Received: 26 January 2026 | Revised: 17 March 2026 | Accepted: 21 April 2026 | Published online: 7 May 2026

REVIEW

Medinformatics
2026, Vol. 00(00) 1–11
DOI: 10.47852/bonviewMEDIN62029207

AI Consensus Validation in Biomedical Research: A Review, Conceptual Framework, and Future Directions

Tan Aik Kah[1,*]

[1] *Clinique d'ophthalmologie, Normah Medical Specialist Centre, Malaysia*

Abstract: Introduction: The exponential growth of biomedical research has strained traditional peer review, creating delays and inconsistencies. Artificial intelligence (AI), particularly large language models (LLMs), offers potential for scalable, reproducible assessment of scientific content. Current tools remain siloed and lack a framework to leverage agreement among diverse models as a validation signal. Methods: We conducted a narrative and conceptual review of literature (2021 onward) from PubMed, arXiv, and IEEE Xplore, focusing on AI-assisted peer review, LLM evaluation, and validation frameworks. We propose AI Consensus Validation (AICV), where agreement among diverse LLMs serves as an early indicator of clarity, novelty, relevance, and conceptual soundness. Results: Our review identifies key gaps in the current landscape, including epistemic homogeneity, fragmentation, and opacity. AICV shifts the paradigm from single-model prediction to multi-model consensus, using convergence as a marker of epistemic robustness. We present the operational workflow of AICV and illustrate its application as a proof-of-concept with three biomedical abstracts, demonstrating its capacity to differentiate submissions based on convergence patterns. Discussion: AICV could be integrated into journal pre-screening, grant triage, and researcher feedback. Challenges include model biases, explainability gaps, and suppressing disruptive ideas. We outline a phased validation roadmap to translate AICV from concept to a trusted tool. Conclusion: AICV represents a promising middle path between fully manual peer review and opaque automation. Leveraging epistemic diversity across LLMs may accelerate biomedical innovation while preserving rigor and creativity if implemented with attention to bias, transparency, and human oversight.

Keywords: artificial intelligence, peer review, biomedical informatics, machine-assisted validation, LLM consensus

1. Introduction

The validation of scientific ideas remains a cornerstone of biomedical progress, yet the primary mechanism for this validation—peer review—is increasingly strained. With over 3 million biomedical articles published annually and submission volumes rising, the traditional model of manual, expert-driven review faces critical bottlenecks: prolonged decision timelines, inconsistent reviewer availability, and inherent subjectivity that can delay the translation of promising innovations into clinical practice [1, 2]. In fast-moving fields such as oncology, genomics, and AI-enhanced diagnostics, where the half-life of knowledge is short and clinical stakes are high, these delays are more than administrative—they represent missed opportunities for therapeutic advance and public health impact [3]. Concurrently, artificial intelligence (AI) has evolved from a specialized analytical tool into a transformative force across the biomedical research lifecycle. Large language models (LLMs) in particular now demonstrate surprising competency in parsing complex scientific text, evaluating argumentative coherence, and even gauging novelty relative to existing literature [4, 5]. While no AI system can replicate human clinical insight, ethical judgment, or domain-specific intuition, these technologies offer something fundamentally new: scalable, reproducible assessment of written scientific content. This capability invites a provocative, timely question: if multiple independently developed AI systems converge in their evaluation of a research idea, could that consensus itself serve as an early, objective signal of the idea's intellectual merit? Emerging tools already apply AI to elements of the editorial process—detecting plagiarism, checking statistical consistency, or suggesting relevant reviewers [6]. However, most operate as isolated utilities rather than integrated frameworks. More critically, few leverage the epistemic diversity of multiple AI systems as a deliberate validation strategy. Where existing approaches often seek to optimize the prediction of a single model, a consensus-based framework would shift focus from individual accuracy to collective agreement—a philosophical realignment from mechanistic explanation to observable, reproducible alignment among distinct AI "perspectives" [7]. This conceptual narrative review therefore proposes and examines AI Consensus Validation (AICV), a conceptual framework in which agreement among diverse LLMs—each trained on different data, developed by different institutions, and fine-tuned under distinct paradigms—is used to evaluate research abstracts across defined dimensions: clarity, novelty, scientific relevance, and conceptual

***Corresponding author:** Tan Aik Kah, Clinique d'ophthalmologie, Normah Medical Specialist Centre, Malaysia. Email: tanak@normah.com

soundness. Convergence among models is treated not as a predictive endpoint but as a marker of epistemic robustness, offering a reproducible, low-cost signal that could complement human expertise. Despite growing interest in AI-assisted peer review, a critical gap remains in how these technologies are conceptualized and operationalized. Most existing approaches focus on optimizing the performance of a single model to approximate human reviewer judgments, emphasizing predictive accuracy rather than epistemic robustness. As a result, current systems often suffer from epistemic homogeneity, limited transparency, and poor reproducibility across platforms, while offering little insight into how different AI systems might independently evaluate the same scientific idea. To date, no widely articulated framework has systematically leveraged agreement among multiple, independently developed LLMs as a validation signal for research quality. This review addresses this gap by proposing AICV, a conceptual framework that shifts the paradigm from single-model prediction to multi-model consensus. Instead of treating AI as a surrogate reviewer, AICV interprets convergence among diverse LLMs as an early indicator of epistemic robustness, providing a scalable and reproducible complement to human evaluation. By synthesizing the current landscape of AI-assisted peer review and demonstrating a proof-of-concept application, this work advances the field by introducing a structured, consensus-based approach to machine-assisted research assessment, offering a potential pathway toward more transparent, interoperable, and scalable validation systems in biomedical publishing. The objectives of this review are threefold: to systematically survey the current landscape of AI-assisted peer review and research assessment in biomedicine, to introduce and conceptually ground the AICV framework with biomedical examples, and to critically examine implementation pathways, ethical risks, and future research directions for machine-assisted validation. By synthesizing recent literature, illustrative case analyses, and ethical policy considerations, this work seeks to inform editors, funders, researchers, and AI developers about a middle path forward—one that harnesses machine intelligence to augment rather than automate scientific critique, potentially accelerating innovation while preserving the creative, human core of scientific discovery.

2. Review Methodology

This article constitutes a narrative and conceptual review with the primary objective of synthesizing the current landscape of AI-assisted peer review in biomedicine, identifying key gaps, and proposing a novel framework (AI Consensus Validation, or AICV) in response. The methodology was designed to provide a comprehensive, scholarly foundation for this proposal, balancing breadth with thematic depth in a rapidly evolving field.

