Zero-shot reasoning for simulating scholarly peer-review

Khalid M. Saqr*

College of Engineering and Technology

Arab Academy for Science, Technology, and Maritime Transport

Alexandria 1029– EGYPT

Abstract

The scholarly publishing ecosystem faces a dual crisis of unmanageable submission volumes and unregulated AI, creating an urgent need for new governance models to safeguard scientific integrity. Current human-led peer review lacks a scalable, objective benchmark, making editorial processes opaque and difficult to audit. Here we investigate a deterministic simulation framework that provides the first stable, evidence-based standard for evaluating AI-generated peer review reports. Analyzing 352 peer-review simulation reports, we identify consistent system state indicators that demonstrate its reliability. First, the system is able to simulate calibrated editorial judgment, with 'Revise' decisions consistently forming the majority outcome (>50%) across all disciplines, while 'Reject' rates dynamically adapt to field-specific norms, rising to 45% in Health Sciences. Second, it maintains unwavering procedural integrity, enforcing a stable 29% evidence-anchoring compliance rate that remains invariant across diverse review tasks and scientific domains. These findings demonstrate a system that is predictably rule-bound, mitigating the stochasticity of generative AI. For the scientific community, this provides a transparent tool to ensure fairness; for publishing strategists, it offers a scalable instrument for auditing workflows, managing integrity risks, and implementing evidence-based governance. Our framework repositions AI as an essential component of institutional accountability, providing the critical infrastructure to maintain trust in scholarly communication.

Keywords: Knowledge, xPeerd, Generative AI, Peer Review, Simulation.

Introduction

Scholarly peer review is strained by a structural mismatch between submission volume, reviewer availability, and editorial capacity. The surge in submissions, now accelerated by easy-to-use generative AI tools, coincides with fewer available experts, increasing desk rejections, renewed interest in pre-prints, and the growth of predatory outlets [1]. Editors and publishers anticipate a shift toward "human-assisted" review, where AI accelerates synthesis and administrative load while humans retain judgment [2, 3]. However, governance lags: most policies address author-side AI, not reviewer-side use, creating an integrity risk if undisclosed tools influence assessments [4]. Retractions tied to undisclosed generative-AI usage signal enforcement will tighten [5]. Together, scale pressures and policy gaps frame a paradox: the same technology poised to stabilize peer review can also erode it if deployed without explicit controls.

Evidence across domains maps both promise and limits. On the upside, AI assistants can reduce meta-review time and broaden coverage without sacrificing human control, yet users report concerns about over-reliance, trust, and agency [6]. Editorial and workflow proposals show end-to-end pipelines for triage, acceptance prediction, and reviewer recommendation, with encouraging but not definitive performance [7–9].

*ORCID: 0000-0002-3058-2705. Email: k.saqr@aast.edu

Field narratives in medicine and OBGYN outline practical gains—query formulation, screening, reference harmonization, and structured critique—paired with explicit safeguards and disclosure norms [10–12]. Still, benchmarked predictivity for pre-publication quality remains modest and context-dependent: correlations are weak-to-moderate and vary by venue and input granularity [13]. Controlled evaluations of automatic reviewing find LLMs deliver passable decisions on constrained tasks yet falter at long-document reasoning, consistent zero-shot scoring, and human-like criticality [14]. Reviewer-facing detection of AI prose hovers near chance overall and improves only with seniority, highlighting practical limits of gatekeeping-by-detection [15].

Risk signals cohere across settings. Hallucination is persistent, reference fabrication occurs, and some readers over-trust AI outputs even when errors are detectable [16, 17]. In medicine, quality and safety metrics for patient-facing answers are generally good but uneven; readability is often too high; rare reference hallucinations still appear [18]. Importantly, AI-generated scholarly text can clear blinded review with major revisions, proving both its usefulness and the need for provenance and accountability [19]. Evaluations that grade models "recommended," "usable with high caution," or "unusable" by domain reinforce that deployment must be scoped and audited, not generalized by hype [20]. Editorial viewpoints converge: AI will likely amplify existing incentives unless journals reconfigure around dialogic, community-centered knowledge practices that keep humans responsible for scholarly communication and content publishing [3]. The emerging consensus is pragmatic rather than utopian: use AI to compress latency and surface evidence, but bind its outputs to transparent, human-verified, and auditable procedures [21–23].

