Making Adequate Use of Generative Language Models in Scientific Research

Lázaro Bustio-Martínez[a,*], Vitali Herrera-Semenets[b], Claudia Feregrino-Uribe[c], Jan van den Berg[d]

[a]*Department of Engineering Studies for Innovation, Universidad Iberoamericana Ciudad de México, Prol. Paseo de Reforma ♯ 880, Lomas de Santa Fe, Ciudad de México, 01219, México.*

[b]*Advanced Technologies Application Center, 7[a] ♯ 21812 e/ 218 y 222, Rpto. Siboney, Playa, Havana, 12200, Cuba.*

[c]*National Institute for Astrophysics, Optics and Electronics, Luis Enrique Erro No 1, Santa María Tonantzintla, Puebla, 72840, México.*

[d]*Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.*

References

*Corresponding author

Email addresses: lazaro.bustio@ibero.mx (Lázaro Bustio-Martínez), vherrera@cenatav.co.cu (Vitali Herrera-Semenets), cferegrino@inaoep.mx (Claudia Feregrino-Uribe), j.vandenberg@tudelft.nl (Jan van den Berg)

Making Adequate Use of Generative Language Models in Scientific Research

Abstract

The rapid proliferation of Large Language Models (LLMs) has led to their widespread adoption across scientific disciplines. However, a growing number of academic publications rely solely on *off-the-shelf* solutions applied to familiar tasks, without a well-elaborated methodological innovation, theoretical grounding, or critical reflection. This trend has given rise to a form of superficial research in which generative LLMs are used without sound methodological motivation and critical risk-based reflection on their use. Based on these observations, this paper presents a critical examination of observed applications of LLMs in scientific publications, contrasting performance-driven applications with conceptually rigorous studies that integrate LLMs within structured scientific frameworks. This paper starts by providing a concise description of the technical internal working of LLMs and, based on that, some of their limited capabilities. Next, through a review of recent literature, the analysis identifies epistemological risks, structural incentives, and reproducibility challenges that compromise the integrity of scientific practice. The study concludes by proposing guidelines for the responsible and meaningful use of LLMs in research, emphasizing the need for theoretical alignment, methodological transparency, and the preservation of human epistemic agency.

Keywords: large language models, scientific methodology, epistemology of AI, research integrity, generative AI

1. Introduction

The advancement of Large Language Models (LLMs) has marked a milestone in contemporary artificial intelligence. Systems such as OpenAI's GPT-4 (introduced in 2023) Gallifant et al. (2024), Anthropic's Claude Anthropic

(2024), and DeepSeek's suite of open-source models DeepSeek (2025) have impressed with their ability to produce human-like text, answer specialized queries, and even generate source code Jiang et al. (2025).

The versatility of these models, powered by neural networks with hundreds of billions of parameters[1] trained on massive text corpora, has led to their rapid adoption across multiple domains. A prominent example is ChatGPT which reached tens of millions of users within months and triggered widespread debate about its use in various fields.

In academic and scientific contexts, the emergence of LLMs has sparked both enthusiasm and caution. On one hand, they are seen as potential tools to streamline research tasks: drafting reports, automating literature reviews, generating hypotheses from large knowledge bases, among other applications Binz et al. (2025).

On the other hand, concerns are growing regarding their indiscriminate and uncritical use. Members of the scientific community question whether relying on a pretrained model to address complex problems truly contributes to knowledge advancement, or whether it fosters methodological dependency that stifles theoretical innovation.

Many doubts arise regarding the authenticity of the solutions produced by LLMs: do these results reflect genuine understanding of the problem, or are they merely echoes of patterns present in the training data? Prior studies have pointed out that high performance on a task does not imply comprehension (not even superficial, let alone deep), and that an LLM may behave as a "stochastic parrot," reproducing learned correlations without grasping their real meaning Bender et al. (2021). This concern is further reinforced by the fact that LLMs operate exclusively through tokenization and numerical embeddings (see Figures 1 and 2), without any form of semantic representation or reasoning over abstract concepts. As such, they do not possess self-awareness or metacognitive capabilities; they generate output by statistically modeling context-dependent sequences, not by understanding them. Despite their impressive capabilities, these systems remain fundamentally opaque and prone to severe errors when faced with unfamiliar domains, often misleading users into overestimating their cognitive depth.

This epistemological issue (the distinction between generating correct re-

[1]Parameters in LLMs are the numerical values the model learns during training, which determine how it processes and generates text.

sponses and truly understanding them) is crucial when evaluating the impact of LLMs on science Bender et al. (2021). Similarly, recent events such as the brief public release and subsequent withdrawal of Meta's Galactica model, which was intended to assist with scientific writing Chartier-Edwards et al. (2025), have raised alarms. The model was shown to fabricate documents and references, posing a threat to the integrity of the literature if such tools are used without rigorous verification.

These concerns frame the motivation of the present work. This paper presents a critical analysis of how LLMs are currently used in scientific research, focusing on the distinction between superficial applications of pretrained models and their integration within epistemologically grounded methodological frameworks. By synthesizing representative studies across disciplines, it identifies the systemic pressures that promote tool-centric publishing at the expense of scientific rigor.

The present study makes five principal contributions to current debates on the role of LLMs in scientific research. First, it introduces a structured analytical framework that distinguishes between superficial applications of LLMs, characterized by direct deployment without methodological innovation, and epistemically integrated uses that embed these systems within theory-driven research designs. Second, it develops a risk management perspective grounded in epistemology, identifying specific vulnerabilities such as "hallucination," opacity, and the displacement of human interpretation, all of which threaten the reliability and interpretability of LLM-assisted scientific outputs. Third, the analysis presents empirical evidence of a widespread trend toward methodological superficiality and discursive homogenization in LLM-related publications, revealing systemic distortions in knowledge production that are driven by academic incentives and technological hype. Fourth, this study articulates normative criteria for responsible LLM integration, emphasizing methodological transparency, domain-specific alignment, and the preservation of human epistemic agency as conditions for legitimate scientific use. Fifth, it introduces a structured evaluation matrix comprising six epistemic and methodological axes. This tool translates the proposed normative criteria into an operational instrument that can guide authors, editors, and reviewers in assessing the scientific legitimacy of LLM-based research, both retrospectively and prospectively.

Together, these contributions reframe the use of generative models not as a purely technical decision but as an epistemic and institutional challenge that demands critical governance and evaluative tools.

By combining critical diagnosis with concrete recommendations, this article intervenes in an urgent debate about the epistemic role of generative AI in science, aiming to support a more reflective, rigorous, and conceptually coherent adoption of these technologies.

The remainder of this article is structured as follows: Section 2 introduces the computational and conceptual foundations of LLMs, providing essential context for their scientific evaluation. Section 3 outlines the qualitative methodology used to conduct the literature review and structure the analytical framework. Section 4 presents a critical review of recent academic publications involving LLMs, emphasizing the distinction between superficial applications and conceptually grounded research. Section 5 explores the broader implications of these findings, including systemic incentives, evaluation norms, and the necessity of rigorous integration protocols; it also introduces a risk management perspective in subsection 5.1. Finally, Section 6 synthesizes the main insights and articulates normative recommendations for the responsible and meaningful incorporation of generative AI into scientific research.

2. Technical Foundations of Large Language Models

Understanding the epistemic and methodological implications of LLMs requires a minimal technical overview of their internal functioning. This section outlines the fundamental computational principles underlying contemporary generative models, with emphasis on the representational strategies, training mechanisms, and architectural components that shape their behavior.

2.1. Tokenization and Embedding

The processing pipeline in LLMs begins with *tokenization*, which converts natural language text into discrete units called *tokens* Rajaraman et al. (2025). Unlike traditional word-based models, modern LLMs rely on subword tokenization schemes (*e.g.*, byte pair encoding), allowing them to handle a wide range of linguistic variations. In English, typical vocabularies include approximately 50,000 distinct tokens. Upon receiving a user prompt, the model tokenizes the input and maps each token to a high-dimensional numerical vector, known as an *embedding* Truong et al. (2024).

Figure 1 illustrates this process in a simplified example. Each input sentence is first converted into a sequence of token indices based on a vocabulary lookup table. These indices are then mapped to low-dimensional continuous

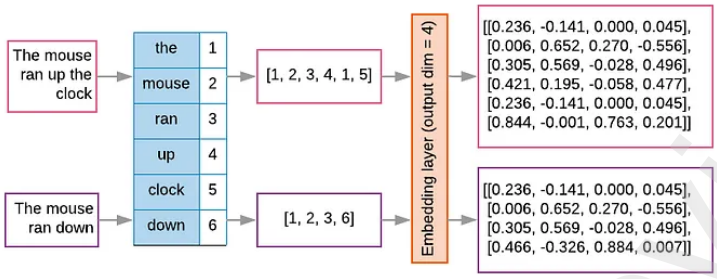


Figure 1: Simplified illustration of tokenization and embedding. Each token is mapped to an index and then projected into a continuous vector space by an embedding layer. This representation forms the input matrix used by subsequent model components Rathinapandi (2023).

vectors via an embedding layer. Although actual LLMs operate in much higher-dimensional spaces (*e.g.*, 12,288 dimensions in GPT-3.5), the figure uses a four-dimensional embedding space for illustrative purposes.

