Pharmacophore

ISSN-2229-5402

Journal home page: http://www.pharmacophorejournal.com

LARGE LANGUAGE MODELS AS PHARMACEUTICAL KNOWLEDGE WORKERS: CAPABILITIES, FAILURE MODES, AND EVIDENCE REQUIREMENTS ACROSS SCIENTIFIC TASKS

Luis Benitez[1]*, Andrea Gomez[2], Jazmin Franco[3], Roberto Cardozo[4]

1. ***Department of LLMs as Pharmaceutical Knowledge Workers, Faculty of Pharmacy, National University of Asunción, Asunción, Paraguay.***
2. ***Department of Capabilities and Failure Modes of LLMs, Faculty of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil.***
3. ***Department of Evidence Requirements for LLM Use, Faculty of Pharmacy, National University of Caaguazú, Caaguazú, Paraguay.***
4. ***Department of Scientific Task Evaluation for LLMs, Faculty of Pharmacy, National University of La Plata, La Plata, Argentina.***

ARTICLE INFO

Received:
18 May 2025
Received in revised form:
20 August 2025
Accepted:
23 August 2025
Available online:
28 August 2025

Keywords: Large language models, Pharmaceutical knowledge work, Scientific evidence synthesis, Molecular assistance, Hallucination, Source provenance

ABSTRACT

Large language models are increasingly positioned as pharmaceutical knowledge workers capable of searching literature, extracting evidence, summarizing documents, organizing knowledge, proposing hypotheses, assisting molecular design, generating code, and coordinating scientific tools. Yet the evidentiary meaning of these capabilities remains unclear. Strong benchmark performance, fluent explanations, or successful demonstrations in constrained environments do not necessarily establish reliable scientific reasoning, experimental value, workflow benefit, or readiness to influence pharmaceutical decisions. This critical state-of-the-field review evaluates large language models according to the scientific tasks they perform, the evidence units used to assess them, and the consequences of their errors. The review develops a proposed task taxonomy spanning evidence processing, knowledge organization, hypothesis generation, molecular assistance, computational support, workflow orchestration, and decision support. It also distinguishes output validity, source grounding, task-valid performance, expert-comparator evidence, prospective workflow validation, and decision-consequence evidence as separate levels of support that should not be treated as interchangeable. Across the literature, the strongest evidence concerns bounded text-processing and tool-mediated tasks, whereas evidence for autonomous scientific judgment, experimentally confirmed molecular contribution, and routine pharmaceutical deployment remains limited. Recurring limitations include hallucinated content, citation failure, data leakage, shortcut learning, hidden substitution of easier tasks for intended scientific constructs, inadequate source attribution, and under-specified human oversight. The central contribution is a role-based evidentiary interpretation of pharmaceutical language-model use in which validation requirements increase with autonomy, irreversibility, and decision consequence. The proposed synthesis is conceptual rather than empirically validated and should be used to structure evaluation, not to certify systems. Reliable adoption will require source-auditable outputs, task-specific comparators, prospective workflow studies, explicit escalation rules, configuration traceability, and governance that treats the model as one component of a broader scientific system.

To Cite This Article: Benitez L, Gomez A, Franco J, Cardozo R. Large Language Models as Pharmaceutical Knowledge Workers: Capabilities, Failure Modes, and Evidence Requirements across Scientific Tasks. Pharmacophore. 2025;16(4):102-11. https://doi.org/10.51847/IwEKFH04Ct

Introduction

Artificial intelligence is already embedded across target identification, molecular representation, property prediction, lead optimization, clinical development, and evidence interpretation. The emergence of large language models extends this trajectory by placing a general natural-language interface over scientific documents, databases, computational tools, and increasingly agentic workflows. The relevant frontier therefore lies at the intersection of established pharmaceutical machine-learning pipelines, AI-enabled changes in drug-design practice, and foundation-model architectures capable of transferring across heterogeneous tasks [1–3]. This convergence creates a new category of system: not merely a predictor trained for one

Corresponding Author: Luis Benitez; Department of LLMs as Pharmaceutical Knowledge Workers, Faculty of Pharmacy, National University of Asunción, Asunción, Paraguay. E-mail: luis.benitez@una.py.

endpoint, but a language-mediated component that can retrieve, transform, connect, and generate scientific information. Describing such systems as pharmaceutical knowledge workers is useful only when the term is treated functionally. It refers to their participation in knowledge-intensive tasks, not to an assumption that they possess expert understanding, professional accountability, or independent scientific authority.

The central evidentiary problem is that apparent capability depends strongly on how the task is constructed and evaluated. Medical language models can perform well on question-answering benchmarks while clinician assessment still identifies deficiencies in factuality, reasoning, bias, and potential harm [4]. This divergence is especially important in pharmaceutical science, where a plausible textual answer may be separated by several validation layers from an actionable conclusion. A model may correctly retrieve a named entity yet misrepresent the relation between entities; summarize an article fluently while omitting a decisive limitation; propose a chemically plausible structure that cannot be synthesized; or recommend an apparently relevant mechanism without evidence that the mechanism is causal, experimentally confirmed, or therapeutically useful. Consequently, benchmark performance must be interpreted as evidence about behavior under a defined test condition, not as a general statement about scientific competence.