2.1. Search strategy and information sources

To capture the breadth of relevant literature, a multi-source search strategy was employed, drawing on both indexed databases and preprint repositories. Electronic searches were conducted in PubMed to identify biomedical and clinical informatics publications, IEEE Xplore to access technical literature on AI and machine learning systems, and arXiv to capture the most recent, cutting-edge research and large-scale evaluations in machine learning and natural language processing applied to science. Searches were restricted to publications from January 2021 onward, reflecting the period coinciding with the widespread availability and application of advanced LLMs in scholarly contexts. The core search strategy combined terms related to the technology ("artificial intelligence," "large language model"), the process ("peer review," "research assessment"), and the domain ("biomedical," "clinical"), ensuring comprehensive coverage across disciplinary boundaries. In addition to peer-reviewed sources, gray literature was examined, including white papers from major academic publishers such as Elsevier and Springer Nature, as well as reports from research funding agencies, to identify implementation trends and policy discussions relevant to AI-assisted research evaluation.

2.2. Study selection and eligibility criteria

The literature selection process was guided by the objective of constructing a coherent, evidence-based narrative, and eligibility criteria were applied to ensure relevance and rigor. Studies were included if they described or empirically evaluated AI or LLM tools for the substantive assessment of research content, such as scoring novelty or checking methodological soundness; if they presented conceptual papers, editorials, or frameworks proposing structures for automated or augmented scientific review; or if they offered empirical studies or critical analyses examining the gaps, biases, ethical challenges, or implementation barriers of machine-assisted evaluation. Exclusion criteria encompassed purely technical descriptions of model architectures without an evaluative or applied component, non-substantive opinion pieces or news articles, and tools focused exclusively on pre-submission language editing or post-publication dissemination without a review component.

2.3. Data extraction and synthesis approach

Given the conceptual and heterogeneous nature of the literature, a formal meta-analysis was not feasible. Instead, a thematic narrative synthesis approach was adopted. This synthesis was informed by a targeted review of key publications and preprints that met the inclusion criteria, selected for their relevance in illustrating the evolving capabilities, implementations, and critical debates in AI-assisted peer review. Key information from included sources—including the AI approach, application context, reported strengths/limitations, and identified risks—was extracted and organized iteratively. Emerging themes were used to structure the review (Sections 3 and 4), focusing on (1) categorizing existing AI tools in the peer review workflow, (2) evaluating the capabilities and limitations of LLMs in research assessment, and (3) synthesizing the persistent conceptual, technical, and ethical gaps in the field. This synthesis directly informs the rationale for the AICV framework proposed in Section 5.

3. Current Landscape of AI in Peer Review

3.1. Overview of AI integration in scholarly communication

The integration of AI into scholarly workflows is no longer speculative but operational, with tools now embedded at multiple stages of the research lifecycle. In the context of peer review and manuscript assessment, AI applications generally fall into three overlapping categories: (1) pre-submission assistance (e.g., language polishing, formatting checks), (2) editorial management (e.g., reviewer recommendation, plagiarism detection), and (3) content evaluation (e.g., preliminary scoring, consistency

checking) (Table 1) [8]. This review focuses primarily on the third category—systems that assess the substantive content of research submissions—as these most directly inform the proposed AICV framework.

Commercial and open-source platforms have emerged, often developed through publisher-academic partnerships. For instance, automated reviewer matching systems leveraging natural language processing have been implemented since the early 2020s, with studies documenting their transformative potential [9]. Plagiarism detection tools such as Crossref's iThenticate have become standard in scholarly publishing, with validation studies reporting >95% accuracy in identifying text similarity [10]. More recently, publishers have piloted AI tools that move beyond plagiarism detection to assess argumentative coherence and methodological soundness, demonstrating potentials while maintaining editorial standards [6].

3.2. Large language models in research assessment

The advent of LLMs such as GPT, Gemini, and Claude has expanded the potential scope of AI-assisted evaluation. Unlike earlier rule-based or machine learning systems trained on narrow datasets, LLMs can generate contextual, nuanced appraisals of scientific text. Recent studies suggest LLMs can assess clarity, novelty, and logical consistency with surprising reliability, sometimes achieving inter-rater agreement comparable to that among human reviewers [3, 11, 12]. For example, Liang et al. (2023) found that GPT4's feedback overlapped with human reviewers at rates similar to those between two human reviewers, indicating that LLMs can provide useful and reliable scientific feedback [13].

However, these evaluations remain largely experimental and are not yet integrated into production editorial systems. Key limitations include a lack of transparency in scoring rationales [14], susceptibility to training data biases [15], and limited validation in real-world, high-stakes settings such as clinical trial reporting or grant adjudication [16, 17]. Moreover, most implementations rely on a single LLM provider, raising concerns about vendor lock-in, reproducibility, and epistemic homogeneity [18].

3.3. Identified gaps and conceptual limitations

Despite increasing activity in the development of AI-assisted review tools, significant gaps persist between their potential and current implementation. Many systems remain fragmented and siloed, being developed for specific platforms or publishers, which limits interoperability and creates inconsistent standards across the scholarly ecosystem. A further challenge lies in opacity and explainability, as numerous tools function as "black boxes," providing scores or recommendations without clear rationales, thereby undermining trust and reducing their utility for author feedback. Validation efforts are often narrowly focused on technical performance metrics such as accuracy or speed, with limited attention to broader impacts on editorial workload, reviewer behavior, or publication bias. Epistemic homogeneity also poses a risk, since most tools rely on a single model or dataset, amplifying embedded biases and lacking mechanisms for "second opinions" from diverse AI systems. Finally, the absence of robust ethical and governance frameworks leaves a vacuum in areas such as data privacy, algorithmic accountability, and appeal mechanisms for AI-driven decisions. Collectively, these limitations highlight the need for a framework that is interoperable, transparent, validated in real-world contexts, epistemically diverse, and ethically grounded, thereby motivating a paradigm shift in how AI is integrated into scientific review.

4. Synthesis: Critical Gaps and the Need for a New Paradigm

The preceding review of the current landscape reveals significant activity and promise in applying AI, particularly LLMs, to biomedical peer review. However, a synthesis of these developments highlights several interconnected conceptual and technical gaps that limit the field's maturation from experimental tools to trusted, scalable infrastructure.