These embeddings encode semantic proximity: tokens with similar meanings are represented by vectors that occupy adjacent positions in the latent space. The result is a matrix of shape $n \times m$, where n is the number of tokens and m the embedding dimension.

2.2. Positional Encoding and Input Representation

Since token embeddings do not encode sequential information, models introduce an additional component: *positional encoding*. This mechanism assigns a vector to each token position, capturing the order of words in a sequence.

Figure 2 illustrates this process. Each input token is first converted into a semantic embedding vector by the embedding model, while its corresponding position in the sequence is separately mapped to a positional encoding vector. These two vectors are then combined through element-wise addition to form a composite representation that preserves both the token's meaning and its location in the input sequence. This augmented matrix serves as the input to the core architecture of the model Vaswani et al. (2017).

2.3. Transformer Architecture

The fundamental computational structure of LLMs is the *transformer* architecture, introduced in Vaswani et al. (2017). Transformers consist of

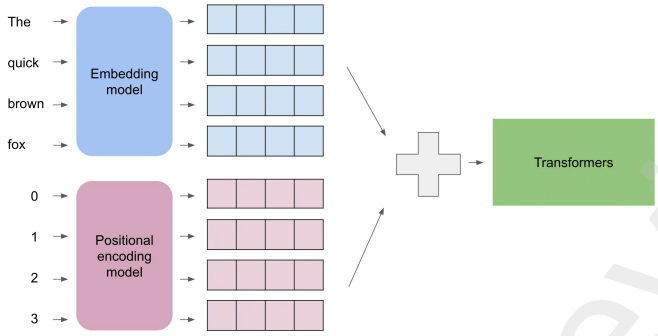


Figure 2: Element-wise addition of semantic embeddings and positional encodings. Each input token is first mapped to a vector by the embedding model, and each token position is separately encoded. The resulting vectors are combined to produce the input matrix used by the Transformer layers Samynathan (2024).

stacked blocks containing multi-head self-attention mechanisms and feed-forward neural networks. Attention layers compute weighted relationships among all tokens in the input, capturing syntactic, semantic, temporal, and even pragmatic dependencies. These attention weights are implemented through query, key, and value matrices, enabling the model to modulate its focus across the sequence.

Figure 3 provides a schematic overview of the original transformer architecture. The model is structured into two main components: the *encoder* (left), responsible for processing the input sequence through repeated applications of self-attention and feedforward layers; and the *decoder* (right), which generates the output sequence by attending to both previous outputs and the encoder's representations. While this encoder–decoder configuration is used in the original formulation for tasks such as machine translation, modern autoregressive LLMs employ only the decoder stack, consisting of masked multi-head self-attention and feedforward layers.

Transformer variants differ in configuration: the original model used an encoder–decoder setup, while autoregressive LLMs like GPT-3 and GPT-4 employ stacks of decoder-only blocks (96 in the case of GPT-3) with a total of 175 billion parameters. Each layer transforms the input representation through a sequence of matrix operations and non-linear activations, propagating context through the network.

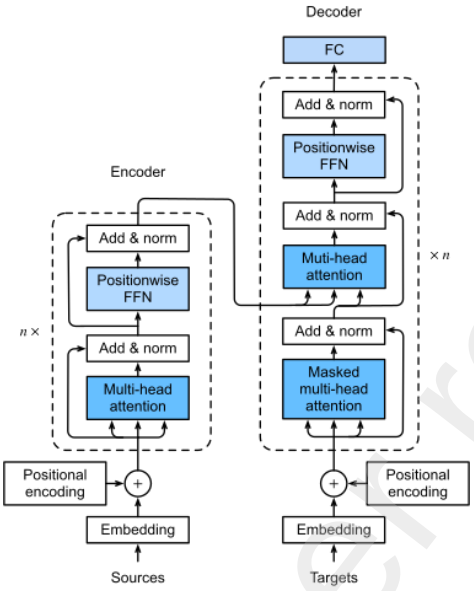


Figure 3: Structure of the original Transformer architecture as proposed in Vaswani et al. (2017). The *encoder* (left) processes the input sequence using self-attention and feed-forward layers, while the *decoder* (right) generates the output sequence through masked attention and encoder-decoder attention mechanisms.

2.4. Training Procedure

Training an LLM involves adjusting the parameters of the transformer using massive corpora of text. These corpora typically consist of billions or trillions of tokens drawn from heterogeneous sources such as books, encyclopedias, news outlets, open-access journals, code repositories, and general web content. The size and diversity of the training data are crucial to ensure broad linguistic coverage and enable the model to generalize across different domains and discourse styles. The objective is to predict the next token in a sequence, given its preceding context. During training, the model iteratively compares its token predictions against ground truth sequences. When incorrect, gradients are computed via backpropagation, and weights are updated accordingly using optimization algorithms. Over time, the model internalizes statistical regularities across diverse linguistic domains.

Notably, this process does not involve grounding in external reality: models learn exclusively from text, without direct perception feedback. The associations formed are statistical rather than semantic, and model outputs reflect correlations learned from data rather than conceptual understanding

Bender et al. (2021).

2.5. Conceptual Limitations

From a cognitive and epistemological standpoint, LLMs differ fundamentally from human reasoning systems. While humans interpret language through reference to lived experience and shared world knowledge, LLMs operate solely on token sequences represented as numbers. This distinction is captured by the semiotic triangle, or *triangle of reference* Ogden and Richards (1923), which illustrates the gap between linguistic signs, mental concepts, and real-world referents. LLMs lack this triangulation entirely; their outputs are disconnected from empirical observations. Consequently, several limitations follow:

- LLMs do not reason over abstract concepts; they generate outputs by modeling statistically plausible continuations of prior tokens.
- LLMs lack a model of the physical world and have no understanding of their internal architecture.
- Their outputs depend entirely on the quality and distribution of training data. When prompted beyond that distribution, LLMs may conclude invented facts, fabricate information, or produce inconsistent responses.

These limitations have direct implications for their use in scientific contexts. As subsequent sections argue, treating LLMs as epistemic agents or theory-generating instruments risks conflating statistical pattern recognition with genuine understanding. Recent work by Apple researchers Apple (2025) further reinforces this diagnosis: their study systematically evaluated a family of so-called *Large Reasoning Models* using synthetic cognitive puzzles with controlled complexity, and found that model performance collapses as reasoning demands increase. Even when models appear to succeed under moderate task complexity, they fail under minimal perturbations, suggesting an absence of abstract reasoning capacity. This evidence supports the view that current LLMs, regardless of parameter size or prompt engineering, lack the structural mechanisms required for genuine cognitive operations. A clear grasp of their technical and conceptual limitations is therefore essential to avoid methodological error and ensure responsible integration in research.

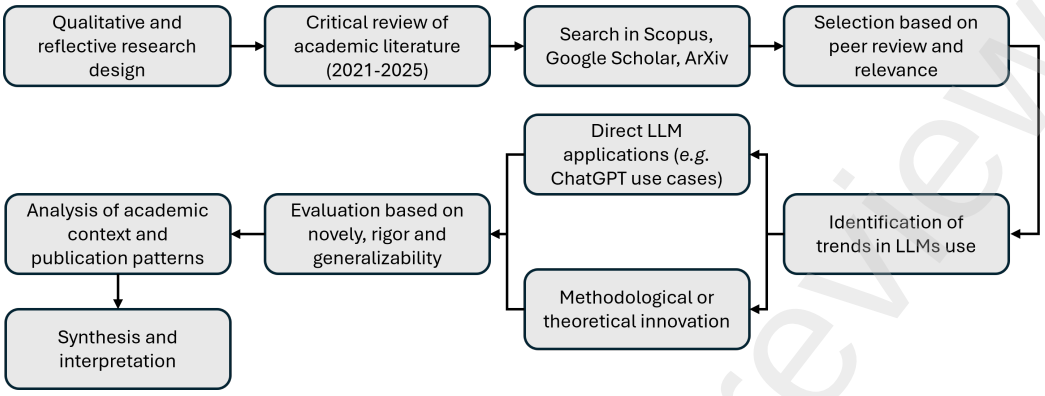


Figure 4: Visual representation of the methodological workflow followed in this study.

3. Methodology

This study adopts a critical qualitative and reflective evaluation framework grounded in a hermeneutic approach. Rather than aiming for statistical generalization, the analysis focuses on interpretive assessment of how LLMs are conceptually and methodologically integrated into scientific research. This method is appropriate for uncovering implicit assumptions, epistemic orientations, and structural patterns that underlie scholarly practices. It aligns with traditions in critical science studies and research evaluation that emphasize the importance of normative and conceptual scrutiny over purely descriptive metrics. The methodological process is summarized in Figure 4, which shows the steps followed in the research process.