The pharmaceutical application landscape is also heterogeneous. Contemporary reviews describe language-model use across literature analysis, biomedical text processing, molecular representation, clinical information, drug development, and operational assistance, but the underlying evidence varies substantially in task definition, source access, model configuration, comparator choice, and validation depth [5]. A literature-search assistant, a molecule-optimization interface, a coding agent, and a treatment-support system may all be described as applications of the same model class, although their error mechanisms and acceptable evidence thresholds differ fundamentally. Collapsing these uses into a single category obscures whether a reported result demonstrates linguistic transformation, retrieval, prediction, reasoning, tool use, or merely reproduction of patterns already present in training data.

This review therefore evaluates large language models as pharmaceutical knowledge workers through three linked questions: what task is the system actually performing, what evidence supports the claimed capability, and what consequence follows if the output is wrong. Its objective is not to rank models or catalogue applications one by one. Instead, it develops a proposed task taxonomy and a corresponding evidentiary interpretation that distinguish demonstrated capability from plausible utility, experimental confirmation, translational value, and routine pharmaceutical readiness. The review treats source grounding, construct validity, expert comparison, workflow performance, and decision consequences as separate analytical dimensions. These dimensions form the basis for identifying where the available evidence is comparatively mature, where claims remain benchmark-dependent, and what additional validation would be required before language-model outputs could responsibly influence scientific or development decisions.

Review Scope, Task Taxonomy, and Critical Appraisal Approach

The technological frontier considered in this review was defined as the state of peer-reviewed evidence available by the end of 2025. The review approach combined transparent scoping logic, systematic reporting discipline, and explicit appraisal of design and reporting weaknesses in comparative artificial-intelligence studies [6–8]. Eligible evidence was required to contribute directly to at least one pharmaceutical knowledge-work task, failure mechanism, evaluation dimension, or governance requirement. The unit of analysis was not the named model alone, because the same underlying model can produce materially different behavior when paired with different prompts, retrieval systems, tools, databases, interfaces, or human reviewers. The primary unit was therefore the model-enabled task configuration: the model version, information source, prompt or instruction structure, available tools, expected output, comparator, evaluation method, and intended downstream use considered together.

The proposed task taxonomy separates four broad forms of pharmaceutical knowledge work. The first is evidence processing, comprising search, screening support, extraction, summarization, and document transformation. The second is knowledge organization, including entity normalization, relation structuring, ontology alignment, evidence linking, and graph-supported representation. The third is scientific production assistance, encompassing hypothesis generation, molecule-related interaction, coding, calculations, protocol drafting, and tool mediation. The fourth is workflow and decision participation, in which the model coordinates multiple steps, prioritizes alternatives, recommends actions, or contributes directly to decisions. These categories are analytically useful but are not empirically validated classes. A single system may occupy several categories, and transitions between them can introduce new risks. For example, an extraction system becomes a decision-support component when its structured output is used without independent review to exclude a safety signal or prioritize a development candidate.

Critical appraisal was organized around claim–evidence fit rather than model novelty. For every application, the review asks whether the reported evidence establishes output correctness, source support, performance on the intended construct, robustness to altered inputs, comparison with appropriate experts, function within an actual workflow, or improvement in a consequential decision. These forms of evidence are cumulative only in a logical sense; evidence at one level does not establish the next. A high-quality summary does not prove complete evidence retrieval, an expert-comparable answer does not prove workflow benefit, and a successful computational proposal does not prove experimental or therapeutic value. The proposed appraisal therefore distinguishes model confidence from calibrated uncertainty, data availability from data suitability, association from mechanism, prediction from explanation, and conceptual readiness from operational deployment. **Table 1** organizes the evidence, constructs, and interpretation boundaries needed to develop review scope, task taxonomy, and critical appraisal approach within the article's central argument.

Table 1. Evidence domains, core questions, scientific requirements, and interpretation boundaries for Review scope, task taxonomy, and critical appraisal approach

