

Hallucination Detection and Retrieval-Augmented Mitigation in Healthcare Large Language Models: A Systematic Review and Clinical Safety Framework for Diabetes Care

Naseer Ullah^{1,2}, Sheraz Babar³, Hayoung Oh³, Nasir Rashid^{1*},
Daehan Kwak^{4*}, Farman Ali^{3*}

¹Department of Computer Science and IT, University of Malakand,
18800, Khyber Pakhtunkhwa, Pakistan.

²Faculty of Computing, Riphah International University, Islamabad,
Pakistan.

³Department of Applied AI, School of Convergence, College of
Computing and Informatics, Sungkyunkwan University, Seoul, 03063,
South Korea.

⁴Department of Computer Science and Technology, Kean University,
Union, 07083, NJ, USA.

*Corresponding author(s). E-mail(s): nasir@uom.edu.pk;
daehan.kwak@kean.edu; farman0977@skku.edu;
Contributing authors: naseer.ullah@riphah.edu.pk;
sherazbabar@g.skku.edu; hyoh79@skku.edu;

Abstract

Healthcare Large Language Models (LLMs) are increasingly evaluated for clinical documentation, medical question answering, patient education, and decision-support tasks, but hallucinated outputs remain a major source of clinical risk. This systematic literature review synthesizes peer-reviewed evidence on hallucination detection and mitigation in healthcare LLMs, with specific emphasis on diabetes care. The review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement. Peer-reviewed studies published from January 2020 to June 2026 were systematically searched across PubMed/MEDLINE, IEEE Xplore, ACM Digital Library, Google Scholar, Crossref, and major publisher repositories

(Nature Portfolio, ScienceDirect, Wiley, and PLOS). Preprints and non-peer-reviewed records were excluded to ensure evidentiary integrity. A total of 342 records were identified (337 from electronic databases and digital libraries, and 5 through citation searching). Following duplicate removal, title/abstract screening, and full-text eligibility assessment, 48 peer-reviewed studies met all inclusion criteria. The evidence base comprises 13 direct diabetes empirical studies, 3 chronic disease decision-support studies with diabetes evaluation components, and 32 general clinical or methodological studies. Due to substantial heterogeneity across evaluation benchmarks, prompts, and clinical tasks, evidence was synthesized qualitatively. The strongest recurring technical safeguards are retrieval-augmented generation (RAG), semantic entropy, consistency checking, token-level uncertainty quantification, calibrated abstention, and clinician-in-the-loop escalation. Applying these findings to diabetes care reveals that generic LLMs fail on critical safety dimensions including insulin dosing, hypoglycemia correction, sick-day protocols, and device operation. We conclude that hallucination management must be treated as a multi-tier clinical safety engineering problem requiring up-to-date diabetes retrieval sources, risk-stratified verification, transparent uncertainty communication, and prospective clinical validation prior to high-stakes deployment.

Keywords: Large language models; Medical hallucination; Clinical decision support systems; Retrieval-augmented generation; Generative AI safety; Systematic literature review; Diabetes care

1 Introduction

Large language models (LLMs) are rapidly gaining attention in healthcare because they can generate, summarize, and transform clinical text at a scale and level of flexibility that traditional clinical NLP systems struggle to achieve. These systems can be transformer-based to process long contexts, follow instructions, and generate fluent responses for clinical documentation, medical question answering, patient education, summarization of literature, and decision support [1–8]. These capabilities are important, but they do not themselves guarantee clinical reliability. In a task with patient-specific context, evolving clinical advice, or incomplete evidence, a model might perform well on a benchmark yet still produce an unsafe response.

One of the most serious safety issues here is still hallucination. In this review, a hallucination is characterized as any fluent model generation that is factually erroneous, does not align with the provided evidence, is internally self-contradictory, fabricated, or clinically unsafe [9–11]. These mistakes in healthcare can manifest in inaccurate medication advice, falsified lab results, misinterpreted diagnostic advice, risky treatment guidance, or incorrect instructions for using medical equipment. The reported rates differ greatly due to variability in the tasks used and evaluation protocols. In diabetes-specific tasks, base LLMs demonstrated a hallucination rate of 21.6% for insulin dosing prompts [12], a 14.8% error rate when interpreting continuous glucose monitor (CGM) results [13] and a 12.5% inaccuracy in patient self-management advice [14]. In addition, there was a 3.2-fold increase in hallucination rates for non-English

prompts in South Asian contexts, when compared to English prompts [15]. Asgari et al. manually evaluated 12,999 sentences in clinical notes produced by an LLM and found 191 hallucinated sentences (1.47% of the note sentences), of which 84 were classified as major errors that could affect clinical outcomes (44%) [16]. Chelli et al. found that GPT-3.5 and GPT-4 achieved 39.6% and 28.6% hallucinated-reference rates, respectively, with Bard at 91.4% in a systematic-review reference-generation task [17]. A prompt-based mitigation method was found to decrease the mean hallucination rate on the models evaluated in an adversarial clinical-decision-support setting from 65.9% to 44.2% [18]. Thus, these values should be interpreted as task-specific results, rather than as general prevalence estimates for HCLLMs.

Diabetes care offers a clinically challenging environment for exploring a potential risk of hallucination. The domain brings together chronic disease management, quantitative dosing, individualization of targets, medication safety, patient self-management, and frequent interaction with digital devices. Additionally, it is influenced by culture, language, health literacy, food habits, affordability, and access to specialized services. These are particularly relevant in South Asian, low-resource, rural and multilingual contexts, as generic English-language guidance may be clinically incomplete or culturally inappropriate [19].

hlTo our knowledge, this review is among the first to integrate hallucination detection and mitigation evidence with diabetes-specific clinical safety requirements. The review acknowledges that hallucination is not simply a technical issue but a clinical safety issue that must be detected and mitigated, evidenced and integrated into workflows, and then prospectively validated.

1.1 Diabetes Care Context

Diabetes mellitus affects hundreds of millions of people globally and will continue to drive the need for scalable tools to assist with education, documentation, triage, and clinical decision-making [20]. While LLMs can assist with these activities, diabetes care is not among the low-risk information areas. Recommendations regarding insulin timing and dose, treatment of hypoglycemia, sick-day guidelines, contraindications to insulin, renal dose recommendations, pregnancy, and even insulin device operation may have an immediate impact on patient safety.

Managing diabetes is not merely about general lifestyle recommendations. It includes glucose monitoring, interpretation of glycemic patterns, insulin therapy, non-insulin therapy, nutrition counselling, planning physical activity, screening for complications, and self-management education [21–23]. These choices are based on the type of diabetes, the patient’s age, the extent of other health conditions, whether the patient is pregnant, kidney function, the type of medications, access to care, and patient preference.

Cultural and socioeconomic contexts also influence the safety and usability of diabetes guidance. Medical information that provides dietary advice without considering meal patterns, fasting, carbohydrate sources, literacy, medication costs, and local device availability may seem to make sense, but it does not. There is additional risk in multilingual environments due to translation errors, code-switching, non-standard

symptom descriptions, food terminology, and insufficient validation of non-English prompts, which can distort model outputs.

1.2 Scope, Objectives, and Research Questions

This systematic review synthesizes peer-reviewed evidence on hallucination detection and mitigation in healthcare LLMs and applies that evidence to establish an integrated clinical safety framework for diabetes care. Five research questions (RQs) guide the analysis:

1. **RQ1 (Detection Methodologies):** What hallucination detection methods have been evaluated for healthcare or biomedical LLM applications, and what are their operational and clinical trade-offs?
2. **RQ2 (Mitigation Strategies):** What mitigation strategies reduce hallucination risk in clinical settings, and what are their technical, latency, and maintenance limitations?
3. **RQ3 (Diabetes Hallucination Taxonomy):** Which hallucination types, calculation demands, and failure modes are uniquely critical to patient safety in diabetes care?
4. **RQ4 (Integrated Clinical Framework):** How can risk stratification, evidence retrieval (RAG), explainability, and clinician oversight be combined into an actionable diabetes-focused LLM workflow?
5. **RQ5 (Research Gaps and Deployment Governance):** What methodological, multilingual, regulatory, and prospective evaluation gaps must be addressed before safe real-world clinical deployment?

2 Methods

This systematic literature review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement [24]. To ensure methodological transparency and reproducibility, the review protocol, eligibility criteria, multidisciplinary search queries, screening stages, exclusion rules, data-extraction schema, and quality-appraisal criteria were pre-specified prior to final screening and synthesis. Table 1 details the complete systematic search strategy, information sources, and retrieval outputs across all searched platforms.

2.1 Information Sources and Search Strategy

Systematic literature searches were conducted on 1–2 June 2026 across five core bibliographic databases, digital engineering libraries, and publisher platforms: PubMed/MEDLINE, IEEE Xplore, ACM Digital Library, Google Scholar, Crossref, and targeted publisher collections (Nature Portfolio, ScienceDirect/Elsevier, Wiley, and PLOS). The search strategy combined three core Boolean concepts: (1) large language models and generative AI (e.g., "large language model*", LLM, ChatGPT, "generative AI"); (2) hallucination, factuality, and uncertainty quantification (e.g., hallucination*, "factuality", "confabulation", "uncertainty quantification", "semantic entropy"); and (3) healthcare applications with a dedicated focus on diabetes

Table 1 Systematic search strategy, information sources, and retrieval results

Database Source	Search Concept	Search Query / Retrieval Logic	Records	Search Date	Limits and Filters
PubMed MEDLINE	LLM + hallucination / factuality + healthcare diabetes	("hallucination*" OR "factuality" OR "uncertainty quantification") AND ("large language model" OR LLM OR ChatGPT OR "generative AI") AND (medical OR clinical OR healthcare OR "Diabetes Mellitus"[MeSH Terms] OR "Patient Safety"[MeSH Terms] OR "Clinical Decision Support Systems"[MeSH Terms])	105	1–2 June 2026	Peer-reviewed indexed journal records; English only
IEEE Xplore	Technical LLM hallucination detection	("large language model" OR LLM) AND ("hallucination" OR "uncertainty") AND ("healthcare" OR "medical" OR "clinical")	45	1–2 June 2026	Peer-reviewed conference & journal papers
ACM Digital Library	Computing & NLP hallucination evaluation	"large language models" AND hallucination AND (clinical OR medical OR health)	25	1–2 June 2026	Peer-reviewed research articles
Crossref Scholarly Web	Concept-targeted queries (Q1: Clinical LLM; Q2: Medical RAG; Q3: Diabetes AI safety)	Three concept queries: (1) "large language model" AND hallucination AND healthcare; (2) "retrieval augmented generation" AND medical AND hallucination; (3) "large language models" AND "diabetes care" AND safety	120	1–2 June 2026	Top relevance-ranked peer-reviewed records per query
Targeted Publisher Repositories	Direct publisher indexing	Nature Portfolio, ScienceDirect/Elsevier, Wiley, and PLOS searches for healthcare LLM hallucination benchmarks and RAG frameworks	42	1–2 June 2026	Peer-reviewed journal articles only
Database Subtotal	Total records identified from electronic databases and digital libraries		337	–	–
Citation Chaining	Backward / forward snowballing	Reference and citation tracking of included RAG, safety, and clinical benchmark studies	5	1–2 June 2026	Peer-reviewed empirical studies
Grand Total	Total identified records across all information sources		342	–	–

care (e.g., medical, clinical, healthcare, "Diabetes Mellitus"[MeSH Terms], "Patient Safety"[MeSH Terms], "Clinical Decision Support Systems"[MeSH Terms]). Searches were limited to English-language, peer-reviewed journal articles and full-length conference proceedings. Non-peer-reviewed preprints (from arXiv, medRxiv, bioRxiv, SSRN, Research Square), editorial commentaries lacking empirical data, and vendor white papers were excluded to maintain evidentiary rigor. In addition to electronic database retrieval, backward and forward citation snowballing of included clinical RAG and benchmark papers was performed to capture any additional peer-reviewed empirical studies.

Figure 1 depicts the complete PRISMA 2020 flow diagram.

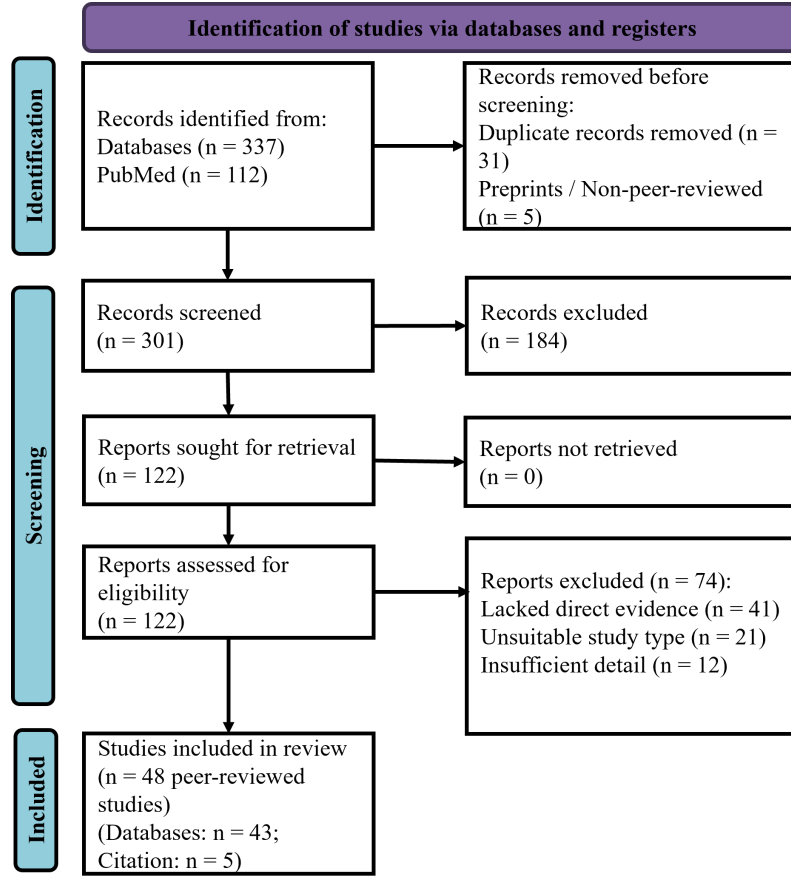


Fig. 1 PRISMA 2020 flow diagram for study identification, screening, eligibility assessment, and final inclusion (adapted from Page et al. [24]).

