Review Article

Can Large Language Models Be Trusted with Pharmacy Work? A Systematic Review of Accuracy, Safety, and Clinical Use

Daniel Fischer[1]*, Laura Meier[1], Stefan Koch[2], Thomas Braun[3]

[1]Department of LLM Trust and Pharmacy Work, Faculty of Pharmacy, University of Freiburg, Freiburg, Germany. [2]Department of AI Accuracy and Clinical Safety, Faculty of Pharmacy, Technical University of Munich, Munich, Germany. [3]Department of LLM Evaluation in Pharmacy Practice, Faculty of Pharmacy, University of Kiel, Kiel, Germany.

Abstract

Large language models (LLMs) can generate medication information, counselling content, interaction assessments, and pharmacotherapy suggestions, but answer quality alone does not establish safe or useful pharmacy practice. To determine how accurately and safely LLM-based systems perform pharmacy-relevant work, what clinical-use evidence exists, and which task-, model-, comparator-, and setting-specific boundaries constrain claims of trustworthiness. A protocol-driven systematic review was conducted in accordance with PRISMA 2020. PubMed/MEDLINE, publisher and Crossref-indexed records, and backward and forward citation searches were examined through 3 August 2026. Eligible studies were peer-reviewed Q1 journal articles evaluating an LLM or LLM-enabled system on a medication-work task. Two reviewers independently screened records, extracted study and model characteristics, harmonized outcomes, and appraised risk of bias, reproducibility, and applicability. Findings were synthesized narratively by pharmacy task and evaluation design because outcomes were unsuitable for meta-analysis. Forty-six records were identified, four duplicates were removed, 42 records were screened, 32 full-text reports were assessed, and 20 studies were included. Evidence was dominated by retrospective, cross-sectional, simulated-case, and output-rating designs. Performance varied with task complexity, model and version, prompting, grounding, reference standard, assessor, and scoring definition. Retrieval augmentation and customization improved selected output measures, but hallucination, omission, inconsistency, weak referencing, and unsafe or insufficiently qualified recommendations remained. Real-world questions or patient-derived data rarely represented prospective workflow implementation, and no included study established improved patient outcomes or safe autonomous pharmacy deployment. LLM trustworthiness in pharmacy is conditional rather than general. Current evidence supports bounded, task-specific evaluation under professional verification, not unsupervised clinical delegation. Prospective, externally validated, harm-sensitive studies are required.

Keywords: Systematic review, Digital pharmacy, Artificial intelligence, Pharmacy practice, Medication safety, Evidence synthesis

Introduction

Large language models are increasingly evaluated for medication information, prescription review, drug-related-problem identification, interaction checking, counselling, and pharmacotherapy support. The medication-harm literature nevertheless remains heterogeneous and concentrated in drug-interaction, decision-support, and pharmacovigilance tasks; a recent scoping review found no prospective studies within its search window through October 2024 [1].

This imbalance matters because a fluent response may be correct on a narrow question while remaining incomplete, poorly referenced, insensitive to patient context, or unsafe when inserted into a pharmacy workflow.

Accuracy-only summaries can therefore misstate what has been demonstrated. Across broader patient-safety and adverse-drug-event literature, technical performance does not by itself establish safer care, and medication-related AI must be evaluated against defined harm domains and use cases [2-4]. A model may match a reference answer yet fail to recognize uncertainty, decline an unsafe request, identify a contraindication, explain its evidential basis, or support appropriate escalation to a pharmacist. An error involving a low-consequence factual item also does not carry the same

Address for correspondence: Daniel Fischer, Department of LLM Trust and Pharmacy Work, Faculty of Pharmacy, University of Freiburg, Freiburg, Germany.
daniel.fischer@pharmazie.uni-freiburg.de
Received: 21 March 2026; **Accepted:** 16 July 2026

How to cite this article: Fischer D, Meier L, Koch S, Braun T. Can Large Language Models Be Trusted with Pharmacy Work? A Systematic Review of Accuracy, Safety, and Clinical Use. Arch Pharm Pract. 2026;17(3):10-8. https://doi.org/10.51847/WHAwvGP2Pm

implications as an incomplete recommendation involving polypharmacy, pregnancy, organ impairment, or treatment modification.

This review asked: among pharmacy-relevant medication tasks performed for patients, pharmacists, or other health professionals, how accurately and safely do LLM-based systems perform; what level of clinical-use evidence exists; and which task-, model-, comparator-, and setting-specific boundaries constrain claims of trustworthiness? Medicines information, counselling, prescription and drug-related-problem review, interaction detection, deprescribing, medication instructions, uncertainty and refusal, pharmacist comparison, and prospective or real-world workflow evidence were examined. Drug discovery, molecular design, education-only examinations, perception surveys without performance outcomes, non-generative prediction models, and tasks without a medication-work endpoint were outside scope. Trustworthiness was treated as a layered judgment involving output validity, safety behaviour, reproducibility, human mediation, workflow applicability, and patient consequence rather than as a single accuracy score.

Review Design, Questions, and Protocol

Protocol and Registration

An a priori protocol specified the review questions, eligibility criteria, search concepts, screening units, extraction fields, appraisal domains, synthesis groups, and sensitivity analyses. Identification, screening, eligibility, inclusion, and synthesis were reported in accordance with PRISMA 2020 [5]. The protocol was not prospectively registered in a public registry; this is reported as a transparency limitation and not represented as registration.

