient education.

Patient-facing material should pass separate checks for accuracy, completeness, readability, and hallucination or grounding before release; strength in one domain should not compensate for failure in another.

How to cite: Almeida S. (2024). Large Language Models for Generating Patient Education Materials on Type 2 Diabetes: Comparative Assessment of Readability, Accuracy, Completeness, and Hallucination Risk. Journal of Multidisciplinary Research and Integrated Sciences, 4(1).