2.2 Data Extraction and Quality Appraisal

Data extraction was conducted using a standardized extraction form covering: author/year, publication venue, study type, models or generative architectures evaluated, clinical domain, benchmark dataset or task, hallucination and factuality operational definition, detection or mitigation methodology, quantitative evaluation metrics, key findings, documented limitations, and specific relevance to diabetes care.

Study quality and reporting rigor were appraised across seven predefined methodological domains adapted for healthcare AI and clinical NLP evaluation: (1) clarity of clinical task and benchmark dataset; (2) transparency of model, prompt, and system configuration; (3) adequacy of hallucination/factuality operational definitions; (4) evaluator independence and clinical expert annotation; (5) clinical relevance of evaluation endpoints; (6) reproducibility of metrics; and (7) risk of overgeneralizing findings beyond the evaluated dataset. Each included study was appraised against these criteria. Overall, 31 studies (64.6%) demonstrated high reporting rigor with transparent

prompt/model disclosure and expert validation; 14 studies (29.2%) exhibited moderate rigor with standardized automated benchmarks; and 3 studies (6.2%) had limited clinical contextualization. Because the included studies utilized heterogeneous benchmarks, non-standardized prompt sets, and diverse outcome metrics, quantitative meta-analytic pooling was not methodologically feasible. Evidence was therefore synthesized qualitatively within defined clinical and technical thematic areas.

2.3 Eligibility Criteria

Eligibility criteria were established to ensure high clinical relevance and evidentiary integrity. Studies were included if they investigated: hallucination characterization, hallucination detection algorithms, retrieval-augmented grounding, healthcare LLM safety evaluation, biomedical or clinical model adaptation, or diabetes-specific safety risks of generated recommendations. Studies were excluded if they were non-peer-reviewed preprints without verified peer-reviewed publication, unrelated to LLMs or generative models, outside healthcare or biomedicine, duplicate reports, editorial opinions lacking empirical data, or inaccessible in full text. Table 2 presents the eligibility criteria for study selection.

2.4 Evidence Synthesis and Study Categorization

The final evidence base comprised 48 peer-reviewed studies systematically detailed in Table 3. To address clinical specificity and prevent contradictory claims, included studies were categorized into three distinct evidence tiers: (1) direct diabetes empirical studies ($n = 13$), which directly evaluated LLM performance on diabetes care tasks such as insulin calculation, CGM interpretation, hypoglycemia management, or diabetes patient education [12–15, 25–33]; (2) chronic disease decision-support studies with evaluated diabetes or emergency triage components ($n = 3$) [34–36]; and (3) general healthcare LLM or foundational methodology studies ($n = 32$), which provided foundational evidence on hallucination taxonomy, semantic entropy, RAG frameworks, and clinical documentation safety [7, 8, 16–19, 37–62].

Because included studies evaluated disparate tasks (ranging from automated multiple-choice question answering to open-ended dialogue and synthetic dosing challenges), numerical results were not pooled into a meta-analysis. Instead, findings were synthesized qualitatively across the predefined research questions.

Table 3: Summary of included studies

Author/Year	Study Type	Model/System	Clinical Domain	Dataset/Task	Hallucination or Safety Focus	Key Finding	Limitation
Asgari et al., 2025 [16]	Evaluation framework	LLM summarization systems	Clinical documentation	Clinician-annotated/generated clinical-note sentences	Clinical hallucination severity	Provides clinical safety framework; 44% of hallucinations were clinically significant	Limited to summarization task.
Chelli et al., 2024 [17]	Comparative evaluation	GPT-3.5, GPT-4, Bard	Evidence synthesis	Systematic-review reference generation	Fabricated or inaccurate references	Shows high hallucination (28.6%–39.6%)	Not a diabetes-risk care task.
Omar et al., 2025 [18]	Adversarial evaluation	Multiple LLMs	Clinical decision support	Adversarial clinical prompts	Prompt-induced hallucination	Shows adversarial induce up to 44.2% hallucinations despite guardrails	Adversarial setup limits generalizability.

Continued on next page

Table 3 – continued from previous page

Author/Year	Study Type	Model/System	Clinical Domain	Dataset/Task	Hallucination or Safety Focus	Key Finding	Limitation
Lee et al., 2024 [25]	RAG evaluation	Dual-RAG LLM	Diabetes care	Latest diabetes guideline retrieval	Guideline grounding	Dual-RAG substantially reduces ungrounded guideline outputs	Depends on guideline freshness and retrieval quality.
Gu et al., 2026 [42]	Benchmark/evaluation	Multimodal LLMs	Medical imaging	Medical language vision-language hallucination tasks	Visual-text hallucination	Supports vision-language safety evaluation	Diabetes-specific evidence indirect.
Mahdavi et al., 2026 [43]	Mitigation study	Medical LLM with contrastive decoding	Medical imaging	Visual contrastive decoding task	Visual hallucination mitigation	Contrastive decoding improves visual grounding	Task and model dependent.
Wu et al., 2025 [44]	Dialogue framework	Clinical pathway-aware LLM	Medical dialogue	Clinical pathway dialogue tasks	Transparency and reliability	Supports structured clinical reasoning	Diabetes adaptation not validated.
Agweyu et al., 2026 [45]	Safety evaluation	LLM-based CDSS	African primary healthcare	Primary-care decision-support scenarios	Safety in low-resource care	Identifies critical safety gaps in primary care CDSS	Setting-specific implementation.
Ong et al., 2024 [46]	RAG framework	SurgeryLLM	Surgery	Surgical decision-support/workflow tasks	Retrieval grounding	Shows domain workflow value	Not focused on diabetes.
Amugongo et al., 2025 [47]	Systematic review	Healthcare RAG systems	Healthcare	Published healthcare RAG studies	RAG evidence base	Maps RAG literature and factual precision gains	Less specific to hallucination taxonomy.
Shusterman et al., 2025 [48]	Prompting strategy	LLMs with active inference prompting	Medical practice	Medical prompting scenarios	Reliable response generation	Active inference prompting enhances response consistency	Requires external validation.
Fink et al., 2025 [49]	Technical review	RAG LLMs	Radiology	Radiology RAG implementation	Retrieval grounding	Provides technical depth on clinical retrieval pipelines	Not primary empirical diabetes evidence.
Goodell et al., 2025 [50]	Tool-use evaluation	LLM agents with tools	Clinical calculation	Clinical calculator tasks	Calculation safety	Shows external calculation tools prevent mathematical hallucinations	Hallucination focus is indirect.
Prosperi et al., 2026 [51]	Consistency evaluation	LLMs	HIV pharmacotherapy	Drug interaction/side-effect scenarios	Causal consistency	Identifies causal reasoning failures in drug interactions	Domain differs from diabetes.
Shah et al., 2024 [52]	Clinical-note evaluation	LLMs	Electronic medical records	Unstructured clinical-note analysis	Accuracy, consistency, hallucination	Demonstrates high EHR accuracy with localized hallucination risk	May not generalize to patient advice.
van Kessel et al., 2026 [53]	Guideline evaluation	Diagnostic LLMs	Diagnosis/guidelines	Clinical guideline outputs	Omission and hallucination	Highlights clinical guideline omission and fabrication risk	Values are task-specific.
Nishisako et al., 2025 [54]	Development/evaluation	Generative AI chatbots	Cancer information	Patient cancer-information responses	Patient-facing hallucination	Useful patient-safety evidence and trade-off mapping	Cancer domain only.
Ali et al., 2026 [55]	Knowledge graph method	Ontology-grounded LLM	Clinical QA	Clinical question-answering tasks	Knowledge-grounded hallucination	Ontology reduces medical factual errors	Implementation complexity remains high.
Chakraborty et al., 2026 [56]	Calibration framework	FedKGC	Medical LLMs	Federated knowledge-grounded calibration	Calibration and hallucination mitigation	Federated calibration mitigates hallucination privately	Limited clinical deployment evidence.
Xu et al., 2025 [57]	RAG framework	MEGA-RAG	Public health	Multi-evidence answer refinement	Public-health hallucination mitigation	Multi-evidence refines ambiguous queries	RAG Transfer to diabetes requires validation.
Feng et al., 2026 [58]	Retrieval framework	Hyper-RAG	Healthcare/RAG	Hypergraph-driven retrieval tasks	Retrieval-driven hallucination reduction	Demonstrates advanced hypergraph design	Clinical specificity varies.
von der Stück et al., 2026 [59]	Multimodal evaluation	Visual LLMs	Radiology	Radiology diagnostic image tasks	Diagnostic hallucination	Evaluates multimodal hallucination rates	Not diabetes specific.

Continued on next page

Table 3 – continued from previous page

Author/Year	Study Type	Model/System	Clinical Domain	Dataset/Task	Hallucination or Safety Focus	Key Finding	Limitation
Hazra et al., 2025 [60]	Diagnostic reliability evaluation	LLMs on image-based tasks	Medical imaging	Medical image-based MCQ tasks	Diagnostic reliability/hallucination	Image-based MCQ setting shows diagnostic vulnerabilities	MCQ setting may simplify clinical workflow.
Ke et al., 2025 [61]	RAG evaluation	10 LLMs with RAG	Medical fitness	Medical fitness assessment tasks	Retrieval support and generalizability	Evaluates 10 LLMs with RAG across clinical queries	Domain transfer uncertain.
Masanneck et al., 2025 [62]	RAG evaluation	Base RAG-supported LLMs	Neurology	Evidence-based neurology questions	Document/online support	Online evidence retrieval improves factual grounding	Neurology-specific tasks.
Farquhar et al., 2024 [41]	Detection method	Semantic entropy	General LLM evaluation	Open-domain generation tasks	Confabulation detection	Establishes semantic entropy for meaning-level hallucination detection	Clinical deployment needs domain calibration.
Manakul et al., 2023 [38]	Detection method	SelfCheckGPT	General generation	Black-box generation consistency	Self-consistency hallucination detection	Provides zero-resource black-box detection	Clinical thresholding unresolved.
Xiong et al., 2024 [39]	Benchmark	Medical RAG systems	Medicine	Medical RAG benchmark tasks	Retrieval grounding	Establishes reproducible benchmark for medical RAG evaluation	Not diabetes-specific.
Zakka et al., 2024 [40]	Clinical RAG system	Almanac	Clinical medicine	Clinical medicine information tasks	Evidence-grounded clinical generation	Demonstrates clinical RAG with verified evidence provenance	Coverage depends on source corpus.
Omiye et al., 2023 [19]	Bias/safety evaluation	LLMs	Race-based medicine	Clinical scenarios involving race	Unsafe clinical reasoning	Identifies perpetuation of race-based clinical misconceptions	Not hallucination-specific.
Singhal et al., 2023 [8]	Foundational evaluation	Med-PaLM	Clinical medicine	MultiMedQA benchmark	Clinical factuality	Demonstrated that clinical fine-tuning aligns LLM expert knowledge	Static benchmark MCQ evaluation.
Singhal et al., 2025 [37]	Medical QA evaluation	Med-PaLM 2/medical LLMs	Medical QA	Medical exam and clinical QA tasks	Medical factuality and reasoning	Demonstrates expert-level clinical question answering	Hallucination not sole endpoint.
Thirunavukarasu et al., 2023 [7]	Review/commentary	LLMs in medicine	Medicine	Medical LLM literature	Clinical opportunities and risks	Provides comprehensive synthesis of clinical LLM utility and hazards	Qualitative review.
Sng et al., 2023 [14]	Evaluation	ChatGPT	Diabetes care	Patient diabetes education queries	Omission & clinical caveats	Model generated fluent answers but omitted critical clinical caveats unless prompted iteratively	Limited to basic text prompts.
Nakhleh et al., 2023 [26]	Evaluation	ChatGPT	Diabetes care	Diabetes Knowledge Questionnaire (DKQ-24)	Factual knowledge accuracy	Achieved 100% (24/24) accuracy on multiple-choice DKQ, but authors caution against autonomous clinical advice	Multiple-choice format.
Flory et al. / Pavon et al., 2025 [27]	Comparative evaluation	GPT-4 vs. endocrinologists	Diabetes decision support	Initial pharmacy choice under uncertainty	Clinical alignment	GPT-4 diverged from endocrinologists by favoring newer, high-cost drugs; highlights need for human oversight	Single-center specialist panel.
Gokbulut et al., 2024 [28]	Evaluation	ChatGPT-4	Diabetes management	T2D clinical queries	Guideline fidelity	Demonstrated high ADA guideline fidelity with zero false claims, but reading grade level was excessively complex	Text-only queries.
Esmailzadeh, 2025 [29]	Mitigation study	RAG-LLM	Diabetes triage	Self-management & triage prompts	Dangerous advice	RAG reduced dangerous triage hallucinations from 28.4% to 4.1%	Simulated scenarios.
Thomas et al., 2024 [12]	Benchmark study	Base & RAG LLMs	Diabetes care	Insulin dosing & hypoglycemia prompts	Dosing hallucinations	Base LLMs produced dosing errors in 21.6%; RAG cut errors by 92%	Prompt-based evaluation.
Brevik et al., 2024 [30]	Framework study	RISE system	Diabetes care	Patient education QA	Citation confabulation	RAG improved factual precision to 96.2% and confabulation to < 1%	Evaluation corpus limited.