4.1. Dominance of the single-model paradigm and epistemic homogeneity

The predominant approach involves optimizing a single LLM to predict human reviewer scores or decisions [9, 13]. This paradigm, while computationally straightforward, inherits critical flaws: it amplifies the specific biases embedded in that model's training data [15, 19] and lacks a mechanism for epistemic diversity. A tool built on a single model provides, in effect, a monolithic

Table 1. Taxonomy of AI tools in scholarly peer review workflows

Category	Description	Common tasks	Example tools/services	Evidence level	Refs.
Pre-submission assistance	AI tools that help authors prepare manuscripts before submission	Language editing, formatting compliance, reference management, figure checking	Grammarly, Writefull, Paperpal, journal-specific checkers	Commercially established, user-reported benefits	[8]
Editorial management	Systems that assist journal staff in managing the review process	Reviewer matching, plagiarism detection, conflict identification, workflow routing	Editorial Manager AI modules, Crossref Similarity Check, proprietary publisher systems	Industry implementation, limited independent validation	[8, 9]
Content evaluation	Tools that assess the substantive quality of research submissions	Preliminary scoring, methodological soundness checks, novelty detection, consistency verification	PLOS ONE AI screener, custom LLM implementations, meta-reanalysis tools	Pilot studies, academic research, experimental	[8, 11]

"opinion," unable to simulate the plurality of perspectives that underpin robust scientific critique [7]. This creates a risk of vendor lock-in and reduces the reproducibility of assessments across different technical platforms [18].

4.2. Fragmentation and the lack of interoperable frameworks

Current AI applications remain siloed within specific publisher platforms or exist as isolated utilities (e.g., plagiarism checkers, reviewer matchers) [6, 8]. There is a notable absence of an interoperable framework that can integrate multiple assessment dimensions (clarity, novelty, soundness) from diverse AI agents into a coherent evaluative signal. This fragmentation impedes the development of consistent standards and benchmarks for AI-assisted review across the scholarly ecosystem.

4.3. The opacity-utility trade-off in AI assessment

Many existing systems function as "black boxes," providing scores or recommendations without transparent, actionable rationales [14, 19]. This opacity severely limits their utility for providing constructive author feedback and undermines editorial and researcher trust. While some rubric-guided approaches are emerging [12], the field lacks a framework that systematically links quantitative scores to interpretable qualitative justifications in a standardized output format.

4.4. Narrow technical validation versus holistic workflow impact

Evaluation studies frequently focus on narrow technical metrics such as accuracy or speed in replicating human judgments [11, 13]. There is insufficient investigation into the broader sociotechnical impact of these tools on editorial workload, reviewer behavior, publication bias, or the diversity of scientific ideas that reach publication. A framework is needed that is designed from the outset to measure not just prediction fidelity but its effect on the entire research evaluation workflow.

4.5. Converging toward a consensus-based paradigm

These gaps—epistemic homogeneity, fragmentation, opacity, and narrow validation—collectively point toward a necessary paradigm shift. The next evolutionary step for AI-assisted review is not a more accurate single model but a framework that leverages agreement and disagreement among diverse, independently constituted AI systems. Such a framework would shift the validation signal from *individual model accuracy* to collective, inter-model consensus, directly addressing the need for robustness, transparency, and a computational analogue of peer critique. This synthesis logically motivates the introduction of the AICV framework.

5. The AICV Framework: Conceptual Foundations and Operational Workflow

5.1. Paradigm shift: From single-model prediction to multi-model consensus

Contemporary applications of AI in research assessment predominantly rely on individual model optimization, wherein a single algorithm is trained to predict human reviewer decisions or quality scores. This approach, while computationally efficient, inherits the limitations of monolithic systems: susceptibility to embedded training biases, lack of epistemic diversity, and opacity in decision rationales. AICV proposes a paradigm shift, reconceptualizing validation not as a problem of predictive accuracy but as a question of inter-model agreement. When multiple independently developed LLMs—trained on distinct corpora and architected under different paradigms—converge in their qualitative assessment of a scientific idea, that consensus itself constitutes a novel form of epistemic signal. This approach aligns with principles of robustness in complex systems, where agreement among diverse agents often indicates stability and reliability.

5.2. Operational architecture of AICV

The operational architecture of the AICV framework is organized into a structured three-phase workflow designed to maximize transparency, reproducibility, and interpretability. In the first phase, dimension specification and scoring protocols are established, with research abstracts evaluated across four explicitly defined dimensions—clarity, novelty, scientific relevance, and conceptual soundness—selected for their relevance to early-stage manuscript triage and their capacity for reliable machine assessment. Each dimension is scored on a five-point Likert scale (1 = very poor, 5 = excellent) and accompanied by a brief textual justification, thereby facilitating both quantitative aggregation and qualitative analysis; these dimensions were chosen because they recur most consistently across editorial guidelines and peer review rubrics, though alternative or expanded sets remain possible.

The second phase involves independent multi-model querying, in which a standardized prompt (see Appendix A) is submitted identically to multiple LLMs, such as GPT-4, Gemini, and Copilot, with each model queried in a new session to prevent cross-contamination. Models are selected to maximize epistemic diversity, varying by training data, developer, and fine-tuning objectives, thereby ensuring independence of perspective.

The third phase focuses on convergence quantification and interpretation, with agreement categorized into explicit agreement (identical or adjacent scores within a one-point difference), implicit agreement (divergent scores but aligned qualitative

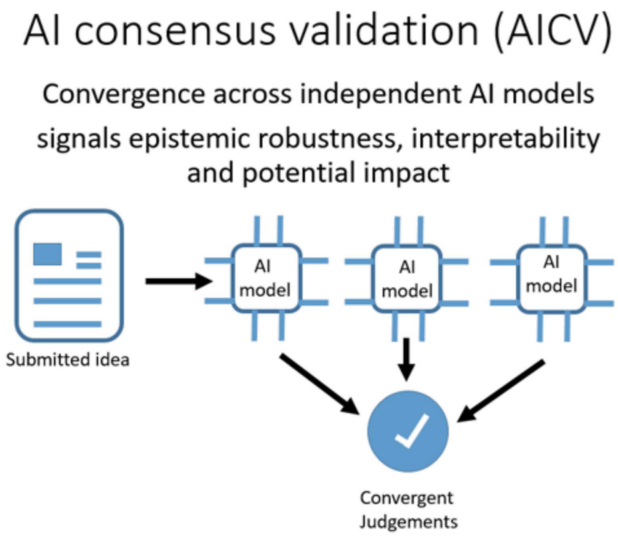

Figure 1. The AI Consensus Validation (AICV) workflow

rationale), and disagreement (two-point or greater divergence with incompatible rationale). A convergence metric is then calculated as the percentage of dimensions exhibiting explicit agreement, with outcomes classified as strong convergence (≥80%), moderate convergence (60–79%), or weak convergence (<60%).

