Journal of Documentation

125

Agentic publications: redesigning scientific publishing in the age of thinking large language models

Roberto Pugliese and George Kourousias
Elettra Sincrotrone Trieste SCpA, Trieste, Italy, and
Francesco Venier and Grazia Garlatti Costa
University of Trieste, Trieste, Italy

Received 30 July 2025
Revised 9 October 2025
Accepted 11 October 2025

Abstract

Purpose – This paper introduces the concept of "Agentic Publication (AP)," a novel large language model (LLM)-driven framework designed to complement traditional scientific publishing by transforming papers into interactive knowledge systems that address challenges created by exponential growth in scientific literature.

Design/methodology/approach – Our architecture integrates structured data (knowledge graphs and metadata) with unstructured content (text and multimedia) through retrieval-augmented generation and multi-agent verification. The system provides interfaces for humans and artificial agents, offering narrative explanations alongside machine-readable outputs. Implementation leverages vector databases for semantic search, knowledge graphs for structured reasoning and collaborative verification agents.

Findings – Our proof-of-concept demonstration showcases multilingual interaction, Application Programming Interface (API) accessibility, continuous knowledge flow and structured knowledge representation. The framework enables dynamic updating of knowledge, synthesis of new findings and customizable detail levels.

Practical implications – The system is a powerful companion for researchers navigating complex knowledge landscapes, offering tailored information access across disciplines while addressing ethical considerations through automated validation, expert oversight and transparent governance.

Originality/value – The AP represents a transformative approach to scientific communication by creating responsive knowledge synthesis systems while maintaining scientific rigor. Integrating multi-agent verification with traditional publishing pathways creates a more efficient, accessible and collaborative research ecosystem, particularly valuable in interdisciplinary fields.

Highlights

(1) The Agentic Publication transforms static scientific papers into dynamic, interactive knowledge systems powered by LLMs
(2) The architecture combines structured and unstructured data with retrieval-augmented generation (RAG) and multi-agent verification processes
(3) The framework provides distinct interfaces for humans and artificial agents, enabling both narrative explanations and machine-readable outputs
(4) Ethical considerations are addressed through automated validation, expert oversight and transparent governance principles
(5) A proof-of-concept demonstrates a practical implementation while preserving compatibility with traditional scientific publishing workflows

Keywords Agentic publications, LLM-driven framework, Scientific publishing, Knowledge systems, Retrieval-augmented generation, Multi-agent verification, Interactive knowledge dissemination

Paper type Research article

1. Introduction

Modern scientific publishing, centered on peer-reviewed journal articles, has driven knowledge dissemination for centuries. However, there is growing recognition that the traditional model is strained and increasingly inadequate (Hughes and Van Heerden, 2024).

Journal of Documentation
Vol. 82 No. 7, 2026
pp. 125-149
Emerald Publishing Limited
e-ISSN: 1758-7379
p-ISSN: 0022-0418
DOI 10.1108/JD-07-2025-0207

Publication bias and editorial priorities filter what research enters the public domain, while the fundamental format – static documents written for humans to read – has remained unchanged since the 17th century, even as the volume of literature grows exponentially and becomes impossible for any individual to fully process (Hu *et al*., 2024; Bucur *et al*., 2022).

This crisis extends beyond format limitations to encompass fundamental systemic challenges in knowledge validation and dissemination.

While the traditional peer-reviewed publishing system has provided essential quality control and institutional credibility for centuries, serving as the backbone of scientific knowledge validation (Kelly *et al*., 2014), it faces mounting challenges that compromise its effectiveness in contemporary research environments. The exponential growth of scientific literature – with nearly 2 million articles published annually and annual growth rates exceeding 4% (Bornmann *et al*., 2021) – has created systematic bottlenecks that strain peer review capacity and make comprehensive knowledge synthesis impossible for individual researchers (Tennant and Ross-Hellauer, 2020). This unprecedented growth has led to an already "overstretched" peer review system (Seghier, 2025), creating a cycle where researchers face mounting publication demands while simultaneously experiencing "reviewer fatigue" from increased review burdens (Drozdz and Ladomery, 2024).

The challenge is particularly acute in fast-moving fields like artificial intelligence, where the pace of industry research increasingly outstrips academic publishing's ability to keep up, risking academia's marginalization as a second-tier player in cutting-edge development (Kozak, 2025).

Contemporary challenges are further compounded by the rise of artificial intelligence in academic writing. A growing number of authors now use generative AI to expedite the writing process (Van Noorden and Perkel, 2023), contributing to increased submission volumes that exacerbate existing system pressures. More concerning is the alarming increase in fraudulent manuscripts produced by "paper mills" exploiting AI tools (Van Noorden, 2023), which places additional burden on an already strained verification system. Meanwhile, the availability of qualified reviewers continues to decline, with researchers increasingly declining review invitations due to time constraints and competing professional demands (Willis, 2016), raising questions about the long-term sustainability of traditional peer review in some fields (Künzli *et al*., 2022).

Beyond scalability issues, the traditional system exhibits documented structural limitations including access barriers through commercial paywalls that create geographic and economic inequalities (Weingart, 2025), publication delays spanning months or years that impede rapid knowledge dissemination, and systematic biases in peer review that can disadvantage research based on author characteristics rather than scientific merit (Lee *et al*., 2013; Manchikanti *et al*., 2015). Research has identified multiple bias categories affecting manuscript evaluation, including prestige bias favoring authors from renowned institutions, geographic bias disadvantaging non-Anglophone researchers and publication bias preferring statistically significant results over null findings (Haffar *et al*., 2019). Additionally, traditional peer review proves limited in detecting methodological errors and ensuring reproducibility, as evidenced by rising retraction rates and the ongoing reproducibility crisis (Emile *et al*., 2022). Studies show that reviewers identify only a fraction of deliberately introduced flaws in manuscripts (Drozdz and Ladomery, 2024).

These converging pressures – exponential growth in submissions, declining reviewer availability, AI-driven increases in both legitimate and fraudulent manuscripts – suggest that while artificial intelligence may exacerbate some traditional publishing challenges, it may also offer part of the solution for creating more efficient and scalable knowledge validation systems (Seghier, 2025; Kozak, 2025). The urgency of adaptation is underscored by the risk that scholarly publishing may be sidelined entirely unless it accelerates its processes while preserving quality control (Kozak, 2025). These documented challenges do not diminish the fundamental value of rigorous quality control mechanisms that traditional publishing provides. Rather, they highlight opportunities for complementary approaches that preserve

essential peer oversight while addressing structural limitations through technological innovation – a need that becomes particularly acute in interdisciplinary fields where researchers must synthesize knowledge across multiple domains and methodological traditions.

Meanwhile, advances in artificial intelligence and particularly LLMs offer a tantalizing opportunity to reimagine how scientific knowledge is shared.

Since the release of ChatGPT in late 2022, awareness and use of LLMs in academia have surged from a niche curiosity to mainstream: a recent survey found that 80% of scientists had used AI chatbots like ChatGPT in their work (Jen and Hj Salam, 2024; Daykan and O'Reilly, 2023; Hughes and Van Heerden, 2024). Initial debates focused on using LLMs as writing assistants to polish or generate papers, raising concerns about plagiarism and accuracy (Lin, 2023; Hughes and Van Heerden, 2024). However, researchers now look beyond AI-assisted writing toward more transformative possibilities (Ahaley *et al.*, 2023; Jen and Hj Salam, 2024; Daykan and O'Reilly, 2023). Some authors claim that next-generation LLMs (guided by human experts) could radically disrupt the entire scientific inquiry and knowledge-sharing system, effectively replacing conventional journals and papers (Ahaley *et al.*, 2023; Hughes and Van Heerden, 2024; Kozak, 2025). Instead of engaging solely with static documents, researchers and interested stakeholders could interact with dynamic knowledge repositories that integrate published findings and provide contextual information responsively. For social scientists, this might include up-to-date economic indicators, policy analyses and management research syntheses, enabling more informed decision-making through timely access to relevant knowledge. Similarly, natural scientists could benefit from real-time integration of experimental results across laboratories, creating opportunities for accelerated discovery through improved knowledge coordination. In this vision, the traditional paper could soon appear outdated (Hughes and Van Heerden, 2024).

Building upon these opportunities and addressing the gaps in scientific dissemination, this manuscript proposes an LLM-powered architecture as a potential complement to the traditional scientific paper, which we term an Agentic Publication (AP). An AP is a scientific artifact enhanced with embedded artificial agents powered by LLMs, which dynamically integrate, verify and disseminate knowledge. This approach has implications across scientific disciplines, including both natural and social sciences, where the need for dynamic knowledge synthesis is particularly acute.

Unlike traditional papers, APs are interactive, continuously updatable and accessible to both human users and other machines. Our goal is outlining a system – composed of LLMs, tools and intelligent agents – that can support scientists to create, store, process and interactively disseminate scientific knowledge on demand. We outline the structure of this system, including its core components (knowledge representation, query interface, updating and verification modules), a potential technology stack and implementation strategies. We also discuss how the system can present information differently to human users versus artificial agents and examine the associated ethical implications. By detailing this architecture and its feasibility, we aim to illustrate a path toward a more accessible, up-to-date and interactive model of scientific communication (Hughes and Van Heerden, 2024) open and responsive, benefiting researchers and society in ways the static journal article alone cannot.

2. Proposed system architecture

2.1 System architecture overview

The proposed system is centered on a LLM continually enriched with scientific findings. We term this concept "Agentic Publication" as it includes AI agent features and supports automated research processes. Figure 1 shows the system's conceptual overview, encompassing new research information flow and knowledge base querying.