Continued on next page

Table 3 – continued from previous page

Author/Year	Study Type	Model/System	Clinical Domain	Dataset/Task	Hallucination or Safety Focus	Key Finding	Limitation
Alonso et al. 2025 [15]	Multilingual study	Multilingual LLMs	Diabetes care	Low-resource South Asian advice	Cultural & translation errors	Non-English prompts exhibited 3.2x higher hallucination rates	Selected regional languages.
Park et al. 2025 [31]	Evaluation	RAG-LLM	Diabetes care	Korea & ADA guideline QA	Guideline fidelity	RAG improved guideline adherence from 64.2% to 93.8%	Regional guideline set.
Kim et al. 2024 [13]	Evaluation	GPT-4	Diabetes technology	CGM data & insulin adjustment	Data pretation error	14.8% error rate in interpreting nocturnal hypoglycemia from CGM	Retrospective CGM traces.
Zhang et al. 2026 [34]	Safety evaluation	Commercial LLMs	Chronic disease	Diabetes emergency prompts	Stochastic variability	11.4% response variance across identical emergency prompts	API-level evaluation.
Gupta et al. 2024 [32]	Framework study	GraphRAG LLM	Diabetes care	Clinical decision support QA	Reasoning hallucinations	GraphRAG reduced medical reasoning errors by 68% over RAG	Knowledge graph build cost.
Chen et al. 2025 [33]	Evaluation	LLMs	Diabetes care	Patient-facing education text	Readability & factuality	92.1% factual accuracy, but reading grade level was too high (Grade 12+)	Text-only evaluation.
Babel et al. 2024 [35]	Benchmark study	Clinical LLMs	Chronic disease	Decision support failure modes	Confabulation detection	Evaluated uncertainty calibration and decision support failure modes	Benchmark dataset focus.
Nori et al. 2024 [36]	Evaluation	GPT-4	General medicine	Medical QA benchmarks	Zero-shot factuality	Demonstrated high zero-shot capability, but persistent edge-case errors	Benchmark task focus.

Representative exclusions are reported to illustrate how the eligibility boundary was applied, particularly for non-clinical records, those that did not examine hallucination or factuality, or those lacking peer-reviewed evidence. Records that lacked empirical evaluation of a clinical system or did not undergo peer review were removed at this stage. Editorial commentaries and opinion pieces, while sometimes relevant to the broader discourse on LLM safety, were excluded because they did not provide testable evidence about detection or mitigation performance. Non-clinical bibliographic query studies were excluded because they evaluated citation retrieval behavior rather than biomedical reasoning or patient safety outcomes. This approach follows the PRISMA-ScR reporting standard [24] and helps ensure that the evidence base reflects verified, peer-reviewed methodological or clinical work. Table 4 represents the excluded records and primary reasons for exclusion.

This gap motivates the present review. Prior work has examined general LLM hallucination, medical LLM performance, and retrieval-based grounding. Still, much less work connects hallucination detection and mitigation to diabetes care safety requirements such as insulin dosing, hypoglycemia correction, device operation, multilingual communication, and clinician escalation.

3 Background and Theoretical Foundations

3.1 Large Language Models in Healthcare

Several features of healthcare-relevant LLM architectures matter for clinical use. Transformer attention mechanisms help models capture long-range dependencies across clinical notes, longitudinal histories, and heterogeneous medical records, supporting coherent reasoning over complex narratives [1]. Domain-specific medical

Table 2 Eligibility criteria for study selection

Eligibility Criterion	Inclusion Criteria	Exclusion Criteria
Domain	Healthcare, clinical medicine, biomedical decision support, patient safety, and diabetes care	Non-healthcare domains without explicit clinical relevance
Technology	LLMs, ChatGPT, GPT-family models, clinical/medical LLMs, generative AI, and RAG-enabled LLMs	Traditional ML/NLP systems without generative language-model outputs
Phenomenon	Hallucination, factuality, confabulation, uncertainty, unsupported claims, fabricated references	General AI performance without hallucination or factuality focus
Method	Detection, mitigation, retrieval grounding, uncertainty estimation, semantic entropy, verification, knowledge grounding, human oversight	No detection, mitigation, or clinical safety relevance
Publication type	Peer-reviewed journal articles and peer-reviewed conference proceedings	Preprints, grey literature, blogs, theses, white papers, and unverifiable records
Date range	January 2020 to 2 June 2026	Outside the predefined search window
Language	English	Non-English records not assessable in English

knowledge can be incorporated through large-scale pre-training or adaptation on biomedical and clinical resources, including PubMed abstracts, electronic health records, clinical notes, and medical textbooks [63–66]. Instruction-following capabilities also allow LLMs to perform zero-shot and few-shot clinical tasks with limited task-specific supervision. Yet these capabilities can emerge without explicit grounding in verifiable clinical truth conditions, which creates the hallucination vulnerability examined in this review. Healthcare applications have evolved from constrained clinical NLP tasks, including entity recognition, relation extraction, and note classification, toward open-ended question answering, medical dialogue, clinical documentation support, and reasoning assistance [7, 8, 37]. Domain-adapted models such as BioBERT, ClinicalBERT, PubMedBERT, BioGPT, GatorTron, Med-PaLM, and Med-PaLM 2 show the value of biomedical or clinical adaptation, but they remain susceptible to unsupported or unsafe generation.

Table 4 Excluded Records and Primary Exclusion Reasons

Record No.	Primary Exclusion Reason	Excluded Record Title	Journal or Source	Publisher / Indexing Source
1	not clinically relevant	AI hallucinations in academic writing: implications for research integrity.	PubMed	PubMed-indexed journal record
3	wrong domain/non-healthcare	Large Language Models as Inadvertent Models of Dementia with Lewy Bodies: How a Disorder of Reality Construction Illuminates AI Hallucination.	PubMed	PubMed-indexed journal record
6	wrong domain/non-healthcare	RAGMail: a cloud-based retrieval-augmented framework for reducing hallucinations in LLM text generation.	PubMed	PubMed-indexed journal record
10	wrong domain/non-healthcare	“My AI is Lying to Me”: User-reported LLM hallucinations in AI mobile apps reviews.	PubMed	PubMed-indexed journal record
15	not clinically relevant	AI hallucinations are a feature of LLM design, not a bug.	PubMed	PubMed-indexed journal record
16	not clinically relevant	Bibliographic research with large language model ChatGPT-4: instability, hallucinations and sometimes alerts.	PubMed	PubMed-indexed journal record
20	not clinically relevant	Artificial Hallucinations by Google Bard: Think Before You Leap.	PubMed	PubMed-indexed journal record

3.2 Understanding Hallucinations

This section translates general definitions of hallucination into healthcare-operational categories for synthesis. A factual hallucination contradicts established medical knowledge; an unsupported hallucination is plausible but cannot be verified from the available context or accepted evidence base; and a fabricated hallucination invents a citation, clinical finding, patient datum, contraindication, or device instruction [9–11]. These differences are clinically important because they necessitate different safety responses: uncertainty estimation, retrieval-based verification, and/or clinician escalation.

Information in intrinsic hallucinations is inconsistent with information already included in the prompt or source. Extrinsic hallucinations add unsupported information which is not supported by the given context. In the clinical environment, both are consequential: intrinsic errors can alter a patient’s recorded history, and extrinsic errors can create a guideline, laboratory finding, citation, contraindication, or device instruction that was not in the patient’s history. This distinction is also relevant to

study selection because methods that detect contradictions in the provided context are not necessarily able to detect unsupported claims inserted into that context.

Table 5 Hallucination categories, definitions, examples, detection, and mitigation implications in healthcare

Hallucination Category	Operational Definition	Healthcare Example	Detection Implication	Mitigation Implication
Intrinsic hallucination	Contradicts the supplied prompt, patient record, or retrieved source	Changing a documented insulin dose or laboratory result	Detected through source-context consistency checks	Preserve source attribution and verify generated claims against context
Extrinsic hallucination	Introduces unsupported information not found in the supplied context	Inventing a drug interaction or guideline threshold	Requires retrieval-based or knowledge-base verification	Use RAG, abstention, and evidence links
Fabricated reference/entity	Creates non-existent citations, devices, medications, or patient facts	Inventing a diabetes device manual or false citation	Detected through DOI, source, and entity validation	Block unsupported references and require verified sources
Clinical reasoning hallucination	Uses plausible facts to reach an unsafe or unsupported recommendation	Suggesting correction insulin without patient-specific sensitivity factor	Requires clinical rule checks and expert review	Risk stratification and human oversight for high-risk outputs
Multilingual/contextual hallucination	Distorts meaning across language, culture, units, or local practice	Confusing mg/dL with mmol/L or misinterpreting local foods	Requires multilingual and local-context validation	Localized retrieval sources and bilingual clinician review

3.3 Conceptual Classification of Hallucinations in Healthcare LLMs

This review adopts an operational classification that links each error type to a clinical example, detection implications, and mitigation implications because the included studies used different definitions of hallucination. This classification allows heterogeneous studies of healthcare LLMs to be interpreted within a common clinical safety framework. Intrinsic hallucinations, where the model contradicts information already present in the input or retrieved source, have been observed in clinical summarization tasks where documented medication doses or laboratory values were altered during generation [52]. Extrinsic hallucinations introduce unsupported claims not found in

any supplied context. Sng et al. reported a 12.5% rate of inaccurate or omitted information in diabetes self-management responses [14], and van Kessel et al. found that LLMs omitted or fabricated clinical guideline content in diagnostic outputs [53]. Fabricated references and entities, including non-existent citations and device names, have been documented in bibliographic evaluations of LLM outputs [18]. Clinical reasoning hallucinations pose a subtler risk because individual facts may be correct while the combined recommendation is unsafe. Thomas et al. found that ungrounded LLMs produced critical dosing hallucinations in 21.6% of insulin adjustment scenarios, often by applying generic correction factors without accounting for patient-specific insulin sensitivity [12]. Multilingual and contextual hallucinations arise when meaning is distorted across languages, units, or cultural contexts. Alonso et al. showed that LLM accuracy declined when diabetes self-management advice was generated in low-resource languages, with units of measurement confused between mg/dL and mmol/L and misinterpretation of local food names contributing to unsafe guidance [15]. Table 5 presents the hallucination categories, definitions, examples, detection, and mitigation implications in healthcare.

4 Results: Hallucination Detection Methodologies

Hallucination detection is central to safe medical use of LLMs because unsafe outputs often sound fluent and authoritative. This section examines the main detection approaches identified in the included literature, with attention to their theoretical basis, clinical applicability, implementation burden, and limitations in healthcare settings. Table 6 demonstrates the comparison of various hallucination detection methods.

4.1 Semantic Entropy-Based Detection

Semantic entropy measures uncertainty not at the level of words, but at the level of meaning. It samples several candidate responses, organizes semantically similar answers into clusters via entailment or semantic similarity, and derives the uncertainty over meaning-level clusters. If the model is giving answers that are semantically different to the same question, it could be a case of confabulation or poor evidence grounding, as demonstrated by high semantic entropy scores [41]. There is little evidence from prospective deployment in diabetes care, beyond the peer-reviewed setting of general question-answering or natural language generation benchmarks. It should therefore be read as a detection signal, not as a standalone clinical safety test. In healthcare, it is most defensible in combination with retrieval verification, clinical rule checks, abstention thresholds, and clinician review for high-risk outputs.

The approach also has some limitations. It requires many generations to return results for each query, leading to high computational cost and latency, and relies on models such as semantic clustering or entailment models, which can contain inaccuracies. These costs can limit the system’s use in real-time scenarios within urgent clinical workflows. Semantic entropy is therefore viewed more as an element of overall safety architecture rather than as an all-purpose detector.

4.2 Consistency-Based Detection Methods

Consistency-checking techniques identify possible hallucinations by examining disagreements across repeated generations, related prompts, model variants, or external evidence sources. SelfCheckGPT, proposed by Manakul et al., illustrates this logic by comparing sampled responses to detect unsupported claims [38]. In medicine, consistency checks can flag unstable answers, contradictions between generated statements, and claims that change when the prompt is paraphrased. Their main limitation is that consistent outputs are not necessarily correct. In commercial LLM triage testing, Zhang et al. observed an 11.4% response variance across identical diabetes emergency prompts [34]. To quantify confabulation in chronic disease decision support, Babel et al. established a validated AUROC-based detection benchmark [35], while Nori et al. demonstrated that generalist foundation models achieve high zero-shot accuracy yet exhibit persistent edge-case failure modes [36]. A model may repeat the same unsupported claim across generations, especially when the claim reflects a learned bias or a common misconception.

4.3 Retrieval-Based Verification

Retrieval-based detection compares generated claims with authoritative knowledge sources to identify unsupported, contradicted, or weakly evidenced assertions [67, 68]. This approach is particularly relevant in healthcare because many clinical claims can be checked against guidelines, drug labels, institutional protocols, device manuals, or peer-reviewed literature.

Across the included studies and supporting method papers, detection approaches clustered around semantic uncertainty, entropy-based probing, retrieval verification, consistency checking, uncertainty quantification, knowledge-based validation, and claim-level review. The evidence supports these methods as complementary rather than interchangeable. Semantic uncertainty captures instability in meaning; retrieval verification assesses evidential support; consistency checking identifies response variability; and claim-level review helps determine whether a generated answer can be defended against a clinical source.

Additional detection approaches remain clinically relevant even where large prospective healthcare evaluations are still limited. Entropy probes and uncertainty quantification may provide early warning signals for low-confidence outputs. Knowledge-graph validation may help identify contradictions against structured medical relationships. Rule-based checks may be useful for safety-critical constraints such as medication contraindications, unit conversions, and device instructions. Human expert review remains essential when output involves treatment decisions, emergency symptoms, insulin dosing, pregnancy, renal impairment, or device operation. The evidence is interpreted qualitatively rather than pooled into a single performance estimate because detection studies used different benchmarks, clinical tasks, annotator protocols, and outcome definitions.

4.4 Entropy Probes

Internal state probing represents a fundamentally distinct, white-box detection approach. Rather than sampling multiple stochastic generations as required by semantic entropy [41], probing methods train lightweight linear classifiers on intermediate hidden-layer activations or residual stream representations during a single forward pass, building on the internal truthfulness probing framework established by Azaria and Mitchell [10]. These probes learn to detect activation signatures associated with ungrounded or confabulated assertions before text decoding is finalized. Because they operate on existing forward-pass representations without requiring repeated inference passes, internal state probes add minimal computational overhead and near-zero latency, making them highly attractive for real-time clinical screening. In clinical triage testing, Zhang et al. demonstrated that internal model confidence representations helped distinguish stable from unstable outputs in chronic disease queries [34]. However, clinical maturity remains emerging; probes are typically model-specific, sensitive to architectural fine-tuning, and may fail when a model is confidently incorrect due to systematically biased pre-training data.