These thresholds are intentionally heuristic and illustrative, intended to demonstrate how convergence patterns may be interpreted in a conceptual framework rather than to represent validated statistical decision rules. The metric functions as a reproducible, summarizable indicator of machine consensus, while the accompanying justifications provide insight into the reasoning behind scores.

5.3. Philosophical and theoretical underpinnings

AICV is conceptually grounded in social epistemology, particularly Longino's [7] emphasis on epistemic diversity as a foundation for objective scientific judgment. Longino argues that transformative criticism within a diverse scientific community—comprising varied methodological commitments, theoretical backgrounds, and evaluative standards—is essential for robust knowledge production. AICV transposes this principle into the computational domain: diverse LLMs, each representing a distinct "training heritage" and architectural paradigm, constitute a synthetic community of critical agents. Their consensus is not merely an aggregation of predictions but a form of machine-enabled critical dialogue.

This framework also engages with debates on reproducibility in computational science. Where single-model approaches struggle with explainability, AICV shifts the focus from why a single black box reached a conclusion to whether multiple black boxes independently align. This represents a move from mechanistic interpretability to outcome-based robustness—a form of reproducibility across models rather than within a single model's internal logic.

5.4. Disagreement as diagnostic insight

AICV explicitly values model disagreement as a source of diagnostic insight rather than system failure. Significant inter-model divergence (e.g., in the novelty or conceptual soundness dimensions) often highlights aspects of an abstract that are ambiguous, underdeveloped, or legitimately contentious. The accompanying justifications allow editors or researchers to identify specific points of uncertainty—for example, whether a methodological assumption is perceived as innovative or unsupported. In this capacity, AICV functions not as an autonomous decision-maker but as a structured sensemaking aid, surfacing latent ambiguities for human expert attention.

This capacity to differentiate between high-consensus and low-consensus submissions enables more nuanced triage. High-consensus abstracts may be expedited for in-depth human review, while low-consensus submissions could be flagged for additional scrutiny or developmental feedback prior to formal peer review.

6. Illustrative Results of the AICV Proof-of-Concept Analysis

The following illustrative analyses present preliminary results demonstrating how the AICV framework can differentiate biomedical submissions based on convergence patterns among independently developed LLMs. These results are intended as conceptual proof-of-concept demonstrations rather than empirical validation.

6.1. Methodological approach to a hypothesis-generating case illustration

To empirically ground the AICV framework and demonstrate its operational feasibility, three published biomedical abstracts were selected to represent a spectrum of research types, methodological rigor, and potential impact. These were chosen to reflect diverse domains within biomedicine: (1) a pilot AI imaging study [20], (2) a large-scale cardiovascular clinical trial [21], and (3) a rare disease case report [22]. These categories were chosen to illustrate how AICV may perform across research designs that differ in methodological structure, evidence strength, and typical editorial evaluation criteria. Each abstract was subjected to the AICV protocol using three leading LLMs: Microsoft Copilot, OpenAI ChatGPT (GPT-4), and Google Gemini (v. 1.5 Pro). The full prompting strategy, raw model outputs, and scoring justifications are available in Appendix A. The purpose of this illustration is not validation but to concretely demonstrate operational feasibility.

6.2. Case analysis and convergence patterns

The results, summarized in Table 2, reveal distinct convergence profiles across the three abstracts, illustrating AICV's capacity to differentiate between submissions based on machine-discernible qualities of clarity, novelty, relevance, and soundness.

6.2.1. Case 1: AI imaging pilot study

Gelman and Abràmoff [20]: A pilot study of deep learning-based monocular depth estimation from fundus photographs.

All three LLMs awarded closely aligned high scores for clarity (4/5), novelty (4–5/5), and relevance (4–5/5), resulting in explicit agreement across these dimensions. Conceptual soundness received uniformly moderate scores (3/5), with models citing the limited sample size (n = 30) and suboptimal image processing success rate (47%) as factors constraining robustness. The resulting convergence was strong (100%), indicating that the abstract was perceived by all models as clearly articulated, methodologically innovative, and clinically relevant, albeit with acknowledged methodological limitations typical of a pilot investigation.

6.2.2. Case 2: Large-scale cardiovascular trial

Lok et al. [21]: Fish-oil supplementation and cardiovascular events in patients receiving hemodialysis.

This abstract received perfect or near-perfect scores for clarity (5/5), relevance (5/5), and conceptual soundness (5/5) from all models, reflecting its rigorous double-blind randomized controlled design, clear reporting of statistical outcomes (HR 0.57), and direct clinical implications. Novelty scores were slightly lower (3–4/5) but still within explicit agreement, with models noting that while omega-3 supplementation is not novel per se, its application in the high-risk hemodialysis population represents a meaningful contextual advance. Convergence was again strong (100%), underscoring how AICV can reliably identify well-executed, high-impact clinical research.

6.2.3. Case 3: Rare disease case report

Yang et al. [22]: Subconjunctival choristoma: Description and surgical management of two pediatric patients.

Table 2. AICV convergence analysis for three biomedical abstracts. This table illustrates a hypothetical AICV output structure demonstrating how multi-LLM consensus may be quantified

Abstract (citation)	Dimension	Copilot	ChatGPT	Gemini	Agreement	Convergence
Abstract 1 Gelman and Abràmoff [20]	Clarity	4	4	4	Explicit	Strong (100%)
	Novelty	4	4	5	Explicit	
	Relevance	5	4	4	Explicit	
	Conceptual soundness	3	3	3	Explicit	
Abstract 2 Lok et al. [21]	Clarity	5	5	5	Explicit	Strong (100%)
	Novelty	4	3	4	Explicit	
	Relevance	5	5	5	Explicit	
	Conceptual soundness	5	5	5	Explicit	
Abstract 3 Yang et al. [22]	Clarity	3	3	4	Explicit	Weak (50%)
	Novelty	2	2	4	Disagreement	
	Relevance	3	3	3	Explicit	
	Conceptual soundness	3	4	5	Disagreement	

This case abstract revealed a more heterogeneous evaluation pattern. While clarity (3–4/5) and relevance (3/5) showed explicit agreement, novelty (2–4/5) and conceptual soundness (3–5/5) exhibited disagreement (≥2-point divergence). Notably, Gemini awarded higher novelty (4/5) and soundness (5/5) scores than Copilot or ChatGPT, citing the rarity of osseous choristoma and the logical diagnostic workflow as strengths. This divergence may reflect differing model sensitivities to rarity versus methodological innovation or variations in training data related to case report valuation. The overall convergence was weak (50%), highlighting how AICV can surface submissions that may be polarizing or require more nuanced human interpretation.