The architecture (see Figures 2 and 3) comprises four interacting components: (1) a *Knowledge Representation Layer* storing scientific information in human-readable and

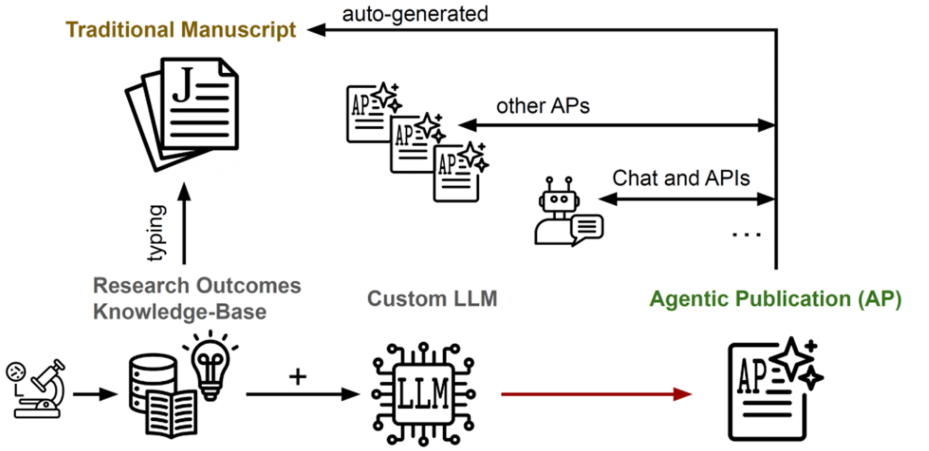

Figure 1. Simplified overview of how research outcomes yield in a traditional manuscript and an Agentic Publication (AP). The knowledge-based including structured and unstructured content (text) is coupled with a suitable AP LLM considering the AP properties described in this manuscript (red arrow), resulting in an AP. APs can be used for basic "chat with your paper" but also provide APIs for computational access (inc. to the data). APs can interact with other APs, but also auto-generate print-ready versions of traditional manuscripts. These auto-generated manuscripts are outputs of the AP for human consumption and accessibility, but the AP itself remains the primary citable entity

machine-interpretable forms; (2) an *Interactive Query Interface* (Application Programming Interface [API], chat or voice-based) for user questions and responses; (3) *Dynamic Updating Mechanisms* that ingest new findings based on query analysis; and (4) *Verification and Governance Processes* ensuring knowledge base quality and integrity. These components are supported by advanced LLMs, agents, databases and integration APIs.

While Figure 1 illustrates that APs can auto-generate print-ready manuscripts, these should be understood as outputs derived from the AP rather than as the AP itself. An AP consists of the curated knowledge base, author analyses and embedded agentic mechanisms with version tracking. From this foundation, multiple manuscript variants can be produced according to author-defined requirements such as style, language or length. To ensure consistency, authors may pre-approve the generation prompt or provide pre-generated, curated versions stored within the AP. In this way, the AP remains an inherently dynamic and interactive system, while also enabling the export of fixed-form documents for human consumption. Auto-generated manuscripts should not be regarded as peer-reviewed, citable works in their own right; rather, the AP is the primary referential entity, with generated documents serving as associated resources that support accessibility and dissemination.

2.2 Knowledge representation layer

Scientific knowledge representation across domains requires formats suitable for both LLM and algorithm consumption, accommodating varied methodologies of different fields.

We propose a hybrid knowledge store with complementary components working as an integrated system. At its foundation lies unstructured content comprising full research texts, including methodology descriptions, observations and results, along with associated artifacts such as figures, tables and datasets. The content is indexed in vector databases for semantic search, enabling LLM retrieval of relevant passages. Semantic indexing captures conceptual relationships despite terminology differences (Hughes and Van Heerden, 2024). Each knowledge entry carries important metadata, including authors, publication date and domain keywords to provide essential context.

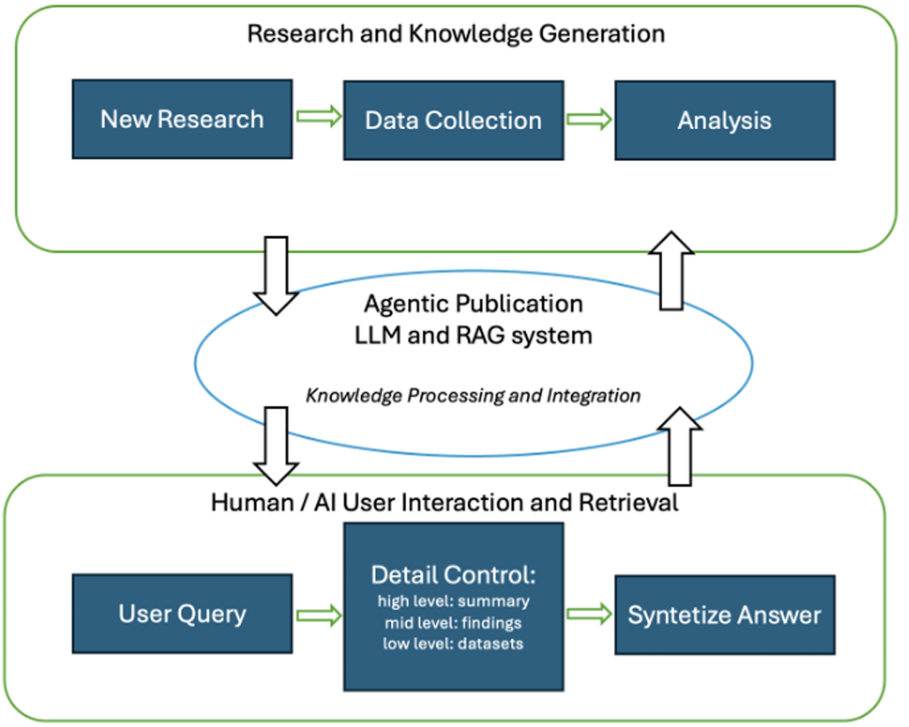

Figure 2. Conceptual illustration of the proposed Agentic Publication (inspired by the "PLOS-LLM" model by Hughes and Van Heerden, 2024). New research and knowledge generation steps feed data into the LLM-centric system, while user queries retrieve synthesized answers. The system allows users to zoom in and out on the level of detail – from high-level summaries (headlines/abstracts) to granular data (complete datasets). An interactive loop based on interaction log analysis helps keep knowledge updated and easily accessible

Working in parallel with this unstructured foundation, structured knowledge components store key facts and relationships in knowledge graphs (Wang and Shi, 2025) or relational databases. This layer employs ontologies to define entities ranging from chemical compounds to business models, enabling machine-interpretable encoding and consistency checks across studies examining similar relationships. Other scientific data formats (HDF) can be supported too, in case via integration tools.

Building on semantic publishing work (Bucur *et al.*, 2022), our system uses LLMs to bridge structured and unstructured realms, interpreting data while populating knowledge graphs.

Multi-modal data support addresses scientific knowledge, including datasets, code, images and audio/video recordings. Multi-modal LLMs can accept and generate multiple data types. The architecture incorporates data repositories linking supplementary materials with textual descriptions, enabling detailed responses and verification against source data. Vision-language models may interpret figures and experimental diagrams, enhancing knowledge integration across disciplines (Dias *et al.*, 2023).

Provenance tracking documents evolution with transparency, preserving original LLMs, reviewer comments, revision history and updates. This strengthens scientific accountability and reproducibility.

In addition to datasets and results, APs are designed to explicitly embed the authors' professional reasoning: interpretations, critiques, links to other fields and reflections on limitations. This "authorial reasoning layer" ensures that the AP is not merely *data* + *LLM*, but

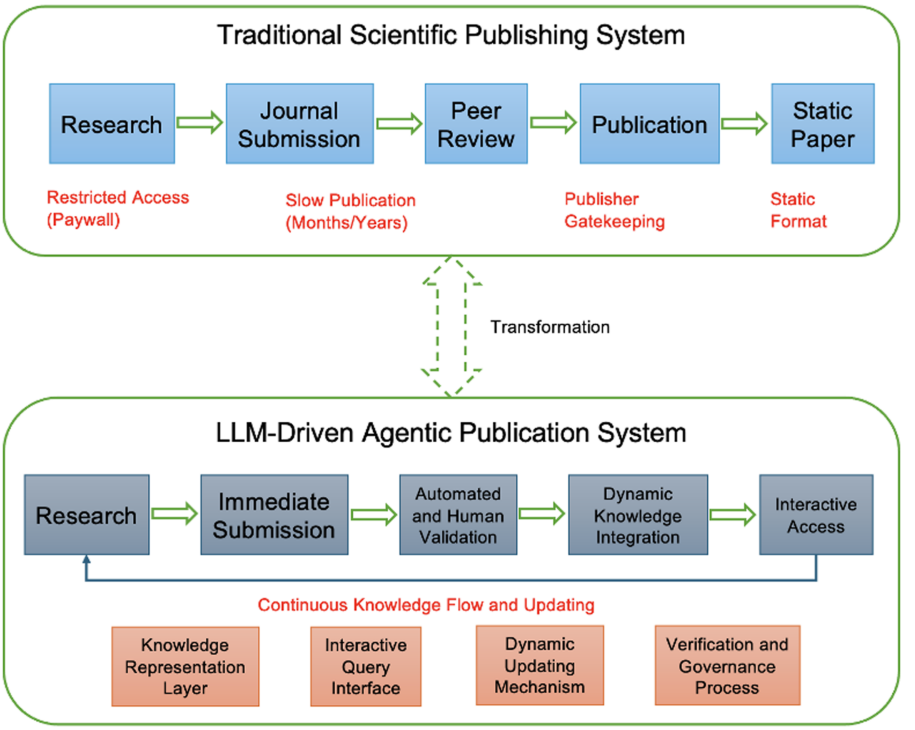

Figure 3. Scientific knowledge workflow: Traditional versus LLM-driven agentic publication system

rather *data + LLM + authorial analysis*. Authors can also provide reasoning profiles or bias prompts that guide the system's responses in transparent and reproducible ways. For example, an author might specify: *"all LLM reasoning should be compatible with Epicurean views," "all responses should remain consistent with the author's previous work"* or *"ignore string-theory approaches when formulating answers."* These profiles are optional, versioned and signed (e.g. ORCID-linked), so that readers and downstream systems can explicitly distinguish between evidence-based outputs and author-guided perspectives. In this way, opinions and viewpoints remain first-class scholarly contributions in APs, complementing the underlying evidence and making reasoning both auditable and reusable.

The architecture's design incorporates upgradability as a fundamental principle, giving an AP inherent potential for continual improvement and upgrading. Upgradability allows seamless integration of new LLMs, enhancing analytical capabilities and engagement beyond original constraints, ensuring long-term relevance.

By combining these, the knowledge representation layer serves as system memory and truth source. It provides redundancy (text for flexibility; structured data for precision), minimizing errors. Scalable infrastructure will support millions of entries using existing "big data" and datalake technologies (Huang, 2021).

2.3 Interactive Query Interface

Users interact with the knowledge system through an intuitive conversational interface, enhancing static article utility (Li and Zhang, 2023). Users pose questions in natural language – *What are the latest findings on X?"* or *"Show me data supporting claim Q"* – and receive tailored answers synthesized from relevant knowledge in the same language of the request.