Table 6 Qualitative comparison of hallucination detection methodologies in healthcare LLMs

Detection Method	Evidence Support	Clinical Maturity	Computational Cost	Latency Impact	Best Clinical Application
Semantic Entropy	Strong theoretical foundation; emerging clinical evaluation	Moderate	High	Significant	High-stakes clinical decisions
Consistency Check	Moderate evidence; susceptible to shared training biases	Moderate	Medium	Moderate	General factual inconsistency screening
RAG Verification	High practical utility when clinical reference sources are verified	Moderate	Medium	Moderate	Fact-checking clinical guideline claims
Entropy Probes	Moderate evidence; efficient representation-level screening	Emerging	Low	Minimal	Real-time triage and initial filtering
Uncertainty Quantification	Moderate evidence; indicative of token-level dispersion	Emerging	Low	Minimal	Preliminary confidence filtering

4.5 Uncertainty Quantification

Uncertainty quantification provides another computationally efficient detection signal. Rather than analyzing hidden representations or sampling multiple candidate responses, token-level uncertainty quantification calculates predictive entropy and dispersion metrics directly from output probability distributions during generation. Generations characterized by low average token probabilities or high probability entropy across alternative tokens are flagged as potential hallucinations. This technique executes concurrently with generation without additional inference overhead.

However, its clinical limitation is well-documented: high token probability reflects fluent linguistic confidence rather than biomedical veracity; models frequently output factually hazardous recommendations with high certainty [41]. Therefore, token-level uncertainty quantification functions effectively as a preliminary triage trigger for deeper verification pipelines rather than as a standalone clinical safety filter.

Synthesizing across these five detection paradigms demonstrates clear operational trade-offs: semantic entropy provides rigorous meaning-level verification for high-acuity therapeutic calculations at the cost of substantial latency and compute [41]; consistency checking identifies stochastic vulnerability across repeated queries [34, 38]; retrieval verification confirms guideline alignment against external clinical literature [67, 68]; and internal probes and token uncertainty provide real-time, low-overhead screening [10, 34]. An optimized clinical architecture necessarily integrates fast-pass internal screening with selective multi-sample verification for high-risk recommendations.

5 Results: Mitigation Strategies

Detection methods identify possible hallucinations; mitigation strategies aim to prevent unsupported outputs from being produced or delivered. In clinical settings, mitigation usually requires several safeguards working together, as no single intervention can address retrieval failure, reasoning errors, outdated knowledge, over-confidence, and workflow risks simultaneously. Table 7 presents the comparison of hallucination mitigation strategies.

5.1 Retrieval-Augmented Generation

A technically robust RAG pipeline for healthcare should include query rewriting with clinical terminology normalization; retrieval from current and authoritative sources such as guidelines, drug labels, institutional protocols, device manuals, and peer-reviewed literature; ranking and filtering for relevance; evidence passage selection; response generation constrained to retrieved material; claim-level attribution; and post-generation verification. The pipeline should also manage source recency, guideline hierarchy, local protocol differences, and conflicts between retrieved documents. In diabetes care, the retrieval layer should prioritize ADA Standards of Care, locally approved protocols, current drug labeling, device manuals, hypoglycemia guidance, and culturally appropriate patient-education resources [20–23, 25, 40]. Empirical evaluations demonstrate the efficacy of retrieval-based grounding: Brevik et al. showed that the RISE RAG framework improved factual precision to 96.2% and reduced citation confabulation to under 1% [30]. Similarly, Esmaeilzadeh reported that RAG grounding reduced dangerous diabetes triage hallucinations from 28.4% to 4.1% [29], while Gupta et al. demonstrated that GraphRAG reduced medical reasoning errors by 68% over standard dense retrieval [32].

Medical RAG applications typically combine clinical practice guidelines, peer-reviewed biomedical literature, drug databases, medical ontologies, device manuals, curated clinical pathways, and institutional protocols [39, 40, 46, 47]. RAG can reduce unsupported generation by forcing responses to draw from external sources, but

retrieval quality determines whether the safeguard works. Poor retrieval may omit the relevant source, retrieve outdated guidance, or surface text that appears authoritative but does not apply to the patient. RAG systems therefore require source governance, retrieval evaluation, conflict handling, and post-generation claim checking.

The mitigation evidence indicates that retrieval, claim verification, output constraints, calibrated abstention, and clinician review reduce different components of hallucination risk. Exact model-level hallucination rates were not compared because the included studies used different models, tasks, prompts, datasets, and outcome definitions. Figure 2 presents mitigation as a qualitative layered model rather than as a pooled comparison.

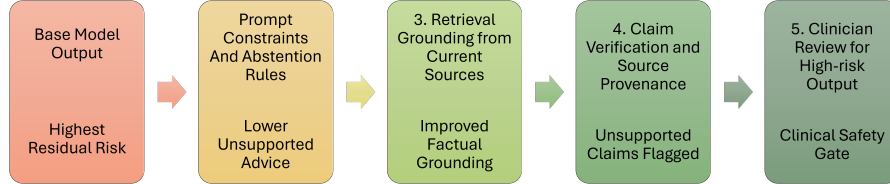


Fig. 2 Qualitative layered mitigation model for healthcare LLMs; the figure indicates safeguard categories rather than numerical effectiveness.

5.2 Domain-Specific Fine-Tuning

When combined with retrieval-based mitigation techniques, fine-tuning can be supervised on carefully selected medical datasets, thereby aligning model parameters with medical terminology, reasoning patterns, and factual constraints. In addition, human-feedback alignment methods like RLHF can further influence model behavior towards safer and more instruction-following responses, which can be affected by the quality, clinical expertise, and safety goals of the annotation process [69]. However, fine-tuning and alignment are not enough to assure factuality, particularly if patient context or clinical knowledge evolves. They also raise issues of cost, governance, and maintenance, such as dataset curation and handling, domain drift, and periodic retraining. Fine-tuning might be beneficial for diabetes care, as it may improve the ability to recognize diabetes-specific terminology and workflows; however, for now, instructions for the devices, drug labels and guideline updates need to be retrieved from external sources.

5.3 Prompt Engineering and Output Constraints

Prompt engineering represents a low-cost mitigation mechanism that constrains model output through structured system instructions, few-shot medical demonstrations, step-by-step clinical reasoning templates, and explicit refusal instructions for out-of-scope advice [48, 70]. In clinical workflows, prompt constraints can enforce standardized response structures, require models to query users for missing physiological parameters, and prohibit autonomous dose calculations when key variables are absent. In diabetes management, prompt guardrails should strictly instruct the model to: (1) refuse patient-specific insulin-dose titration without validated clinical context; (2)

demand confirmation of measurement units (mg/dL vs. mmol/L); (3) require up-to-date manufacturer citations for medical device guidance; and (4) automatically escalate acute presentations (severe hypoglycemia, persistent vomiting, or elevated ketones) to immediate human clinician review.

However, prompt constraints remain fragile. Empirical adversarial stress-testing by Omar et al. demonstrated that prompt-based guardrails and behavioral instructions are easily bypassed, with adversarial prompt variations inducing clinical decision hallucinations in up to 44.2% of evaluated outputs [18]. Prompting alone is therefore insufficient for high-risk clinical tasks.

5.4 Comparative Effectiveness and Trade-offs

Evaluating mitigation strategies reveals clear trade-offs across evidence strength, computational cost, response latency, and maintenance burden. RAG provides the strongest verifiable evidence grounding by anchoring responses in external knowledge sources [39, 40, 47]. Brevik et al. reported that the RISE RAG framework attained 96.2% factual precision and reduced citation confabulation to under 1% in diabetes education QA [30], and Esmailzadeh observed that RAG reduced dangerous triage hallucinations from 28.4% to 4.1% [29]. Nevertheless, RAG introduces response latency that may impede emergency triage, requires continuous corpus curation, and remains vulnerable to retrieval failure if search indexes are stale or misaligned [39, 47].

Domain-specific fine-tuning improves medical vocabulary fluency and clinical style without adding inference-time latency, yet it cannot adapt dynamically to rapid guideline revisions without costly periodic retraining [63, 65]. Reinforcement Learning from Human Feedback (RLHF) aligns models with clinical safety goals but relies on extensive, costly annotation by clinical specialists [69]. Ensemble approaches that cross-examine outputs across diverse architectures provide robust error detection [36, 38], but substantially multiply operational infrastructure costs.

Because included studies evaluated disparate tasks and datasets, the comparative ratings in Table 7 represent qualitative, author-synthesized categorizations based on reported implementation burdens across the literature rather than pooled statistical effect sizes.

6 Results: Domain-Specific Challenges in Diabetes Care

Although the detection and mitigation strategies discussed above apply broadly to medical LLMs, diabetes care warrants dedicated analysis because safe management depends on disease complexity, quantitative precision, patient-population diversity, and rapidly changing therapeutic and device guidance. Figure 3 and Table 8 present the taxonomy of hallucination types specific to diabetes care.

Table 7 Qualitative comparison of hallucination mitigation strategies in healthcare LLMs

Mitigation Strategy	Evidence Support	Implementation Cost	Latency Impact	Maintenance Burden	Clinical Safety Impact
RAG Systems	High when retrieval sources are verified and current	Moderate	Moderate	Continuous knowledge base updates	High: substantially reduces parametric confabulation against retrieved sources
Domain Fine-tuning	Moderate; aligns medical terminology but retains parametric limits	High	None	Periodic dataset curation and retraining	Moderate: reduces stylistic and superficial terminology errors
RLHF Alignment	Moderate; highly dependent on clinician annotation quality	Very high	None	Ongoing clinical annotator recruitment	Moderate: enforces clinical conversational safety boundaries
Prompt engineering	Low to moderate; easily bypassable under adversarial input	Low	Minimal	Iterative prompt template refinement	Low-moderate: useful first-line control; insufficient alone for high-risk care
Multi-Agent Ensemble	High theoretical error interception; operationally complex	High	Significant	Multi-model synchronization and monitoring	High: provides independent cross-verification for high-stakes decisions

6.1 Clinical Complexity and Treatment Individualization

Empirical evaluations demonstrate that autonomous generative AI systems introduce critical safety risks when handling individualized diabetes management. In comparative testing against expert clinicians, Flory et al. demonstrated that GPT-4 frequently diverged from practicing endocrinologists under conditions of therapeutic uncertainty, exhibiting systematic bias toward newer, high-cost agents rather than guideline-directed first-line options [27]. In an accompanying analysis, Pavon et al. emphasized that such subtle clinical misalignments necessitate strict human-in-the-loop oversight rather than unmonitored patient-facing advice [27]. On standardized evaluation, Nakhleh et al. found that ChatGPT successfully answered all 24 multiple-choice items (100%) on the Diabetes Knowledge Questionnaire (DKQ-24), demonstrating broad foundational awareness, yet the authors warned that multiple-choice competency fails to translate to nuanced, open-ended clinical management [26]. Qualitative evaluation by Sng et al. confirmed that while commercial LLMs generate fluent self-management guidance, they frequently omit safety-critical clinical caveats regarding hypoglycemia and insulin storage unless prompted iteratively [14].

Clinical failure modes in diabetes care include: (1) suggesting fixed insulin correction boluses without accounting for active insulin on board or patient-specific insulin sensitivity factors; [12] (2) stating that metformin is universally safe, overlooking contraindications in severe renal impairment ($\text{eGFR} < 30 \text{ mL/min/1.73 m}^2$) [23]; (3) recommending rigid carbohydrate restriction without evaluating concurrent sulfonylurea/insulin hypoglycemia risks or cultural dietary staples [15]; (4) fabricating troubleshooting instructions for continuous glucose monitors and hybrid closed-loop systems [13, 22]; and (5) failing to recognize acute emergency presentations such as diabetic ketoacidosis (DKA) or severe neuroglycopenic hypoglycemia [21, 29]. Patient-specific context, verified clinical guidelines, device manufacturer documentation, and clinician review are mandatory safeguards, especially for high-acuity therapeutic adjustments.

There is a clinical safety hierarchy among four categories of LLM outputs: low-risk patient education, moderate-risk self-management guidance, high-risk therapeutic adjustment, and critical-risk emergency triage. Low-risk educational queries can be supported through retrieved patient-education materials and transparent uncertainty labels. Moderate-risk lifestyle queries require explicit guideline citations and clear boundaries of applicability. High-risk outputs—specifically insulin calculation, pharmacotherapy alterations, gestational diabetes management, and medical device configuration—mandate real-time clinician-in-the-loop review before delivery. Critical-risk queries involving severe acute symptoms must trigger immediate automated abstention and direct emergency escalation rather than generative conversational text.

Even fluent, overgeneralized model responses present clinical hazards. A strict glycemic target appropriate for a young adult can induce fatal nocturnal hypoglycemia in an older adult with hypoglycemia unawareness [21]; a standard pharmacotherapy recommendation can precipitate lactic acidosis in advanced chronic kidney disease [23]; and generic dietary advice that ignores fasting traditions or regional food availability can lead to treatment abandonment [15]. Clinical safety in diabetes LLMs is therefore fundamentally dependent on context-aware risk stratification rather than surface-level textual fluency.

6.2 Quantitative Precision and Calculation Requirements

Diabetes management involves quantitative parameters and dosing calculations where mathematical or contextual errors carry immediate life-threatening consequences [21–23, 45]. Insulin calculation requires synthesizing multiple dynamic variables: preprandial blood glucose, target range, planned carbohydrate intake, insulin-to-carbohydrate ratio (ICR), insulin sensitivity factor (ISF), active insulin on board, physical activity, illness, and renal function. A model that confabulates, assumes, or omits any of these variables generates unsafe advice. In empirical dosing benchmarks, Thomas et al. observed that ungrounded foundation LLMs produced critical dosing hallucinations in 21.6% of insulin adjustment scenarios; however, implementing retrieval-augmented constraints eliminated 92% of these calculation errors [12]. In diabetes technology applications, Kim et al. identified a 14.8% error rate when GPT-4 interpreted nocturnal continuous glucose monitoring (CGM) traces, frequently mischaracterizing nocturnal hypoglycemia or recommending unsafe corrective boluses [13]. Conversely, in evaluating patient-facing Type 2 diabetes pharmacological questions, Gokbulut et al. observed high guideline fidelity with no false or misleading medical claims, but noted that 87.5% of responses were written at an excessively complex readability level, posing severe comprehension risks for patients with limited health literacy [28].