Review Questions

The primary question addressed performance, safety, and clinical-use maturity for LLM-enabled pharmacy work. Secondary questions examined variation by task, setting, model and version, prompt, grounding strategy, comparator, reference standard, outcome definition, assessor, and validation level. Search sources, dates, strategies, deduplication rules, supplementary searching, and selection decisions were documented using PRISMA-S principles [6]. Records, reports, and studies were treated as distinct units, although each included study was represented by one report.

Eligibility Criteria

Eligible reports were peer-reviewed journal articles published from 2017 through 3 August 2026, had a verifiable DOI, appeared in a journal verified as Q1 in a relevant category using the latest completed ranking available before the search freeze, and provided extractable empirical evidence for a pharmacy-relevant medication task. Eligible systems were general-purpose, customized, retrieval-augmented, or otherwise LLM-enabled generative systems. Conventional machine-learning prediction systems without generative language functionality were excluded.

Pharmacy-Task Taxonomy and Model Eligibility

The prespecified taxonomy comprised medicines-information answering; prescription review; medication-therapy management and drug-related-problem identification; adverse-drug-reaction recognition or causality support; drug–drug and drug–herb interaction assessment; deprescribing; patient counselling; personalized medication instructions; and refusal, uncertainty, clarification, or escalation behaviour. Questions could be simulated, curated, literature-derived, collected from practice, or generated from patient data, but these input sources were not treated as equivalent evidence levels.

Prediction-oriented evaluations were assessed for participants, inputs, outcomes, analysis, applicability, and AI-specific risks using relevant PROBAST+AI domains [7]. Studies approaching clinical decision support or workflow use were additionally examined for intended users, workflow integration, human factors, safety reporting, and prospective evaluation features reflected in DECIDE-AI [8]. These frameworks organized appraisal; they did not validate an LLM or convert output-level testing into clinical-effectiveness evidence.

Eligibility, Definitions, and Information Sources

Operational Definitions

General medical knowledge performance was distinguished from pharmacy-task performance because success on broad clinical benchmarks cannot be treated as direct evidence of medication-work safety [9]. LLM output was defined as generated natural-language content produced in response to a pharmacy-relevant prompt, with or without retrieval, customization, or structured prompting. Fluency was not considered evidence of correctness because plausible medical output can coexist with factual error, unsupported reasoning, and uncertain clinical reliability [10].

Accuracy comprised agreement with a prespecified reference standard or expert judgment. Completeness, relevance, consistency, citation quality, harmfulness, and uncertainty behaviour were extracted separately. Clinical-use evidence required more than answer scoring, such as intended-user interaction, prospective workflow testing, implementation measures, treatment decisions, or patient outcomes. Human-rated usefulness of generated decision-support suggestions was treated as an intermediate output rather than evidence of acceptance, deployment, or patient benefit [11]. Conventional machine learning, generative AI, conversational systems, and LLM-based decision support were kept analytically distinct [12].

Information Sources and Search

PubMed/MEDLINE was searched through 3 August 2026. Publisher and Crossref-indexed records were searched to verify titles, journal records, and DOIs and to identify eligible reports not retrieved through the principal bibliographic search. Search concepts combined “large language model,”

LLM, ChatGPT, GPT-4, and generative AI with pharmacy, pharmacist, medication, drug information, pharmacotherapy, prescription, counselling, drug interaction, medication safety, accuracy, hallucination, uncertainty, refusal, clinical use, workflow, and validation. Limits were English language, journal article, and publication from 2017 through the search date.

Supplementary searching comprised backward and forward citation checking from the principal medication-harm review, key pharmacy evaluations, and methodological guidance. DOI-first deduplication was followed by normalized title, author, and year comparison.

Dual Screening and Data Extraction

Two reviewers independently screened titles and abstracts and then assessed full texts. Disagreements were resolved through discussion and consensus, and full-text exclusion reasons were assigned from a prespecified hierarchy.

Using a structured form, two reviewers extracted study design, setting, users or population, question or data source, sample size, model and version, access date where reported, prompting and repetition strategy, retrieval or customization, comparator, reference standard, assessor expertise and blinding, outcome definitions, numerical results, safety findings, reproducibility information, prospective workflow features, patient outcomes, subgroup or equity reporting, and author-stated limitations. Extracted information was reconciled against the source report.

Outcome Harmonization

Results were retained in their original denominators, scales, and measurement forms. No common accuracy construct was imposed across binary correctness, rubric scores, Likert ratings, agreement measures, completeness assessments, or preference outcomes. Safety findings were classified by potential consequence, including harmful recommendations, missed contraindications or interactions, unsupported treatment changes, omissions, and failures to qualify or escalate. Refusal was distinguished from evasion, omission, generic disclaimers, requests for clarification, and calibrated expressions of uncertainty.

Table 1 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for eligibility, definitions, and evidence-extraction framework for the review Can Large Language Models Be Trusted with Pharmacy Work? A Systematic Review of Accuracy, Safety, and Clinical Use.

Table 1. Eligibility, definitions, and evidence-extraction framework for the review Can Large Language Models Be Trusted with Pharmacy Work? A Systematic Review of Accuracy, Safety, and Clinical Use.