The interface design incorporates several essential features that enhance scientific knowledge accessibility. Conversational Q&A stands at the core of this system, supporting iterative dialogue with follow-up questions and clarification requests. The LLM contextual understanding enables natural conversation flow, offering greater flexibility than keyword searches by integrating information across sources and tailoring explanations to specific questions.

Another crucial aspect is adjustable detail functionality, often called "zoom," which provides varying information levels for different users. Users can access succinct summaries, abstract-style overviews, detailed discussions with evidence or raw experimental data (Hughes and Van Heerden, 2024). For instance, policymakers might need consensus conclusions, while experts require underlying data inspection. This dynamic adjustment replaces static paper text with a responsive presentation tailored to user needs.

The system further enhances knowledge delivery through multi-modal responses that leverage diverse knowledge base data. When asked about trends, the system generates graphs from datasets, analyzes images and integrates tables, images or audio with text responses, making knowledge consumption more engaging (Cardon, 2023).

User-friendly input mechanisms further lower usage barriers, including voice interaction through speech-to-text, enabling hands-free operation valuable for practitioners in field settings. Web and mobile platform availability democratizes scientific knowledge access (Li and Zhang, 2023).

Finally, context and personalization features enhance the user experience by maintaining user profiles, adjusting explanations based on background – simpler language for students, technical detail for experts – while remembering previous interactions to avoid redundancy.

Behind this interface, the LLMs and associated agents do the heavy lifting, interpreting the question, fetching relevant knowledge and composing an answer. Answers include references or links to supporting knowledge base data. For instance, the interface might show in-line citations that users can click to inspect the original study or data source that backs each statement. Combining natural language with ease with scholarly rigor (sources, evidence on demand), the interactive interface can make consuming scientific knowledge across disciplines convenient and trustworthy.

2.4 Agents in an Agentic Publication

An AP can also be described as an agent in an agent ecosystem. In the context of LLMs, an agent refers to an AI system that leverages an LLM as its "brain" to autonomously or semi-autonomously perform complex tasks, make decisions and interact with users or other systems. Unlike a standalone LLM – which simply generates text responses to inputs – an LLM agent can plan actions, use external tools and maintain memory across interactions, enabling it to handle multi-step workflows and dynamic environments.

Key characteristics of LLM agents include autonomy, planning, tool use, memory, reasoning and persona/profiling. It is, anyhow, reasonable to think about an agent ecosystem, where scientists can create new AP agents by uploading their specific knowledge and discoveries. The ecosystem also includes non-AP agents that are dedicated to specific bookkeeping tasks to support the functioning and integrity of the whole AP ecosystem. In the rest of the paper when appropriate, we will specify if the presented features will be implemented by the AP agent or by these non-AP ecosystem agents.

As it is now evident, APs can include multiple agents, each fulfilling distinct roles. For example, a primary authoring agent may synthesize content, a verification agent can cross-reference claims, and a reviewer agent may simulate critical peer review. This modular design fosters collaborative intelligence and robust internal checks, improving trust and scalability.

2.5 Dynamic updating mechanisms

A cornerstone of our proposed system is that it remains up-to-date, reflecting the evolving nature of scientific knowledge across all fields (Klami *et al.*, 2024). Rather than the multi-

month or multi-year cycles of traditional publishing and literature reviews, this architecture allows scientific knowledge to be continuously refreshed in real time.

The foundation of our approach relies on continuous knowledge ingestion, where researchers submit findings to the system as soon as they are available, rather than waiting for complete, polished papers. Once a study's methodology is approved and experiments are conducted, the methods and initial results can be uploaded to the AP ecosystem as a new agent (Hughes and Van Heerden, 2024). This approach can even precede formal analysis – early hypotheses, intermediate results or registered trial protocols can be fed in, so the model knows "something is in progress." When final results are obtained, authors upload those along with datasets and analyses. These parallels open notebook science principles, where the knowledge base is incrementally built as research progresses, coupled to an intelligent model.

Supporting this continuous flow are automated pipelines and APIs that facilitate ingestion by accepting various input formats. Researchers might fill structured templates for their studies or upload preprints for system parsing. Integration with existing databases is crucial – connecting to arXiv or PubMed to automatically pull new content daily. Journal platforms could evolve to interface directly with the AP ecosystem, pushing content upon publication. Web crawlers can scan repositories for relevant updates, with all inputs time-stamped and labeled for version control.

To ensure reliability, each submission triggers a validation routine before full integration. This involves automated checks executed by bookkeeping agents (plagiarism detection, consistency verification and statistical review) and routing to human experts when needed. Only after passing this validation gate is knowledge considered verified in the system, ensuring the model's knowledge base remains reliable and does not accumulate unchecked errors.

The core architecture incorporates model updating processes occurring in two ways. A RAG approach immediately incorporates new facts by referencing the latest information from the AP knowledge repository at query time. However, periodic fine-tuning of the LLM improves fluency and reasoning on new data – the model might be updated nightly or weekly on new content. This dual approach balances currency and performance while mitigating context window limitations (Hu *et al.*, 2024).

A powerful feature is the AP system's capacity for automated synthesis and updates, where aggregated knowledge updates automatically with each new piece of evidence. Meta-analyses could re-run as new data arrives, updating consensus answers and recomputing summary statistics. This means reviews never go out of date – a stark contrast to static articles obsolete upon publication (Hughes and Van Heerden, 2024). The system maintains uncertainty estimates that update as evidence accumulates.

Finally, the architecture addresses scalability requirements as the knowledge base grows. It might be partitioned by domain or distributed globally, with specialized AP subsystem instances for different fields sharing backbones while receiving domain-specific fine-tuning. These strategies ensure responsiveness even as thousands of studies are integrated.

In summary, dynamic updating mechanisms transform scientific knowledge dissemination from a slow, discontinuous process into a live, continuous and responsive cycle. New knowledge is ingested promptly, validated and made queryable, with the model's outputs evolving as knowledge evolves (Klami *et al.*, 2024). This closes the gap between discovery and dissemination, accelerating insight spread through scientific communities.

2.6 Verification and quality control processes

Any system disseminating scientific knowledge must uphold accuracy, credibility and trustworthiness standards. While traditional peer review serves as a quality gatekeeper, it has well-documented flaws and delays. In our LLM-agent architecture, verification is equally crucial, given AI's propensity to confidently generate incorrect information if unchecked. We propose a multi-layered verification process combining human expertise and automated agents.

Central to our verification framework is human expert oversight through "AI-assisted peer review." When submitting a new AP, the system identifies relevant experts using its knowledge graph, inviting them to evaluate submissions (Hughes and Van Heerden, 2024). Experts interact with submissions via natural language queries: "How was the sample size chosen? Are there similar studies to compare?" The AI-assisted review surfaces issues quickly by comparing new results to related literature and highlighting conflicts for reviewer attention. Reviewers provide feedback and scores rating rigor, novelty and validity. Only studies meeting credibility thresholds are fully integrated, with others tagged with caution or lower confidence scores.

This process is transparent: reviews and scores are logged and attached to knowledge entries for public viewing, unlike traditional hidden peer review. Over time, reputation systems could develop where contributors and reviewers have track records, incentivizing quality contributions. Human oversight remains vital to catch nuances AI might miss (Hughes and Van Heerden, 2024).

Automated validation agents enhance verification through specialized AI performing specific checks. Evidence Validator LLMs cross-check factual claims against existing knowledge, while statistical checkers re-calculate results and flag analysis errors like p-hacking. Provenance tracers verify citations and guard against plagiarism. Multiple agents critique submissions in seconds, functioning as exhaustive reviewers that check everything from mathematics to consistency with known physics (Hughes and Van Heerden, 2024).

The system implements continuous error monitoring at query time through secondary processes that verify responses before display. When the LLM produces summaries, retrieval steps pull original sources to ensure accuracy. If stating "Study X found Y," the AP system double-checks that Study X indeed reported Y, mitigating hallucinations and building user trust.

A critical aspect of our verification approach involves the calibration of uncertainty as part of quality control, which attaches confidence scores to answers based on source agreement, data quality and conflicting evidence. When questions lack sufficient data, this is explicitly stated rather than guessed, paralleling how human authors acknowledge inconclusive evidence (Hughes and Van Heerden, 2024).

Complementing these verification mechanisms are safety and bias checks that scan for problematic content. When data overwhelmingly represents specific regions or demographics, potential skew is noted.

Therefore, maintaining quality in an LLM-driven knowledge repository requires reimagining peer review as integral and adaptable to different scientific disciplines. By combining human wisdom with artificial vigilance, the system ensures reliable knowledge dissemination. While LLMs may increasingly catch errors autonomously (Binz *et al.*, 2025), human judgment remains crucial for nuanced evaluations (Klami *et al.*, 2024; Hughes and Van Heerden, 2024).

2.7 Technology stack and integration

Building the above system is a significant engineering challenge, but many building blocks exist today. Here, we suggest a plausible technology stack and frameworks to implement the proposed architecture, leveraging state-of-the-art tools.

At the center of our implementation architecture stands the LLM, forming the core intelligence of the system. This could be based on transformer architectures with capabilities on par with GPT-4 or similar models like Gemini 2.5, Claude 3.7, given their demonstrated performance on diverse knowledge tasks (Freire *et al.*, 2024). Organizations might choose an open-source LLM (such as LLaMA 3, DeepSeek, . . .) and fine-tune it on scientific text to create a domain-optimized model. Techniques like LoRA (Low-Rank Adaptation) can efficiently update the model on new data without full retraining, allowing continuous learning. While closed APIs like OpenAI's GPT-4 could rely on retrieval augmentation exclusively,

open models offer more flexibility for direct tuning and on-premises deployment necessary for data privacy. A hierarchy of models might be employed: a large general model for understanding queries and generating answers, with smaller specialized models for specific domains like code execution or image analysis.

Supporting the LLM, a robust retrieval and database layer combines vector databases and traditional databases to implement semantic search and handle diverse data storage needs. Tools like FAISS, Pinecone or Weaviate can store high-dimensional text embeddings, allowing similarity search to fetch relevant contexts in milliseconds. For structured metadata and knowledge graphs, graph databases like Neo4j could store relationships and enable complex queries. Traditional SQL databases can handle tabular experimental results. The system includes an indexing pipeline: when new data arrives, embeddings are generated and inserted into vector databases, structured facts are extracted for graph databases, and raw files are stored in distributed systems with links.