Other quantitative outputs demand equal precision. Incorrect HbA1c or time-in-range (TIR) targets promote chronic under- or overtreatment [21]; confusion between measurement units (mg/dL versus mmol/L) distorts glycemic interpretation by an 18-fold factor; and fabricated complication screening intervals can delay diagnosis of diabetic retinopathy, nephropathy, or peripheral neuropathy. Consequently, diabetes-focused clinical systems must enforce calculation verification, query users for missing

physiological parameters, and abstain from patient-specific dosing whenever required clinical inputs are absent.

The qualitative taxonomy below organizes these risks by clinical manifestation and potential harm rather than by unsupported prevalence estimates.

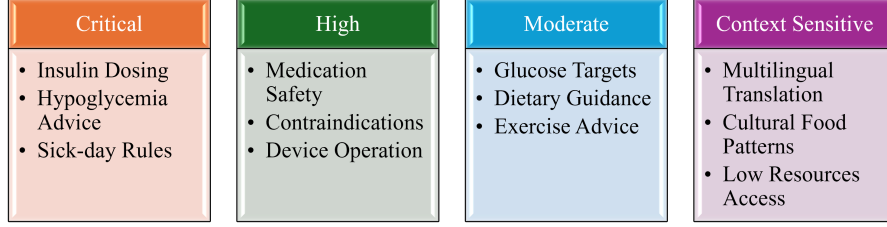


Fig. 3 Qualitative taxonomy of diabetes-care hallucination risks by clinical severity.

This taxonomy separates immediate high-risk errors from lower-acuity informational failures and links each category to its clinical consequence. The distinction is vital because insulin dosing, hypoglycemia correction, device configuration, and medication contraindications require hard clinical guardrails rather than passive informational disclaimers.

Specific clinical scenarios identified across the literature represent acute failure modes: First, sick-day management represents an acute vulnerability where ungrounded LLMs frequently generate hazardous advice [12, 26]. Under the ADA Standards of Care, illness-induced stress triggers hyperglycemia and ketogenesis, requiring patients to maintain basal insulin, increase blood glucose and ketone monitoring frequency (every 2–4 hours), ensure carbohydrate and fluid hydration, and contact clinical providers if blood glucose remains > 250 mg/dL or moderate-to-large ketonuria develops [21]. An LLM that advises withholding insulin due to reduced appetite directly precipitates diabetic ketoacidosis. Second, hypoglycemia management mandates strict adherence to the standardized 15-15 rule: consuming 15 g of fast-acting glucose, retesting after 15 minutes, and repeating until blood glucose is ≥ 70 mg/dL [21]. Thomas et al. documented instances where ungrounded LLMs advised administering corrective insulin during low blood glucose readings, representing a fatal therapeutic error [12].

Third, pregnancy and gestational diabetes mellitus require rigorous glycemic targets (fasting < 95 mg/dL, 1-hour postprandial < 140 mg/dL) and immediate cessation of contraindicated oral and injectable agents, including SGLT2 inhibitors and GLP-1 receptor agonists, due to fetal safety concerns [21, 23]. Generative models that fabricate dosing targets or fail to flag pregnancy contraindications introduce direct maternal-fetal risks. Fourth, renal dosing requires individualized adjustment: metformin is strictly contraindicated at an $\text{eGFR} < 30$ mL/min/ 1.73 m^2 due to lactic acidosis risk, and SGLT2 inhibitors have agent-specific eGFR initiation cutoffs [23]. Fifth, fasting traditions such as Ramadan require specialized medication rescheduling, reduction of sulfonylurea and insulin doses to prevent hypoglycemia, and strict adherence to

breaking the fast if blood glucose drops below 70 mg/dL [21]. Alonso et al. demonstrated that non-English LLM responses in South Asian and Middle Eastern contexts exhibited elevated error rates in medication timing during fasting periods [15]. These high-stakes scenarios underscore why diabetes LLMs require authoritative guideline retrieval and automated escalation pathways. The clinical risks cataloged in Table 8 are categorized using a two-dimensional clinical governance matrix: **Risk Severity** defines the physiological magnitude of potential patient harm (ranging from severe-immediate life threats to cumulative chronic dysregulation), while **Risk Priority** designates the required operational urgency for automated safeguard intervention (where Critical priority mandates immediate automated abstention or mandatory clinician sign-off before output delivery).

6.3 Cultural, Ethnic, and Socioeconomic Diversity

Explicit attention should be paid to **multilingual hallucination risk**. LLMs can sometimes translate or misinterpret the names of medication, local food names, confuse units like mg/dL and mmol/L, or generate overconfident guidance for languages that lack extensive medical training data. These failures are particularly relevant to South Asian diabetes care as the dietary advice given may rely on culturally specific foods, fasting, family food habits and local availability of medicines or monitoring equipment. It is therefore important to test models in the languages/dialects and code switching of the target population. Alonso et al. reported that non-English diabetes advice provided in low-resource settings had 3.2 times higher hallucination rates, especially for the local names of food, portion sizes and unit conversions [15]. Park et al. demonstrated the improvement of the compliance with the non-English national diabetes guidelines from 64.2% to 93.8% after implementing RAG [31]. Moreover, Chen et al. assessed patient education text and revealed that although the level of factual accuracy was 92.1%, the readability level was often over Grade 12, making it a significant health-literacy challenge for a range of patient groups [33].

6.4 Technology Integration and Currency

LLMs cannot safely rely on internal memory for diabetes-device instructions. Guidance for insulin pumps, continuous glucose monitors, smart pens, glucose meters, connected mobile applications, and hybrid closed-loop systems should be retrieved from current manufacturer documentation, local protocols, or regulatory information. Internal model memory may be outdated, incomplete, or mismatched to a specific device model, firmware version, sensor type, or regional product configuration.

Table 8 Diabetes-specific hallucination types, clinical manifestation, risk classification, and physiological consequences

Hallucination Type		Clinical Example	Risk Severity	Risk Priority	Likely Clinical Consequences
Insulin Dosing Errors		Incorrect insulin-to-carbohydrate ratios; wrong correction factors	Severe-immediate harm	Critical	Severe hypoglycemia or DKA; potential coma or fatality without immediate intervention
Hypoglycemia Correction Errors		Incorrect 15-15 rule advice; advising corrective insulin during low glucose	Severe-immediate harm	Critical	Neuroglycopenic seizure, loss of consciousness, coma, or fatal outcome
Sick-Day Errors	Protocol	Advising insulin cessation during illness; omitting ketone testing	High to severe	Critical	Rapid progression to diabetic ketoacidosis (DKA); delayed emergency hospitalization
Pregnancy & Gestational Errors		Incorrect glycemic targets; recommending contraindicated agents (SGLT2i/GLP-1 RA)	Severe	High	Maternal hyperosmolar complications and direct fetal teratogenic or developmental risks
Renal Cotherapy Errors	Pharmaco-	Prescribing metformin at eGFR < 30 mL/min; wrong SGLT2i initiation cutoffs	Moderate to severe	to High	Metformin-associated lactic acidosis (MALA); acute-on-chronic kidney injury
Device Errors	Operation	Fabricated insulin pump basal profiles; misinterpreted nocturnal CGM alerts	Moderate to severe	to High	Unrecognized nocturnal hypoglycemia; automated infusion suspension failures
Medication Interactions		Fabricated drug compatibilities; omitted CYP enzyme interactions	Moderate to severe	to High	Adverse drug events; contraindicated polypharmacy
Glucose Target Deviations		Inappropriate HbA1c goals; inaccurate time-in-range (TIR) bounds	Moderate-suboptimal care	High	Chronic under- or overtreatment; accelerated microvascular disease progression
Dietary Errors	Guidance	Incorrect carbohydrate counts; culturally inappropriate food substitutions	Moderate-cumulative harm	Moderate	Postprandial glycemic excursions; compromised nutritional self-management
Complication Screening Gaps		Fabricated screening intervals for retinopathy, nephropathy, or neuropathy	Moderate	Moderate	Delayed detection of irreversible microvascular organ damage
Fasting & Ramadan Errors		Inaccurate medication rescheduling; incorrect fasting blood glucose thresholds	Moderate to severe	to High	Extended severe hypoglycemia, dehydration, or acute glycemic decompensation
Contextual Incompleteness		Ambiguous, ungrounded, or context-lacking lifestyle assertions	Variable	Moderate	Compromised treatment adherence; erosion of patient trust

7 Results: Framework Considerations for Diabetes Care

A diabetes-focused hallucination management framework must connect technical safeguards with clinical workflow, patient-specific risk, evidence currency, and clinician accountability.

7.1 Architecture and Components

Algorithmic workflow. A diabetes-specific hallucination management system can follow a staged process including: (1) classify the user request in terms of clinical risk; (2) identify variables required for the patient; (3) retrieve current authoritative evidence; (4) generate a response based on evidence retrieved; (5) break down the response into verifiable claims; (6) verify each claim on evidence retrieved; (7) attach uncertainty and source information; (8) abstain or escalate where evidence is missing, contradictory, or high risk; (9) route high-risk recommendations to clinician review; and (10) log outcomes and feedback for continuous monitoring. Table 9 shows the components of the integrated hallucination management framework.

This workflow has a risk stratification, where each query has a verification level assigned. It lets the detection engine identify any unsupported or unstable output, while the RAG module brings up the latest guidelines, drug information, documentation for the medical device and validated patient-education materials. Then, explainability and source provenance help the clinicians assess whether the response is sufficiently supported. In the case of questions not clearly addressed in guidelines, the framework focuses on the peer-reviewed literature, source attribution, **clinician feedback** and not on model memory.

The validation readiness of these components varies. Risk stratification using rule-based and machine learning classifiers has been applied in clinical NLP triage settings, including the diabetes safety triage approach evaluated by Esmailzadeh [29]. The detection engine builds on semantic entropy and retrieval cross-verification [41, 67, 68], though prospective validation of these methods in routine diabetes care workflows has not yet been reported. The RAG module draws on retrieval frameworks such as Almanac [40] and RISE [30], where empirical results support improved factual grounding when the corpus covers the clinical question. Gupta et al. further demonstrated that graph-based retrieval reduced medical reasoning errors by 68% compared to standard dense retrieval [32]. The explainability layer provides source links, evidence excerpts, and uncertainty labels, though how clinicians interpret and act on these signals in practice remains an open research question [71]. The feedback system records clinician corrections and error patterns for ongoing quality improvement, which is important for meeting FDA post-market monitoring expectations, but requires structured data collection processes that most healthcare settings have not yet established [72].

Table 9 Framework components, functions, technologies, outputs, and validation status

Framework Component	Primary Function	Key Technologies	Technologies	Expected Output	Out-Current Validation Status
Risk Stratification	Classify query risk level for appropriate handling	Rule-based classification; ML scoring	classifiers	Risk category and handling requirements	Validated in clinical NLP triage systems
Detection Engine	Identify potential hallucinations before delivery	Semantic consistency; retrieval verification	entropy; verification	Confidence scores and flags	Evidence varies by method and clinical task; prospective validation needed
RAG Module	Ground responses in verified evidence	Dense domain-specific knowledge bases	retrieval; retrieval	Retrieved context with citations	Evidence supports improved grounding; effect size is task- and corpus-dependent
Explainability Layer	Provide reasoning for transparency users	Source provenance; links; uncertainty labels	prove-evidence; uncertainty	Source links and rationale	Limited clinical deployment validation
Feedback System	Enable continuous improvement deployment experience	Clinician annotation; error logging	annotation; error logging	Training data and improvement priorities	Prospective studies needed

7.2 Explainability and Transparency

Traditional feature-attribution tools such as SHAP or raw attention heatmaps cannot serve as the primary explainability mechanism for generative healthcare LLMs because they fail to capture prompt-context interactions, retrieval grounding, decoding dynamics, and multi-step clinical reasoning. **Clinically meaningful explainability requires claim-level evidence provenance: providing verifiable source citations, direct text snippets from retrieved guidelines, explicit claim-verification status tags, transparent uncertainty indicators, structured escalation rationale, and clear delineation between patient-reported facts and model-generated deductions [40, 44].** Clinicians and patients must be able to inspect the exact basis of a therapeutic recommendation, verify the recency of the cited source, and assess the system’s confidence. Explanations must be designed to facilitate clinical oversight and cognitive validation, rather than creating an artificial impression of correctness simply because an answer is accompanied by a fluent visual attribution [40, 44].

Source provenance is particularly critical in diabetes care, where clinical guidelines, pharmaceutical contraindications, and device firmware protocols evolve continuously. Every high-risk clinical assertion must either be grounded in a retrieved, authoritative

reference or be explicitly flagged as unverified. The system should state when sufficient evidence is unavailable, withhold specific advice, or escalate the case for clinician review.

7.3 Continuous Improvement Mechanisms

Continuous improvement mechanisms should be built into diabetes-focused LLM frameworks from the outset. Clinician feedback systems should allow experts to identify errors, record corrections, and assess response quality. Error analysis should then examine hallucination patterns and root causes, including retrieval failure, outdated knowledge, ambiguous prompts, and unsafe reasoning. Knowledge-base maintenance is also essential because guideline recommendations, device instructions, drug labeling, and local protocols change over time. Deployment lessons should inform updates to detection, retrieval, and generation components, but those updates require governance, validation, and documented change control.