Review element	Operational definition	Eligibility or decision rule	Data field	Rationale	Bias or ambiguity risk	Planned synthesis use
LLM-enabled system	Generative system producing medication-related language output	Include general-purpose, customized, or retrieval-augmented LLMs; exclude non-generative prediction tools	Model, version, access date, grounding	Separates LLM evidence from broader AI categories [12].	Version drift and opaque updates	Stratification by model configuration and reporting
Pharmacy work	Medication task performed, supported, or verified by pharmacy professionals	Require a direct medication-work endpoint	Task, intended user, setting	General benchmarks are not direct pharmacy evidence [9].	Overlap with medicine and education	Synthesis by task taxonomy
Input source	Real, retrospective, curated, or simulated question, case, or patient data	Include but label the source explicitly	Population, case source, complexity	Input provenance affects applicability	"Real-world" may still describe offline testing	Separate source realism from implementation
Comparator or standard	Pharmacist, physician, guideline, database, product information, expert panel, or rubric	Require an identifiable standard or assessor process	Standard, expertise, adjudication, blinding	Correctness depends on comparator quality	Expert agreement may conceal shared error	Interpret findings within comparator strength
Accuracy	Agreement with the study's stated reference standard	Preserve the original metric and denominator	Correctness, score, agreement	Broad knowledge success is insufficient [9].	Binary labels can conceal omissions and severity	Report within compatible metric families
Completeness and relevance	Coverage of necessary content and responsiveness to the question	Extract separately from accuracy	Omissions, directness, relevance, references	Fluent output may remain incomplete or unsupported [10].	Rubrics and thresholds differ	Identify recurrent omission patterns
Safety	Harmful advice, missed contraindication or interaction, unsupported change, or failed escalation	Require described adjudication	Error type, severity, harmfulness, escalation	Harm must be explicitly defined	Insensitive testing may fail to detect harm	Synthesize by potential consequence

Refusal and uncertainty	Decline, qualification, clarification request, or escalation	Distinguish from omission, evasion, and generic caveats	Refusal, confidence, clarification, escalation	Safe behaviour includes recognizing limits	Calibration metrics were rarely prespecified	Narrative cross-task synthesis
Clinical-use evidence	Intended-user, workflow, decision, or patient-outcome evaluation	Output ratings alone do not qualify	Prospective design, workflow measure, decision or patient outcome	Usefulness is an intermediate outcome [11].	Simulated realism may imply effectiveness	Evidence ladder from output testing to outcomes
Applicability and implementation	Transferability to intended users, setting, workload, oversight, and monitoring	Judge against the proposed intended use	Workflow, oversight, monitoring, external validation	Early evaluation requires workflow and human-factor reporting [8].	Technical studies may imply readiness	State explicit practice non-inferences
Equity	Differential performance by language, literacy, demographic, or clinical subgroup	Extract when tested; otherwise record as not assessed	Subgroups and differential outcomes	Missing subgroup evidence limits generalizability	Aggregate scores may hide variation	Map evidence gaps without imputing fairness

Eligibility and extraction rules determine what counts as evidence; they do not establish that any system is trustworthy.

Search, Selection, Extraction, Appraisal, and Synthesis

Search and Selection Procedures

Search outputs were entered into a record-level log containing a unique identifier, source, search date, title, DOI, duplicate status, title-and-abstract decision, retrieval status, full-text decision, exclusion reason, and final synthesis status. Deduplication used DOI matching first and normalized bibliographic comparison second. Independent screening decisions were reconciled before progression. Full texts were sought for every record retained after title-and-abstract screening.

Extraction and Synthesis

Extraction was independently completed and reconciled against each report. When outcomes were incompletely reported, the available numerator, denominator, scale, or qualitative judgment was retained without deriving unreported performance values. Because tasks, models, prompts, reference standards, assessors, and outcomes were noncommensurable, meta-analysis was not undertaken. Narrative synthesis prespecified task groups, evaluation-design groups, direction and consistency of findings, limitations, and evidence displays in accordance with SWiM [13].

Model-development and validation reports were expected to disclose data provenance, participants or cases, outcomes, performance measures, and evaluation design sufficiently for appraisal, using relevant TRIPOD+AI principles where applicable [14]. These criteria were applied modularly: conversational answer-rating studies were not judged as prediction models, but absence of the model version, prompt, access date, reference standard, or external evaluation remained a reproducibility or applicability concern.

Risk-of-Bias and Applicability Appraisal

A review-specific appraisal combined applicable PROBAST+AI domains with DECIDE-AI and TRIPOD+AI concepts. Domains covered selection and input-data risk; reference-standard quality and independence; assessor expertise and blinding; prompt and model reproducibility; outcome validity; repeated-query consistency; external validation; prospective workflow testing; patient outcomes; and subgroup or equity assessment.

AI-versus-clinician comparisons were scrutinized for design symmetry, comparator quality, reporting completeness, external validation, and risk of bias before equivalence or superiority claims were accepted [15]. Overall judgments were structured reviewer assessments rather than regulatory or clinical determinations.