6.3. Interpretation and implications for editorial workflow

These illustrative cases demonstrate that AICV can produce interpretable, differential signals across varied research formats. High-convergence abstracts (Cases 1 and 2) may be prioritized for expedited editorial handling, as their core attributes are consistently recognized across multiple AI evaluators. Low-convergence abstracts (Case 3) do not necessarily indicate low quality; rather, they highlight submissions where AI judgment is divided—often reflecting inherent tensions in how novelty or soundness is defined across research paradigms (e.g., incremental clinical trials vs. rare case documentation). Such abstracts may benefit from additional scrutiny or tailored reviewer selection to ensure domain-appropriate evaluation.

Importantly, the justifications provided by each model (Appendix A) offer editorial insight beyond the numerical scores. For example, in Case 3, Gemini's rationale emphasized the educational value of rare case documentation, whereas ChatGPT focused on the lack of generalizable findings. This qualitative layer allows editors to understand why models disagreed, informing more nuanced handling decisions.

6.4. Limitations of the illustrative analysis

These examples are intended as proof-of-concept demonstrations, not as validation of AICV's predictive accuracy. The small sample size (n = 3), non-blinded selection of abstracts, and use of only three LLMs constrain generalizability. Furthermore, the abstracts were evaluated in isolation, without comparison to human reviewer scores or eventual publication outcomes. These limitations underscore the need for the large-scale, controlled validation studies outlined in Section 8. Future work could extend this demonstration through sensitivity analyses examining how convergence patterns change with different model selections, scoring rubrics, or evaluation dimensions.

7. Risks, Bias, and Ethical Considerations

7.1. Algorithmic bias and the perpetuation of scientific inequities

The deployment of AI systems in research assessment carries the inherent risk of amplifying and institutionalizing existing biases within the scientific literature [19, 23, 24]. LLMs are trained on historical corpora that reflect systemic inequities in publication, including gender, geographic, and institutional biases [25–30]. Studies have demonstrated that AI tools can undervalue research from low- and middle-income countries, disfavor novel methodologies outside mainstream paradigms, and penalize disciplinary jargon or writing styles that deviate from Western academic norms [31]. In the context of AICV, consensus may reflect convergent bias rather than intellectual merit—a dangerous scenario in which multiple models agree because they have been trained on the same skewed data, not because an idea is scientifically sound.

Mitigation Strategies. To counteract this, AICV implementations must prioritize epistemic diversity in model selection, incorporating models trained on regionally specific, open-access, and multidisciplinary corpora. Continuous bias auditing should be mandated, comparing AICV scores against human reviewer decisions across demographic and methodological categories. Furthermore, human-in-the-loop oversight remains essential for borderline or high-stakes decisions, ensuring that algorithmic outputs are interpreted within appropriate cultural and disciplinary contexts.

7.2. Opacity and the challenge of explainability

AICV, like many AI-assisted systems, operates with a degree of opacity. While individual model justifications offer some insight, the underlying reasoning processes of LLMs remain poorly understood [32]. This "black-box alignment" poses significant challenges for accountability, especially when consensus leads to editorial decisions that affect researchers' careers and scientific advancement. If an abstract is deprioritized based on low AICV convergence, authors deserve a comprehensible explanation—not merely a score aggregation.

Mitigation Strategies. Future iterations of AICV should integrate explainable AI (XAI) techniques tailored for multi-model consensus systems. This could include attention mapping to highlight text segments most influential in scoring or contrastive explanations that clarify why one abstract received higher convergence than another. Journals and funders adopting such tools must establish clear policies for transparent reporting, including disclosure of model versions, prompts, and scoring thresholds used in evaluation.

7.3. Suppression of novelty and disruptive innovation

A well-documented limitation of LLMs is their tendency to favor ideas that resemble existing literature, potentially penalizing truly novel or paradigm-shifting research [33]. This is particularly concerning in biomedicine, where transformative advances often emerge from unconventional approaches or high-risk hypotheses. An overreliance on AI consensus could inadvertently filter out disruptive science that does not conform to established patterns, stifling innovation and reinforcing intellectual silos.

Mitigation Strategies. AICV should be designed as a triage aid, not an autonomous gatekeeper. Implementation frameworks must include "novelty override" mechanisms, where submissions with low consensus but high human editorial interest are fast-tracked for expert review. Additionally, AICV could incorporate a disruption detection module that flags abstracts with highly divergent novelty scores for special consideration, recognizing that disagreement among models may signal an idea that is ahead of its time.

7.4. Ethical governance and accountability frameworks

The integration of AI into peer review raises profound ethical questions regarding responsibility, consent, and recourse [34–36]. Who is accountable if an AICV system erroneously rejects a groundbreaking study? How should researchers be informed that their work has been evaluated—even in part—by AI? Current scholarly norms and guidelines are ill-equipped to address these questions, underscoring the need for explicit governance structures. To address these concerns, we recommend the use of FAIR-Guidelines for AI-Assisted Review (Findable, Accessible, Interoperable, Responsible) [37], which provide a structured framework for ethical oversight.

7.4.1. Disclosure requirements

Transparency is foundational to trust in scholarly communication [35]. Mandatory disclosure of AI tool use in editorial decision letters ensures that authors are fully aware of the role machine assessment played in the evaluation of their work. Such disclosure should specify the dimensions assessed, the models employed, and the extent to which AI outputs influenced editorial decisions. This measure not only promotes accountability but also empowers authors to contextualize feedback and respond appropriately.

7.4.2. Appeal mechanisms

Fairness requires that authors retain the right to challenge AI-assisted evaluations. Clear, standardized pathways for appeal should be established, allowing authors to request human reevaluation if they believe the AI assessment was flawed or biased. These mechanisms must be accessible, timely, and impartial, ensuring that recourse is not merely symbolic but genuinely protective of scholarly integrity. Appeals also provide valuable feedback loops for improving AI systems by highlighting cases where machine judgments diverge from human expertise.