To coordinate various processes, orchestration and agent frameworks like LangChain, Haystack or CrewAI provide abstractions for chaining LLM calls with retrieval and other actions. For instance, a chain handles queries: user question → LLM parses question → retrieval tool fetches relevant chunks → LLM composes answers. If answers require visualization or analysis, the system could use agents with Python execution tools to load datasets and produce charts. The multi-agent scenario can be coordinated via controller scripts or shared memory systems for agent communication (Hu *et al.*, 2024).

The user experience is delivered through a front-end interface implemented as web or mobile applications. Modern frameworks (React, Vue) could build chat UIs with features like citation highlighting, result filtering and data visualization outputs. For voice interaction, integration with speech-to-text services and text-to-speech for replies can be added. The front-end communicates with backends via REST or WebSocket APIs, with an emphasis on UX design to present complex scientific answers in an understandable manner.

For external connectivity and system expansion, APIs expose the system's knowledge for external tools and agents to consume. Other AI systems might query this knowledge base via APIs supporting natural language queries and returning structured results. Integration with ID systems like DOIs for datasets and ORCID for authors helps maintain links to the broader research ecosystem. Additionally, connecting to existing scientific infrastructure (clinical trial registries and data repositories) via their APIs can enrich the system's content and reliability.

Behind the scenes, a continuous deployment and scaling approach based on cloud microservices architecture hosts these components. LLMs run on GPU servers while databases operate on scalable clusters. Kubernetes can manage services and scale them based on load – during heavy query periods, more instances spin up, while ingestion services scale when batches of new studies arrive. Logging and monitoring services are critical, with every answer logged along with sources used to audit model behavior and detect systematic errors.

To ensure ethical use, model alignment and safety incorporate an AI ethics layer that could implement tools to detect toxic or biased content (Koçak, 2024). This includes intermediate filters checking outputs and adversarial testing suites that regularly probe the system with corner cases or misinformation to refine model responses.

Traditional publications historically required minimal resources, limited to basic web hosting. In contrast, APs have substantially higher cloud resource requirements, necessitating sophisticated infrastructures capable of storing and executing LLMs, maintaining extensive metadata and dynamically responding to user interactions. The required computational power scales with concurrent readers, introducing significant hosting considerations. Frameworks such as the European Open Science Cloud (EOSC) are ideally suited for hosting APs. Specialized EOSC Agentic Publication Nodes can be established by major research institutions to make necessary hardware resources available to the research community. The Italian synchrotron Elettra Sincrotrone Trieste is currently experimenting with such infrastructure based on its Data Lake ecosystem. The demonstration AP accompanying this

paper is hosted through Elettra's cloud infrastructure, expected to yield valuable insights into infrastructure costs, sustainability and scalability.

Therefore, while the overall system is ambitious, it can be constructed by integrating existing technologies: LLMs for language understanding, vector search for knowledge retrieval (Hu *et al.*, 2024), knowledge graphs for structured reasoning and orchestrators for tool use. Early prototypes might start with simpler versions and progressively add complex features as components mature in reliability.

3. From concept to prototype: implementing the Agentic Publication system

We now outline how one would implement a prototype of this LLM-agent-based knowledge dissemination system, highlighting considerations for diverse scientific fields. The process can be broken down into several stages, from data ingestion to user interaction.

3.1 Knowledge ingestion pipeline

The first step involves assembling the knowledge base through integrated processes that collect and structure research content.

The pipeline begins with data acquisition, where we gather scientific documents from repositories (arXiv, PubMed, Web of Science and Scopus) using tools like Semantic Scholar API or CrossRef for DOI content. This possibility may be useful at the beginning to facilitate system adoption.

A parallel submission portal allows researchers to directly upload manuscripts or research details through web forms.

Once data is acquired, parsing and processing transform raw documents into structured content using NLP techniques. Documents are partitioned into sections (abstract, methods and results) and further into manageable chunks. We can employ PDF parsing libraries or LaTeX source when available to extract text and references, followed by cleaning procedures to remove artifacts and convert specialized notation to plain text.

Alongside content extraction, metadata extraction identifies critical contextual information – title, authors, affiliations, keywords, publication venue, date and references – using specialized libraries or DOI-based services. This metadata populates structured databases and enables proper citation formatting.

To enable semantic search capabilities, embeddings indexing converts content chunks into machine-understandable representations using pre-trained models or sentence-transformers fine-tuned on scholarly data. These vectors are stored in vector databases with pointers to original sources, enabling semantic similarity search based on meaning rather than keywords.

For more structured knowledge representation, knowledge graph population creates semantic relationships through named entity recognition (NER) to identify entities (diseases, chemicals and instruments) and relation extraction to detect relationships ("X treats Y"). The LLM can extract structured claims using prompts like "List main claims as triples (subject, relation and object)" with results linked to existing ontology terms for consistency.

Finally, multimedia handling addresses nontextual components by storing original files while generating thumbnails for images and extracting captions as indexable text. Datasets in common formats are loaded into databases or have their schemas recorded for future queries.

This pipeline operates continuously to incorporate new submissions and updates, ensuring current knowledge base maintenance. AI Agents can be programmed and used to implement this pipeline.

3.2 LLM integration and training

With data in place, a possible next step is integrating the core LLM through interconnected processes that transform a general model into a specialized scientific knowledge system.

The journey begins with initial training to adapt the model to scientific content. If starting from a pre-trained model (like GPT-3/4), one might fine-tune it on curated and diverse domain data, ensuring representation across scientific disciplines. This could involve supervised fine-tuning on question–answer pairs derived from the corpus or instructing it on knowledge usage by providing examples of how to extract information from papers and answer questions. When fine-tuning data are sparse, an alternative is few-shot prompting at runtime, which allows the system to learn from limited examples while maintaining flexibility. This step may become less important as the quality of the LLMs increases.

Once trained, implementing a retrieval-augmented setup creates the crucial connection between the LLM and knowledge base. This mechanism ensures that before answering queries, the system performs vector search on indexed knowledge. For example, using the LangChain framework, create a QA chain: user query → embed query → similarity search top N passages from vector database → feed those passages plus the question into the LLM prompt as context → get answer. This step is crucial for grounding LLM answers in the specific literature and preventing hallucination. The prompt template should instruct the model to quote sources and use retrieved text to justify answers, maintaining scientific rigor.

To extend capabilities beyond simple question answering, tool use configuration enables the LLM to perform specialized functions when needed. This may use predefined prompt formats (like the ReAct framework or OpenAI function calling) where the model outputs special tokens to call tools such as calculators or plot generators. For instance, if the query is numerical ("What's the average result across these studies?"), the model might call a calculator tool with values extracted from sources, then incorporate the result into its answer. This functionality allows the system to perform computations and generate visualizations that enhance response value. New interoperability standards like the Model Context Protocol (MCP) by Anthropic simplify such integrations by allowing models to interface with diverse tools and environments seamlessly (Hou *et al.*, 2025). MCP helps ensure consistent context handling across agent interactions, enhancing performance in multi-agent AP ecosystems.

For complex queries requiring multiple perspectives or verification, agent collaboration establishes communication channels between specialized components. If implementing multiple agents (like a verifier agent), it is important to configure them to communicate with each other. For example, the main answering agent could pass its response to a verifier agent with knowledge base access for comments or corrections, returning the final answer only after verification. This can be orchestrated through simple loops: generate answer → verify → if verifier finds issues, revise answer. Such collaborative architecture mimics the peer review process in science, enhancing reliability. Inter AI agent communication protocols like the Agent2Agent Protocol (A2A) developed by Google and Microsoft, or the Agent Communication Protocol (ACP) developed by IBM Research, may facilitate this step.

Before deployment, rigorous testing evaluates the system's performance and safety. Begin with known questions that can be verified manually, checking if answers are correct and if provided sources match the content. This is an iterative process: if the model makes mistakes (e.g. attributing results to the wrong papers), adjust the prompting, improve retrieval by increasing retrieved passages or adding re-ranking steps, or add fine-tuning examples addressing specific error modes. It is important to ensure the model doesn't return sensitive or inappropriate content. Thorough testing builds confidence in the system's reliability before it faces real-world scientific inquiries.

3.3 User interface development

Developing the front-end requires thoughtful interface design, balancing functionality with accessibility.

At the core of this interface lies the chat interface, which features a simple chat window with message history for context maintenance and follow-up questions. Citation display

makes references like "(Smith *et al.*, 2021)" clickable with complete reference details in tooltips. Rich media content – images, graphs and model-generated plots – renders seamlessly within chat flow for enhanced comprehension.

To accommodate varying information needs, the system should provide detail level controls for detail level through intuitive UI elements such as sliders or buttons that adjust between "summary" and "detail" views. These trigger different prompts requesting brief answers or comprehensive responses with methodological information. While users can request detail levels through natural language, visible UI cues enhance the discoverability of advanced functionality.

Accessibility considerations lead us to incorporate voice input/output capabilities that expand interaction beyond text through microphone input with speech transcription and text-to-speech conversion for responses. This enhances accessibility and enables hands-free operation in laboratory or field environments.

Beyond human interfaces, the system should implement an agent API that enables programmatic interaction through REST endpoints (e.g. POST/query), accepting JSON payloads with questions and authentication credentials. Responses return structured data including answers and references, allowing external software and AI agents to consume knowledge programmatically. Comprehensive documentation facilitates integration by developers and other systems. Of course, protocols like A2A or ACP, or other emerging standards, can be adopted instead.

3.4 Dynamic update and feedback loop

Maintaining system updates and improvements requires comprehensive interconnected components. Scheduled updates implement schedulers (cron jobs), triggering ingestion pipelines for new content daily or weekly. Real-time processes handle direct author submissions with automated re-training or index updates following data ingestion. Continuous integration tools run ingestion or other forms of bookkeeping scripts with monitoring systems catching process failures.

User feedback integration allows answer rating and issue flagging (thumbs up/down, problem flags) with mechanisms capturing domain-specific feedback. This provides valuable retraining input – frequently flagged incorrect answers guide investigation and model improvement, potentially revealing misunderstood, nuanced questions requiring additional training data.

Ensuring optimal functioning requires robust performance monitoring logs, metrics including response time, queries answered and retrieval success rates. Query distribution tracking provides strategic insights – frequently asked topics may indicate content gaps or needs for dedicated sub-models.