Large Language Models (LLMs) behave as zero-shot reasoners [24]. This has rapidly matured into a diverse field of research. Current work moves beyond demonstrating this initial capability to refining its methodology and expanding its application to increasingly complex domains. Methodological advancements seek to improve the quality and structure of the reasoning process, with techniques like "Plan-and-Solve" prompting designed to overcome the logical and calculation errors of earlier methods [25]. Simultaneously, research is democratising these abilities beyond the largest models, showing that instruction tuning with chain-of-thought rationales can imbue smaller models with strong reasoning capabilities [26].

The application of this zero-shot paradigm now extends well beyond text-based tasks. In multimodal domains, LLMs are being leveraged for visual reasoning by translating images into effective textual prompts [27], improving question prompts to activate latent reasoning [28], and using divergent reasoning to capture fuzzy semantics in image retrieval [29]. Furthermore, LLMs are being integrated into embodied agents, acting as zero-shot human simulators for human-robot interaction [30], as commonsense engines for robotic navigation [31], and even as zero-shot reward models for reinforcement learning [32]. This trend culminates in applications in specialized, high-stakes domains such as medicine, where collaborative multi-agent frameworks enhance zero-shot reasoning for complex diagnostic challenges [33]. Collectively, this research demonstrates a clear trajectory from proving a concept to engineering robust, adaptable reasoning systems for diverse real-world problems.

This paper presents a zero-shot reasoning framework that simulates scholarly peer review while encoding normative constraints for integrity, transparency, and contestability. We formalize reviewer beliefs, evidence aggregation, critique generation, and decision rules under proper scoring and deontic guardrails, and we design interfaces that require disclosure, citation verification, and human adjudication. The mathematical model aims to convert today's ad hoc tool use into a principled, auditable process that measurably reduces turnaround time without degrading validity. Subsequent sections specify the formal state space, evidence operators, critique acts, loss and reward functions, and protocol-level guarantees for safety and accountability.

Assumptions, Framework, and Model

xPeerd is a zero-shot reasoning agent designed to execute a constrained Bayesian–argumentation decision process with deontic guards and optional game-theoretic simulation. The agent (i) grounds every assertion in manuscript evidence, (ii) performs defeasible argument evaluation, (iii) issues a decision under explicit norms, and (iv) can simulate multi-round editorial dynamics and double-blind reviewers.

It is important to clarify the nature of this framework. The mathematical model presented herein serves as the core rule-based reasoning process used to simulate peer-review. The system is designed to be procedurally reproducible: given the same manuscript and review task, the same set of constraints, evaluation

criteria, and logical pathways are deterministically invoked. This rule-bound architecture mitigates the stochasticity inherent in unconstrained generative models. While minor textual variations in the generated output are possible, the core evaluative judgments and the resulting decision category remain stable. The following equations, therefore, represent the formal logic that the underlying language model is constrained to approximate, rather than a direct numerical computation.

Ontology and Tasks

A manuscript is a triple

$$M \triangleq \langle C, E, P \rangle, \tag{1}$$

where C is a finite set of claims, E are evidential units (text spans, figures, tables), and $P \subset \mathbb{N}$ is a page index. The review task is

$$\tau \in \{\text{HCReview}, \text{DAReview}, \text{confReview}, \text{PRR}, \text{DBReviewSim}\}. \tag{2}$$

Assumptions and Operators

A1. Epistemic domain: $X \triangleq C \cup \{h\}$ where h is a binary fabrication hypothesis.

A2. Belief state b is a probability function over X; prior b_0 given before observing M.

A3. Page-anchored observation map obs : $(M, p) \mapsto E_p \subseteq E$ with $p \in P$.

A4. Likelihood $\mathcal{L}(o \mid x)$ defined for some $(o, x) \in E \times X$; otherwise apply belief revision.

A5. Deontic vocabulary: obligations $O(\cdot)$ and forbiddances $F(\cdot)$ constrain admissible actions.