Reproducibility appraisal recorded unavailable prompts, unreported model snapshots or access dates, changing proprietary systems, inaccessible data, incomplete scoring instructions, and absent analytic code. These omissions can prevent independent reconstruction even when headline performance values are available [16]. A study could consequently have low concern for answer scoring within its sample but high concern for reproducibility or applicability.

Planned Sensitivity Analyses

Sensitivity analyses were qualitative because no pooled estimate was calculated. Findings were re-examined by general-purpose versus customized or retrieval-augmented systems; simulated versus practice-derived questions or patient data; factual versus patient-specific reasoning tasks; pharmacist or validated-resource comparators versus less independent rubrics; blinded versus unblinded assessment; single versus repeated prompts; and adequate versus incomplete reporting of model version and prompt. Conclusions were also checked after excluding studies whose principal endpoint was preference or perceived usefulness rather than factual or safety performance.

Protocol Deviations

No substantive eligibility, outcome-domain, or appraisal deviations occurred after screening began. Quantitative synthesis was conditional on sufficiently comparable tasks, outcomes, and denominators; that condition was not met, so the prespecified narrative pathway was used. The scope did not expand to education-only benchmarks, drug discovery, non-generative AI, or perception studies. Narrative synthesis was selected because heterogeneity precluded defensible pooling, not because narrative synthesis itself established clinical validity.

Search and Selection Results

Study-Selection Flow

The searches identified 46 records: PubMed/MEDLINE, 29; Crossref or publisher searches, 10; and citation chasing, seven. Four duplicates were removed, leaving 42 records screened. Ten were excluded, and all 32 reports sought were retrieved and assessed. Twelve full texts were excluded: secondary review rather than primary study (n=4), education or examination benchmark outside pharmacy-work outcomes (n=3), not verifiably Q1 in a relevant category (n=1), conference or meeting abstract only (n=2), adjacent specialty outside operational pharmacy-work scope (n=1), and indexed after the prespecified search freeze (n=1). Twenty reports representing 20 studies entered narrative synthesis; none entered quantitative synthesis.

Figure 1 presents the completed review-selection pathway in a black-and-white PRISMA 2020 flow diagram for study selection, documenting identification, deduplication, screening, eligibility, exclusions with reasons, and final inclusion using the approved record-accounting values.

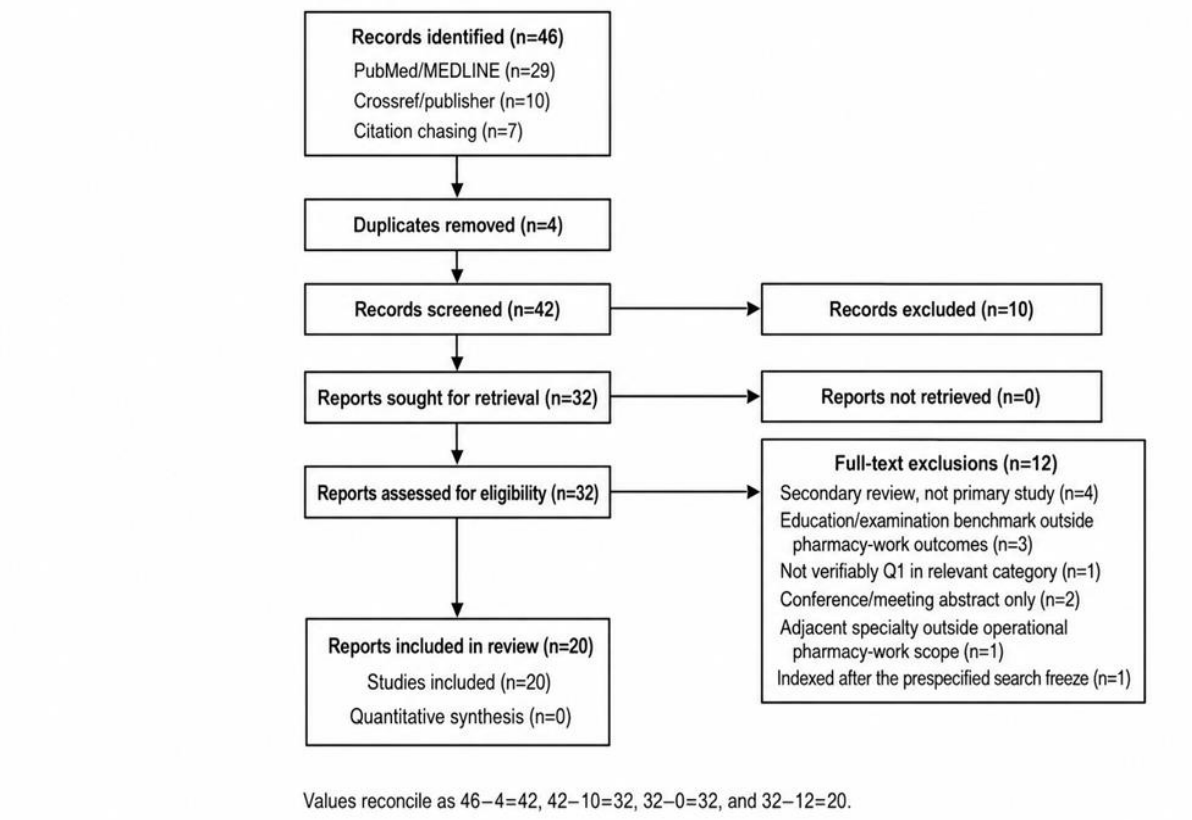

Figure 1. PRISMA 2020 Flow Diagram for Study Selection

Study, Model, and Task Characteristics

The evidence ranged from multi-domain clinical-pharmacy tasks and medication-therapy-management cases to real drug-information questions, with heterogeneous comparators and scores [17-19]. Repeated prompting showed that accuracy and consistency were separable [20]. Every study was retrospective, controlled, simulated, or output based; none evaluated a live prospective pharmacy workflow, patient outcomes, or convincing external validation.