As the system gains adoption, scaling considerations address growing usage through service replication, load balancers for APIs, multiple LLM container instances and autoscaling policies. Caching common queries improves efficiency while requiring careful design to handle knowledge updates that change answers, potentially using short TTL or caching only unchanged underlying data.

Throughout the implementation, iterative testing and refinement recognize this as a continuously improving system rather than a finished product. Initial user studies guide development by observing scientist interactions, similar to studies testing LLM tools for workers who appreciated quick retrieval while valuing human expertise (Freire *et al.*, 2024). Such feedback informs further development, including explanatory features or verification tightening based on trust concerns.

The result is a working prototype where users receive coherent answers with sources, new papers influence responses, and the system maintains conversational interaction through continuous improvement.

3.5 Agentic Publication demo

To examine and support the implementation's feasibility, a limited demo has been developed and deployed. This demo implements a single AP without any AI bookkeeping agent and is not meant to demonstrate all the features of a complete AP system, since it has already been established that this requires targeted analysis and professional software engineering. This manuscript aims to provide the concepts, architecture and insights on the AP concept. That said, the authors coded with RAG principles a limited single AP-like version with as the knowledge-base (kb) the very same content of this manuscript. The AP demo is accessible and has been assigned a DOI (https://doi.org/10.34965/agenticpublication.3567a) (Pugliese, 2025), and its features include a *landing page* that presents the paper metadata and a description of the different features currently available. The current version of the demo has been implemented as a Chatbot based on the VoiceFlow platform and implements a set of workflows that define the basic behavior of the AP.

Voiceflow is a comprehensive platform for designing, prototyping and deploying conversational AI applications. The platform enables the creation of voice and chat interfaces through a visual, no-code interface that allows developers and non-technical users alike to build complex conversational flows. A key feature of Voiceflow is its integrated knowledge base system that supports RAG capabilities, allowing conversational agents to access and leverage external information sources to provide more accurate and contextually relevant responses. Voiceflow supports integration with various natural language understanding services and voice assistants, including Amazon Alexa, Google Assistant and custom solutions. The platform features collaborative tools for team development, testing frameworks for conversation validation and analytics capabilities to measure user engagement. It also allows the integration of other specific agents in the workflows of the conversational AI application or agent.

As a result, a first important feature of the system is the capability to respond to user interactions (i.e. chat with your paper) via chat or voice and in multiple languages, thus reducing the language barrier. Figure 5 reports the demo Q&A interface. The different workflows currently implemented (Figure 4) allow the user to download the full version of the paper (this paper), a visual presentation version of the paper, the associated datasets (in this case, the paper references, with metadata, abstract and summaries), a visual representation of the paper (the architecture, a mind map, . . .).

Voiceflow allows interaction via well-defined REST APIs, using different programming languages. This permits the interaction with an AP by AI agents and other APs.

All the interactions of the readers with the AP are logged, and a periodical revision of the interactions is examined by the authors who can decide to improve the paper by adding content and responding to the readers' requests. This permits the implementation of the feedback mechanism in the architecture (the arrows pointing up in Figure 2) and stimulates new data collection and new research.

This limited demo provides an early, limited yet important insight into the feasibility of LLM integration to a specific knowledge base customized as an alternative to a print-ready publication with additional features than those found in generic LLMs.

4. Human vs. artificial agent knowledge consumption

This system is unique because it serves two kinds of consumers: human users (researchers, students, practitioners or the interested public) across diverse scientific fields and artificial agents (software systems or algorithms that utilize scientific knowledge). These two audiences have different needs and ways of processing information. We propose tailoring the knowledge representation and access modalities accordingly. We have in mind that when research agents interact with each other, they can form a positive feedback loop of innovation and improvement. Collaboration, specialization, automation and recursive improvement can lead to an intelligence explosion and accelerate progress rapidly, potentially outpacing human

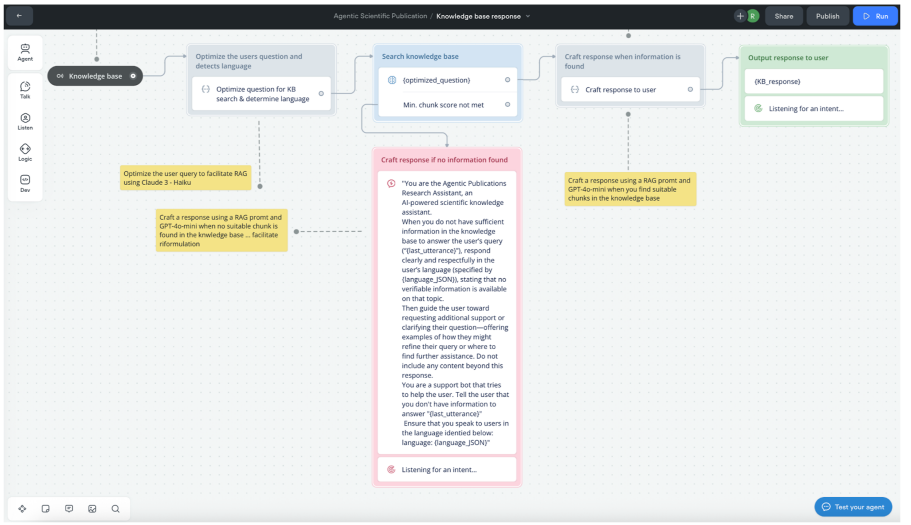

Figure 4. The Voiceflow development environment, a workflow to respond to the reader's queries using RAG. A high-resolution version is available at: https://doi.org/10.34965/I60500

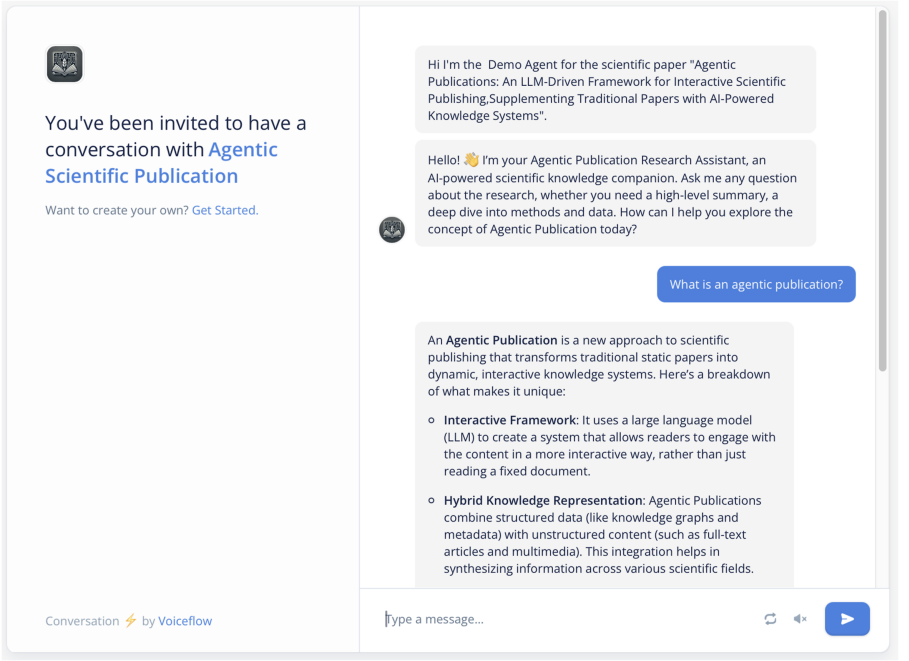

Figure 5. The Agentic Publication Q&A user interface in action. A high-resolution version is available at: A high-resolution version is available at: https://doi.org/10.34965/I60500

capabilities. Agents themselves have among their possibilities the capability to adapt to the specific audience, and this is not only a matter of user interface (chat or API).

4.1 Knowledge dissemination for human users

For human users, the priority is clarity, context and usability. The system functions as a knowledgeable assistant or interactive textbook, adapting to user expertise levels.

Narrative explanations form the foundation of effective knowledge exchange. Humans understand through explanations, analogies and narratives, so the LLM should present answers with sufficient context and background adapted to specific scientific domains. Complex questions might begin with brief concept explanations before proceeding to specific findings rather than assuming prior knowledge. Tone adjusts appropriately – formal for seasoned scientists, pedagogical for students – while always aiming to educate rather than output raw facts.

Although many examples in this paper focus on experimental sciences, APs are equally suited to disciplines where reasoning, interpretation and narrative are the primary scholarly contributions. In the social sciences and humanities, the AP knowledge base is constructed not from experimental datasets but from textual corpora, historical archives, theoretical frameworks and critical arguments. Crucially, the author's intellectual contribution – interpretation, synthesis, critique and perspective – is preserved as a core layer of the AP, not as an optional add-on. Authors can define reasoning profiles that explicitly encode their standpoint (e.g. *"interpret this passage through a deconstructionist lens," "align responses with the author's prior work on postcolonial theory"* or *"exclude realist approaches in international relations analysis"*). In this way, the AP makes reasoning itself transparent, auditable and reusable, while ensuring that authorial judgment, originality and interpretive stance remain central to scholarly credit. Verification mechanisms in this context operate not by recalculating statistics, but by checking fidelity to sources, consistency across arguments and by highlighting alternative interpretations. Thus, APs can capture and disseminate the intellectual reasoning that lies at the heart of social sciences and humanities scholarship while making the author's interpretive role more explicit and enduring than in traditional static publications.

Building upon clear explanations, interactive exploration capabilities enhance learning by allowing users to explore why and how. Beyond initial answers, users can ask follow-ups like "Why do these studies disagree?" or "How reliable is this data?" The system provides drilling-down options, suggesting "Would you like to see key studies supporting this conclusion?" or "See data breakdown contributing to this answer." These mimics read references in review papers or examine appendix data but in a more guided manner.

Complementing textual information, visualization and summarization tools help users grasp complex concepts efficiently. On-the-fly visual summaries include bar charts for comparisons or line plots for trends over time, accompanied by clear captions. The interface offers click-based summaries where users highlight concepts for concise definitions, helping those unfamiliar with jargon catch up without leaving the interface.

The foundation of user adoption rests on trust and transparency mechanisms that reveal the system's reasoning process. Every factual statement should be traceable to sources through hover text or footnotes. "Show how this answer was derived" options reveal consultation of multiple studies with human-friendly descriptions like "This answer is based on five studies published between 2018 and 2021, which consistently found X." The system produces mini "methods sections" describing answer aggregation, addressing AI black-box concerns.