The analysis begins with a critical review of academic literature published between 2022 and 2025. This specific time frame was chosen because it captures the period following the public release and widespread adoption of LLMs. These years represent a turning point in the visibility and accessibility of LLMs within the scientific community. By focusing on this window, this study aims to assess how researchers have responded to the widespread availability of these models, both in terms of experimental usage and conceptual integration.

The literature search was conducted using Google Scholar, Scopus, and arXiv. Peer-reviewed journal articles and conference proceedings were prioritized. Preprints were considered only if they had a demonstrable impact on academic discourse (*e.g.*, through citations or adoption in subsequent studies). A limited number of non-academic sources were included only when

directly cited in the scholarly literature or when they contributed substantively to the historical reconstruction of model releases.

From the collected materials, common patterns in how LLMs are used were identified. Two main categories of research were distinguished:

1. Studies centered on direct LLM application: This category comprises works that primarily focus on applying an existing model to a specific domain task without modifying the model or proposing methodological innovations. Such studies often emphasize performance metrics, but rarely engage with deeper theoretical or epistemological questions. Their main goal is to validate the application of LLMs in a specific task, rather than to produce new knowledge or methodological advancements.

2. Studies driven by methodological development: This category encompasses research that, while possibly involving LLMs, is structured around the proposal of new methods, algorithms, or theoretical frameworks. These contributions typically seek to advance understanding, improve interpretability, or explore hybrid systems that integrate LLMs with symbolic reasoning, domain-specific logic, or task-specific optimization techniques.

Each study was evaluated using three main criteria selected based on established principles in scientific methodology that emphasize conceptual contribution, generalizability, and procedural rigor as core indicators of research validity. They align with normative standards in empirical science, including external validity Shadish et al. (2002), reproducibility and methodological transparency OECD (2015), and the advancement of theory through novel insights Kuhn (1962). This evaluative framework enables a systematic appraisal of LLM-related studies in terms of their epistemic substance rather than tool-centric performance.

1. Whether it introduced new conceptual, methodological, or technical insights relevant to the scientific problem it addressed.

2. The extent to which its findings could be generalized beyond the immediate context or dataset.

3. Whether it adhered to standards of scientific rigor, including the use of appropriate baselines, validation methods, and transparency in reporting.

In addition to the content of each study, the surrounding academic context was examined. This included analyzing publication patterns, the frequency of similar submissions within specific conferences or journals, and the influence of prevailing research incentives, such as pressure to publish rapidly or demonstrate technical novelty.

The resulting analysis does not aim to quantify trends in a statistical sense, but rather to develop a structured and theoretically informed reflection on current research practices. By contrasting different forms of engagement with LLMs, this study seeks to contribute to a deeper understanding of their role in contemporary scientific discourse.

Given the heterogeneity of the reviewed studies and the absence of a shared theoretical baseline for evaluating LLM-based research, this study does not adopt a hypothesis-testing paradigm. Instead, it formulates a set of guiding questions to structure the analysis and enable comparative evaluation across cases.

These questions are not merely descriptive; they aim to uncover the epistemic orientation, methodological rigor, and scientific relevance of different integration strategies. Specifically, the analysis is guided by the following inquiries:

- To what extent are LLMs being used as instruments for task automation versus components of conceptually grounded scientific workflows?
- What methodological patterns characterize each usage type, and what epistemic risks do they entail?
- Under what conditions can the integration of LLMs contribute to scientific advancement rather than superficial innovation?

These questions inform the classification criteria presented above and shape the analytical distinction between revised studies. Rather than seeking generalizable metrics, the aim is to elucidate patterns of use that bear consequences for scientific validity, interpretability, and long-term knowledge production.

To operationalize the qualitative analysis, two core analytical categories were defined based on how LLMs are incorporated into research practices:

- Direct application: Refers to studies that use pretrained LLMs as *off-the-shelf* tools for performing familiar tasks (*e.g.*, summarization, exam

solving, information extraction) without modifying the underlying scientific methodology or introducing conceptual innovation.

- Conceptual integration: Denotes research efforts where LLMs are embedded within structured methodological, theoretical, or technical frameworks. These studies often aim to extend or transform existing scientific practices by aligning model capabilities with epistemic goals, domain-specific constraints, or structured workflows.

This dichotomy enables a structured comparison of the epistemological and methodological roles that LLMs play across diverse fields. The categories were applied consistently throughout the review and form the basis for the classification presented in Section 4.

To anticipate the analytical synthesis developed in Section 5, this study also formulated a complementary set of evaluative criteria derived from the observed patterns in LLM integration. These six criteria (methodological justification, theoretical alignment, experimental transparency, interpretive control, epistemic risk management, and conceptual contribution) were extracted inductively through iterative reading of the selected studies and validated against normative principles in scientific methodology. Rather than being imposed a priori, they emerged as operational dimensions that consistently distinguished robust from superficial uses of generative models. They are introduced here to clarify the analytical scaffolding used in subsequent sections and are consolidated into an evaluative matrix in Section 5.2.

3.1. Corpus Selection Protocol

To ensure the transparency and reproducibility of the literature review process, a structured protocol was used to identify and select the set of studies analyzed in this work. The procedure consisted of the following steps:

1. Sources consulted: Primary searches were conducted on Google Scholar, Scopus, and arXiv between January 2022 and May 2025. Supplementary material was retrieved from cross-references found within selected articles.

2. Search terms: Queries combined terms such as "LLMs," "large language models," "ChatGPT," "GPT-4," "scientific research," "methodology," "integration," and "AI-assisted science". Boolean combinations were used to increase coverage, e.g., "ChatGPT AND scientific writing" or "LLM AND systematic review".

3. Inclusion criteria:

   (a) The article must describe a scientific research process in which an LLM was directly used or critically analyzed.

   (b) The LLM must be a general-purpose model, not a small domain-specific transformer.

   (c) The paper must present empirical results, a technical framework, or a conceptual critique that contributes to understanding the use of LLMs in research contexts.

   (d) The document must be dated between 2022 and 2025.

4. Exclusion criteria:

   - Articles that merely speculate on potential uses of LLMs without concrete implementation.
   - Papers centered on education or ethics without empirical components related to scientific practice.
   - Preprints with no citation record, community uptake, or methodological contribution.

5. Selection process: After removing duplicates, titles and abstracts were screened for relevance. Full texts were reviewed to determine compliance with inclusion criteria. A final corpus of 20 studies was retained for structured analysis.

This typological distinction between direct and structured uses provides the analytical basis for evaluating the empirical corpus. In the next section, this framework is applied to a selected set of LLM-based studies, beginning with their distributional patterns and bibliometric properties.

4. Literature Review

This review examines the current integration of LLMs within scientific research, based on the analytical framework detailed in Section 3. Instead of compiling a descriptive inventory of publications, it categorizes scholarly output into two distinct types: (i) studies that employ pretrained LLMs to perform established tasks without methodological innovation, and (ii) studies that embed LLMs within broader conceptual or technical frameworks to

generate new scientific knowledge. This structure enables a systematic evaluation of the roles LLMs play in scientific inquiry and the degree to which they contribute to epistemic advancement.

4.1. Bibliometric Profile of the Reviewed Corpus

This section provides a bibliometric characterization of the studies analyzed in this review. The aim here is to identify patterns related to publication dates, disciplinary focus, access models, etc., thereby situating the corpus within broader scientific publishing trends. Figure 5 show the results of the bibliometric study.

Figure 5a shows the distribution of publications by year. The corpus spans from 2022 to 2025, with a marked concentration in 2024, which accounts for 14 of the 20 reviewed studies. This suggests an intensification of publication activity during that year, though causal factors cannot be determined from bibliographic metadata alone.

In terms of disciplinary distribution, the articles span a broad range of domains (Figure 5b). The most represented areas are education and science policy, each accounting for three studies. Health informatics appears in two studies. All other domains (including medicine, law, and neuroscience among others) are represented by a single publication. This indicates a wide but shallow disciplinary spread across the reviewed corpus.

Regarding publication venues, the articles are distributed across a wide range of journals, with a small number of outlets appearing multiple times (Figure 5c). The most frequent channels are the "Journal of Biomedical Informatics," "Digital Medicine," and "Frontiers in Research Metrics and Analytics," each contributing two articles. Other contributions are spread across academic journals associated with major publishers such as Springer, Nature, Elsevier, and Cambridge, each represented by a single article. The inclusion of a corporate outlet, "Apple Education Reports," reflects the participation of private-sector entities in the discourse on LLMs.

Figure 5d shows the distribution of publications by editorial group. The most represented publisher is Springer, followed by Elsevier and Nature Portfolio. Other contributions are distributed among Cambridge University Press, PLOS, Apple, Frontiers, Taylor & Francis, and additional specialized outlets. Figure 5e categorizes the type of publication channel: the majority of studies were published in academic journals, with one appearing in an open-access repository and one in a corporate outlet.