Evidence-Base Characteristics

Medicines-Information Accuracy

Medication-question studies combined directness, accuracy, completeness, relevance, references, and expert judgment differently [21]. Factual benchmarks did not test patient-specific reasoning [22]. Practice questions showed task-dependent performance insufficient for replacement claims [23], while cross-model rankings changed by scenario and dimension [24].

Prescription and Drug-Related-Problem Performance

Controlled prescription and drug-related-problem studies suggested task support but remained simulated and complexity-sensitive [18].

Table 2 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for characteristics and classification of the evidence included in the review Can Large Language Models Be Trusted with Pharmacy Work? A Systematic Review of Accuracy, Safety, and Clinical Use.

Table 2. Characteristics and classification of the evidence included in the review Can Large Language Models Be Trusted with Pharmacy Work? A Systematic Review of Accuracy, Safety, and Clinical Use.

Evidence category	Study or source design	Population or setting	AI system or intervention	Comparator or reference standard	Outcome or construct	Validation level	Key limitation
Multi-task and management	Controlled or simulated cases [17, 18, 20]	Clinical-pharmacy cases	General-purpose LLMs	Pharmacists or expert solutions	Accuracy, consistency, drug-related problems, recommendations	Internal output testing	Small sets; no workflow or outcomes
Drug-information service	Retrospective practice questions [19, 21]	Academic or hospital services	ChatGPT snapshots	Pharmacists, literature, reference answers	Correctness, completeness, relevance, references	Practice-derived output testing	Single services; unstable versions
Factual and pharmacology knowledge	Exploratory benchmarks [22, 23]	Factual or practice questions	General-purpose chatbots	Answer keys or clinical pharmacologists	Accuracy and usefulness	Internal benchmark	Complex reasoning untested
Cross-model clinical scenarios	Mixed-methods comparison [24]	Four structured scenarios	Eight generative systems	Expert rubric	Six quality dimensions	Cross-model internal comparison	Rapid version change; no implementation

Patient-Counselling Performance

Package-insert customization improved selected over-the-counter counselling measures, but reliability remained incomplete and patients were not tested [25]. Community-pharmacy scenarios suggested possible support while leaving diverse patients and high-consequence decisions untested [26]. For selected real inquiries, GPT-4-level answers approached pharmacist ratings but remained variable and required verification [27]. No study measured comprehension, adherence, behaviour, or harm prospectively. Omissions and unsupported or fabricated references remained recurrent [19].

Refusal and Uncertainty Behaviour

Appropriateness was higher for basic public questions than hospital-professional questions, and inappropriate outputs often lacked specificity [28]. Retrieval augmentation improved simulated medication instructions without establishing comprehension or benefit [29]. Search-integrated chatbots could remain incomplete, difficult to read, or unsafe despite favourable overall ratings [30]. Refusal, clarification, calibration, and escalation were rarely prespecified; absence of overt harm did not establish safe uncertainty behaviour.

Table 3 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for methodological quality, applicability, and evidence gaps in the review Can Large Language Models Be Trusted with Pharmacy Work? A Systematic Review of Accuracy, Safety, and Clinical Use.

Table 3. Methodological quality, applicability, and evidence gaps in the review Can Large Language Models Be Trusted with Pharmacy Work? A Systematic Review of Accuracy, Safety, and Clinical Use.

Appraisal domain	Observed evidence pattern	Methodological concern	Applicability consequence	Direction of potential bias	Evidence needed	Interpretive boundary
Sampling and standards	Curated or single-service questions; variable experts and rubrics [19, 21]	Limited representativeness and non-equivalent standards	Scores cannot be pooled	Often optimistic	Multisite sampling and independent adjudication	Agreement is not correctness
Reproducibility	Versions, prompts, and repeated outputs varied [20]	Results may not reconstruct	Snapshot findings age rapidly	Either direction	Versioned prompts, dates, repetitions, code	One snapshot cannot define a model family
Outcome validity	Accuracy dominated over completeness and harm [22]	Omissions may be hidden	High scores may coexist with unsafe answers	Optimistic	Harm-weighted multidomain outcomes	Accuracy is not trustworthiness

Grounding and uncertainty	Grounding improved selected outputs [25, 29]; calibration was rarely tested [28]	Task-specific gains and unsafe confidence	Generalization unknown	Optimistic	External validation and standardized refusal tests	Grounding is not guaranteed fidelity
Clinical applicability	No live workflow or patient outcomes	Output testing dominates	Benefit and harm remain unknown	Strongly optimistic if inferred	Prospective human-factors and outcome studies	No autonomous-use inference

Prospective and Real-World Evidence

For 70 real pharmacotherapy queries, blinded evaluators rated physician responses higher in quality and factual correctness than ChatGPT-3.5 outputs [31]. A retrieval-supported system improved drug-related-problem identification in controlled cases across 16 specialties but did not show autonomous effectiveness [32]. Interaction prediction from hospitalized-patient data remained retrospective output validation requiring established resources and clinical adjudication [33]. Real-world inputs were not equivalent to prospective implementation, and no study established improved patient outcomes.