Recognizing the diversity of user needs, adaptability and personalization features tailor experiences to individual preferences. Clinicians might prefer patient outcome contexts while basic scientists want mechanistic insights. User profiles or preference settings enable customization, with the system inferring values from interactions, proactively including statistical details for users who frequently drill into such information. Educational settings might feature "quiz me" modes or interactive dialogues for topic learning.

Ultimately, the focus remains on communication for human consumption, making vast scientific knowledge accessible and comprehensible through natural language generation and thoughtful UI design that aligns with human learning and inquiry patterns.

4.2 Knowledge access for artificial agents

Despite the selected API or interaction protocol (A2A, ACP or new standards), artificial agents (such as other AP, other AI systems, software tools or robots) require knowledge in more structured and unambiguous forms that they can process effectively, while still capturing essential contextual information. Several design considerations enable effective machine-to-machine knowledge transfer.

Machine-readable formats represent the foundation of effective agent communication. In addition to natural language narratives, all information in the knowledge base should be accessible via structured data formats. This includes the knowledge graph of scientific claims, which agents can query via SPARQL or Cypher, and numerical data in standardized formats (CSV, JSON). When an agent queries the system, instead of receiving a paragraph, it might receive a JSON response with fields like answer_summary, supporting_studies (list of DOIs) and data_points (actual numeric values). For example, an agent asking "What is the band gap of material X according to the latest research?" could get a response object containing a numeric value range, units and references. This structured answer allows the agent to directly use the data without extra parsing.

To ensure precision and consistency, ontologies and identifiers are critical in knowledge representation. To avoid ambiguity common in natural language, the system should map knowledge to shared ontologies and use unique identifiers for entities. For instance, while humans might see "Vitamin B12," agents would benefit from a CHEBI ID for B12 or a PubChem ID. Linking a disease to ontologies like MeSH or ICD codes ensures consistency if a disease is mentioned. A knowledge graph can serve this role by linking different names to the same node (synonym resolution) so that agents querying "cobalamin" and "Vitamin B12" get identical results. This aligns with FAIR data principles, emphasizing unique identifiers and comprehensive metadata for machine usability.

For efficient access, API endpoints for agents provide specialized interfaces that directly serve machine needs. We can create specialized endpoints like a/facts endpoint, where agents can request something like */facts ? subject = CompoundX&relation = affectsandobject = DiseaseY* to get all known facts connecting CompoundX and DiseaseY (with references and confidence scores). Another example is a/data endpoint where agents can fetch raw datasets of particular studies by ID, enabling external analysis. Other systems (like data visualization tools or hypothesis generation algorithms) can seamlessly draw on the knowledge base by offering these granular access points.

The knowledge ecosystem can be enriched through agents as contributors participating in knowledge creation. Not only will agents consume knowledge, but they might also contribute back (Saito and Tsukiyama, 2024). For instance, a data mining agent might scan the latest chemistry preprints and add new entries automatically, or an agent could run meta-analyses on data subsets and insert results. To facilitate this, the system's input API should accept submissions from authorized agents, treating them similarly to human submissions with appropriate validation. Over time, a network of specialized agents could surround the central LLM, each feeding it curated information from different domains.

Supporting dynamic workflows, real-time reasoning capabilities address time-sensitive knowledge needs. Some agents might query the system as part of larger reasoning pipelines. For example, while diagnosing a patient, an AI physician system might query the AP system for the latest research on a rare disease. It needs answers quickly and in a format that can integrate with patient data. This demands that responses to agent queries are concise, formalized and reliably structured. Optimizing the API for performance (caching common queries and prioritizing computational resources for high-frequency machine requests) becomes necessary when serving agents at scale.

Error handling and robustness mechanisms must be integrated into the system design to ensure reliable interactions. When dealing with agents, misunderstandings that humans might spot won't be noticed by consuming programs. So, the system must provide, along with answers, measures of confidence or validity. Agents can then decide to trust the data or seek

alternatives. For example, the API could include a confidence_score field or a warnings list (like "conflicting evidence present"). This way, agents can be programmed to handle uncertain answers appropriately. The system should practice defensive communication when interacting with machines, anticipating that outputs might be taken at face value.

In summary, we transform the richly detailed, human-facing knowledge into a precise, codified knowledge base interface for artificial agents. The same underlying data support both layers; the presentation and access method differ. By catering to machines, we ensure knowledge is truly computationally accessible, allowing AI systems to participate in scientific workflows directly (from literature analysis to experimental design) (Eger *et al.*, 2025). The traditional paper, which "remains mostly inaccessible to automated approaches" (Bucur *et al.*, 2022), is transcended by a format where knowledge is as legible to computers as it is to people.

5. Ethical considerations

Augmenting scientific publishing across all disciplines, including social sciences, with an AI-driven system raises a host of ethical and societal considerations (Koçak, 2024; Fornalik *et al.*, 2024; Ros and Samuel, 2024; Ajiye and Omokhabi, 2025; Yousaf, 2025). It is imperative to address these proactively to ensure the system advances knowledge sharing responsibly. In the following, we discuss key ethical issues and how our proposed model can mitigate them. As APs gain autonomous decision-making capabilities, ethical oversight must evolve beyond data fairness and bias detection. Responsible design should include boundaries on agent autonomy, especially in scientific fields where misinterpretation or overreach could have real-world consequences. Embedding value-aligned goals into agents ensures they remain instruments of support, not decision-makers beyond their scope.

5.1 Bias and fairness

Every dataset and model has biases, and LLMs trained on scientific literature reflect literature biases, including research focus disparities, gender/racial biases in study populations, underrepresentation of specific communities and publication bias toward positive results.

Promoting fairness requires actively monitoring and correcting biases through diverse knowledge base incorporation – research from global regions, multiple languages and varied publication venues, including fewer mainstream journals. Bias detection algorithms and agents can periodically analyze system responses, identifying systematic favoritism toward certain assumptions. User feedback mechanisms regarding biased outputs can create continuous improvement cycles. Development requires diverse stakeholders, including ethicists, under-represented community representatives, and globally distributed domain experts, to align system behavior with inclusive values.

The goal is "designing out" current system inequities – if certain voices are marginalized in publishing, the AI system could surface their contributions more visibly rather than amplifying existing bias. This requires continuous vigilance mechanisms rather than one-time fixes.

5.2 Accuracy, misinformation and hallucinations

LLMs sometimes hallucinate, generating plausible but incorrect statements, which is dangerous in scientific contexts. Meta's Galactica model exemplified this risk, confidently producing incorrect information and fake citations before withdrawal.

Safeguards include retrieval-augmented approaches, grounding answers in actual literature, always providing sources, and defaulting to "evidence is inconclusive" when uncertain rather than fabricating results. Verification agents catch hallucinations – reference verifiers flag non-existent citations for correction using only real, retrieved references.

Transparency and traceability enable error tracking through source backing and audit trails, logging all answers and generation processes. Users can flag incorrect answers for system maintainer analysis. The AP system should avoid absolute terms when unwarranted, including

phrases indicating consensus degree for nuanced rather than misleading definitive statements when science remains uncertain.

5.3 Transparency and trust

Scientists and the public require high transparency and accountability from AI knowledge systems. Open processes document inclusion criteria, model training methods and verification algorithms for community inspection. Open science project approaches – with model weights, code and knowledge base openly accessible – allow independent evaluation and community-driven improvement. Regular performance reports publish system metrics, creating accountability through measurable indicators.

User education encourages a critical approach to AI answers, treating them with the same scrutiny as single-paper claims. Interface reminders suggest verifying critical results from original sources, fostering healthy, informed skepticism during the transition period of learning to work with AI tools.

5.4 Ethical use and abuse

Manipulation concerns include submitting misleading "research" to skew answers toward particular agendas. Robust verification agents and source vetting provide primary defense, heavily trusting sources passing rigorous checks while avoiding status-quo bias by considering innovative research from less prominent sources. Content safeguards should detect and refuse dangerous or unethical requests.

For privacy considerations, the system should handle unpublished results or patient data through appropriate confidentiality measures and aggregation techniques, preventing re-identification. Tiered access levels keep general knowledge open while only securing sensitive data for authorized individuals.

Intellectual property and credit address researchers' concerns about recognition and ownership through clear contribution attribution and citation-like credit systems, and tracking when contributed results helps answer queries. Legal frameworks focus initially on open-access and author-submitted content to avoid copyright violations.

LLM limitations include hallucination susceptibility, limited context windows, lack of genuine comprehension, static knowledge and training data biases. Mitigation requires a rigorous validation protocol, agents and explicit sourcing requirements for reliability in scientific contexts.

5.5 Governance and control

Given its potential power as a science access method, who controls this knowledge system platform is critical. Governance models include international consortium collaboration (universities, libraries and academies), nonprofit stewardship, avoiding profit motives conflicting with open knowledge and community involvement through user councils or transparent feedback mechanisms.

The system must serve the scientific community and society, not corporate interests. Funding sources include public research grants treating it as critical infrastructure, philanthropy, consortium models with university contributions or commercial API tiers subsidizing free academic/public use while avoiding pay-to-play disparities.

5.6 Alignment with scientific values

The AI must align with scientific values: truth-seeking, skepticism, openness and rigor. This means accurately representing consensus and uncertainty rather than favoring exciting results, acknowledging ignorance when evidence is weak, and demonstrating intellectual honesty.

Facilitating critical thinking involves sometimes answering questions with questions or suggesting experiments to engage users in scientific inquiry rather than positioning AI as a

knowledge oracle. When errors occur, the system should correct them forthrightly and proactively notify users of updates, a service traditional journals rarely provide.

New AI-discovered knowledge requires rigorous scrutiny and human testing, maintaining healthy synergy between artificial and human intelligence. The system must carefully handle value-laden research where findings intersect with social, political or ethical considerations unresolvable through data alone.

The ethical framework is equal to the technical framework in importance. Embedding responsible AI principles – bias mitigation, transparency, accountability and human oversight – aims to build platforms accelerating knowledge sharing equitably and trustworthily while augmenting rather than distorting human intellect and scientific processes.

6. Challenges and future directions

Implementing the proposed AP system for scientific knowledge dissemination is undoubtedly ambitious. There are significant challenges to overcome and exciting opportunities for future development. We discuss some of the main hurdles and how the system might evolve.