Evidence domain	Eligible evidence or method	Reported capability or claim	Strength of support	Reproducibility or bias concern	Unresolved gap
Review scope	Transparent eligibility criteria, traceable literature selection, explicit frontier definition, and claim-level evidence mapping	The review represents the relevant technological frontier and distinguishes directly applicable from peripheral evidence	Strong methodological support when selection and exclusions are auditable	Opaque screening decisions, selective inclusion of favorable demonstrations, and rapid publication turnover	Standard approaches for updating fast-changing language-model evidence without destabilizing prior conclusions
Task taxonomy	Task decomposition based on inputs, outputs, tools, users, and downstream decisions	Different pharmaceutical applications can be evaluated according to the work actually performed rather than the model label	Moderate-to-strong conceptual support across heterogeneous application studies	Category overlap and relabelling of the same behavior as search, reasoning, generation, or decision support	Empirical testing of whether the proposed categories predict distinct error patterns or validation needs
Model-enabled task configuration	Documentation of model version, prompt, retrieval source, tools, interface, comparator, and use context	Performance belongs to the configured system rather than to the base language model in isolation	Strong conceptual support	Undisclosed prompts, changing proprietary models, undocumented tool versions, and irreproducible retrieval states	Stable reporting standards for dynamic, multi-component pharmaceutical systems
Evidence-processing capability	Annotated extraction sets, retrieval audits, source-linked summaries, omission analysis, and expert review	The system can identify, transform, or condense scientific information	Moderate support for bounded tasks	Curated datasets, limited source diversity, annotation ambiguity, and reliance on surface-overlap metrics	Prospective evaluation in live pharmaceutical literature and document-review workflows
Knowledge organization	Ontology mapping, entity resolution, relation extraction, evidence graphs, and provenance-preserving integration	The system can structure heterogeneous biomedical knowledge	Moderate support for organization; weaker support for biological truth	Schema mismatch, stale relations, extraction error propagation, and loss of source lineage	Methods that preserve uncertainty, contradiction, and temporal change within organized knowledge
Scientific production assistance	Hypothesis-generation exercises, molecule-design cases, code execution, tool use, and expert adjudication	The system can generate or operationalize scientific candidates and procedures	Emerging and task-dependent	Memorization, hidden human scaffolding, invalid objectives, silent code errors, and selective case reporting	Prospective design–make–test cycles and reproducible computational workflow studies
Expert comparison	Blinded scoring, domain-expert review, realistic cases, and multidimensional error analysis	The system performs similarly to or differently from experts on a defined task	Informative but insufficient alone	Comparator selection, artificial cases, narrow rubrics, and disagreement among experts	Evaluation of teams, escalation behavior, error interception, and consequences of disagreement
Workflow and decision evidence	Shadow-mode deployment, prospective workflow studies, human-factors evaluation, and decision-consequence analysis	The system improves efficiency, quality, safety, or scientific decisions in practice	Limited for most pharmaceutical roles	Automation bias, verification burden, workflow adaptation, institutional confounding, and underreported failures	Multi-site prospective evidence with predefined outcomes, error recovery, and post-deployment monitoring
Readiness interpretation	Integrated assessment of consequence, autonomy, reversibility, oversight, and evidence maturity	A task configuration is suitable for a bounded role under specified controls	Proposed conceptual synthesis rather than a validated certification framework	Risk that qualitative categories are interpreted as formal approval or universal maturity levels	Validation of role-specific readiness criteria across pharmaceutical organizations and tasks

Scientific Search, Extraction, Summarization, and Knowledge Organization

Scientific search and extraction are often discussed together, but they answer different questions. Search concerns whether the relevant evidence is found, whereas extraction concerns whether specified information is correctly identified within the retrieved material. Domain-adapted language representations have improved biomedical named-entity recognition, relation extraction, and question answering relative to general-domain baselines [9]. These results support the use of language models as components of evidence-processing systems, particularly when the target fields and annotations are well defined. They do not establish complete literature coverage, correct interpretation of all retrieved studies, or suitability for downstream decisions. A search assistant may return relevant articles while missing contradictory evidence; an extraction model may correctly identify a drug and an adverse event while misclassifying the relation between them. Evaluation must therefore separate retrieval sensitivity, ranking quality, extraction correctness, source diversity, and the consequences of omitted or mischaracterized evidence.

Generative biomedical models extend extraction by producing relational statements, answers, and narrative summaries. Domain-specific generative pretraining can improve biomedical text-mining and generation tasks, but success on these benchmarks does not itself demonstrate that generated claims are grounded in inspectable sources [10]. Comparable considerations apply to models trained on longitudinal clinical records: specialized pretraining can improve selected clinical language tasks, while transfer remains conditional on documentation practices, institutional data distributions, and the provenance of the underlying records [11]. In pharmaceutical knowledge work, this distinction is consequential because outputs frequently combine published evidence, internal reports, structured databases, and tacit assumptions. A system may produce a coherent synthesis even when its components originate from incompatible populations, assay conditions, document versions, or evidentiary standards. Reliable use therefore requires traceability from each material claim to the source passage or structured record from which it was derived.

Knowledge organization adds another layer by converting extracted material into structured entities, relations, ontologies, and evidence networks. Biomedical knowledge graphs can integrate heterogeneous information, but their scientific value depends on entity resolution, schema design, ontology alignment, temporal updating, and provenance preservation [12]. A graph edge may represent a reported association, a predicted interaction, an experimentally supported mechanism, or a curated clinical relation; treating these edge types as interchangeable would create false evidentiary equivalence. Summarization systems face a parallel problem. Their evaluation should address factual consistency, completeness, source attribution, and clinically or scientifically important omissions rather than linguistic fluency alone [13]. Accordingly, the review interprets search, extraction, summarization, and knowledge organization as an evidence chain rather than a single capability. Each stage can improve efficiency, but each can also propagate or amplify upstream error. The most defensible near-term role is therefore assisted evidence processing in which retrieved sources remain inspectable, structured outputs remain auditable, and expert reviewers can identify omissions, contradictions, and inappropriate inferential transitions before the information is used in pharmaceutical reasoning.