Risk-of-Bias Findings

Across studies, selection, reference-standard, reproducibility, outcome-validity, and applicability concerns were common.

Discussion: Principal Findings and Interpretation

Trustworthiness was conditional on task, version, prompt, grounding, comparator, scoring rule, and consequence. Across deprescribing and interaction tasks, agreement or detection accuracy did not guarantee complete, consistent, clinically appropriate recommendations [34-36]. Evidence was strongest for bounded tasks with explicit standards and weakest for patient-specific counselling, treatment modification, uncertainty management, workflow, equity, and outcomes.

Figure 2 presents the original conceptual synthesis for evidence-synthesis framework for large language model accuracy, safety, and clinical use in pharmacy, showing how the article's principal components, evidence relationships, uncertainties, and decision boundaries are connected.

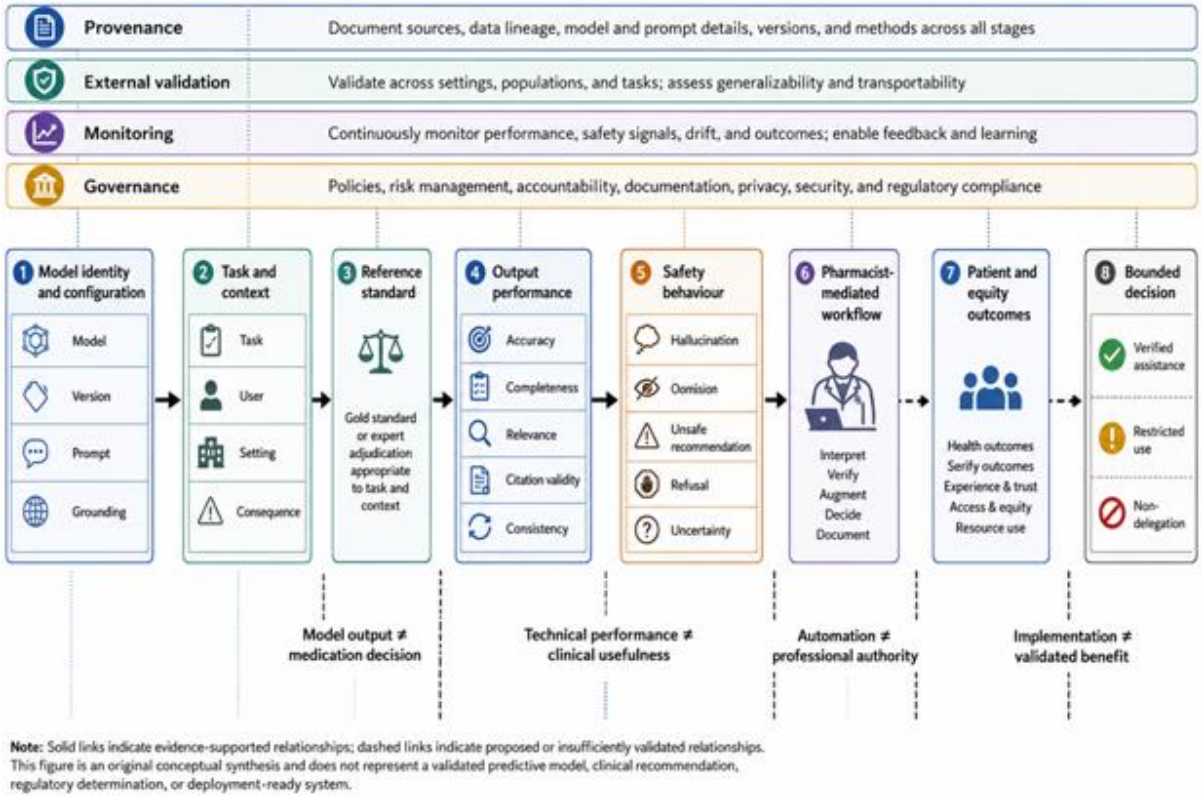

Figure 2. Evidence-Synthesis Framework for Large Language Model Accuracy, Safety, and Clinical Use in Pharmacy

Implications, Strengths, and Limitations

Implementation can create automation bias, deskilling, workflow distortion, and new failure pathways despite favourable average performance [37]. Superficial explanations or citations do not establish faithful or safe reasoning [38]. Responsible clinical AI requires explicit intended use, representative data, external validation, monitoring, human oversight, and harm-aware governance [39].

Table 4 organizes the evidence, constructs, relationships, uncertainties, and boundary conditions required for integrated synthesis, uncertainty, and decision implications for the review Can Large Language Models Be Trusted with Pharmacy Work? A Systematic Review of Accuracy, Safety, and Clinical Use.

Table 4. Integrated synthesis, uncertainty, and decision implications for the review Can Large Language Models Be Trusted with Pharmacy Work? A Systematic Review of Accuracy, Safety, and Clinical Use.