7.4.3. Data privacy protections

Respecting data privacy is critical in maintaining ethical standards. Submitted manuscripts must not be retained or repurposed for model training without explicit author consent, as such practices risk intellectual property violations and undermine trust in editorial systems [38]. Strong safeguards should be implemented to ensure compliance with international data protection regulations, including the General Data Protection Regulation and equivalent frameworks, while also providing authors with clear information about how their data is stored, processed, and secured.

7.4.4. Independent auditing

Independent third-party auditing is essential to ensure that AICV systems operate fairly and effectively. Regular audits should evaluate bias, accuracy, and equity across diverse research domains, with findings made publicly available to promote accountability. Auditing processes should include both technical validation (e.g., reproducibility of outputs, error rates) and ethical review (e.g., fairness across demographic groups, avoidance of systemic bias). By embedding external oversight, the scholarly community can safeguard against unchecked algorithmic influence and reinforce confidence in AI-assisted peer review [36].

7.5. Equity and access in global research ecosystems

The adoption of advanced AI tools in publishing and funding review risks exacerbating the digital divide between well-resourced and under-resourced institutions [31]. Commercial LLM APIs incur costs, and technical expertise is required for implementation and interpretation. Without deliberate effort, AICV could become a tool that further centralizes scholarly authority in wealthy regions and institutions.

Mitigation Strategies. Open-source implementations of AICV, coupled with partnerships with academic consortia in low-resource settings, are essential. Funding agencies and publishers should subsidize access to AI review tools for journals and institutions in developing regions. Training programs for editors and reviewers on interpreting AI outputs should be globally accessible and multilingual.

7.6. Philosophical and normative implications

Beyond technical and ethical risks, the integration of AI consensus into validation processes challenges fundamental norms of scientific evaluation. Peer review has traditionally been a humanistic, deliberative process rooted in expertise, judgment, and scholarly dialogue. Replacing or supplementing this with machine consensus risks commodifying scientific critique,

reducing nuanced evaluation to scalable metrics optimized for efficiency [39]. The scientific community must consciously decide whether and how to preserve the irreplaceable elements of human judgment—including intuition, ethical reasoning, and the capacity to recognize genius in unexpected forms.

8. Implementation Roadmap and Future Research

8.1. A Phased pathway to integration

The translation of AICV from a conceptual framework into a validated, trusted tool requires a structured, evidence-based implementation pathway. We propose a three-phase roadmap designed to balance innovation with rigorous evaluation, ensuring that each stage addresses technical, ethical, and practical dimensions of deployment.

8.1.1. Phase 1: Pilot validation and benchmarking

The first phase focuses on establishing baseline performance metrics and calibrating scoring thresholds against human judgment. A curated dataset of more than 500 biomedical abstracts with known peer review outcomes (accepted, revised, rejected) from multiple journals would serve as the foundation. AICV would be applied using four to five diverse LLMs, with convergence metrics and qualitative justifications recorded. Inter-LLM agreement rates would be quantified to evaluate consensus consistency, while deviations would inform threshold refinement. In parallel, blinded human reviewers would evaluate the same abstracts using an identical rubric. Agreement between AICV outputs and human scores would be analyzed using intraclass correlation coefficients and Fleiss' kappa, providing quantitative measures of reliability. Additional pilot metrics could include editorial triage efficiency—time saved per abstract—and sensitivity in detecting manuscripts requiring major revisions. Deliverables from this phase include performance benchmarks, optimized scoring thresholds, and an open-access dataset to enable community validation and replication.

8.1.2. Phase 2: Prospective field trials in real workflows

The second phase evaluates AICV's utility in live editorial and funding review environments. Partnerships with journals such as PLOS ONE and BMJ Open, as well as funders like the Wellcome Trust and NIH pilot programs, would allow integration of AICV as an optional triage module. A/B testing would compare workflows with and without AICV pre-screening, measuring outcomes such as time-to-decision, reviewer workload, and editorial satisfaction. Inter-LLM agreement patterns and their correlation with editorial decisions would be tracked as a key performance indicator, providing empirical evidence of model reliability and practical utility. Longitudinal monitoring would track the impact of AICV on publication diversity, citation equity, and the visibility of accepted papers. Deliverables include real-world feasibility data, workflow integration guidelines, and stakeholder feedback reports, providing evidence of practical utility and acceptance.

8.1.3. Phase 3: Scaling and governance framework development

The final phase emphasizes responsible scaling and governance. Reporting standards would be developed to ensure transparency in AICV use, including documentation of model versions, prompts, and convergence thresholds. Equity and bias mitigation protocols would be established, supported by regular algorithmic audits to monitor fairness across disciplines and demographics. Interdisciplinary dialogue among AI researchers, editors, ethicists, and policymakers would be fostered through workshops and consensus conferences, ensuring that governance structures reflect diverse perspectives. Deliverables include a governance white paper, open-source toolkits for implementation, and policy guidelines for journals and funders, collectively laying the foundation for ethical, scalable adoption of AICV.

8.2. Critical research priorities

Advancing the field of machine-assisted research assessment requires the prioritization of several critical research questions that address both methodological and ethical dimensions. Domain-specific validation remains essential, as the performance of AICV may vary across biomedical subfields such as clinical trials, bioinformatics, and public health, raising the question of whether field-specific fine-tuned models are necessary. Longitudinal impact must also be examined to determine whether the use of AICV in manuscript triage influences the diversity of published research, citation equity, or the career trajectories of early-stage researchers. Explainability and trust represent another priority, with research needed to identify which XAI methods most effectively communicate multi-model consensus to editors and authors and how such transparency affects adoption. Cost-benefit analysis is equally important, requiring systematic evaluation of the computational, environmental, and financial costs of deploying AICV at scale and comparison of these costs against reductions in human reviewer workload and accelerated publication cycles. Finally, novelty detection poses a conceptual challenge, as future work must explore whether AICV can be augmented to better recognize and protect genuinely disruptive ideas and whether "anti-consensus" signals should be developed to flag high-risk, high-reward submissions. Collectively, these priorities define a research agenda aimed at ensuring that AICV evolves as a robust, equitable, and trustworthy complement to human scientific judgment.