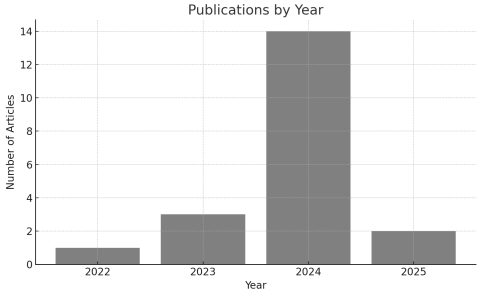
Finally, a lexical analysis of article titles reveals recurring linguistic patterns. As shown in the wordcloud in Figure 5f, the most frequent terms include "LLMs," "ChatGPT," "GPT," "education," and "scientific," reflecting dominant themes in the reviewed corpus. The presence of phrases such as "LLMs in..." suggests a tendency toward framing LLMs as tools embedded in specific disciplinary contexts. However, no consistent use of expressions like "using GPT-4 to..." or "ChatGPT and..." was observed. These lexical trends provide a surface-level indication of topic focus, but further analysis would be needed to assess narrative convergence or epistemic uniformity.

While the bibliometric profile provides a descriptive overview of publication patterns, disciplinary spread, and access models, it does not by itself address the epistemic quality or methodological depth of LLM-related research. To move beyond surface-level trends, the following subsections present a qualitative analysis of how LLMs are integrated into scientific workflows. This interpretive evaluation builds upon the contextual insights offered by the bibliometric data, but shifts the focus toward the conceptual rigor, theoretical framing, and research relevance of selected studies. The transition from quantitative description to qualitative critique allows for a more comprehensive understanding of the current landscape of generative model usage in science.

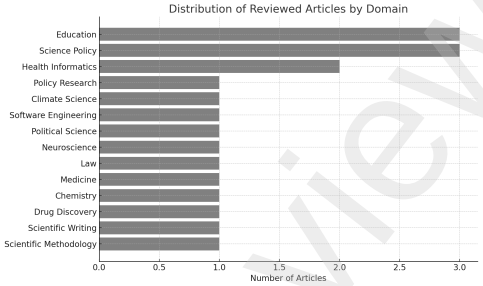
4.2. Applications of Pretrained LLMs to Established Tasks

Unlike earlier models designed for narrow applications, current LLMs (often referred to as *foundation models*) can perform a wide range of tasks such as summarization, translation, question answering, and code generation, without requiring task-specific fine-tuning. This versatility stems from their exposure to heterogeneous training data, including encyclopedic, academic, and web-based sources. However, their outputs remain constrained by the statistical properties and biases of this data, which can lead to asymmetric performance across languages or the reinforcement of social stereotypes.

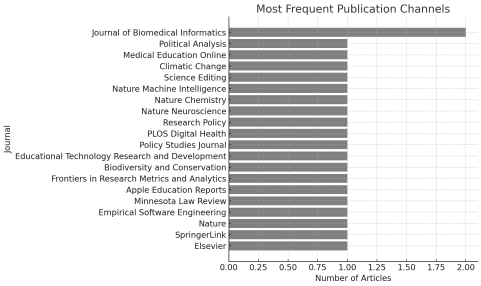
Initial academic attention focused on LLMs' ability to emulate human performance in formal assessments. In education, studies have explored whether systems like ChatGPT can answer multiple-choice questions, compose essays, or solve classroom problems Hosseini et al. (2023). Bordt et al. Bordt and von Luxburg (2023) evaluated ChatGPT's performance in an undergraduate computer science exam using blind grading, finding that the model passed but without demonstrating conceptual understanding. Sparrow et al. Sparrow and Flenady (2025) further criticize such uses as epistemi-



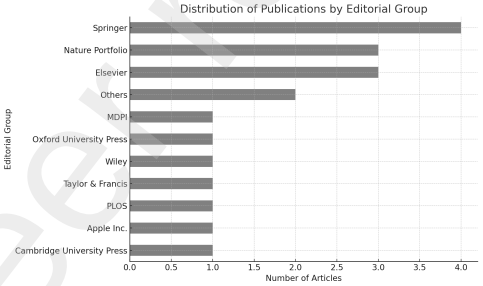
(a) Annual distribution of reviewed publications (2022–2025).



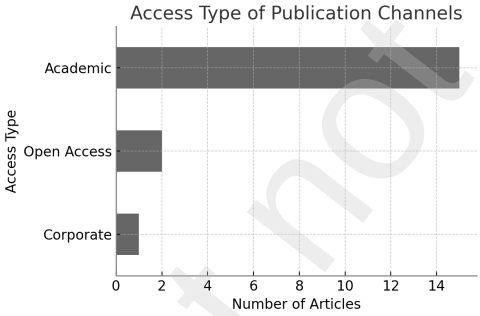
(b) Distribution of reviewed articles by domain.



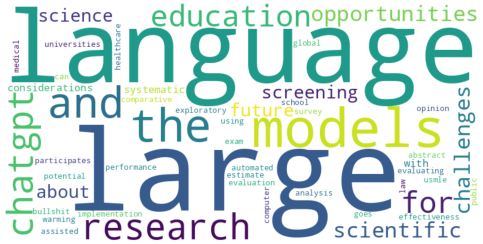
(c) Most frequent publication channels among the reviewed studies.



(d) Distribution of publications by editorial group.



(e) Access type of publication channels: academic, open-access, and corporate.



(f) Wordcloud of the 50 most frequent terms in article titles.

Figure 5: Bibliometric visualizations of the reviewed corpus.

cally shallow, arguing that LLMs lack the intentionality and moral reasoning essential for pedagogical authority. In medicine, models like GPT-4 have passed the United States Medical Licensing Examination (USMLE) without domain-specific training Kung et al. (2023). Legal studies report similar re-

sults: Choi et al. Choi et al. (2022) found that ChatGPT could pass law school exams, though with performance below the human average.

These cases illustrate the appeal of LLMs for standardized evaluation tasks, where performance metrics often overshadow concerns about conceptual understanding Apple (2025). This pattern extends to other domains. In climate science, LLMs have been used to transform structured datasets into accessible summaries for public audiences Lee et al. (2024). This is consistent with broader trends in the ML literature, where models are increasingly applied to convert complex scientific information into human-interpretable formats, albeit with limited epistemic innovation Rolnick et al. (2022). In software engineering, Hou et al. Hou et al. (2024) identified widespread use of LLMs for code generation, bug fixing, and question answering, yet noted a lack of attention to robustness or interpretability. In political science, Kyuwon et al. You et al. (2023) employed GPT-based systems to extract structured information from legislative texts, automating a task previously dependent on manual coding.

Further empirical studies confirm the prevalence of instrumental applications. Li et al. Li et al. (2024) and Landschaft et al. Landschaft et al. (2024) used off-the-shelf models like GPT-4 and Claude to automate abstract screening in systematic reviews and policy research, reporting high recall but minimal methodological adaptation. These systems were applied without model modification, prompt engineering, or reframing of the underlying task. Galli et al. Galli et al. (2025) provide a detailed methodological overview of LLM-based abstract screening in biomedical systematic reviews, emphasizing the practical benefits of zero-shot classification while underscoring the need for rigorous prompt engineering and human oversight. Despite the technical sophistication of their proposed workflow, the approach maintains the established structure of systematic reviews without introducing conceptual reconfiguration, thereby reinforcing the instrumental pattern observed across domains. In science policy, Fecher et al. Fecher et al. (2025) conducted a Delphi study involving 72 experts to assess the systemic implications of LLM adoption. While participants acknowledged potential benefits for efficiency, the study highlighted widespread concerns about superficial applications, epistemic opacity, and the erosion of quality control mechanisms. These findings suggest that the inclusion of LLMs often serves to signal innovation rather than substantively reconfigure scientific practice. Mammides et al. Mammides and Papadopoulos (2024) further examined the role of LLMs in interdisciplinary research through an expert-based analysis. Their findings

indicate that LLMs are predominantly used as auxiliary instruments, applied to tasks such as translation, summarization, or idea generation, without modifying the epistemic goals or methodological structure of the research. The study emphasizes that, although potentially useful, these applications remain operational in nature and do not involve conceptual or theoretical integration.

Collectively, these examples typify a category of research that treats LLMs as interchangeable tools within established analytic workflows. While often effective at automating labor-intensive processes, such uses seldom engage with underlying scientific theories, methodological innovation, or epistemological critique.

4.3. Innovative Frameworks Leveraging LLM Capabilities

In contrast to instrumental deployments, a growing body of research incorporates LLMs within structured scientific frameworks designed to extend their functionality beyond surface-level automation. One such development is Google's Titans architecture, which introduces a cognitively inspired memory system that separates short-term, long-term, and persistent memory components. This design allows models to handle sequences exceeding two million tokens and dynamically retain information based on prediction errors, resembling surprise-driven learning in human cognition Behrouz et al. (2024). Such architectural enhancements improve context management, inference-time adaptation, and long-range generalization, indicating a shift toward more flexible and cognitively aligned model design.

Concrete implementations of these ideas appear across disciplines. In neuroscience, Luo et al. Luo et al. (2025) report that BrainGPT, a domain-specific LLM trained exclusively on curated neuroscience literature, surpassed expert researchers in predicting experimental outcomes. Using the BrainBench benchmark, BrainGPT achieved approximately 86% accuracy compared to 63% for human experts, illustrating that a calibrated, forward-looking framework can enable LLMs to contribute to scientific inference.