Hypothesis Generation, Molecular Assistance, Coding, and Workflow Orchestration

Large language models can support hypothesis generation by connecting literature, structured resources, computational tools, and user-defined questions. Their most visible extension is agentic operation, in which a model selects tools, decomposes objectives, generates intermediate instructions, and uses outputs from one stage to initiate another. A system integrating literature search, synthesis planning, code execution, laboratory hardware, and chemical experimentation has demonstrated that an LLM can coordinate selected chemical research tasks under substantial technical scaffolding [14]. This evidence establishes bounded orchestration capability, not general autonomous pharmaceutical competence. The demonstrated performance depends on tool availability, access permissions, interface design, error handling, and the constraints imposed by the developers. These dependencies are part of the evaluated system and should not be attributed solely to the language model.

Tool augmentation can improve performance when a task requires chemical databases, calculations, or specialist software unavailable to a language-only model. Chemistry-oriented agents can route questions to external tools and incorporate the resulting outputs into subsequent responses [15]. This architecture may reduce some failures caused by attempting calculations or database recall directly through probabilistic text generation. It also introduces new failure points: an agent may select the wrong tool, formulate an invalid query, misread a returned value, overlook units or boundary conditions, or continue after an execution failure. Evaluation should therefore measure tool selection, parameter construction, execution validity, output interpretation, and recovery from failure as separate capabilities. A correct final answer cannot reveal whether unsafe or irreproducible intermediate operations occurred.

Hypothesis generation should similarly be separated from hypothesis confirmation. Language-model translation between drug molecules and indication descriptions may assist the generation or ranking of molecule–indication associations [16]. Such associations can be useful for exploratory triage, particularly when they expose connections that a scientist can investigate through source review and mechanistic analysis. They do not demonstrate that the suggested relationship is novel, causal, biologically meaningful, or therapeutically actionable. Interactive molecule-optimization systems extend this functionality by translating natural-language objectives into proposed structural modifications [17]. Their computational outputs may satisfy selected validity or property criteria while remaining unevaluated for synthetic accessibility, assay behavior, pharmacokinetics, formulation constraints, toxicity, manufacturability, or clinical relevance. Molecular assistance should therefore be interpreted as candidate generation under specified objectives rather than evidence of successful drug design.

Language models may also lower the practical barrier between scientists and specialist computational systems. A conversational interface can construct reward functions, configure molecular-generation tools, and support iterative design cases without requiring the user to manipulate every software parameter directly [18]. This form of mediation may be valuable for coding, workflow prototyping, and communication across disciplinary boundaries. Nevertheless, interface accessibility is distinct from scientific validity. Generated code requires executable tests, dependency capture, unit checking, input validation, and review of the underlying scientific assumptions. The review therefore proposes that hypothesis generation, molecule assistance, coding, and orchestration be assessed as composite workflows. The evidence record should identify which operations were produced by the model, which were performed by external tools, which were manually corrected, and which conclusions received independent computational or experimental confirmation.

Hallucination, Citation Failure, Leakage, and Hidden Task Substitution

Hallucination is best understood as a family of failures rather than a single property of generated text. It can include statements inconsistent with a provided source, claims unsupported by any inspectable evidence, invented entities or relations, and answers that are linguistically appropriate but incompatible with the task context [19]. In pharmaceutical settings, these forms have different consequences. An invented explanatory sentence in a preliminary draft may be detectable through editing, whereas an unsupported contraindication, assay interpretation, or mechanistic claim could alter a consequential decision. Evaluation should therefore classify hallucinations by source relationship, scientific function, detectability, and potential consequence. A global factuality score may obscure whether errors cluster in the claims that matter most.

Citation failure is a related but independent problem. Language models may produce references that do not exist or combine plausible author, journal, title, and date elements into incorrect bibliographic records [20]. Even an authentic reference can be used misleadingly if it does not support the attached claim, reports a different population, or presents a prediction as experimental evidence. Retrieval augmentation can improve access to documents, but the presence of retrieved material does not ensure that the model interpreted or cited it correctly. Pharmaceutical evidence workflows therefore require at least three verification layers: bibliographic resolution, confirmation that the cited passage supports the specific claim, and appraisal of whether the source is methodologically appropriate for the intended inference. Citation presence should never be treated as a proxy for evidentiary grounding.

Data leakage can make apparent model capability difficult to interpret. Information from evaluation examples may enter model development through pretraining corpora, prompt selection, fine-tuning data, repeated benchmark exposure, or preprocessing performed before data partitioning. Such leakage can inflate performance and contribute to irreproducible machine-learning claims [21]. The challenge is particularly acute for closed language models whose training data are incompletely disclosed and for scientific benchmarks assembled from publicly available literature. Prospective or time-separated test sets, private challenge sets, locked evaluation protocols, and explicit model-access dates can reduce some forms of contamination. They cannot always prove that a model has never encountered related material, so residual uncertainty should be acknowledged rather than hidden behind benchmark scores.