8.3. Toward an ecosystem of open tools and standards

The responsible evolution of AI in peer review depends on collaboration, transparency, and adherence to open science principles, requiring the development of an ecosystem of tools and standards that can be scrutinized, adapted, and replicated. Open-source implementations of AICV would enable independent evaluation and foster community-driven refinement, while shared benchmark datasets encompassing diverse research types and outcomes would provide a common foundation for comparative assessment. Interoperability standards are equally critical, ensuring that AICV tools can integrate seamlessly with existing editorial management systems such as Editorial Manager and ScholarOne, thereby reducing barriers to adoption. Finally, international consortia under the auspices of organizations such as the World Health Organization or UNESCO could provide ethical and equitable guidance for global deployment, establishing governance structures that safeguard privacy, accountability, and inclusivity. Collectively, these measures would advance the creation of a transparent, interoperable, and ethically grounded infrastructure for machine-assisted research evaluation.

8.4. Concluding the roadmap

The journey toward AI-augmented peer review is not merely technical—it is socio-technical, requiring continuous dialogue

among developers, users, and stakeholders. By proceeding with deliberate validation, transparent reporting, and inclusive governance, the biomedical community can harness tools like AICV to alleviate systemic bottlenecks while safeguarding the integrity, creativity, and equity of scientific discourse. Beyond its immediate conceptual contribution, AICV also reframes how machine intelligence might be integrated into scientific validation workflows. Rather than positioning AI as a surrogate reviewer tasked with replicating human judgment, the framework leverages the epistemic diversity of independently developed models as a source of evaluative signal. In this sense, AICV introduces a shift from algorithmic prediction to algorithmic consensus, suggesting that reproducible convergence across heterogeneous AI systems may provide a complementary indicator of intellectual robustness. If further validated, this paradigm could inform future editorial triage systems, grant review pipelines, and large-scale research prioritization efforts, particularly in fields where the volume of submissions exceeds the capacity of traditional peer review.

9. Conclusion

The accelerating volume and complexity of biomedical research demand innovative approaches to validation that complement, rather than replace, the essential role of human expertise. This conceptual review has proposed and examined AICV, a framework that leverages agreement among diverse LLMs as an early, reproducible signal of a research idea's clarity, novelty, relevance, and conceptual soundness. Through a structured analysis of the current AI-in-peer-review landscape, illustrative case applications, and a critical appraisal of ethical and operational risks, this review positions AICV as a promising middle path between fully manual peer review and opaque, monolithic automation.

AICV is not a speculative future technology but a feasible, implementable approach built on existing LLM capabilities. Its strength lies in epistemic diversity—the deliberate use of multiple, independently developed AI systems to simulate a form of machine-enabled critical dialogue. When these systems converge, they provide a robust, transparent signal that can help editors, funders, and researchers triage submissions efficiently. When they disagree, they surface points of ambiguity or innovation worthy of deeper human scrutiny.

However, the integration of AI into scientific validation carries profound responsibilities. Risks of bias amplification, novelty suppression, and ethical governance gaps must be addressed through rigorous validation, transparent reporting, and inclusive policymaking. The phased implementation roadmap outlined in this review provides a structured pathway from pilot studies to scaled deployment, emphasizing continuous evaluation and stakeholder engagement.

As biomedical informatics and AI continue to converge, tools like AICV represent more than technical solutions—they invite a reimagining of how scientific critique is structured, scaled, and democratized. By adopting such frameworks with caution, transparency, and a steadfast commitment to scientific creativity, the research community can harness AI not as a replacement for human judgment but as a catalyst for a more efficient, equitable, and innovative future for biomedical discovery. By introducing the concept of AICV, this work contributes a new theoretical lens through which AI-assisted research evaluation can be understood. Rather than optimizing individual model performance, the framework highlights the potential value of convergence among diverse AI systems as an indicator of epistemic robustness—an idea that has not yet been systematically explored in biomedical peer review.

Ethical Statement

This article is a review of previously published literature. It does not involve any studies with human participants or animals conducted by the authors. Therefore, no ethical approval or informed consent was required. All data discussed in this review are derived from sources that had already obtained the appropriate ethical approvals in their original studies.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

This article is a review of previously published literature. No new data were generated or analyzed in this study. All information supporting the findings of this review is available within the cited references and in Appendix A.

Author Contribution Statement

Tan Aik Kah: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

Supplementary Information

Appendix A is available at https://doi.org/10.47852/bonviewMEDIN62029207.

References

[1] Kovanis, M., Porcher, R., Ravaud, P., & Trinquart, L. (2016). The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLOS ONE*, *11*(11), e0166387. https://doi.org/10.1371/journal.pone.0166387

[2] Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., Nanni, F., & Coates, J. A. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biology*, *19*(4), e3000959. https://doi.org/10.1371/journal.pbio.3000959

[3] Bornmann, L., Haunschild, R., & 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*, *8*(1), 224. https://doi.org/10.1057/s41599-021-00903-w

[4] Si, C., Yang, D., & Hashimoto, T. (2024). Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *arXiv Preprint:2409.04109*.

[5] Hubert, K. F., Awa, K. N., & Zabelina, D. L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. *Scientific Reports*, *14*(1), 3440. https://doi.org/10.1038/s41598-024-53303-w

[6] Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review*, *8*(1), 4. https://doi.org/10.1186/s41073-023-00133-5

[7] Longino, H. E. (2022). What's social about social epistemology? *The Journal of Philosophy*, *119*(4), 169–195. https://doi.org/10.5840/jphil2022119413

[8] Brainard, J. (2023). Journals take up arms against AI-written text. *Science*, *379*(6634), 740–741. https://doi.org/10.1126/science.adh2762

[9] Zhuang, Z., Chen, J., Xu, H., Jiang, Y., & Lin, J. (2025). Large language models for automated scholarly paper review: A survey. *Information Fusion*, *124*, 103332. https://doi.org/10.1016/j.inffus.2025.103332

[10] Foltýnek, T., Meuschke, N., & Gipp, B. (2020). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys*, *52*(6), 1–42. https://doi.org/10.1145/3345317

[11] Shin, H., Tang, J., Lee, Y., Kim, N., Lim, H., Cho, J. Y., . . . , & Kim, J. (2025). Automatically evaluating the paper reviewing capability of large language models. *arXiv Preprint:2502.17086*.