Synthesis domain	Evidence supporting the finding	Conflicting or absent evidence	Heterogeneity or overlap issue	Confidence boundary	Practice implication	Research priority
Drug information	Bounded factual success [22]	Completeness and referencing failures [19]	Rubrics and snapshots differed	Narrow tasks only	Verify authoritative sources	Versioned multisite benchmarks
Review and drug-related problems	Controlled support potential [18, 32]	No workflow or outcomes	Complexity and problem definitions varied	Low for clinical benefit	Verified assistance only	Prospective pharmacist-mediated trials
Counselling	Grounding improved content [25, 29]	Comprehension, equity, and harm untested	Products, languages, and contexts differed	Low beyond tested prompts	Do not delegate	Patient-centred studies
Interactions and deprescribing	Some detection or agreement [34, 35]	Recommendations varied [36]	Databases and thresholds differed	Task-specific	Pharmacist adjudication	Harm-weighted external validation
Uncertainty and governance	Qualifications sometimes appeared [28]; oversight is required [37, 39]	Calibration and deployment benefit absent	Refusal and caveats overlapped	Very low	Restrict and monitor use	Challenge sets and surveillance

These are review-derived boundaries, not approval or deployment determinations.

Strengths were record-level accounting, dual review, explicit evidence-level distinctions, and harm-sensitive appraisal. Limitations were changing proprietary systems, language and Q1 restrictions, heterogeneous standards, absent pooling, incomplete reporting, publication bias, and possible favourable-question selection. The review characterizes evidence available through 3 August 2026.

Conclusion

Large language models cannot presently be considered generally trustworthy for pharmacy work. Selected systems performed some bounded tasks under controlled conditions, but results varied by task, model, prompt, grounding, comparator, and scoring method. Hallucinations, omissions, unsupported recommendations, inconsistency, weak referencing, and inadequately tested uncertainty remain important limitations.

Real-world questions or patient data did not constitute prospective implementation evidence. No included study established improved patient outcomes, safe autonomous operation, universal applicability, or deployment readiness. The defensible current role is bounded assistance with authoritative-source checking and accountable pharmacist verification. Prospective studies should evaluate representative workflows, consequential errors, calibrated refusal and escalation, comprehension, equity, human factors, external validity, monitoring, and clinical outcomes.

ACKNOWLEDGMENTS: None
CONFLICT OF INTEREST: None
FINANCIAL SUPPORT: None
ETHICS STATEMENT: None

References

1. Ong JCL, Chen MH, Ng N, Elangovan K, Tan NYT, Jin L, et al. A scoping review on generative AI and large language models in mitigating medication related harm. NPJ Digit Med. 2025;8(1):182. doi:10.1038/s41746-025-01565-7
2. Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Rui A, et al. The potential of artificial intelligence to improve patient safety: A scoping review. NPJ Digit Med. 2021;4(1):54. doi:10.1038/s41746-021-00423-6
3. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: Systematic literature review. JMIR Med Inform. 2020;8(7). doi:10.2196/18599
4. Syrowatka A, Song W, Amato MG, Foer D, Edrees H, Co Z, et al. Key use cases for artificial intelligence to reduce the frequency of adverse drug events: A scoping review. Lancet Digit Health. 2022;4(2). doi:10.1016/S2589-7500(21)00229-6
5. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372. doi:10.1136/bmj.n71
6. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. 2021;10(1):39. doi:10.1186/s13643-020-01542-z
7. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro CL, Dhiman P, et al. PROBAST+AI: An updated quality, risk-of-bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388. doi:10.1136/bmj-2024-082505
8. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-33. doi:10.1038/s41591-022-01772-9
9. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. doi:10.1038/s41586-023-06291-2
10. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233-9. doi:10.1056/NEJMsr2214184
11. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated suggestions from ChatGPT to optimize