In chemistry, Boiko et al. Boiko et al. (2023) developed Coscientist, a GPT-4–driven robotic laboratory infrastructure that autonomously designs, plans, and executes complex chemical protocols, from literature mining to physical experimentation with minimal human oversight. In biomedical informatics, Toufiq et al. Toufiq et al. (2023) used GPT-4 to prioritize candidate genes in gene–disease networks, identifying latent molecular relationships subsequently validated by experts. In drug discovery, Song et al. Song

et al. (2025) describe PharmaSwarm, a coordinated multi-agent system based on GPT models that formulates and tests research hypotheses, resulting in confirmed compound–target associations via graph reasoning and pathway integration.

Other studies explore how LLMs can support the research writing process. Procko et al. Procko et al. (2023) proposed a structured workflow that delegates technical drafting tasks, such as generating abstracts or outlining sections, to LLMs, while retaining interpretive and argumentative responsibility for human authors. Wang et al. Wang et al. (2024) introduced the Language Symbolic Programs (LSP) framework, which allows LLMs to generate intermediate symbolic steps interpretable by humans and executable by machines. This design enhances generalization and transparency, enabling integration of LLMs within scientific workflows that require procedural clarity and traceable inference. Lissack Lissack and Meagher (2024) adds a governance perspective, arguing for institutional mechanisms that preserve authorship, accountability, and scientific agency in LLM-mediated knowledge production.

Despite their potential, such frameworks face critical limitations. LLMs lack semantic understanding and function solely through statistical correlations between tokens, without mechanisms for causal inference or conceptual reasoning Bender et al. (2021); Cuskley et al. (2024). A closely related problem is the occurence of what authors termed "hallucination" Navigli et al. (2023); Shanahan (2024); Binz et al. (2025), which poses serious risks in scientific applications due to the generation of plausible but factually incorrect content. The failure of Galactica illustrates these dangers, as the system produced authoritative-sounding outputs that were often inaccurate or misleading Chartier-Edwards et al. (2025).

These examples demonstrate that, when embedded within disciplined methodological frameworks, LLMs can serve as tools for scientific discovery rather than mere task automation. By aligning their integration with domain-specific knowledge and structured oversight, such approaches illustrate a pathway toward epistemically meaningful applications. However, their success depends not only on technical innovation, but also on critical awareness of the models' limitations and the institutional conditions under which they are deployed.

4.4. Critical Assessment and Field-Level Patterns

While several studies demonstrate conceptually grounded uses of LLMs, other research exhibits limitations that raise concerns about their epistemic and methodological value. Three critical issues are especially prominent: (i) systemic bias, (ii) lack of reproducibility, and (iii) superficiality at scale.

LLMs trained on large-scale internet data are prone to reproducing and amplifying societal biases. Empirical studies show that these models respond differently to prompts that vary only in demographic details, reinforcing stereotypes or skewing outcomes in areas such as healthcare and social science Hosseini et al. (2023); Navigli et al. (2023). Ullman et al. Ullman (2023) further demonstrate that LLMs fail systematically on trivial modifications to reasoning tasks, such as Theory of Mind evaluations, despite passing the original version. These failures highlight that apparent competence often masks brittle pattern-matching, rather than genuine understanding. Such distortions and fragilities undermine claims of neutrality, generalization, and fairness, especially in high-stakes or cognitively demanding research settings.

Reproducibility remains an unresolved challenge. Many state-of-the-art models are closed systems whose training data, architectural details, and hyperparameters are proprietary, preventing independent verification and undermining foundational principles of scientific transparency Pineau et al. (2021). Even in the case of open models, such as DeepSeek's R1, slight variations in prompt structure or model versioning can lead to inconsistent outputs. Furthermore, revised publications frequently omit critical details about prompt engineering, filtering criteria, or evaluation procedures, further hindering replication efforts.

The prevalence of superficial applications is further supported by recent bibliometric research. Gencer Gencer and Gencer (2025) documents an exponential rise in LLM-related healthcare publications: from a single article in 2021 to 337 in 2023, with an additional 238 recorded by mid-2024. The majority of these contributions focus on automating documentation or producing summaries, without introducing innovations. A keyword analysis revealed dominant clusters centered on "ChatGPT" and "scientific writing," reflecting a tendency toward implementation rather than conceptual advancement. These patterns are corroborated by Kobak et al. Kobak et al. (2025), who conducted a lexical analysis of PubMed abstracts and found that over 13% of 2024 articles exhibit stylistic markers of LLM-generated content, particularly in title structure and vocabulary. According to the authors, this trend points

to a growing homogenization of scientific language and a potential erosion of epistemic diversity.

To synthesize the preceding analysis, Tables 1 and 2 present representative studies discussed in this review. Table 1 includes works that apply LLMs directly to established workflows without altering methodological foundations, while Table 2 highlights conceptually informed integrations where LLMs are embedded within structured scientific frameworks.

Table 1: Direct use of LLMs in scientific research (2021–2025).

Author(s)	Year	Domain	Use Type	Critical Commentary
Hosseini et al. Hosseini et al. (2023)	2023	Education	Direct	Tests ChatGPT on educational tasks; prioritizes scores over pedagogy.
Sparrow et al. Sparrow and Flenady (2025)	2025	Education	Direct	Critiques factual accuracy without conceptual grounding.
Kung et al. Kung et al. (2023)	2023	Medicine	Direct	Evaluates GPT-4 on USMLE; no domain adaptation.
Choi et al. Choi et al. (2022)	2022	Law	Direct	Assesses exam performance; lacks legal reasoning.
Apple Apple (2025)	2024	Education	Direct	Focuses on metrics; neglects depth.
Fecher Fecher et al. (2025)	2024	Science Policy	Direct	Superficial LLM use to suggest novelty.
Mammides Mammides and Papadopoulos (2024)	2024	Science Policy	Direct	Superficial LLM use to suggest novelty.
Li et al. Li et al. (2024)	2024	Health Informatics	Direct	Abstract screening with GPT-4; effective but unchanged methodology.
Landschaft et al. Landschaft et al. (2024)	2024	Policy Research	Direct	Title/abstract review; no conceptual framing.
Lee et al. Lee et al. (2024)	2024	Climate Science	Direct	Simulates opinion; shows bias by demographics.
Hou et al. Hou et al. (2024)	2024	Software Engineering	Direct	Surveys 400+ studies; little focus on robustness.
Kyuwon et al. You et al. (2023)	2024	Political Science	Direct	Automates info extraction; no theoretical advance.
Galli et al. Galli et al. (2025)	2025	Health Informatics	Direct	Describes zero-shot GPT-4 screening in reviews; maintains traditional workflow structure.

Table 2: Conceptual integration of LLMs in scientific research (2021–2025).

Author(s)	**Year**	**Domain**	**Use Type**	**Critical Commentary**
Luo et al. Luo et al. (2025)	2024	Neuroscience	Conceptual	BrainGPT outperforms experts; domain-specific corpus.
Boiko et al. Boiko et al. (2023)	2023	Chemistry	Conceptual	GPT-4 automates experimental design in robotics.
Toufiq et al. Toufiq et al. (2023)	2024	Biomedical Informatics	Conceptual	Gene prioritization validated by experts.
Song et al. Song et al. (2025)	2024	Drug Discovery	Conceptual	GPT-based agents generate verifiable hypotheses.
Procko et al. Procko et al. (2023)	2023	Scientific Writing	Conceptual	LLMs draft structure; humans handle interpretation.
Lissack Lissack and Meagher (2024)	2024	Science Policy	Conceptual	Calls for governance to safeguard scientific agency.
Wang et al. Wang et al. (2024)	2024	Scientific Methodology	Conceptual	Introduces LSP framework: interpretable symbolic steps for traceable inference.

As Tables 1 and 2 illustrate, most direct-use cases involve routine task automation without theoretical engagement, while conceptually grounded implementations, although comparatively rare, demonstrate deeper epistemic integration. This contrast underscores the need for robust evaluative criteria to differentiate between superficial adoption and substantively grounded innovation.

5. Towards Responsible LLM Use

The preceding analysis reveals a structural asymmetry in how LLMs are currently integrated into scientific research. While some studies demonstrate conceptually grounded and methodologically rigorous applications, a large proportion exhibit superficial uses driven by performance metrics, tool availability, or academic expediency. This section addresses both the systemic incentives that facilitate such patterns and the normative conditions necessary for responsible LLM integration.

Several structural factors help explain the dominance of direct-use implementations. In highly competitive research environments, speed and output volume are frequently prioritized over conceptual depth. Pretrained models such as GPT-4 offer a low-cost entry point for producing publishable results without requiring the design of novel methodologies or multi-stage experimental frameworks. This makes them particularly attractive in time-constrained contexts such as graduate theses or conference-driven publication cycles. The relative ease of deploying a general-purpose model can displace more demanding forms of scientific reasoning.