6.1 Challenges

Ensuring accuracy across millions of ingested claims presents the fundamental challenge, as even tiny error rates could allow many faulty statements through. Near-zero tolerance for critical errors is essential, especially in medicine, requiring continuous AI refinement and tiered approaches with extra human scrutiny for high-stakes information.

Infrastructure demands create substantial hurdles through expensive GPU/TPU requirements, ample storage and bandwidth needs. Unlike traditional publishing's human labor and low-tech distribution, this requires cutting-edge hardware and engineering. Scaling for global usage may hit technical limits, particularly inference speed bottlenecks where lengthy answers frustrate users expecting instant results. Defining an appropriate governance model for a decentralized global system remains an open challenge. Decisions are needed on how to coordinate independent nodes, ensure interoperability, and establish sustainable funding and management structures for a shared scientific infrastructure. Cost sustainability requires hybrid funding models combining government, private, and volunteer computing resources.

Given centuries of entrenched practices, researcher adoption remains uncertain despite potential perfect functionality. Scholar resistance persists since career advancement is tied to traditional publications. Cultural shifts must recognize knowledge base contributions academically, while initial skepticism requires demonstrating reliability through validations comparing AI syntheses to expert reviews. Modified incentive structures acknowledging AI-era contributions represent crucial non-technical challenges.

For APs to gain acceptance, authorship and attribution must remain central. In our model, the AP itself is the primary citable entity, with versioned, frozen snapshots (each with its own DOI) available to satisfy traditional committees requiring fixed artifacts. Authorship is anchored in ORCID identities and CRediT roles, recorded in the AP's provenance ledger to capture diverse contributions including data, code, analysis, authorial reasoning, and verification. This approach preserves traditional forms of recognition while making attribution more transparent and machine-verifiable than in static papers. Beyond citations, APs can support extended impact metrics: for instance, recorded "intelligent interactions" (e.g. verified queries, downloads, API calls), endorsements from expert agents or human reviewers, and evidence of reuse across fields. Together, these metrics form an AP Impact Profile, analogous to citation counts or journal impact factors, but more flexible and reflective of real scientific engagement. By combining traditional identifiers with richer digital provenance, APs ensure that contributing scientists can advance their careers at least as effectively as through conventional publications, while also benefiting from new, auditable measures of scholarly influence.

A common concern is that AI systems may displace scientists as the primary thinkers in society, relegating them to providers of raw data while machines generate the reasoning. APs are designed explicitly to counter this scenario. In our model, the author's intellectual contribution—thinking, interpretation, and reasoning—remains central and irreplaceable. The AP makes these contributions more visible rather than less: every interpretation, critique, or perspective added by the author becomes a signed, versioned component of the AP's reasoning layer. Authors can even encode their interpretive stance through reasoning profiles (e.g. *"all responses should be compatible with E. views," "align answers with the author's prior work," "exclude string-theory perspectives"*). These profiles ensure that human reasoning guides the LLM, not the other way around. Far from diminishing scholarly thought, APs amplify it: they preserve intellectual arguments alongside data, make reasoning auditable, and allow future readers and systems to engage with the author's thought process directly. In this way, APs safeguard the scientist's role as a thinker while extending the reach, transparency, and influence of their reasoning.

Coexistence with traditional publishing is fundamental. Double-reporting raises canonicity questions when papers and AI submissions diverge. Citation protocols require permanent identifiers, while journal policies must address whether AI systems constitute prior publication. Gradual hybrid models where journals adopt systems internally or become data providers may facilitate adoption.

Some knowledge types resist current AI representations. Mathematical proofs, theoretical arguments, and visual insights may be unsuitable for text-based processing, while systems struggle with symbolic reasoning and qualitative analysis. The risk exists of losing science's narrative quality, discovery stories, and reasoning insights that well-written papers provide. Overcoming limitations requires sophisticated mathematical representation and ensuring systems handle arguments beyond facts.

A further challenge concerns the reliance on manuscripts and preprints as input to APs. Similar to traditional preprint servers, early-stage versions may contain incomplete, inconsistent, or even erroneous information. This risk is amplified in APs, since models might internalize and propagate such content. One technical question is whether models should be "retrained" whenever a revised or peer-reviewed version becomes available. While continual learning approaches (e.g. fine-tuning with LoRA, retrieval-augmented updates) make partial revision feasible, fully removing prior information from a statistical model remains technically difficult and is not equivalent to training from scratch on final, peer-reviewed material. This limitation must therefore be acknowledged explicitly.

To mitigate the issue, we propose introducing a dedicated **versioning agent** that records and manages successive versions of each AP, from initial manuscript through peer-reviewed publication. New versions would be ingested with rigorous prompting protocols (e.g. "this is the latest version, base reasoning on this"; "this revision corrects x, y, z identified in v2.3"; "this version adds Section 5 in response to reviewer comments"). By explicitly encoding temporal and revision metadata, the system can privilege the most recent, peer-reviewed version in reasoning while retaining access to earlier versions for transparency and provenance. Such version control requires integration with established databases and identifier systems beyond the LLM itself, but is technically feasible and aligns with practices in software and data versioning.

Although the field is still nascent and few systematic studies exist on the long-term effects of training on preprints, acknowledging this risk and outlining practical safeguards is critical. We see this as an important avenue for future targeted research on maintaining reliability and trustworthiness in agentic publishing ecosystems.

6.2 Future directions

Enhanced reasoning and hypothesis generation could transform the system from passive Q&A to an active scientific contributor. As vast knowledge accumulates, the system could propose

insights and identify research opportunities by scanning for anomalies or under-explored connections, alerting scientists: "No one has tested compound X in context Y, though related data suggests promise." This enables AI-driven discovery through agents running atop the knowledge base to generate vetted hypotheses, closing loops where AI helps create knowledge, not just disseminate it. In social sciences, systems could assist qualitative analysis by identifying themes in interview transcripts or aid theory building by highlighting inconsistencies between theories and empirical findings.

Personalized knowledge curation could provide each user with tailored AI researchers using the global knowledge base, but adapting to specific interests. Researchers could subscribe to topics for proactive development notifications or receive draft paper suggestions for relevant citations and related work. This hyper-personalization makes vast scientific information individually navigable through user profiles and advanced recommendation systems.

Multilingual and cross-disciplinary expansion addresses science's global nature. LLMs enable truly multilingual systems that ingest papers in multiple languages and offer Q&A in those languages, breaking down barriers. Cross-disciplinary synthesis could connect computer science algorithms to biological needs by noticing analogous problems. Discovery modes deliberately pulling seemingly unrelated information could spur creative thinking more effectively than siloed human approaches.

Real-time experimental data integration represents an exciting frontier where lab instruments feed data directly into knowledge systems for immediate LLM analysis and interpretation suggestions. This blurs experiment-publication lines through self-documenting experiments immediately connected to prior knowledge, supporting longer-term "AI Scientist" visions of automated research pipelines from data collection to dissemination.

Robust evaluation frameworks become essential as systems mature, since traditional precision/recall metrics prove insufficient. Future benchmarks for "AI literature review quality" could simulate assistance with real review papers rated by human experts. Continual evaluation, including stress-testing with adversarial inputs, will guide improvements in this emerging "AI scientific knowledge evaluation" field.

Legal and policy frameworks must evolve to formalize system roles through institutional guidelines for research workflow integration. Grant agencies might require result deposits upon project completion, universities could train effective system use and citation, while journals might evolve into knowledge base validators rather than separate article producers.

Automatic verification methods and agents analogous to software unit testing could systematically check LLM outputs against factual assertions and logical consistency criteria, alerting potential inaccuracies. AP cross-talk enables multiple publications to communicate as independent or unified agents, facilitating dynamic dialogue and collaborative reasoning across domains. Agentic-to-classical conversion capabilities transition interactive publications to print formats tailored by user preferences for style, language, and detail level.

Dataset interaction transforms static supplements into dynamic analytical tools where integrated LLMs enable direct querying, visualization, and computation, empowering readers to explore insights actively and pursue new research questions.

This vision requires incremental realization through interdisciplinary collaboration among AI experts, domain scientists, librarians, ethicists, and others while carefully managing human elements and maintaining scientific quality standards. The potential transformation holds immense promise for creating platforms where knowledge flows more freely, potentially enabling any researcher to instantly consult human scientific knowledge as easily as a conversation.

Declaration of generative AI and AI-assisted technologies in the writing process

While preparing this work, the authors used Local Gwen 32 B to do text editing, improvement, and summarization. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the published article.

Acknowledgments
The authors gratefully acknowledge the constructive comments and suggestions provided by the anonymous reviewers, which have significantly contributed to improving the clarity and quality of this work.

References

Ahaley, S.S., Pandey, A., Juneja, S.K., Gupta, T.S. and Vijayakumar, S. (2023), "ChatGPT in medical writing: a game-changer or a gimmick?", *Perspectives in Clinical Research*, Vol. 15 No. 4, pp. 165-171, doi: 10.4103/picr.picr_167_23.

Ajiye, O.T. and Omokhabi, A.D.A. (2025), "The potential and ethical issues of artificial intelligence in improving academic writing", *ShodhAI: Journal of Artificial Intelligence*, Vol. 2 No. 1, pp. 1-9, doi: 10.29121/shodhai.v2.i1.2025.24.

Binz, M., Alaniz, S., Roskies, A.L., Aczél, B., Bergstrom, C.T., Allen, C., Schad, D., Wulff, D., West, J.D., Zhang, Q., Shiffrin, R.M., Gershman, S.J., Popov, V., Bender, E.M., Marelli, M., Botvinick, M.M., Akata, Z. and Schulz, E. (2025), "How should the advancement of large language models affect the practice of science?", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 122 No. 5, e2401227121, doi: 10.1073/pnas.2401227121.

Bornmann, L., Haunschild, R. and Mutz, R. (2021), "Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases", *Humanities and Social Sciences Communications*, Vol. 8 No. 1, p. 224, doi: 10.1057/s41599-021-00903-w.

Bucur, C., Kuhn, T., Ceolin, D. and Ossenbruggen, J.V. (2022), "Nanopublication-based semantic publishing and reviewing: a field study with formalization papers", *PeerJ Computer Science*, Vol. 9, e1159, doi: 10.7717/peerj-cs.1159.

Cardon, P.W. (2023), "Searching for the right Metaphors to understand and interrogate the AI age", *Business Communication Research and Practice*, Vol. 6 No. 2, pp. 65-69, doi: 10.22682/bcrp.2023.6.2.65.