Hidden task substitution occurs when a system appears to solve the stated scientific problem by exploiting cues that support an easier, unintended task. Shortcut learning demonstrates that high performance can arise from correlates that do not represent the intended construct [22]. A model asked to identify mechanistic plausibility may instead reproduce frequently discussed associations; a literature assistant may rank familiar journals rather than retrieve the most relevant evidence; and a molecule–indication system may recover well-known relationships from training data rather than generate new hypotheses. These concerns become more consequential in medication-related applications, where the available evidence remains heterogeneous and prospective testing is limited [23]. The review therefore treats leakage, shortcut learning, and hidden task substitution as construct-validity failures. Evaluation must ask not only whether the answer is correct, but what information was available, what strategy could have produced the answer, and whether that strategy would remain valid when superficial cues change.

Evaluation against Experts, Source Evidence, and Decision Consequences

Expert comparison is informative only when expertise and the evaluated task are specified. A broad chemistry evaluation framework has shown that model performance varies across factual knowledge, reasoning, tool use, and comparison with chemists [24]. This multidimensional approach is preferable to a single aggregate score because pharmaceutical expertise is not a unitary construct. A medicinal chemist, information scientist, computational biologist, pharmacometrician, formulation scientist, and safety reviewer bring different knowledge, error sensitivities, and decision responsibilities. The appropriate comparator may therefore be an individual specialist, a multidisciplinary group, a documented workflow, or an evidence database. Model–expert comparison should report disagreement patterns and scientifically important errors rather than focusing only on mean performance.

Evaluation realism also matters. Testing on open-ended longitudinal records can reveal failures in information prioritization, diagnostic reasoning, guideline adherence, laboratory interpretation, and robustness that may be hidden by conventional question-answer formats [25]. Pharmaceutical tasks have analogous complexity. A development decision may depend on incomplete records, conflicting assays, changing product profiles, uncertain evidence quality, and organizational constraints. An evaluation that supplies only the decisive facts can overestimate performance by removing the search and prioritization burden that defines the actual work. Realistic evaluation should preserve irrelevant information, missing data, temporal order, conflicting sources, and opportunities to abstain or escalate.

Source evidence and expert scoring should be examined together. In personalized oncology, language models have generated potentially complementary treatment ideas while remaining below the credibility required to automate expert molecular tumor-board work [26]. Conversely, strong performance on structured reasoning rubrics in selected written cases demonstrates competence on those rubric dimensions without establishing superior real-world practice [27]. These findings illustrate why expert-comparable output, source-grounded output, and useful decision support are overlapping but non-equivalent claims. A system may produce a reasonable option without citing adequate evidence, cite suitable evidence while overlooking patient- or program-specific constraints, or receive a high reasoning score while failing to communicate uncertainty and escalation needs. **Figure 1** presents the pharmaceutical knowledge-worker reliability observatory, showing how the article's key components and boundaries are connected within evaluation against experts, source evidence, and decision consequences.

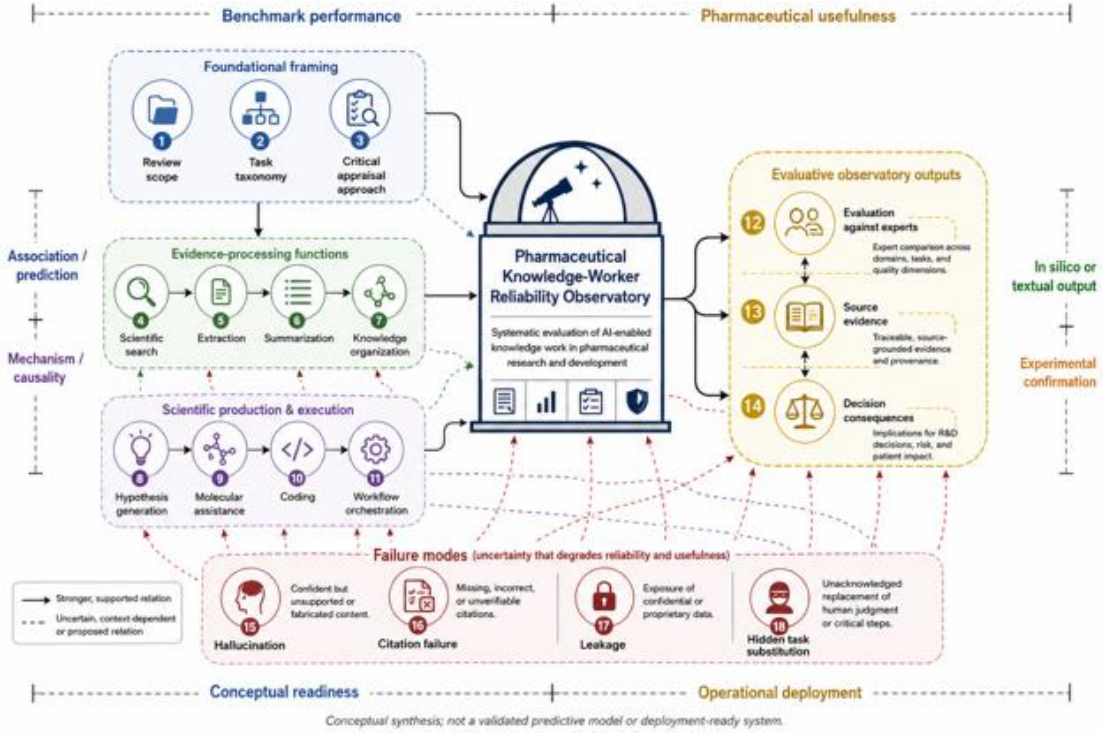

Figure 1. Pharmaceutical Knowledge-Worker Reliability Observatory.