This dynamic is reinforced by the visibility and prestige currently associated with generative AI. Terms such as "ChatGPT" or "GPT-4" function as lexical attractors in academic publishing, increasing the likelihood of citation, funding, and media dissemination. As Gencer's bibliometric study indicates, this incentive structure contributes to the exponential rise of LLM-related publications that focus on automation and surface-level integration Gencer and Gencer (2025). Kobak et al. Kobak et al. (2025) further confirm this trend through large-scale lexical analysis, showing a convergence in title structures and vocabulary associated with LLM studies, suggestive of thematic homogenization.

This pattern has been characterized by critics such as Arawjo as producing "wrapper papers," in which the presence of an LLM becomes the primary contribution, displacing the research question itself Arawjo (2024). While not

inherently invalid, such work risks conflating novelty with tool invocation and contributes to a dilution of scientific standards.

Another contributing factor is the accessibility of LLMs through APIs and public interfaces. Unlike earlier waves of AI research that required specialized infrastructure, current models can be deployed by any user with minimal programming skills. While this democratization has benefits, it also reduces the epistemic cost of experimentation and can encourage unsupervised or uncritical usage. In many cases, essential stages of problem framing, domain alignment, or theoretical contextualization are bypassed in favor of tool demonstration.

Metric-based evaluation further amplifies this tendency. Benchmark scores such as accuracy or recall are often accepted as sufficient evidence of model utility, even when they do not reflect interpretability, conceptual understanding, or robustness. As documented by Binz et al. Binz et al. (2025), and reinforced by Apple Apple (2025), LLMs frequently fail under minor increases in task complexity, revealing that high scores may mask shallow heuristics rather than demonstrate genuine comprehension. Ullman's findings on Theory of Mind perturbations Ullman (2023) similarly illustrate the brittleness of current architectures under minimal semantic shifts.

Taken together, these pressures create an environment in which LLMs are frequently adopted without adequate methodological justification. However, this does not imply that LLM-based research is inherently epistemically weak. Rather, the distinction lies in the conditions under which models are integrated.

The most compelling cases identified in this review embed LLMs within structured, domain-aware workflows that preserve human interpretation and scientific accountability. For instance, Luo et al. Luo et al. (2025) demonstrate that domain-specific fine-tuning, grounded in curated literature and evaluated through predictive accuracy against human experts, enables LLMs to support scientific inference rather than merely automate tasks. Similarly, Song et al. Song et al. (2025) propose a multi-agent framework for hypothesis generation and testing in drug discovery, where model outputs are validated through graph reasoning and pathway analysis. Wang et al. Wang et al. (2024) introduce interpretable symbolic intermediaries that increase the transparency and traceability of LLM reasoning, while Procko et al. Procko et al. (2023) offer a structured workflow for scientific writing that clearly separates generative assistance from argumentative responsibility. These examples illustrate that responsible integration is not only feasible, but epis-

temically productive when grounded in clear theoretical frameworks and interpretive oversight.

While these studies represent significant methodological progress, they do not eliminate the epistemic limitations inherent to LLMs. In the case of BrainGPT Luo et al. (2025), predictive superiority is achieved within a constrained domain using a curated corpus and a task-specific benchmark, limiting generalizability. PharmaSwarm Song et al. (2025) relies on internal validation loops mediated by other LLMs, introducing potential opacity in the verification process and raising concerns about circular evaluation. The LSP framework Wang et al. (2024), although enhancing traceability through symbolic intermediaries, still depends on statistically generated outputs without guaranteeing semantic understanding. These limitations do not negate the value of the contributions, but underscore the need for critical oversight even in conceptually sophisticated integrations.

To consolidate the normative position developed in this section, Table 3 outlines the primary challenges observed in current LLM research practices and the corresponding recommendations derived from the analysis.

As the table indicates, addressing the epistemic risks associated with LLMs requires not only technical guidelines, but also a reassertion of methodological standards and interpretive responsibility. These are not peripheral concerns: they define the boundary between generative output and scientific knowledge. This contrast is not merely qualitative, but also reflected in the distribution of studies analyzed. As shown in Figure 6, the majority of reviewed works (65%) adopt LLMs as direct instruments for automating existing tasks, while only a minority (35%) integrate them into structured conceptual frameworks. This imbalance underscores the structural asymmetry identified throughout the discussion and reinforces the need for normative criteria that distinguish superficial application from epistemically grounded innovation.

The above recommendations outline general principles for the responsible use of LLMs in scientific research. However, to render these principles actionable within institutional and project-level contexts, they must be embedded within a structured framework. One such approach involves the management of epistemic risks, which provides a normative and operational lens through which to assess the potential harms and limitations of generative AI. The next subsection develops this perspective in greater detail.

Table 3: Summary of critical issues and normative responses discussed in Section 5.

Observed Trend or Challenge	**Recommended Scientific Response**
Superficial use driven by publication pressure and tool availability.	Prioritize research questions over model availability; assess whether LLM use is warranted by the problem context.
Prestige incentives promoting "wrapper papers".	Anchor contributions in domain-relevant theory or empirical gaps rather than tool novelty.
Metric-based validation privileging performance over understanding.	Emphasize interpretability, causal modeling, and alignment with theoretical constructs over benchmark scores.
Democratized access without methodological training.	Encourage critical literacy in LLM use; integrate training on epistemic risks and limitations.
Opacity in model use and reporting.	Document all experimental variables (model version, prompt, parameters) to ensure transparency and reproducibility.
Substitution of human interpretation by model output.	Retain researcher responsibility in interpreting, validating, and contextualizing results.
Lack of integration with traditional methods.	Embed LLMs within hybrid designs combining symbolic reasoning, empirical testing, or qualitative analysis.

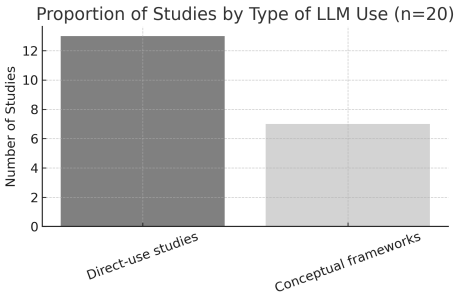


Figure 6: Proportion of studies using LLMs as direct tools versus structured conceptual frameworks.

5.1. Generative AI as a Risk Management Challenge

In light of the limitations documented throughout this study, the integration of LLMs into scientific workflows must be reframed as a problem of epistemic risk management. These risks are not technical malfunctions, but systematic vulnerabilities that threaten the reliability, interpretability, and theoretical coherence of scientific outputs involving LLMs. Three classes of epistemic risk are particularly salient:

a) LLMs generate plausible text devoid of semantic understanding. This can result in "hallucinations," biased completions, or factually incorrect content presented with unwarranted confidence.

b) The reasoning trajectory between prompt and output is often inaccessible. This impedes scrutiny, limits replicability, and weakens the epistemic transparency essential to scientific inquiry.

c) When LLMs are used without theoretical framing or human interpretation, their outputs may replace scientific reasoning instead of supporting it. This displaces methodological accountability with superficial automation.

These risks are not hypothetical. Ullman et al. Ullman (2023) showed that minor perturbations in reasoning tasks cause LLMs to fail systematically, revealing brittleness beneath apparent competence. Apple Apple (2025) demonstrated that models tuned for reasoning collapse under increasing task complexity, indicating that benchmark performance may mask shallow statistical heuristics. These findings align with earlier concerns raised by

Bender et al. Bender et al. (2021) and Cuskley et al. Cuskley et al. (2024) about the limitations of stochastic text generation as a proxy for understanding.

Conversely, some studies show how these risks can be mitigated through careful experimental design. Toufiq et al. Toufiq et al. (2023) employed expert validation to filter outputs from LLM-guided biomedical pipelines. Boiko et al. Boiko et al. (2023) embedded LLMs within a robotic chemistry platform, where outputs were constrained by empirical feedback loops. Wang et al. Wang et al. (2024) enforced interpretability via intermediate symbolic representations aligned with human reasoning steps. These cases do not eliminate risk, but they manage it through structured integration.

This perspective resonates with the broader sociological framing of modern science as embedded in a "risk society" Beck (1992), where technological complexity introduces endogenous threats to knowledge reliability. Under this paradigm, risk is not residual; it is constitutive. Scientific practices involving LLMs must therefore be designed with these structural uncertainties in mind.

The ISO 31000 standard conceptualizes risk as "the effect of uncertainty on objectives" ISO (2018). Applied to LLMs, this implies a shift from tool deployment to uncertainty management. A pragmatic approach to this challenge involves iterative control processes adapted from standard risk frameworks:

1. Define the intended scientific use and clarify epistemic objectives.
2. Identify risks related to bias, "hallucination," opacity, and interpretive ambiguity.
3. Implement structural safeguards: prompt audits, expert validation, task-specific constraints.
4. Document model versioning, prompt design, parameters, and evaluation metrics to support reproducibility.
5. Monitor system behavior and update protocols as models and tasks evolve.

This framework does not reject generative models but calls for their responsible use within hybrid epistemic systems. LLMs are best understood

not as autonomous agents of discovery, but as fallible computational instruments embedded within broader workflows. Their epistemic value depends not on their fluency or scale, but on the rigor of the methods used to constrain, verify, and interpret their outputs.