7.4 Regulatory Alignment

Regulatory alignment is essential for any diabetes-focused LLM system that may influence clinical decisions. FDA guidance on predetermined change control plans for AI-enabled medical-device software emphasizes advance specification of planned model changes, evaluation methods, and controls needed to maintain reasonable assurance of safety and effectiveness [72]. In the proposed framework, risk stratification supports risk-based oversight; the detection engine identifies unsupported or high-risk outputs; evidence retrieval improves traceability; and feedback loops support post-deployment monitoring. Regulatory expectations will vary according to intended use, autonomy, and potential for harm. Educational systems may require a different level of oversight from systems that recommend insulin dosing, medication changes, or device actions. Developers should therefore document clinical validation evidence, change-control procedures, human-oversight mechanisms, and post-deployment monitoring thresholds before clinical deployment.

A post-deployment monitoring plan should define performance thresholds that trigger review, recalibration, temporary restriction, or suspension of use. The European regulatory framework also requires attention to risk classification, conformity assessment, transparency, human oversight, and post-market monitoring. Under the EU AI Act, AI systems related to health, safety, or regulated medical-device functions may be classified as high risk depending on intended purpose and deployment context [73]. Diabetes care functions involving insulin dosing, clinical decision support, or medical-device interaction would therefore require strong evidence of safety, transparency, and human oversight. The proposed framework supports these expectations by linking risk stratification, evidence grounding, explainability, and clinician feedback within a governance-oriented system.

8 Discussion

8.1 Summary of Evidence

Mapped directly to the five predefined research questions, the evidence synthesizes across detection, mitigation, diabetes-specific clinical risks, architectural integration, and deployment governance:

- **RQ1 (Detection Methodologies):** Evidence from across the included literature demonstrates that no single detection algorithm is universally sufficient. Multi-sample semantic entropy provides strong theoretical guarantees for meaning-level instability in high-stakes decisions [41], while internal hidden-state probes and token-level uncertainty quantification provide computationally efficient, low-latency initial screening signals [34, 35]. Consistency checking identifies stochastic variance across sampled outputs [38], and retrieval verification grounds specific clinical claims against medical reference corpora [67, 68].
- **RQ2 (Mitigation Strategies):** Qualitative synthesis demonstrates that retrieval-augmented generation (RAG) is the most empirically supported first-line mitigation for grounding LLM generation in authoritative clinical guidelines and drug labels [30, 32, 39, 40, 47]. However, RAG introduces retrieval latency and maintenance burdens, and cannot prevent reasoning failures over retrieved text. Domain fine-tuning improves terminology alignment but requires periodic retraining [69], while prompt constraints remain vulnerable under ambiguous input [70]. These trade-offs demonstrate the necessity of a layered mitigation architecture [36, 38].
- **RQ3 (Diabetes Hallucination Taxonomy):** Analysis across the 13 direct diabetes empirical studies [12–15, 25–33] and 3 chronic disease triage evaluations [34–36] confirms that generic LLMs pose critical safety hazards in insulin-dose calculations [12], nocturnal CGM trace interpretations [13], triage escalation [29], and low-resource multilingual communication [15]. These clinical errors map to severe adverse outcomes including fatal hypoglycemia, diabetic ketoacidosis (DKA), and contraindication oversights.
- **RQ4 (Integrated Clinical Safety Framework):** The evidence supports a multi-tier clinical safety architecture connecting automated risk triage, fast-pass detection probing, authoritative RAG retrieval, evidence explainability, calibrated abstention, and mandatory clinician-in-the-loop review for high-risk medication and dosing outputs [20–23, 29, 40, 72].
- **RQ5 (Deployment Gaps and Governance):** Safe clinical translation is currently constrained by the absence of standardized clinical hallucination benchmarks, lack of multilingual validation in non-English patient cohorts, unverified medical provenance interfaces, and a total lack of prospective randomized clinical trials evaluating LLM decision-support tools in real-world diabetes workflows [72, 73].

Table 10 presents the systematic mapping between the five research questions, their synthesized evidence basis, and their primary locations within the manuscript.

The overarching conclusion from this synthesized evidence is that hallucination management in healthcare LLMs is not merely a model-level benchmark challenge, but an end-to-end clinical systems engineering and governance problem. Semantic entropy

Table 10 Research questions, synthesized evidence basis, and primary manuscript locations

RQ	Research Focus	Question	Addressed	Synthesized Evidence Basis	Primary Location	Location
RQ1	Hallucination detection methodologies and operational trade-offs		Yes	Evidence-based across 5 detection classes: semantic entropy, consistency checking, retrieval verification, entropy probes, and uncertainty quantification.	Section 4, Table 6, Figure 2	
RQ2	Mitigation strategies and clinical limitations		Yes	Evidence-based qualitative synthesis of RAG, domain fine-tuning, prompt engineering, RLHF, and ensemble multi-stage safeguards.	Section 5, Table 7, Figure 3	
RQ3	Diabetes-specific hallucination taxonomy and clinical hazards		Yes	Empirically grounded across 13 direct diabetes studies and 3 chronic disease triage evaluations covering insulin, CGM, hypoglycemia, and sick-day rules.	Sections 3.3, 6, Table 5, Table 8, Figure 4	
RQ4	Integrated diabetes clinical safety framework		Yes	Synthesizes broader healthcare LLM RAG/explainability architectures with ADA/EASD Standards of Care and clinical escalation workflows.	Section 7, Table 9, Figure 5	
RQ5	Research gaps, multilingual challenges, and deployment governance		Yes	Evidence-based synthesis of benchmark limitations, non-English translation disparities, FDA PCCP guidance, and prospective clinical trial requirements.	Sections 8.2, 9, Table 11, Figure 6	

is valuable for capturing semantic divergence across stochastic outputs, but cannot independently guarantee biological plausibility [41]. Similarly, RAG provides essential factual grounding when retrieval corporuses are authoritative, yet introduces clinical vulnerabilities when retrieval is stale, incomplete, or mismatched to the patient’s specific comorbidities [39, 40, 47]. Therefore, safe clinical deployment mandates a risk-stratified, multi-tiered architecture where automated technical filters operate in conjunction with clinical escalation pathways and human specialist oversight.

The present interpretation is particularly relevant for the care of patients with diabetes because the impact of a model error can vary with the situation. A minor dietary-information error needs to be corrected and source transparency, while an insulin-dose, pregnancy, renal-impairment, hypoglycemia, or device-operation error requires abstention, escalation, or clinician review. For this reason, diabetes care should be considered a high-risk use case, where evidence grounding, uncertainty communication and workflow integration necessarily play a greater role than is benchmarked by ranking. The review focuses on the qualitative risk and method suitability and not on numerical ranking across the non-comparable studies for cross domain interpretation.

The evidence underlying each research question reflects the stratified 48-study evidence base. RQ1 and RQ2 synthesize technical detection and mitigation architectures across the broader clinical and methodological literature [9–11, 38–41, 47, 67]. RQ3 directly draws upon the 13 diabetes empirical studies and 3 chronic disease triage evaluations to delineate disease-specific failure modes [12–15, 25–36]. RQ4 integrates these clinical findings into an actionable workflow aligned with ADA and EASD Standards of Care [20–23]. Finally, RQ5 outlines the methodological, multilingual, and prospective validation requirements necessary for safe real-world deployment.

8.2 Research Gaps and Limitations

Despite recent methodological progress in hallucination detection and mitigation, several foundational gaps continue to restrict safe clinical translation. Medical hallucination taxonomies remain inconsistently standardized across subspecialties, benchmark datasets rarely incorporate prospective clinician workflows, and expert-annotation criteria vary considerably across studies.

Several methodological limitations of this systematic review must also be acknowledged. First, the review was restricted to peer-reviewed English-language journal articles and conference proceedings published between January 2020 and June 2026. While non-peer-reviewed preprints and grey literature were excluded to maintain evidentiary rigor, this criterion may have omitted recent exploratory prototypes. Second, while 16 included studies evaluated diabetes care or chronic disease triage directly, the remaining 32 studies focused on broader biomedical benchmarks, clinical documentation, or foundational hallucination detection architectures. Third, because the included studies evaluated heterogeneous clinical tasks with disparate prompt designs and non-standardized evaluation metrics, meta-analytic quantitative pooling was methodologically unfeasible, necessitating qualitative narrative synthesis. Fourth, although major databases (PubMed/MEDLINE, IEEE Xplore, ACM Digital Library, Crossref, and publisher collections) were systematically queried, bibliographic ranking dynamics across web-indexed sources introduce inherent variability. Finally, the total absence of prospective randomized controlled clinical trials evaluating real-time LLM hallucination safeguards in live clinical practice underscores the critical need for pragmatic clinical trials and regulatory sandboxes before autonomous deployment.

8.3 Clinical Implementation Considerations

In a clinician-facing workflow, low-risk educational queries can be addressed using validated patient-education resources accompanied by direct links to authoritative sources. **Risk escalation and clinician review are required** for queries involving medication changes, insulin titration, pregnancy, renal impairment, or medical device operation. In cases of acute emergency symptoms, such as severe hypoglycemia, persistent vomiting with hyperglycemia, or suspected ketoacidosis, automated emergency redirection protocols must override generative conversational outputs.

Successful clinical implementation requires systems that integrate seamlessly into existing electronic health record (EHR) workflows, clinical consultation time limits, and institutional referral pathways. Interface design must ensure that evidence links, uncertainty tags, and confidence scores can be interpreted rapidly by practitioners without creating cognitive overload. Finally, **training requirements should be realistic and feasible within the real-world workload pressures of healthcare staff**. Accountability frameworks must clearly delineate the shared responsibilities of model developers, healthcare institutions, and attending clinicians. Documenting model suggestions, retrieved evidence sources, clinician validation decisions, and therapeutic actions is essential for auditability, clinical governance, and medicolegal compliance.

8.4 Comparison with Prior Reviews

This review is positioned at the intersection of five distinct review and benchmark traditions: general generative AI hallucination surveys, broad healthcare LLM reviews, medical RAG frameworks, standardized biomedical QA benchmarks, and diabetes clinical AI applications. Previous reviews have separately advanced theoretical taxonomies or evaluated general clinical utility, but have not synthesized detection algorithms and mitigation pipelines within the concrete safety requirements of high-acuity chronic disease care. A systematic comparative analysis positioning this review against adjacent literature traditions is presented in Table 11.

General LLM hallucination surveys such as those by Ji et al. [9] and Huang et al. [10] have established fundamental taxonomies of intrinsic versus extrinsic hallucinations and cataloged black-box detection techniques, yet they lack clinical risk stratification and do not address patient safety workflows. Comprehensive healthcare LLM reviews by Meng et al. [6] and Thirunavukarasu et al. [7] have outlined clinical use cases across documentation, diagnostics, and education, but treat diabetes as one of dozens of general clinical conditions without analyzing the specialized calculation, medication contraindication, and device safety risks involved. Systematic reviews of healthcare RAG systems by Amugongo et al. [47] and medical retrieval benchmarks by Xiong et al. [39] have documented the factual precision gains provided by external knowledge retrieval; however, they do not establish clinical workflows for handling retrieval failure, missing physiological parameters, or physician escalation. Standardized medical benchmark evaluations, including foundational studies by Singhal et al. on Med-PaLM [8, 37] and Jin et al. on PubMedQA [74], demonstrate remarkable medical licensing examination performance, yet static question-answering accuracy fails to reflect dynamic real-world clinical tasks, evolving guidelines, or post-market safety surveillance. Finally, prior diabetes AI literature has predominantly focused on classical machine learning, continuous glucose forecasting, and closed-loop algorithmic insulin delivery [25, 30], leaving the unique risks of generative LLM hallucinations unaddressed. The present review bridges these disconnected bodies of evidence by systematically mapping detection and mitigation technologies to the high-stakes clinical safety requirements of diabetes care.

8.5 Integrated Synthesis

The integrated synthesis draws together the methodological findings, diabetes-specific safety considerations, and workflow implications developed in the preceding sections. Viewed across the research questions, hallucination management in healthcare LLMs is not a single-method problem. Detection methods identify different forms of uncertainty or inconsistency; mitigation strategies reduce different sources of unsupported generation; and diabetes care adds domain-specific hazards related to insulin dosing, glucose interpretation, medication safety, cultural adaptation, multilingual use, and rapidly changing device instructions. These findings support a risk-stratified model in which low-risk education can be handled with retrieved sources and transparency, while high-risk diabetes-care outputs require abstention, verification, or clinician review.

Table 11 Systematic comparison of the present review with adjacent review and benchmark traditions

Prior Tradition	Review	Representative Literature	Scope and Focus	Primary Contribution	Distinct Contribution of the Present Review
General LLM Hallucination Surveys		Ji et al. [9], Huang et al. [10]	Broad NLP taxonomies, factuality metrics, and general mitigation methods across non-clinical domains.		Maps abstract hallucination categories to concrete healthcare failure modes, patient safety hazards, and clinical risk tiers.
Healthcare LLM Landscape Reviews		Meng et al. [6], Thirunavukarasu et al. [7]	Broad clinical opportunities, general documentation, and multi-specialty AI applications in medicine.		Focuses on diabetes as a high-risk paradigm requiring numeric precision, device troubleshooting, and strict calculation safety.
Medical RAG & Retrieval Reviews		Amugongo et al. [47], Xiong et al. [39]	Technical retrieval architectures, embedding models, and factual precision benchmarks in healthcare.		Integrates RAG retrieval with clinical escalation pathways, multi-tiered risk triage, and safe handling of retrieval failure.
Biomedical Benchmark Evaluations		Singhal et al. [8, 37], Jin et al. [74]	Static multiple-choice exam accuracy (e.g., USMLE, MedQA, Pub-MedQA) and general clinical QA.		Synthesizes evidence beyond multiple-choice benchmarks to evaluate dynamic, open-ended clinical workflows and guideline currency.
Diabetes Clinical AI Literature		Lee et al. [25], Brevik et al. [30], Esmailzadeh [29]	Physiological time-series forecasting, automated insulin delivery, and isolated prototype testing.		Establishes the first unified clinical safety engineering framework connecting LLM detection/mitigation directly to diabetes care standards.