The figure is an original conceptual synthesis that organizes the article's central contribution across review scope, task taxonomy, and critical appraisal approach, review scope, task taxonomy, critical appraisal approach, scientific search, extraction, summarization, and knowledge organization, scientific search. Arrows and grouping indicate proposed or evidence-supported relationships rather than measured effect sizes. The figure does not represent a validated predictive model, regulatory determination, clinical recommendation, or deployment-ready system.

Decision consequences constitute the highest evidentiary burden because the performance of the model, the actions of the user, and the surrounding workflow can interact. A systematic review and meta-analysis of human–AI combinations found that combined systems do not automatically outperform the better individual component [28]. Human oversight may fail when users are rushed, overly trusting, unable to access the underlying evidence, or uncertain about which component produced an error. Conversely, a model may add value by widening the option set, standardizing routine checks, or drawing attention to overlooked evidence. The appropriate endpoint is therefore not whether a model resembles an expert in isolation, but whether the configured human–AI system improves a defined task without introducing unacceptable verification burden, error propagation, or harm. Prospective evaluation should measure escalation, disagreement resolution, recovery from failure, and the downstream consequences of acting on or rejecting model outputs.

Role-Based Evidence Requirements for Pharmaceutical Knowledge Work

Role-specific evidence requirements should connect clinical-impact prerequisites, responsible deployment safeguards, and prospective trial-reporting standards rather than treating benchmark accuracy as sufficient [29–31]. The review proposes that requirements increase with four characteristics: the degree of autonomy granted to the system, the consequence of an incorrect output, the reversibility of the resulting action, and the ability of a qualified user to inspect the supporting evidence. A low-consequence drafting tool can be evaluated through output audits and user review. A system that prioritizes safety cases, excludes compounds, recommends dose changes, or initiates laboratory operations requires more representative validation, explicit human authority, and prospective evidence. These are proposed review criteria, not validated certification categories.

Protocol-stage specification is essential when an AI system may influence a consequential intervention. Relevant details include the model configuration, eligible inputs, intended outputs, human–AI interaction, handling of missing information, escalation rules, error management, and analysis plan [32]. Pharmaceutical evaluations should apply the same principle before a system is exposed to operational decisions. A study should state whether users can modify prompts, whether retrieval sources are fixed or dynamic, how model updates are handled, what information is hidden from the system, and which outputs require independent confirmation. Without this pre-specification, favorable results can reflect informal adjustments or selective interpretation that would not be reproducible in routine use.

Before definitive outcome studies, early live evaluation should examine actual users, workflow fit, human factors, system failures, and iterative modification [33]. Shadow-mode deployment may be appropriate when a system produces outputs without influencing decisions, allowing comparison with existing practice and identification of silent failure patterns. Controlled pilot studies can then assess whether users understand system limitations, whether verification demands offset efficiency gains, and whether the interface supports appropriate abstention and escalation. These studies should document not

only successful interactions but also abandoned recommendations, disagreements, overrides, and incidents in which the system or user failed to recognize insufficient evidence.

The resulting role-based interpretation contains four broad readiness boundaries. Evidence-processing assistance may be considered for bounded use when sources remain inspectable and outputs receive sampled or complete audit. Scientific production assistance requires executable or experimentally testable outputs, documented tools, and independent validation of decisive claims. Workflow orchestration requires permission controls, monitoring of intermediate steps, reproducible state capture, and tested recovery from component failure. Decision participation requires prospective evidence that the human–AI configuration improves a relevant decision or outcome under representative conditions. Movement between these roles is not automatic: a system validated for drafting summaries should not inherit readiness for prioritization, recommendation, or autonomous execution merely because the same underlying model is used.

Governance and Research Priorities for Reliable Adoption

Governance should be organized around the application and its consequences rather than assuming that one policy can regulate every use of a general-purpose model. Risk-based oversight of large language models should address high-consequence uses, transparency, non-discrimination, content controls, and the responsibilities attached to specific applications [34]. In pharmaceutical organizations, this implies that an externally hosted drafting assistant, an internal literature system, a molecule-design agent, and a clinical decision component require different access controls, validation records, escalation procedures, and change-management processes. The base model is only one component; retrieval systems, connected databases, tools, interfaces, and local operating procedures also require governance.

Ethical adoption requires more than filtering undesirable language after generation. Responsible use depends on technical safeguards, organizational culture, data stewardship, professional training, user incentives, and the distribution of authority between people and automated components [35]. A system may be technically accurate yet institutionally unsafe if users cannot challenge its outputs, if productivity targets discourage verification, or if uncertainty is hidden behind authoritative wording. Governance must therefore specify who may use the system, for which purposes, with what training, and under whose responsibility. It should also protect legitimate scientific disagreement rather than encouraging users to accept a standardized answer because it was generated consistently.