Scientific research must thus treat LLMs as epistemic accelerators, not as substitutes for conceptual reasoning. Risk-aware design is essential not only to avoid error, but to preserve the integrity of the knowledge production process itself.

5.2. Towards an Operational Framework for the Scientific Evaluation of LLM Usage

The normative recommendations articulated in Section 5, along with the risk-based framing presented in Section 5.1, require systematic mechanisms for implementation. This subsection consolidates the evaluative axes developed throughout the discussion into a formal framework designed to operationalize the criteria proposed. The goal is to translate the theoretical principles outlined above into concrete criteria that can guide peer review, editorial decisions, and research design.

The normative principles outlined in Section 5, along with the evidence discussed in Sections 4.2 to 4.4, lead to the formulation of six operational questions that should guide the evaluation of research involving LLMs:

1. *Is there an explicit methodological justification for the use of the generative model?:* The article must explain why the inclusion of an LLM is necessary to address the research question or solve the problem, beyond mere technical availability.

2. *Is the model aligned with theoretical or methodological frameworks relevant to the domain?:* The integration of the LLM must be contextualized within specific disciplinary approaches, not treated as a generic, decontextualized insertion.

3. *Are the experimental conditions under which the model was used explicitly documented?:* The article must report the model version, prompt design, parameters used, and evaluation procedures in sufficient detail to ensure replicability.

4. *Is human oversight maintained in the interpretation of model outputs?:* A clear distinction must be drawn between automatically generated

content and the scientific interpretation, validation, or inference performed by researchers.

5. *Are the epistemic risks associated with LLMs addressed and mitigated?:* The study must acknowledge and counteract issues such as "hallucination," bias, opacity, or interpretive ambiguity, as discussed in Sections 2.5 and 5.1.

6. *Does the LLM contribute a verifiable conceptual or methodological innovation?:* The value of the study must reside not in the use of the model itself, but in how it enables non-trivial inference, hypothesis generation, or methodological advancement.

These six axes can be translated into an evaluation matrix that enables editorial boards, academic committees, and practitioners to incorporate objective criteria when assessing the quality, originality, and epistemic legitimacy of research involving generative artificial intelligence. This matrix is presented in Table 4.

Table 4 is structured into three columns: (i) the name of each criterion, corresponding directly to the six evaluative questions above; (ii) a condensed formulation of each evaluative question, adapted for operational use; and (iii) a column of indicators that describe concrete conditions under which each criterion can be considered satisfied. These indicators are not exhaustive, but provide a minimal threshold for assessing scientific rigor, interpretability, and epistemic responsibility in LLM-assisted research.

The matrix presented in Table 4 is designed as a practical instrument that can be adopted in both prospective (i.e., during research planning) and retrospective (i.e., during manuscript evaluation) evaluations. In prospective contexts, it provides a checklist for authors and research designers to ensure conceptual alignment, methodological clarity, and interpretive responsibility before model deployment. Retrospectively, it can be used by reviewers, editors, and funding agencies to assess whether submitted work meets minimal scientific standards when incorporating generative models. Unlike narrative recommendations, this tabular instrument supports structured comparison, repeatable judgment, and the institutionalization of evaluative baselines across disciplinary contexts.

Such tools would help reduce the prevalence of what Arawjo describes as "wrapper papers", which are publications in which the presence of an

Table 4: Operational matrix for evaluating the scientific use of LLMs in academic publications.

Criterion	**Evaluative Question**	**Indicators of Compliance**
Methodological Justification	Is the use of the LLM explicitly justified in relation to the research question?	The rationale for using the model is clearly stated; its inclusion is not decorative or superficial.
Theoretical Alignment	Is the use of the LLM embedded within relevant domain-specific frameworks?	Coherence with disciplinary approaches is demonstrated; use is not generic or disconnected.
Experimental Transparency	Are the model version, prompts, parameters, and evaluation conditions documented?	Full replicability is enabled through detailed methodological disclosure.
Interpretive Control	Is scientific interpretation conducted by the authors rather than delegated to the model?	A clear division is maintained between generated content and human-led analysis.
Epistemic Risk Management	Are limitations like hallucination, bias, and opacity addressed and mitigated?	Risks are explicitly acknowledged and safeguards such as validation and filtering are implemented.
Methodological or Conceptual Contribution	Does the LLM enable a substantive scientific innovation or inferential gain?	The model supports novel insight, methodological advancement, or non-trivial inference.

LLM becomes the central contribution, displacing substantive scientific inquiry Arawjo (2024). Establishing baseline criteria prevents the proliferation of methodologically hollow studies published on the basis of superficial technological appeal.

Adopting this operational perspective does not restrict scientific exploration with generative models. Rather, it establishes the minimal conditions under which such exploration can be recognized as legitimate knowledge production, rather than opportunistic deployment of technological resources devoid of scientific grounding.

6. Conclusions

The proliferation of LLMs constitutes one of the most consequential technological developments in recent years. Their ability to generate fluent and contextually plausible language has expanded the scope of computational support across scientific domains, from biomedical research to education and technical writing. However, this review demonstrates that their widespread integration into research practices introduces profound epistemological, methodological, and institutional challenges.

A key finding is the structural asymmetry in how LLMs are currently employed. Most studies use these models as drop-in tools for automating existing tasks, often without theoretical justification, hypothesis framing, or interpretive oversight. This pattern reflects broader academic incentives favoring rapid output over conceptual contribution and reinforces a trend toward superficial innovation. In contrast, rigorous applications embed LLMs within domain-specific workflows, ensure transparency in experimental setup, and maintain human-led interpretation.

Generative models must not be treated as epistemic agents, but as probabilistic instruments whose scientific value depends entirely on their situated use. Their integration demands more than empirical adequacy: it requires alignment with disciplinary frameworks, methodological disclosure, and interpretive control. Several studies reviewed here illustrate such responsible integration. Luo et al. Luo et al. (2025) enhance predictive inference through domain-specific fine-tuning; Wang et al. Wang et al. (2024) impose symbolic transparency via intermediate representations; Procko et al. Procko et al. (2023) delineate human-machine boundaries in scientific writing. These examples show that LLMs can support hypothesis generation, planning, and exposition when embedded in structured, accountable workflows.

This article also advances a normative framework for evaluating LLM-based research. The proposed matrix translates six core criteria—ranging from methodological justification to epistemic risk management—into operational indicators for use in peer review, research planning, and editorial assessment. This contribution responds to the growing institutional need for concrete mechanisms to distinguish legitimate scientific use of generative AI from tool-centric experimentation.

Ultimately, the future role of LLMs in science will not be determined by their capabilities alone, but by the standards with which they are integrated. Their epistemic value resides not in their scale or fluency, but in the rigor

of the frameworks that constrain, interpret, and justify their use. Ensuring this rigor is not a technical matter; it is a scientific imperative.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to improve the grammar and clarity of the manuscript. Following the use of this tool, the authors thoroughly reviewed and edited the content as needed, and take full responsibility for the final version of the publication. This practice is fully aligned with the critical perspective developed throughout this study.

References

Anthropic, 2024. Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family.

Apple, 2025. The Illusion of Thinking: Evaluating Reasoning in Large Language Models. Technical Report. Apple Inc. URL: https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf.

Arawjo, I., 2024. Llm wrapper papers are hurting hci research. URL: https://ianarawjo.medium.com/llm-wrapper-papers-are-hurting-hci-research-8ad416a5d59a. medium article, accessed April 15, 2025.

Beck, U., 1992. Risk Society: Towards a New Modernity. Sage Publications, London, UK.

Behrouz, A., Zhong, P., Mirrokni, V., 2024. Titans: Learning to memorize at test time. URL: https://arxiv.org/abs/2501.00663, arXiv:2501.00663. arXiv preprint arXiv:2501.00663.

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA. p. 610–623. URL: https://doi.org/10.1145/3442188.3445922, doi:10.1145/3442188.3445922.

Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C.T., et al., 2025. How should the advancement of large language models affect the practice of science? Proceedings of the National Academy of Sciences 122, e2401227121. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2401227121, doi:10.1073/pnas.2401227121, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2401227121.

Boiko, D.A., MacKnight, R., Kline, B., Gomes, G., 2023. Autonomous chemical research with large language models. Nature 624, 570–578. URL: https://doi.org/10.1038/s41586-023-06792-0, doi:10.1038/s41586-023-06792-0.

Bordt, S., von Luxburg, U., 2023. Chatgpt participates in a computer science exam. URL: https://arxiv.org/abs/2303.09461, arXiv:2303.09461. arXiv preprint arXiv:2303.09461.

Chartier-Edwards, N., Grenier, E., Goujon, V., 2025. Galactica's disassemblage: Meta's beta and the omega of post-human science. AI & Society 40, 2069–2081. URL: https://doi.org/10.1007/s00146-024-02088-7, doi:10.1007/s00146-024-02088-7.

Choi, J.H., Hickman, K.E., Monahan, A.B., Schwarcz, D., 2022. Chatgpt goes to law school. Journal of Legal Education 71, 387–400. URL: https://jle.aals.org/home/vol71/iss3/2/.