Daykan, Y. and O'reilly, B. (2023), "The impact of artificial intelligence on academic life", *International Urogynecology Journal*, Vol. 34 No. 1661, p. 1661, doi: 10.1007/s00192-023-05613-2.

Dias, F.S., Moroni, A.S. and Pedrini, H. (2023), "Using generative models to create a visual description of climate change", *Arts and Technology*, doi: 10.1007/978-3-031-55319-6_14.

Drozdz, J.A. and Ladomery, M.R. (2024), "The peer review process: past, present, and future", *British Journal of Biomedical Science*, Vol. 81, 12054, doi: 10.3389/bjbs.2024.12054.

Eger, S., Cao, Y., D'Souza, J., Geiger, A., Greisinger, C., Gross, S., Hou, Y., Krenn, B., Lauscher, A., Li, Y., Lin, C., Moosavi, N.S., Zhao, W. and Miller, T. (2025), "Transforming science with Large Language Models: a survey on AI-assisted scientific discovery, experimentation, content generation, and evaluation", arXiv preprint arXiv:2502.05151, doi: 10.48550/arXiv.2502.05151.

Emile, S.H., Hamid, H.K.S., Atici, S.D., Kosker, D.N., Papa, M.V., Elfeki, H., Tan, C.Y., El-Hussuna, A. and Wexner, S.D. (2022), "Types, limitations, and possible alternatives of peer review based on the literature and surgeons' opinions via Twitter: a narrative review", *Science Editing*, Vol. 9 No. 1, pp. 3-14, doi: 10.6087/kcse.257.

Fornalik, M., Makuch, M., Lemanska, A., Moska, S., Wiczewska, M., Anderko, I., Stochaj, L., Szczygiel, M. and Zielińska, A. (2024), "Rise of the machines: trends and challenges of implementing AI in biomedical scientific writing", *Exploration of Digital Health Technologies*, Vol. 2 No. 5, pp. 235-248, doi: 10.37349/edht.2024.00024.

Freire, S.K., Wang, C., Foosherian, M., Wellsandt, S., Ruiz-Arenas, S. and Niforatos, E. (2024), "Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking", *Frontiers in Artificial Intelligence*, Vol. 7, 1293084, doi: 10.3389/frai.2024.1293084.

Haffar, S., Bazerbachi, F. and Murad, M.H. (2019), "Peer review bias: a critical review", *Mayo Clinic Proceedings*, Vol. 94 No. 4, pp. 670-676, doi: 10.1016/j.mayocp.2018.09.004.

Hou, X., Zhao, Y., Wang, S. and Wang, H. (2025), "Model context protocol (mcp): landscape, security threats, and future research directions", arXiv preprint arXiv:2503.23278, doi: 10.48550/arXiv.2503.23278.

Hu, Y., Liu, D., Wang, Q., Yu, C., Ji, H. and Xiong, J. (2024), "Automating knowledge discovery from scientific literature via LLMs: a dual-agent approach with progressive ontology prompting", arXiv preprint arXiv:2409.00054, doi: 10.48550/arXiv.2409.00054.

Huang, J. (2021), "Leveraging big data and machine learning for digital transformation", *Journal of Integrated Design and Process Science*, Vol. 23 No. 3, pp. 1-3, doi: 10.3233/JID190020.

Hughes, R.C. and Van Heerden, A.C. (2024), "PLOS-LLM: can and should AI enable a new paradigm of scientific knowledge sharing?", *PLOS Digital Health*, Vol. 3 No. 4, e0000501, doi: 10.1371/journal.pdig.0000501.

Jen, S.L. and Hj Salam, A.R. (2024), "A systematic review on the use of artificial intelligence in writing", *International Journal of Academic Research in Progressive Education and Development*, Vol. 13 No. 1, pp. 1819-1829, doi: 10.6007/IJARPED/v13-i1/20584.

Kelly, J., Sadeghieh, T. and Adeli, K. (2014), "Peer review in scientific publications: benefits, critiques, & a survival guide", *EJIFCC*, Vol. 25 No. 3, pp. 227-243.

Klami, A., Damoulas, T., Engkvist, O., Rinke, P. and Kaski, S. (2024), "Virtual laboratories: transforming research with AI", *Data-Centric Engineering*, Vol. 5, e19, doi: 10.1017/dce.2024.15.

Koçak, Z. (2024), "Publication ethics in the era of artificial intelligence", *Journal of Korean Medical Science*, Vol. 39 No. 33, e249, doi: 10.3346/jkms.2024.39.e249.

Kozak, M. (2025), "The science publishing manifesto: AI moves fast, science publishing must too", *Journal of Documentation*, Vol. 81 Nos 5-6, pp. 1357-1369, doi: 10.1108/JD-04-2025-0118.

Künzli, N., Röösli, M., Ammann, P., Janssen, N.A., Künzli, N. and Brauer, M. (2022), "The end of peer review in public health sciences?", *Environmental Health Perspectives*, Vol. 130 No. 3, 034501.

Lee, C.J., Sugimoto, C.R., Zhang, G. and Cronin, B. (2013), "Bias in peer review", *Journal of the American Society for Information Science and Technology*, Vol. 64 No. 1, pp. 2-17, doi: 10.1002/asi.22784.

Li, D. and Zhang, Z. (2023), "MetaQA: enhancing human-centered data search using Generative Pre-trained Transformer (GPT) language model and artificial intelligence", *PLoS One*, Vol. 18 No. 11, e0293034, doi: 10.1371/journal.pone.0293034.

Lin, Z. (2023), "Why and how to embrace AI such as ChatGPT in your academic life", *Royal Society Open Science*, Vol. 10, 230658, doi: 10.1098/rsos.230658.

Manchikanti, L., Kaye, A.D., Boswell, M.V. and Hirsch, J.A. (2015), "Medical journal peer review: process and bias", *Pain Physician*, Vol. 18 No. 1, pp. E1-E14.

Pugliese, R. (2025), *Demo of an Agentic Publication on the Concept of Agentic Publications*, Elettra-Sincrotrone Trieste S.C, p. A, doi: 10.34965/AGENTICPUBLICATION.3567A.

Ros, T. and Samuel, A. (2024), "Navigating the AI Frontier: a guide for ethical academic writing", *eLearn*, Vol. 10, doi: 10.1145/3703094.3694981.

Saito, H. and Tsukiyama, T. (2024), "Use of artificial intelligence in manuscript preparation-AI as a Co-author", *The International Journal of Periodontics and Restorative Dentistry*, Vol. 45 No. 3, pp. 1-12, doi: 10.11607/prd.7022.

Seghier, M.L. (2025), "AI-powered peer review needs human supervision", *Journal of Information, Communication and Ethics in Society*, Vol. 23 No. 1, pp. 104-116, doi: 10.1108/jices-09-2024-0132.

Smith, R.C., Winschiers-Theophilus, H., Loi, D., Abreu de Paula, R., Kambunga, A.P., Samuel, M.M. and Zaman, T. (2021), "Decolonizing design practices: towards pluriversality", in Kitamura, Y., Quigley, A., Isbister, K. and Igarashi, T. (Eds), *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA 2021 Article 83 Association for Computing Machinery, doi: 10.1145/3411763.3441334.

Tennant, J.P. and Ross-Hellauer, T. (2020), "The limitations to our understanding of peer review", *Research Integrity and Peer Review*, Vol. 5 No. 1, pp. 1-19, doi: 10.1186/s41073-020-00092-1.

Van Noorden, R. (2023), "The science publishing boom: how AI is changing academia", *Nature*, Vol. 615 No. 7951, pp. 192-194.

Van Noorden, R. and Perkel, J.M. (2023), "AI writing tools could hand scientists the gift of time", *Nature*, 7934, pp. 20-22, pp. 611.

Wang, H. and Shi, Y. (2025), "Knowledge graph combined with retrieval-augmented generation for enhancing LMs reasoning: a survey", *Academic Journal of Science and Technology*, Vol. 14 No. 1, pp. 227-235, doi: 10.48550/arXiv.2503.10677.

Weingart, P. (2025), "The changed publishing culture of science", *European Review*, Vol. 33 No. S1, pp. 1-16, doi: 10.1017/s1062798725000183.

Willis, M. (2016), "Why do peer reviewers decline to review manuscripts? A study of reviewer invitation responses", *Learned Publishing*, Vol. 29 No. 1, pp. 5-7, doi: 10.1002/leap.1006.

Yousaf, M.N. (2025), "Practical considerations and ethical implications of using artificial intelligence in writing scientific manuscripts", *ACG Case Reports*, Vol. 12 No. 2, e01629, doi: 10.14309/crj.0000000000001629.

Further reading

Bennett, L. and Abusalem, A. (2024), "Artificial intelligence (AI) and its potential impact on the future of higher education", *Athens Journal of Education*, Vol. 11 No. 3, pp. 195-212, doi: 10.30958/aje.11-3-2.

Doskaliuk, B., Zimba, O., Yessirkepov, M., Klishch, I. and Yatsyshyn, R. (2025), "Artificial intelligence in peer review: enhancing efficiency while preserving integrity", *Journal of Korean Medical Science*, Vol. 40 No. 7, e92, doi: 10.3346/jkms.2025.40.e92.

Pugliese, R. and Kourousias, G. (2025), *Dataset of Figures for Agentic Publication Manuscript [Data Set]*, Elettra Sincrotrone Trieste, doi: 10.34965/I60500, available at: https://scicompapps.elettra.eu/agenticpub_demo/

Schinkel, M., Paranjape, K. and Nanayakkara, P.W. (2023), "Written by humans or artificial intelligence? That is the question", *Annals of Internal Medicine*, Vol. 176 No. 4, pp. 572-573, doi: 10.7326/M23-0154.

Sriram, N. (2025), "Harmonizing innovation and integrity: ethical perspectives on artificial intelligence (AI) in academic writing", *International Journal of Pharmaceuticals and Health Care Research*, Vol. 13 No. 1, pp. 59-65, doi: 10.61096/ijphr.v13.iss1.2025.59-65.

Thurzo, A., Strunga, M., Urban, R., Surovková, J. and Afrashtehfar, K.I. (2023), "Impact of artificial intelligence on dental education: a review and guide for curriculum update", *Education Sciences*, Vol. 13 No. 2, p. 150, doi: 10.3390/educsci13020150.

Corresponding author
Grazia Garlatti Costa can be contacted at: grazia.garlatticosta@units.it