clinical decision support. J Am Med Inform Assoc. 2023;30(7):1237-45. doi:10.1093/jamia/ocad072
12. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med. 2023;388(13):1201-8. doi:10.1056/NEJMra2302038
13. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ. 2020;368. doi:10.1136/bmj.l6890
14. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385. doi:10.1136/bmj-2023-078378
15. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368. doi:10.1136/bmj.m689
16. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Massive Analysis Quality Control (MAQC) Society Board of Directors, Waldron L, et al. Transparency and reproducibility in artificial intelligence. Nature. 2020;586(7829). doi:10.1038/s41586-020-2766-y
17. Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: A comparative study of ChatGPT and clinical pharmacists. Br J Clin Pharmacol. 2024;90(1):232-8. doi:10.1111/bcp.15896
18. Roosan D, Padua P, Khan R, Khan H, Verzosa C, Wu Y. Effectiveness of ChatGPT in clinical pharmacy and the role of artificial intelligence in medication therapy management. J Am Pharm Assoc (2003). 2024;64(2):422-8. doi:10.1016/j.japh.2023.11.023
19. Morath B, Chiriac U, Jaszkowski E, Deiss C, Nurnberg H, Horth K, et al. Performance and risks of ChatGPT used in drug information: An exploratory real-world analysis. Eur J Hosp Pharm. 2024;31(6):491-7. doi:10.1136/ejhpharm-2023-003750
20. Al-Dujaili Z, Al Omari S, Pillai J, Al Faraj A. Assessing the accuracy and consistency of ChatGPT in clinical pharmacy management: A preliminary analysis with clinical pharmacy experts worldwide. Res Social Adm Pharm. 2023;19(12):1590-4. doi:10.1016/j.sapharm.2023.08.012
21. Grossman S, Zerilli T, Nathan JP. Appropriateness of ChatGPT as a resource for medication-related questions. Br J Clin Pharmacol. 2024;90(10):2691-5. doi:10.1111/bcp.16212
22. van Nuland M, Erdogan A, Acar C, Contrucci R, Hilbrants S, Maanach L, et al. Performance of ChatGPT on factual knowledge questions regarding clinical pharmacy. J Clin Pharmacol. 2024;64(9):1095-100. doi:10.1002/jcph.2443
23. Montastruc F, Storck W, de Canecaude C, Victor L, Li J, Cesbron C, et al. Will artificial intelligence chatbots replace clinical pharmacologists? An exploratory study in clinical practice. Eur J Clin Pharmacol. 2023;79(10):1375-84. doi:10.1007/s00228-023-03547-8
24. Li L, Du P, Huang X, Zhao H, Ni M, Yan M, et al. Comparative analysis of generative artificial intelligence systems in solving clinical pharmacy problems: Mixed methods study. JMIR Med Inform. 2025;13. doi:10.2196/76128
25. Kiyomiya K, Aomori T, Ohtani H. Medication counseling for OTC drugs using customized ChatGPT-4: Comparison with ChatGPT-3.5 and ChatGPT-4o. Digit Health. 2025;11:20552076251323810. doi:10.1177/20552076251323810
26. Shin E, Hartman M, Ramanathan M. Performance of the ChatGPT large language model for decision support in community pharmacy. Br J Clin Pharmacol. 2024;90(12):3320-33. doi:10.1111/bcp.16215
27. Albogami Y, Alfakhri A, Alaqil A, Alkoraishi A, Alshammari H, Elsharawy Y, et al. Safety and quality of AI chatbots for drug-related inquiries: A real-world comparison with licensed pharmacists. Digit Health. 2024;10:20552076241253523. doi:10.1177/20552076241253523
28. Hsu HY, Hsu KC, Hou SY, Wu CL, Hsieh YW, Cheng YD. Examining real-world medication consultations and drug-herb interactions: ChatGPT performance evaluation. JMIR Med Educ. 2023;9. doi:10.2196/48433
29. de Jesus DDR, de Souza Junior AP, de Albergaria ET, Pagano AS, de Oliveira IJR, Dos Santos Dias C, et al. Enhanced LLM-supported instructions for medication use through retrieval-augmented generation. Comput Biol Med. 2025;198(Pt A):111135. doi:10.1016/j.compbiomed.2025.111135
30. Andrikyan W, Sametinger SM, Kosfeld F, Jung-Poppe L, Fromm MF, Maas R, et al. Artificial intelligence-powered chatbots in search engines: A cross-sectional study on the quality and risks of drug information for patients. BMJ Qual Saf. 2025;34(2):100-9. doi:10.1136/bmjqs-2024-017476
31. Krichevsky B, Engeli S, Bode-Boger SM, Koop F, Schulze Westhoff M, Schroder S, et al. Human vs artificial intelligence: Physicians outperform ChatGPT in real-world pharmacotherapy counselling. Br J Clin Pharmacol. 2026;92(3):891-902. doi:10.1002/bcp.70321
32. Ong JCL, Jin L, Elangovan K, Lim GYS, Lim DYZ, Sng GGR, et al. Large language model as clinical decision support system augments medication safety in 16 clinical specialties. Cell Rep Med. 2025;6(10):102323. doi:10.1016/j.xcrm.2025.102323
33. Radha Krishnan RP, Hung EH, Ashford M, Edillo CE, Gardner C, Hatrick HB, et al. Evaluating ChatGPT performance in predicting drug-drug interactions from hospitalized patient data. Br J Clin Pharmacol. 2024;90(12):3361-6. doi:10.1111/bcp.16275
34. Buzancic I, Belec D, Drzaic M, Kummer I, Brkic J, Fialova D, et al. Clinical decision-making in benzodiazepine deprescribing by healthcare providers vs AI-assisted approach. Br J Clin Pharmacol. 2024;90(3):662-74. doi:10.1111/bcp.15963
35. Bischof T, al Jalali V, Zeitlinger M, Jorda A, Hana M, Singeorzan KN, et al. Chat GPT vs clinical decision support systems in the analysis of drug-drug interactions. Clin Pharmacol Ther. 2025;117(4):1142-7. doi:10.1002/cpt.3585
36. Chase A, Most A, Sikora A, Smith SE, Devlin JW, Xu S, et al. Evaluation of large language models' ability to identify clinically relevant drug-drug interactions and generate high-quality clinical pharmacotherapy recommendations. Am J Health Syst Pharm. 2026;83(13). doi:10.1093/ajhp/zxaf168
37. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517-8. doi:10.1001/jama.2017.7797
38. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11). doi:10.1016/S2589-7500(21)00208-9
39. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. Nat Med. 2019;25(9):1337-40. doi:10.1038/s41591-019-0548-6