Cuskley, C., Woods, R., Flaherty, M., 2024. The limitations of large language models for understanding human language and cognition. Open Mind 8, 1058–1083. URL: https://doi.org/10.1162/opmi%5Fa%5F00160, doi:10.1162/opmi_a_00160.

DeepSeek, 2025. Deepseek-vl and deepseek-coder: Advancing open multimodal and code-generation models. https://github.com/deepseek-ai.

Fecher, B., Hebing, M., Laufer, M., Pohle, J., Sofsky, F., 2025. Friend or foe? exploring the implications of large language models on the science system. AI & Society 40, 447–459. URL: https://doi.org/10.1007/s00146-023-01791-1, doi:10.1007/s00146-023-01791-1.

Galli, C., Gavrilova, A.V., Calciolari, E., 2025. Large language models in systematic review screening: Opportunities, challenges, and methodologi-

cal considerations. Information 16. URL: https://www.mdpi.com/2078-2489/16/5/378, doi:10.3390/info16050378.

Gallifant, J., Fiske, A., Levites Strekalova, Y.A., Osorio-Valencia, J.S., Parke, R., Mwavu, R., Martinez, N., Gichoya, J.W., Ghassemi, M., Demner-Fushman, D., McCoy, L.G., Celi, L.A., Pierce, R., 2024. Peer review of gpt-4 technical report and systems card. PLOS Digital Health 3, 1–15. URL: https://doi.org/10.1371/journal.pdig.0000417, doi:10.1371/journal.pdig.0000417.

Gencer, G., Gencer, K., 2025. Large language models in healthcare: A bibliometric analysis and examination of research trends. Journal of Multidisciplinary Healthcare 18, 223–238. URL: https://www.tandfonline.com/doi/full/10.2147/JMDH.S502351, doi:10.2147/jmdh.s502351.

Hosseini, M., Gao, C.A., Liebovitz, D.M., Carvalho, A.M., Ahmad, F.S., et al., 2023. An exploratory survey about using chatgpt in education, healthcare, and research. Plos One 18, e0292216. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0292216, doi:10.1371/journal.pone.0292216.

Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., et al., 2024. Large language models for software engineering: A systematic literature review. ACM Trans. Softw. Eng. Methodol. 33. URL: https://doi.org/10.1145/3695988, doi:10.1145/3695988.

ISO, 2018. ISO 31000:2018 Risk management – Guidelines. Technical Report. International Organization for Standardization. Geneva, Switzerland. URL: https://www.iso.org/standard/65694.html.

Jiang, Q., Gao, Z., Karniadakis, G.E., 2025. Deepseek vs. chatgpt vs. claude: A comparative study for scientific computing and scientific machine learning tasks. Theoretical and Applied Mechanics Letters 15, 100583. URL: https://www.sciencedirect.com/science/article/pii/S2095034925000157, doi:https://doi.org/10.1016/j.taml.2025.100583.

Kobak, D., González-Márquez, R., Ágnes Horvát, E., Lause, J., 2025. Delving into chatgpt usage in academic writing through excess vocabulary.

URL: https://arxiv.org/abs/2406.07016, arXiv:2406.07016. arXiv preprint arXiv:2406.07016.

Kuhn, T.S., 1962. The Structure of Scientific Revolutions. 1st ed., University of Chicago Press.

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., Tseng, V., 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. PLOS Digital Health 2, 1–12. URL: https://doi.org/10.1371/journal.pdig.0000198, doi:10.1371/journal.pdig.0000198.

Landschaft, A., Antweiler, D., Mackay, S., Kugler, S., Rüping, S., et al., 2024. Implementation and evaluation of an additional gpt-4-based reviewer in prisma-based medical systematic literature reviews. International Journal of Medical Informatics 189, 105531. URL: https://www.sciencedirect.com/science/article/pii/S1386505624001941, doi:https://doi.org/10.1016/j.ijmedinf.2024.105531.

Lee, S., Peng, T.Q., Goldberg, M.H., Rosenthal, S.A., Kotcher, J.E., Maibach, E.W., Leiserowitz, A., 2024. Can large language models estimate public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. PLOS Climate 3, 1–14. URL: https://doi.org/10.1371/journal.pclm.0000429, doi:10.1371/journal.pclm.0000429.

Li, M., Sun, J., Tan, X., 2024. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Systematic Reviews 13, 219. URL: https://doi.org/10.1186/s13643-024-02609-x, doi:10.1186/s13643-024-02609-x.

Lissack, M., Meagher, B., 2024. Navigating the future of large language models in scientific research: Opportunities, challenges, and ethical considerations. SSRN Electronic Journal. URL: https://ssrn.com/abstract=4949829, doi:10.2139/ssrn.4949829.

Luo, X., Rechardt, A., Sun, G., Nejad, K.K., Yáñez, F., et al., 2025. Large language models surpass human experts in predicting neuroscience results. Nature Human Behaviour 9, 305–315. URL: https://doi.org/10.1038/s41562-024-02046-9, doi:10.1038/s41562-024-02046-9.

Mammides, C., Papadopoulos, H., 2024. The role of large language models in interdisciplinary research: Opportunities, challenges and ways forward. Methods in Ecology and Evolution 15, 1774–1776. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14398, doi:10.1111/2041-210X.14398.

Navigli, R., Conia, S., Ross, B., 2023. Biases in large language models: Origins, inventory, and discussion. J. Data and Information Quality 15. URL: https://doi.org/10.1145/3597307, doi:10.1145/3597307.

OECD, 2015. Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development. OECD Publishing, Paris. doi:10.1787/9789264239012-en.

Ogden, C.K., Richards, I.A., 1923. The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism. Harcourt, Brace and Company, New York. URL: https://archive.org/details/meaningofmeaning00ogde.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., et al., 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). J. Mach. Learn. Res. 22.

Procko, T., Davidoff, A., Elvira, T., Ochoa, O., 2023. Towards improved scientific knowledge proliferation: Leveraging large language models on the traditional scientific writing workflow. SSRN Electronic Journal. URL: https://ssrn.com/abstract=4594836, doi:10.2139/ssrn.4594836.

Rajaraman, N., Jiao, J., Ramchandran, K., 2025. An analysis of tokenization: Transformers under markov data, in: Proceedings of the 38th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. pp. 1999:1–1999:54.

Rathinapandi, G., 2023. Tokenization vs. embedding: Understanding the differences and their importance in nlp. URL: https://geoffrey-geofe.medium.com/tokenization-vs-embedding-understanding-the-differences-and-their-importance-in-nlp-b62718b5964a. medium, Geoffrey Rathinapandi.

Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., et al., 2022. Tackling climate change with machine learning. ACM Comput. Surv. 55. URL: https://doi.org/10.1145/3485128, doi:10.1145/3485128.

Samynathan, S., 2024. Transformers in large language model. URL: https://medium.com/version-1/transformers-in-large-language-model-2f7d485f50b0. medium, Version 1, publicado el 8 de julio de 2024.

Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston, MA.

Shanahan, M., 2024. Talking about large language models. Commun. ACM 67, 68–79. URL: https://doi.org/10.1145/3624724, doi:10.1145/3624724.

Song, K., Trotter, A., Chen, J.Y., 2025. Llm agent swarm for hypothesis-driven drug discovery. URL: https://arxiv.org/abs/2504.17967, arXiv:2504.17967.

Sparrow, R., Flenady, G., 2025. Bullshit universities: the future of automated education. AI & Society URL: https://doi.org/10.1007/s00146-025-02340-8, doi:10.1007/s00146-025-02340-8.

Toufiq, M., Rinchai, D., Bettacchioli, E., Kabeer, B.S.A., Khan, T., et al., 2023. Harnessing large language models (llms) for candidate gene prioritization and selection. Journal of Translational Medicine 21, 728. URL: https://doi.org/10.1186/s12967-023-04576-8, doi:10.1186/s12967-023-04576-8.

Truong, T., Otmakhova, Y., Verspoor, K., Cohn, T., Baldwin, T., 2024. Revisiting subword tokenization: A case study on affixal negation in large language models, in: Duh, K., Gomez, H., Bethard, S. (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico. pp. 5082–5095. URL: https://aclanthology.org/2024.naacl-long.284/, doi:10.18653/v1/2024.naacl-long.284.

Ullman, T., 2023. Large language models fail on trivial alterations to theory-of-mind tasks. URL: https://arxiv.org/abs/2302.08399, arXiv:2302.08399. arXiv preprint arXiv:2302.08399.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et al., 2017. Attention is all you need, in: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 5998–6008. URL: https://proceedings.neurips.cc/paper%5Ffiles/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wang, R., Si, S., Yu, F., Wiesmann, D., Hsieh, C.J., et al., 2024. Large language models are interpretable learners. URL: https://arxiv.org/abs/2406.17224, arXiv:2406.17224. arXiv preprint arXiv:2406.17224.

You, H., Lee, K., Paci, S., Park, J., Zheng, 2023. Applications of GPT in Political Science Research: Extracting Information from Unstructured Text. Technical Report. Princeton University.