Minimum documentation is required for model claims to be auditable. Reporting frameworks for clinical artificial-intelligence modeling emphasize transparent description of data, preprocessing, model development, evaluation, and reproducibility [36]. Pharmaceutical LLM systems require extensions of this record: model provider and version, access date, prompt templates, system instructions, retrieval configuration, document corpus, tool versions, software environment, user permissions, output filters, and human corrections. A reproducible evidence record should also distinguish temporary exploratory use from validated production use. When any component changes, the organization should determine whether prior validation remains applicable or whether targeted re-evaluation is required.

LLM-specific reporting guidance reinforces the need to document task formulation, prompting, retrieval, data, human oversight, evaluation, and open-science practices [37]. Ethical implementation further requires attention to bias, opacity, professional responsibility, and the redistribution of decision authority [38]. Near-term research should therefore prioritize claim-level provenance, private or prospectively created evaluation sets, error distributions rather than average scores alone, human–AI disagreement, verification burden, version drift, and prospective pharmaceutical workflow studies. Longer-term work should connect model-assisted activity to experimental success, development decisions, safety outcomes, and organizational accountability without assuming that technical performance guarantees benefit. **Table 2** organizes the evidence, constructs, and interpretation boundaries needed to develop governance and research priorities for reliable adoption within the article's central argument.

Table 2. Evaluation, implementation, and research priorities arising from Large Language Models as Pharmaceutical Knowledge Workers Capabilities, Failure Modes, and Evidence

Approach or application	Data and task context	Evaluation practice	Evidence maturity	Translational limitation	Review interpretation
Scientific search and extraction	Published literature, internal reports, structured biomedical records, and predefined extraction fields	Retrieval audits, annotated test sets, source-level verification, omission analysis, and external validation	Source-audited benchmark evidence is feasible; live-workflow evidence remains limited	Missed evidence, ranking bias, corpus restrictions, annotation ambiguity, and domain shift	Appropriate for assisted evidence processing when every consequential output remains traceable to inspectable sources
Summarization and knowledge organization	Single documents, multi-document evidence sets, ontologies, knowledge graphs, and heterogeneous databases	Factual-consistency review, claim–source alignment, completeness assessment, contradiction handling, and provenance checks	Task demonstrations and retrospective comparisons are available	Fluent summaries can conceal omissions; graph structure can preserve incorrect or stale relations	Useful for organization and drafting, but not a substitute for appraisal of source quality or biological validity

Hypothesis generation	Literature-derived associations, molecule–indication relations, mechanistic prompts, and exploratory research questions	Novelty checking, contamination analysis, mechanistic review, negative controls, and expert prioritization	Predominantly conceptual or retrospective computational evidence	Memorization, popularity bias, absent causal support, and lack of experimental confirmation	Outputs should be labelled as candidate hypotheses and should not be presented as mechanisms or validated discoveries
Molecular assistance	Natural-language objectives, molecular representations, generative tools, and property-optimization tasks	Chemical-validity checks, independent property calculations, synthesis assessment, orthogonal assays, and staged design–make–test evaluation	In-silico demonstration to source-audited benchmark maturity	Computational objectives omit many developability, safety, formulation, manufacturing, and therapeutic constraints	May support ideation and interface with specialist tools; pharmaceutical value requires layered experimental confirmation
Coding and computational assistance	Data processing, statistical scripts, simulations, database queries, and specialist scientific software	Executable tests, code review, dependency capture, unit validation, reproducibility checks, and failure-injection tests	Bounded task demonstrations are available	Silent code defects, package drift, invalid assumptions, security risks, and irreproducible environments	Suitable for supervised acceleration when code and outputs are independently tested
Workflow orchestration	Multi-step agents connecting search, databases, code, calculations, hardware, or laboratory systems	Intermediate-state logging, permission controls, tool-routing assessment, recovery testing, red teaming, and shadow-mode evaluation	Selected demonstrations; prospective pharmaceutical workflow evidence is sparse	Cascading failures, unsafe actions, hidden human scaffolding, and unclear component accountability	Does not support general autonomous pharmaceutical research; requires evaluation of the entire configured system
Expert-facing decision support	Scientific prioritization, safety review, development choices, individualized evidence interpretation, or therapeutic options	Multidimensional expert comparison, realistic cases, source audit, abstention testing, human-factors assessment, and prospective decision evaluation	Expert-comparator evidence exists in selected contexts; decision-consequence evidence is limited	Comparator choice, artificial cases, automation bias, verification burden, and absent prospective outcomes	May complement expert work, but expert-like output does not establish decision authority or replacement readiness
Governed routine adoption	Versioned production systems embedded in quality-managed or regulated organizational processes	Change control, access governance, continuous monitoring, incident response, audit trails, periodic revalidation, and outcome surveillance	Proposed readiness endpoint rather than a generally established state	Model updates, vendor dependency, privacy, drift, distributed responsibility, and jurisdictional variation	Routine readiness should be claimed only for a specified application, configuration, user group, and evidence record