[12] Anghel, C., Craciun, M. V., Pecheanu, E., Cocu, A., Anghel, A. A., Iacobescu, P., . . . , & Dragosloveanu, S. (2025). CourseEvalAI: Rubric-guided framework for transparent and consistent evaluation of large language models. *Computers*, *14*(10), 431. https://doi.org/10.3390/computers14100431

[13] Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., . . . , & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, *1*(8), AIoa2400196. https://doi.org/10.1056/AIoa2400196

[14] Sun, Z. (2025). Large language models in peer review: Challenges and opportunities. *Scientometrics*, *130*, 5503–5546. https://doi.org/10.1007/s11192-025-05440-w

[15] Zhu, C., Xiong, J., Ma, J. R., Lu, Z., Liu, Y., & Li, L. (2025). When your reviewer is an LLM: Biases, divergence, and prompt injection risks in peer review. *arXiv Preprint: 2509.09912*

[16] Kuo, S.-M., Tai, S.-K., Lin, H.-Y., & Chen, R.-C. (2025). Automated clinical trial data analysis and report generation by integrating Retrieval-Augmented Generation (RAG) and Large Language Model (LLM) technologies. *AI*, *6*(8), 188. https://doi.org/10.3390/ai6080188

[17] Srinivasan, A., Berkowitz, J., Friedrich, N. A., Kivelson, S., & Tatonetti, N. P. (2025). Large language model analysis of reporting quality of randomized clinical trial articles: A systematic review. *JAMA Network Open*, *8*(8), e2529418. https://doi.org/10.1001/jamanetworkopen.2025.29418

[18] Leung, T. I. (2026). LLMs in peer review—How publishing policies must advance. *JAMA Network Open*, *9*(1), e2552042. https://doi.org/10.1001/jamanetworkopen.2025.52042

[19] Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA Dermatology*, *157*(11), 1362. https://doi.org/10.1001/jamadermatol.2021.3129

[20] Gelman, R., & Abràmoff, M. D. (2023). A pilot study of deep learning-based monocular depth estimation from fundus photographs. *Medinformatics*, *4*(2), 933. https://doi.org/10.47852/bonviewMEDIN42023933

[21] Lok, C. E., Farkouh, M., Hemmelgarn, B. R., Moist, L. M., Polkinghorne, K. R., Tomlinson, G., . . . , & Udell, J. A. (2026). Fish-oil supplementation and cardiovascular events in patients receiving hemodialysis. *The New England Journal of Medicine*, *394*(2), 128–137. https://doi.org/10.1056/NEJMoa2513032

[22] Yang, C. C., Anderson, J., Deem, C., & Tam, E. (2025). Subconjunctival choristoma: Description and surgical management of two pediatric patients. *Cureus*, *17*(12), e99225. https://doi.org/10.7759/cureus.99225

[23] Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. USA: The MIT Press.

[24] Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., & White, J. A. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *NPJ Digital Medicine*, *8*(1), 154. https://doi.org/10.1038/s41746-025-01503-7

[25] Mirza, I., Jafari, A. A., Ozcinar, C., & Anbarjafari, G. (2025). Quantifying gender bias in large language models using information-theoretic and statistical analysis. *Information*, *16*(5), 358. https://doi.org/10.3390/info16050358

[26] Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, 12–24. https://doi.org/10.1145/3582269.3615599

[27] Torres, N., Ulloa, C., Araya, I., Ayala, M., & Jara, S. (2025). A comprehensive analysis of gender, racial, and prompt-induced biases in large language models. *International Journal of Data Science and Analytics*, *20*, 3797–3834. https://doi.org/10.1007/s41060-024-00696-6

[28] Anaweokhai, A. (2025). A systematic framework for investigating algorithmic bias as a social determinant of health in low and middle-income countries. *Journal of Medical Health Research and Psychiatry*, *2*(2), 1–11. https://doi.org/10.70844/jmhrp.2025.2.2.38

[29] Yang, J., Clifton, L., Dung, N. T., Phong, N. T., Yen, L. M., Thy, D. B. X., . . . , & Clifton, D. A. (2024). Mitigating machine learning bias between high-income and low–middle-income countries for enhanced model fairness and generalizability. *Scientific Reports*, *14*(1), 13318. https://doi.org/10.1038/s41598-024-64210-5

[30] Resnik, P. (2025). Large language models are biased because they are large language models. *Computational Linguistics*, *51*(3), 885–906. https://doi.org/10.1162/coli_a_00558

[31] Li, Z., & He, Z. (2026). AI writing tools could lead scholars from low-income countries to erase their own voices. *Nature*, *649*(8097), 555. https://doi.org/10.1038/d41586-026-00121-x

[32] Kamihara, T., Omura, T., & Shimizu, A. (2025). Potential and limitations of large language models for medical literature analysis: A preliminary investigation. *Cureus*, *17*(9), e92590. https://doi.org/10.7759/cureus.92590

[33] Peters, U., & Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, *12*(4), 241776. https://doi.org/10.1098/rsos.241776

[34] Kemal, Ö. (2025). Artificial intelligence in peer review: Ethical risks and practical limits. *Turkish Archives of Otorhinolaryngology*, *63*(3), 108–109. https://doi.org/10.4274/tao.2025.2025-8-12

[35] Porsdam Mann, S., Vazirani, A. A., Aboy, M., Earp, B. D., Minssen, T., Cohen, I. G., & Savulescu, J. (2024). Guidelines for ethical use and acknowledgement of large language

models in academic writing. *Nature Machine Intelligence*, *6*(11), 1272–1274. https://doi.org/10.1038/s42256-024-00922-7

[36] Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. *AI and Ethics*, *5*(3), 3265–3279. https://doi.org/10.1007/s43681-024-00653-w

[37] Huerta, E. A., Blaiszik, B., Brinson, L. C., Bouchard, K. E., Diaz, D., Doglioni, C., . . . , & Zhu, R. (2023). FAIR for AI: An interdisciplinary and international community building perspective. *Scientific Data*, *10*(1), 487. https://doi.org/10.1038/s41597-023-02298-6

[38] Pressman, S. M., Borna, S., Gomez-Cabello, C. A., Haider, S. A., Haider, C., & Forte, A. J. (2024). AI and ethics: A systematic review of the ethical considerations of large language model use in surgery research. *Healthcare*, *12*(8), 825. https://doi.org/10.3390/healthcare12080825

[39] Thelwall, M. (2025). Research quality evaluation by AI in the era of large language models: Advantages, disadvantages, and systemic effects—An opinion paper. *Scientometrics*, *130*(10), 5309–5321. https://doi.org/10.1007/s11192-025-05361-8

How to Cite: Kah, T. A. (2026). AI Consensus Validation in Biomedical Research: A Review, Conceptual Framework, and Future Directions. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN62029207