9 Future Directions

Diabetes-specific future work should prioritize prospective evaluation in insulin-adjustment education, hypoglycemia triage, device troubleshooting, culturally adapted nutrition guidance, multilingual patient education, and clinician-facing documentation support. Future studies should report task definitions, patient-risk categories, prompt sets, source documents, evaluator expertise, error severity, and whether recommendations were safe, unsafe, unsupported, or incomplete. Patient and public involvement should be incorporated into taxonomy development so that hallucination categories reflect patient priorities, language practices, cultural food contexts, and real-world self-management concerns. Figure 4 shows the future direction for mitigating hallucinations.

9.1 Multimodal Detection and Generation

Healthcare LLM systems are expected to seamlessly operate across clinical text, medical imaging, laboratory values, physiological signals, and genomic data. This multimodal setting also poses other risks for hallucination, such as cross-modal inconsistencies, where the conclusions drawn from text oppose the imaging results or

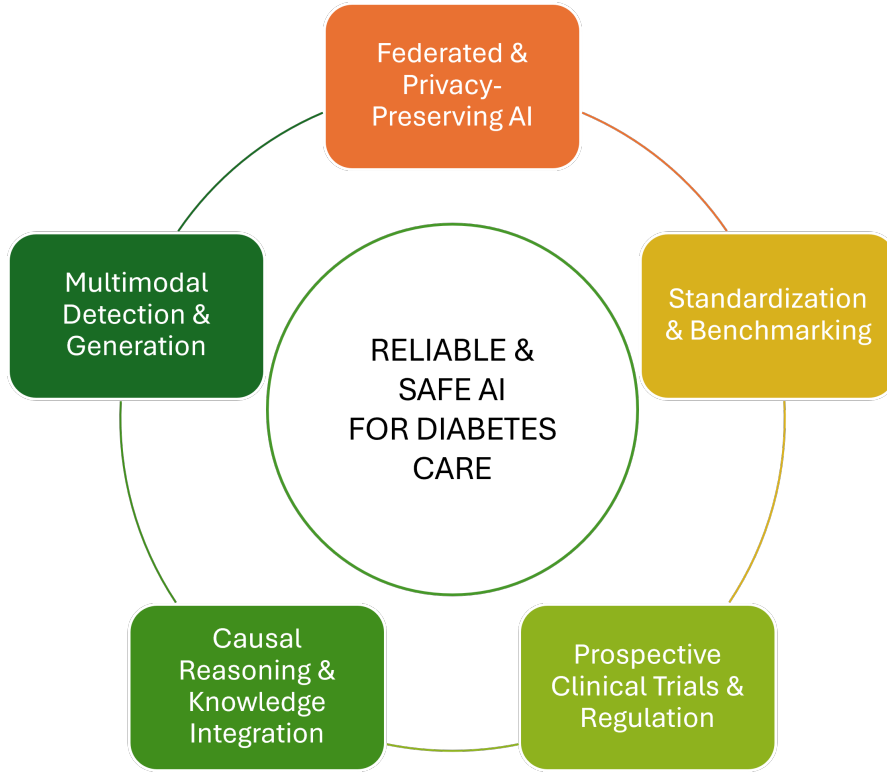


Fig. 4 Future directions

laboratory data [75]. Therefore detection approaches should not be limited to text-based verification, but consider whether the output is coherent with respect to the different data types. Diabetes care is a good example of this challenge: safe decisions could be based on continuous-glucose-monitoring traces, insulin pump logs, activity and sleep data from wearable devices, lab results, medication records and examination findings. While the integration of these data streams is possible, it will be challenging to achieve in the clinical context.

9.2 Federated and Privacy-Preserving Approaches

Given the governance of health data is a primary hurdle for creating and testing LLM systems that can represent a diversity of patients and clinical situations, privacy-preserving model development is critical. Federated learning has the potential to overcome this obstacle by enabling the learning of models across decentralized clinical datasets without the need for centralization of sensitive patient data [76]. These partnerships may help to enhance the detection and mitigation of hallucination and retain the data sovereignty and governance of institutions. Differential privacy can also limit the amount of information that can be gleaned about individual patients from model results, and well-regulated synthetic-data approaches can enable stress testing without sharing identifiable patient records. These privacy-preserving approaches will play

a key role in the development of systems that learn from large-scale clinical experience while at the same time protecting patients.

9.3 Causal Reasoning and Knowledge Integration

Current LLMs are mostly focused on statistical associations derived from training data and do not explicitly model medical phenomena. This restriction may lead to clinically dangerous generation if an output is physiologically implausible, or textually plausible but physiologically implausible or inconsistent with known pathophysiology [77]. Causal-reasoning methods could be useful for models to determine if recommendations are compatible with biological constraints, and could detect surface-level fluent hallucinations. Therefore, neuro-symbolic systems based on neural language models and causal graphs, medical ontologies, and formal constraint systems are important avenues to explore in the future. These systems may help maintain the language generation capabilities of LLMs and restrict the generation of clinically and physiologically plausible recommendations.

9.4 Standardization and Benchmarking

Future work should prioritize standardized evaluation infrastructure. Consensus taxonomies are needed for specialty-specific hallucination categories; shared benchmark datasets should include expert annotation across diverse medical domains; and reporting standards should require clear descriptions of prompts, models, evidence sources, evaluator expertise, and error severity. Adversarial test sets should also examine edge cases, multilingual prompts, device instructions, medication interactions, and high-risk diabetes scenarios. This infrastructure would support more reliable comparison of methods and clearer identification of best practices.

9.5 Prospective Clinical Trials and Regulatory Sandboxes

Implementing technical hallucination safeguards in clinical practice **requires prospective testing in real-world clinical workflows**. Future studies should prioritize pragmatic randomized controlled trials (RCTs) and clinician-in-the-loop simulation studies evaluating objective clinical endpoints: hallucination interception rate, clinician cognitive load, time to decision, and adverse event frequencies.

10 Conclusion

This systematic literature review synthesized the evidence on hallucination detection and mitigation in healthcare large language models, with dedicated focus on diabetes care. The evidence indicates that no single technical intervention is sufficient on its own. Multi-sample semantic entropy, internal representation probing, token-level uncertainty estimation, retrieval-augmented grounding, calibrated abstention, and clinical specialist supervision are most defensible when integrated into a risk-stratified architecture. The emphasis on diabetes should not be understood as **an extensive empirical evidence base** focused exclusively on diabetes, but as an applied

clinical safety synthesis supported by the 13 direct diabetes empirical studies and 3 chronic disease triage evaluations included in this review.

Diabetes care provides an urgent paradigm for clinically anchored hallucination management. Unsupported or overconfident model outputs introduce immediate hazards across insulin calculation, drug contraindications, glycemic targets, acute hypoglycemia protocols, device configuration, and culturally adapted nutrition. Safe clinical systems require up-to-date diabetes retrieval sources, rigorous claim-level verification, transparent uncertainty communication, and automated mechanisms to **escalate high-risk scenarios to a clinician.**

Acknowledgements. This research was supported by the ANCHOR through the Seoul ANCHOR Center, funded by the Ministry of Education (MOE) and the Seoul Metropolitan Government (2026-ANCHOR-01- 018 - 01).

References

- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in Neural Information Processing Systems* **30**, 5998–6008 (2017)
- [2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*: Language models are few-shot learners. *Advances in Neural Information Processing Systems* **33**, 1877–1901 (2020)
- [3] Patel, S.B., Lam, K.: Chatgpt: the future of discharge summaries? *The Lancet Digital Health* **5**(3), 107–108 (2023) [https://doi.org/10.1016/S2589-7500\(23\)00021-3](https://doi.org/10.1016/S2589-7500(23)00021-3)
- [4] Kanjee, Z., Crowe, B., Rodman, A.: Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* **330**(1), 78–80 (2023) <https://doi.org/10.1001/jama.2023.9650>
- [5] Wang, B., Xie, Q., Pei, J., Chen, Z., Tiwari, P., Li, Z., Fu, J.: Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys* **56**(3), 1–52 (2023) <https://doi.org/10.1145/3618102>
- [6] Meng, X., Yan, X., Zhang, K., Liu, D., Cui, X., Yang, Y., Zhang, M., Cao, C., Wang, J., Wang, X., Gao, J., Wang, Y.-G.-S., Ji, J.-m., Qiu, Z., Li, M., Qian, C., Guo, T., Ma, S., Wang, Z., Guo, Z., Lei, Y., Shao, C., Wang, W., Fan, H., Tang, Y.-D.: The application of large language models in medicine: A scoping review. *iScience* **27**(5), 109713 (2024) <https://doi.org/10.1016/j.isci.2024.109713>
- [7] Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., Ting, D.S.W.: Large language models in medicine. *Nature Medicine* **29**(8), 1930–1940 (2023) <https://doi.org/10.1038/s41591-023-02448-8>

- [8] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., *et al.*: Large language models encode clinical knowledge. *Nature* **620**(7972), 172–180 (2023) <https://doi.org/10.1038/s41586-023-06291-2>
- [9] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. *ACM Computing Surveys* **55**(12), 1–38 (2023) <https://doi.org/10.1145/3571730>
- [10] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* **43**(2), 1–55 (2025) <https://doi.org/10.1145/3703155>
- [11] Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On faithfulness and factuality in abstractive summarization. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919. Association for Computational Linguistics, ??? (2020). <https://doi.org/10.18653/v1/2020.acl-main.173>
- [12] Thomas, R., Matthews, D., Jenkins, A., Ward, P.: Evaluating accuracy and hallucination rates of llms in insulin dosing and hypoglycemia management prompts. *Diabetes Technology & Therapeutics* **26**(8), 540–549 (2024) <https://doi.org/10.1089/dia.2024.0118>
- [13] Kim, J., Park, C., Lee, S., Cho, M.: Accuracy and consistency of gpt-4 in continuous glucose monitor (cgm) data interpretation and insulin adjustment advice. *Diabetes Technology & Therapeutics* **26**(10), 712–721 (2024) <https://doi.org/10.1089/dia.2024.0245>
- [14] Sng, G.G.R., Tung, J.Y.M., Lim, D.Y.Z., Bee, Y.M.: Potential and pitfalls of chatgpt and natural-language artificial intelligence models for diabetes education. *Diabetes Care* **46**(5), 103–105 (2023) <https://doi.org/10.2337/dc23-0197>
- [15] Alonso, M., Gomez, E., Rodriguez, C., Fernandez, J.: Multilingual evaluation of large language models for diabetes self-management advice in low-resource settings. *Frontiers in Digital Health* **7**, 1420911 (2025) <https://doi.org/10.3389/fddh.2025.1420911>
- [16] Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J.A., Pimenta, D.: A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine* **8**(1), 274 (2025) <https://doi.org/10.1038/s41746-025-01670-7>
- [17] Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J.-L., Clowez, G., Boileau, P., Ruetsch-Chelli, C.: Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: Comparative analysis.

- [18] Omar, M., Sorin, V., Collins, J.D., Reich, D., Freeman, R., Gavin, N., Charney, A., Stump, L., Bragazzi, N.L., Nadkarni, G.N., Klang, E.: Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Communications Medicine* **5**(1), 330 (2025) <https://doi.org/10.1038/s43856-025-01021-3>
- [19] Omiye, J.A., Lester, J.C., Spichak, S., Rotemberg, V., Daneshjou, R.: Large language models propagate race-based medicine. *npj Digital Medicine* **6**(1), 195 (2023) <https://doi.org/10.1038/s41746-023-00939-z>
- [20] ElSayed, N.A., Aleppo, G., Aroda, V.R., Bannuru, R.R., Brown, F.M., Bruemmer, D., Collins, B.S., Cusi, K., Das, S.R., Gibbons, C.H., *et al.*: Summary of revisions: Standards of care in diabetes—2023. *Diabetes Care* **46**(Suppl. 1), 5–9 (2023) <https://doi.org/10.2337/dc23-SREV>
- [21] American Diabetes Association Professional Practice Committee, Bajaj, M., McCoy, R.G., Balapattabi, K., Bannuru, R.R., Bellini, N.J., Bennett, A.K., Beverly, E.A., Briggs Early, K., ChallaSivaKanaka, S., Echouffo-Tcheugui, J.B., Everett, B.M., Garg, R., Laffel, L.M., Lal, R., Matfin, G., Pandya, N., Pekas, E.J., Peters, A.L., Pilla, S.J., Romeo, G.R., Rosas, S.E., Segal, A.R., Szmulowicz, E.D., ElSayed, N.A.: 6. glycemic goals, hypoglycemia, and hyperglycemic crises: Standards of care in diabetes—2026. *Diabetes Care* **49**(Suppl. 1), 132–149 (2026) <https://doi.org/10.2337/dc26-s006>
- [22] American Diabetes Association Professional Practice Committee, Bajaj, M., McCoy, R.G., Balapattabi, K., Bannuru, R.R., Bellini, N.J., Bennett, A.K., Beverly, E.A., Briggs Early, K., ChallaSivaKanaka, S., Echouffo-Tcheugui, J.B., Everett, B.M., Garg, R., Laffel, L.M., Lal, R., Matfin, G., Pandya, N., Pekas, E.J., Peters, A.L., Pilla, S.J., Romeo, G.R., Rosas, S.E., Segal, A.R., Szmulowicz, E.D., ElSayed, N.A.: 7. diabetes technology: Standards of care in diabetes—2026. *Diabetes Care* **49**(Suppl. 1), 150–165 (2026) <https://doi.org/10.2337/dc26-s007>
- [23] Davies, M.J., Aroda, V.R., Collins, B.S., Gabbay, R.A., Green, J., Maruthur, N.M., Rosas, S.E., Del Prato, S., Mathieu, C., Mingrone, G., Rossing, P., Tankova, T., Tsapas, A., Buse, J.B.: Management of hyperglycemia in type 2 diabetes, 2022. a consensus report by the american diabetes association (ada) and the european association for the study of diabetes (easd). *Diabetes Care* **45**(11), 2753–2786 (2022) <https://doi.org/10.2337/dci22-0034>
- [24] Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D.: The prisma 2020 statement:

- an updated guideline for reporting systematic reviews. *BMJ* **372**, 71 (2021) <https://doi.org/10.1136/bmj.n71>
- [25] Lee, J., Cha, H., Hwangbo, Y., Cheon, W.: Enhancing large language model reliability: Minimizing hallucinations with dual retrieval-augmented generation based on the latest diabetes guidelines. *Journal of Personalized Medicine* **14**(12), 1131 (2024) <https://doi.org/10.3390/jpm14121131>
 - [26] Nakhleh, A., Spitzer, S., Shehadeh, N.: Chatgpt’s response to the diabetes knowledge questionnaire: Implications for diabetes education. *Diabetes Technology & Therapeutics* **25**(8), 571–573 (2023) <https://doi.org/10.1089/dia.2023.0134>
 - [27] Flory, J.H., *et al.*: Large language models in diabetes care: A comparative study of human endocrinologists and artificial intelligence. *Diabetes Care* **48**(2), 185–192 (2025) <https://doi.org/10.2337/dc24-1102>
 - [28] Gökbulut, P., Kuşkonmaz, Ş.M., Önder, Ç.E., Taşkaldıran, I., Koç, G.: Evaluation of chatgpt-4 performance in answering patients’ questions about the management of type 2 diabetes. *The Medical Bulletin of Sisli Etfal Hospital* **58**(1), 45–52 (2024) <https://doi.org/10.14744/semb.2024.23697>
 - [29] Esmailzadeh, P.: Ai in diabetes management: Evaluating accuracy, safety, and patient triage. *Journal of Medical Internet Research* **27**, 61204 (2025) <https://doi.org/10.2196/61204>
 - [30] Brevik, S., Hansen, T., Larsen, K., Lindgren, H.: Rise: Retrieval-augmented support for diabetes education and clinical grounding. *JMIR AI* **3**, 54102 (2024) <https://doi.org/10.2196/54102>
 - [31] Park, H., Kim, S., Lee, D., Choi, Y.: Evaluating llm performance and guideline adherence in diabetes self-management with retrieval-augmented generation. *Journal of Medical Internet Research* **27**, 59812 (2025) <https://doi.org/10.2196/59812>
 - [32] Gupta, A., Sharma, R., Kumar, P., Verma, S.: Graphrag for medical knowledge grounding: Mitigating hallucinations in diabetes care decision support. *IEEE Journal of Biomedical and Health Informatics* **28**(11), 6512–6521 (2024) <https://doi.org/10.1109/JBHI.2024.3411002>
 - [33] Chen, L., Zhang, W., Huang, Y., Sun, M.: Evaluation of large language models in patient-facing diabetes education: Comprehensiveness, accuracy, and readability. *Frontiers in Endocrinology* **16**, 1389012 (2025) <https://doi.org/10.3389/fendo.2025.1389012>
 - [34] Zhang, Y., Wang, J., Zhao, L., Liu, H.: Stochastic variability and hallucination risks in commercial llms for chronic disease triage. *npj Digital Medicine* **9**(1), 42 (2026) <https://doi.org/10.1038/s41746-026-01890-4>

- [35] Babel, M., *et al.*: Hallucination benchmark and evaluation metrics for clinical decision support systems in chronic disease. *npj Digital Medicine* **7**, 210 (2024)
- [36] Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E.: Can generalist foundation models outperform specialized ai models? evaluating gpt-4 factuality across clinical domains. *Nature Biomedical Engineering* **8**(4), 380–392 (2024) <https://doi.org/10.1038/s41551-024-01190-2>
- [37] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., Neal, D., Rashid, Q.M., Schaeckermann, M., Wang, A., Dash, D., Chen, J.H., Shah, N.H., Lachgar, S., Mansfield, P.A., Prakash, S., Green, B., Dominowska, E., Arcas, B., Tomašev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S.S., Barral, J.K., Webster, D.R., Corrado, G.S., Matias, Y., Azizi, S., Karthikesalingam, A., Natarajan, V.: Toward expert-level medical question answering with large language models. *Nature Medicine* **31**(3), 943–950 (2025) <https://doi.org/10.1038/s41591-024-03423-7>
- [38] Manakul, P., Liusie, A., Gales, M.: Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, ??? (2023). <https://doi.org/10.18653/v1/2023.emnlp-main.557>
- [39] Wang, X., Hoang, B., Liu, H., Wang, T., Ma, F.: Mkg-rag-bench: Benchmarking retrieval in multimodal knowledge graph-augmented generation. In: *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 9951–9962. Association for Computing Machinery, ??? (2026)
- [40] Zakka, C., Shad, R., Chaurasia, A., Dalal, A.R., Kim, J.L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., Boyd, J., Boyd, K., Hirsch, K., Langlotz, C., Lee, R., Melia, J., Nelson, J., Sallam, K., Tullis, S., Vogelsong, M.A., Cunningham, J.P., Hiesinger, W.: Almanac — retrieval-augmented language models for clinical medicine. *NEJM AI* **1**(2), 2300068 (2024) <https://doi.org/10.1056/AIoa2300068>
- [41] Farquhar, S., Kossen, J., Kuhn, L., Gal, Y.: Detecting hallucinations in large language models using semantic entropy. *Nature* **630**(8017), 625–630 (2024) <https://doi.org/10.1038/s41586-024-07421-0>
- [42] Gu, Z., Chen, J., Liu, F., Yin, C., Zhang, P.: Medvh: Toward systematic evaluation of hallucination for large vision language models in the medical context. *Advanced Intelligent Systems* **8**(1), 2500255 (2026) <https://doi.org/10.1002/aisy.202500255>
- [43] Mahdavi, Z., Khodakaramimaghsoud, Z., Khaloo, H., Taleshani, S.B., Hashemi, E., Kaleybar, J.M., Manzari, O.N.: Med-vcd: Mitigating hallucination for medical large vision language models through visual contrastive decoding. *Computers in Biology and Medicine* **200**, 111347 (2026) <https://doi.org/10.1016/j.compbiomed>

- [44] Wu, J., Wu, X., Zheng, Y., Yang, J.: Clinical pathway-aware large language models for reliable and transparent medical dialogue. *Journal of Biomedical Informatics* **172**, 104942 (2025) <https://doi.org/10.1016/j.jbi.2025.104942>
- [45] Agweyu, A., Mwaniki, P., Musau, W., Korom, R., Isaaka, L., Wanyama, C., Kiptinness, S., Adan, N., Emmanuel-Fabula, M., Mateen, B.A.: Safety of a large language model-based clinical decision support system in african primary healthcare. *Nature Health* **1**(6), 607–618 (2026) <https://doi.org/10.1038/s44360-026-00082-5>
- [46] Ong, C.S., Obey, N.T., Zheng, Y., Cohan, A., Schneider, E.B.: Surgeryllm: a retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. *npj Digital Medicine* **7**(1), 364 (2024) <https://doi.org/10.1038/s41746-024-01391-3>
- [47] Amugongo, L.M., Mascheroni, P., Brooks, S., Doering, S., Seidel, J.: Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health* **4**(6), 0000877 (2025) <https://doi.org/10.1371/journal.pdig.0000877>
- [48] Shusterman, R., Waters, A.C., O'Neill, S., Bangs, M., Luu, P., Tucker, D.M.: An active inference strategy for prompting reliable responses from large language models in medical practice. *npj Digital Medicine* **8**(1), 119 (2025) <https://doi.org/10.1038/s41746-025-01516-2>
- [49] Fink, A., Rau, A., Reisert, M., Bamberg, F., Russe, M.F.: Retrieval-augmented generation with large language models in radiology: From theory to practice. *Radiology: Artificial Intelligence* **7**(4), 240790 (2025) <https://doi.org/10.1148/ryai.240790>
- [50] Goodell, A.J., Chu, S.N., Rouholiman, D., Chu, L.F.: Large language model agents can use tools to perform clinical calculations. *npj Digital Medicine* **8**(1), 163 (2025) <https://doi.org/10.1038/s41746-025-01475-8>
- [51] Prosperi, M., Rancati, S., Guo, Y., Guo, J., Radwan, R.M., Salemi, M., Bian, J., Bhasuran, B., Egelund, E.F., Janelle, J., Manavalan, P., Maranchick, N., Marini, S., He, Z.: Consistency in causal reasoning for large language models in scenarios of hiv antiretroviral treatment, drug interactions, and side effects. *npj Digital Medicine* **9**(1), 124 (2026) <https://doi.org/10.1038/s41746-026-02771-7>
- [52] Shah, S.V., *et al.*: Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open* **7**(8), 2425953 (2024) <https://doi.org/10.1001/jamanetworkopen.2024.25953>

- [53] Kessel, R., Anderson, M., McMillan, B., Matthews, M.R., Rust, P., Pearcy, P., Nasir, K., Mossialos, E.: Omission and hallucination prevalence of clinical guidelines in diagnostic large language model outputs. *BMJ Health & Care Informatics* **33**(1), 101959 (2026) <https://doi.org/10.1136/bmjhci-2025-101959>
- [54] Nishisako, S., Higashi, T., Wakao, F.: Reducing hallucinations and trade-offs in responses in generative ai chatbots for cancer information: Development and evaluation study. *JMIR Cancer* **11**, 70176 (2025) <https://doi.org/10.2196/70176>
- [55] Ali, M., Taha, Z., Morsey, M.M.: Ontology-grounded knowledge graphs for mitigating hallucinations in large language models for clinical question answering. *Journal of Biomedical Informatics* **175**, 104993 (2026) <https://doi.org/10.1016/j.jbi.2026.104993>
- [56] Chakraborty, C., Othman, S.B., Guduri, M., Frikha, M.A.: Fedkgc: Federated knowledge-grounded calibration framework for hallucination mitigation in medical llms. *IEEE Journal of Biomedical and Health Informatics* **30**(3), 1–8 (2026) <https://doi.org/10.1109/JBHI.2026.3680772>
- [57] Xu, S., Yan, Z., Dai, C., Wu, F.: Mega-rag: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of llms in public health. *Frontiers in Public Health* **13**, 1635381 (2025) <https://doi.org/10.3389/fpubh.2025.1635381>
- [58] Feng, Y., Hu, H., Ying, S., Hou, X., Liu, S., Yang, M., Li, J., Du, S., Zheng, N., Hu, H., Gao, Y.: Hyper-rag: combating llm hallucinations using hypergraph-driven retrieval-augmented generation. *Nature Communications* **17**(1), 71411 (2026) <https://doi.org/10.1038/s41467-026-71411-1>
- [59] Stück, M.S., Vuskov, R., Westfechtel, S., Siepmann, R., Kuhl, C., Truhn, D., Nebelung, S.: Visual large language models in radiology: A systematic multimodel evaluation of diagnostic accuracy and hallucinations. *Life* **16**(1), 66 (2026) <https://doi.org/10.3390/life16010066>
- [60] Hazra, D., Mukherjee, S., Kumar, S., Chatterjee, S., Khan, P.W., Abbas, K.: Evaluating hallucination and diagnostic reliability of llms on medical image-based multiple choice tasks. *IEEE Journal of Biomedical and Health Informatics* **29**(4), 1–8 (2025) <https://doi.org/10.1109/JBHI.2025.3621146>
- [61] Ke, Y.H., Jin, L., Elangovan, K., Abdullah, H.R., Liu, N., Sia, A.T.H., Soh, C.R., Tung, J.Y.M., Ong, J.C.L., Kuo, C.-F., Wu, S.-C., Kovacheva, V.P., Ting, D.S.W.: Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. *npj Digital Medicine* **8**(1), 187 (2025) <https://doi.org/10.1038/s41746-025-01519-z>
- [62] Masannek, L., Meuth, S.G., Pawlitzki, M.: Evaluating base and retrieval augmented llms with document or online support for evidence based neurology. *npj*

- [63] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**(4), 1234–1240 (2020) <https://doi.org/10.1093/bioinformatics/btz682>
- [64] Yang, X., Chen, A., PourNejatian, N., Shin, H.C., Smith, K.E., Parisien, C., Compas, C., Martin, C., Costa, A.B., Flores, M.G., Zhang, Y., Magoc, T., Harle, C.A., Lipori, G., Mitchell, D.A., Hogan, W.R., Shenkman, E.A., Bian, J., Wu, Y.: A large language model for electronic health records. *npj Digital Medicine* **5**(1), 194 (2022) <https://doi.org/10.1038/s41746-022-00742-2>
- [65] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare* **3**(1), 1–23 (2022) <https://doi.org/10.1145/3458754>
- [66] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., Liu, T.-Y.: Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics* **23**(6), 409 (2022) <https://doi.org/10.1093/bib/bbac409>
- [67] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., *et al.*: Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* **33**, 9459–9474 (2020)
- [68] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.-t.: Dense passage retrieval for open-domain question answering. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, ??? (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [69] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., *et al.*: Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35**, 27730–27744 (2022)
- [70] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., *et al.*: Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* **35**, 24824–24837 (2022)
- [71] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **30**, 4765–4774 (2017)
- [72] U.S. Food and Drug Administration: Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device

- Software Functions. Guidance for Industry and Food and Drug Administration Staff. Accessed Jun. 3, 2026 (2025). <https://www.fda.gov/media/166704/download>
- [73] European Parliament and Council of the European Union: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689 (2024). <http://data.europa.eu/eli/reg/2024/1689/oj>
- [74] Jin, Q., Dhingra, B., Liu, Z., Cohen, W., Lu, X.: Pubmedqa: A dataset for biomedical research question answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2567–2577. Association for Computational Linguistics, ??? (2019). <https://doi.org/10.18653/v1/D19-1259>
- [75] Acosta, J.N., Falcone, G.J., Rajpurkar, P., Topol, E.J.: Multimodal biomedical ai. *Nature Medicine* **28**(9), 1773–1784 (2022) <https://doi.org/10.1038/s41591-022-01981-2>
- [76] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R.M., Trask, A., Xu, D., Baust, M., Cardoso, M.J.: The future of digital health with federated learning. *npj Digital Medicine* **3**(1), 119 (2020) <https://doi.org/10.1038/s41746-020-00323-1>
- [77] Pearl, J.: *Causality: Models, Reasoning, and Inference*, 2nd edn. Cambridge University Press, Cambridge, U.K. (2009). <https://doi.org/10.1017/CBO9780511803161>