Conclusion

Large language models can participate meaningfully in pharmaceutical knowledge work, but their value is task-specific and evidentially conditional. The strongest current case is for bounded assistance in searching, extracting, organizing, drafting, coding, and mediating access to specialist tools under transparent human review. The evidence is substantially weaker for autonomous hypothesis validation, molecule selection, workflow control, and consequential scientific or clinical decisions. This review's original contribution is a proposed task-and-evidence interpretation that separates output validity, source grounding, intended-task performance, expert comparison, workflow validation, and decision consequences. It also emphasizes that hallucination, citation failure, leakage, shortcut learning, and hidden task substitution can invalidate apparently strong results at multiple stages. The framework is not a validated maturity instrument and should not be interpreted as regulatory or deployment approval. Its purpose is to make claims more precise: a system should be described according to the work it performs, the evidence that supports that work, the oversight under which it operates, and the consequences of error. Reliable adoption will depend less on assigning a general intelligence label to a model than on constructing source-auditable, reproducible, and governable pharmaceutical systems in which scientific responsibility remains explicit.

Acknowledgments: None

Conflict of interest: None

Financial support: None

Ethics statement: None

References

1. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77. doi:10.1038/s41573-019-0024-5
2. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, et al. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov. 2020;19(5):353-64. doi:10.1038/s41573-019-0050-3
3. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65. doi:10.1038/s41586-023-05881-4
4. 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
5. Zheng Y, Koh HY, Ju J, Yang M, May LT, Webb GI, et al. Large language models for drug discovery and development. Patterns (N Y). 2025;6(10):101346. doi:10.1016/j.patter.2025.101346
6. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467-73. doi:10.7326/M18-0850
7. 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
8. 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
9. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40. doi:10.1093/bioinformatics/btz682
10. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform. 2022;23(6). doi:10.1093/bib/bbac409
11. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. NPJ Digit Med. 2022;5(1):194. doi:10.1038/s41746-022-00742-2
12. Nicholson DN, Greene CS. Constructing knowledge graphs and their biomedical applications. Comput Struct Biotechnol J. 2020;18:1414-28. doi:10.1016/j.csbj.2020.05.017
13. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. NPJ Digit Med. 2023;6(1):158. doi:10.1038/s41746-023-00896-7
14. Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624(7992):570-8. doi:10.1038/s41586-023-06792-0
15. Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. Augmenting large language models with chemistry tools. Nat Mach Intell. 2024;6(5):525-35. doi:10.1038/s42256-024-00832-8
16. Oniani D, Hilsman J, Zang C, Wang J, Cai L, Zawala J, et al. Emerging opportunities of using large language models for translation between drug molecules and indications. Sci Rep. 2024;14(1):10738. doi:10.1038/s41598-024-61124-0
17. Ye G, Cai X, Lai H, Wang X, Huang J, Wang L, et al. DrugAssist: A large language model for molecule optimization. Brief Bioinform. 2025;26(1). doi:10.1093/bib/bbae693
18. Ishida S, Sato T, Honma T, Terayama K. Large language models open new way of AI-assisted molecule design for chemists. J Cheminform. 2025;17(1):36. doi:10.1186/s13321-025-00984-8
19. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):248. doi:10.1145/3571730
20. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep. 2023;13(1):14045. doi:10.1038/s41598-023-41032-5
21. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. doi:10.1016/j.patter.2023.100804
22. Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nat Mach Intell. 2020;2(11):665-73. doi:10.1038/s42256-020-00257-z
23. 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
24. Mirza A, Alampara N, Kunchapu S, Ríos-García M, Emoekabu B, Krishnan A, et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. Nat Chem. 2025;17(7):1027-34. doi:10.1038/s41557-025-01815-x
25. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613-22. doi:10.1038/s41591-024-03097-1
26. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open. 2023;6(11). doi:10.1001/jamanetworkopen.2023.43689
27. Cabral S, Restrepo D, Kanjee Z, Wilson P, Crowe B, Abdulnour RE, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. JAMA Intern Med. 2024;184(5):581-3. doi:10.1001/jamainternmed.2024.0295

28. Vaccaro M, Almaatouq A, Malone TW. When combinations of humans and AI are useful: A systematic review and meta-analysis. Nat Hum Behav. 2024;8(12):2293-303. doi:10.1038/s41562-024-02024-1
29. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2
30. 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
31. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. Nat Med. 2020;26(9):1364-74. doi:10.1038/s41591-020-1034-x
32. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. Nat Med. 2020;26(9):1351-63. doi:10.1038/s41591-020-1037-7
33. 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
34. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6(1):120. doi:10.1038/s41746-023-00873-0
35. Harrer S. Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. EBioMedicine. 2023;90:104512. doi:10.1016/j.ebiom.2023.104512
36. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4. doi:10.1038/s41591-020-1041-y
37. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60-9. doi:10.1038/s41591-024-03425-5
38. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med. 2018;378(11):981-3. doi:10.1056/NEJMp1714229