Epistemic Update with Page Grounding

Given an observation $o_p \in E_p$, update b to b' by

$$b'(x) \;=\; \begin{cases} \dfrac{b(x)\,\mathcal{L}(o_p \mid x)}{\sum_{x' \in X} b(x')\,\mathcal{L}(o_p \mid x')} & \text{if } \mathcal{L}(\cdot \mid x) \text{ defined,} \\ \big(b \star o_p\big)(x) & \text{otherwise,} \end{cases} \qquad x \in X, \tag{3}$$

where $\star$ is an AGM-style revision operator satisfying success, inclusion, and minimal change. Every reported assertion must be page-linked:

$$\forall a \in C : \; \text{Report}(a) \;\Rightarrow\; \exists\, p \in P \; \text{Link}(a, p). \tag{4}$$

Normative Constraints

The minimal guard set is

$$O(\text{scope\_only}), \quad O(\text{ground\_to\_pages}), \quad O(\text{check\_coherence}), \; O(\text{check\_novelty}), \; O(\text{check\_fabrication}), \tag{5}$$

$$F(\text{off\_topic}), \quad F(\text{no\_manuscript}).$$

Execution is admissible only if all active $O(\cdot)$ are satisfiable and no $F(\cdot)$ is violated. Repeated violations trigger refusal text.

Argumentation Layer

Construct a Dung framework

$$G \triangleq (A, R), \qquad A \subseteq C \times P, \quad R \subseteq A \times A, \tag{6}$$

where each $a \in A$ is an atomic, page-anchored proposition, and R encodes attack (and support via standard reductions). For a chosen semantics $\sigma \in \{\text{grounded}, \text{preferred}\}$, compute the accepted extension $\text{Ext}_\sigma(G)$. Map acceptance to credibility weights

$$w : A \to \mathbb{R}, \qquad w(a) \triangleq w_0 + \kappa \cdot \mathbf{1}[a \in \text{Ext}_\sigma(G)], \tag{7}$$

and define a justification operator $J(a) \subseteq E$ returning the minimal evidential set supporting a.

Integrity Risk and Scoring

xPeerd's simulated decision-making is guided by two principal assessment dimensions, which are conceptualized as abstract scores.

First, we define an integrity fraud risk, *rf*, as a qualitative flag activated by heuristic detectors. This is formally abstracted as a posterior probability over a fabrication hypothesis, h:

$$r_f \triangleq \Pr(h = 1|E) = \rho(D_{\text{data}}(E), D_{\text{lang}}(E)), \tag{8}$$

where D_{data} represents a set of checks for data anomalies (e.g., incomplete results, unverifiable claims, statistical inconsistencies) and D_{lang} represents checks for linguistic anomalies associated with academic fraud (e.g., incoherent concepts, excessive jargon, logical fallacies). The aggregator ρ is not a numerical function but a logical operator that flags a high risk if a critical threshold of evidence from these detectors is met.

Second, we define a manuscript score, $S(M)$, as an abstraction of a multi-dimensional qualitative evaluation. It combines assessments of coherence, evidential fit, and methodological validity:

$$S(M) \triangleq \alpha \cdot \text{Coh}(G) + \beta \cdot \text{Fit}(b') + \gamma \cdot \text{MethVal}(M), \quad \alpha, \beta, \gamma \geq 0, \alpha + \beta + \gamma = 1. \tag{9}$$

Here, $\text{Coh}(G)$ evaluates the logical consistency and theoretical alignment of the manuscript's argumentation graph. $\text{Fit}(b')$ assesses the degree to which claims are sufficiently supported by the cited evidence within the document, a process proxied by the system's rule to anchor all critiques to specific page citations. Finally, $\text{MethVal}(M)$ evaluates the appropriateness and reproducibility of the research methods, guided by field-specific verification checklists (e.g., for STEM, social sciences, or humanities) as specified in Equation (20). The weights (α, β, γ) represent the conceptual importance of each dimension in the overall evaluation framework. Set

$$W_{\min} \triangleq \min_{a \in A_{\text{crit}}} w(a). \tag{10}$$

Decision Function

The decision is a tuple or a categorical label:

$$\delta(M, \tau) \in \langle \text{summary}, \ \text{majors}, \ \text{minors}, \ \text{rec} \rangle \quad \text{or} \quad \{\text{Accept}, \text{Revise}, \text{Reject}\}. \tag{11}$$

With thresholds $0 < \theta_1 \leq \theta_2 < 1$ and $\lambda \in \mathbb{R}$,

$$\delta(M, \tau) = \begin{cases} \text{Reject} & \text{if } r_f > \theta_2 \text{ or } W_{\min} \leq \lambda, \\ \text{Revise} & \text{if } \theta_1 \leq r_f \leq \theta_2 \text{ or any } O(\cdot) \text{ unmet}, \\ \text{Accept} & \text{if } r_f < \theta_1 \text{ and } W_{\min} > \lambda. \end{cases} \tag{12}$$

Major issues target integrity, reproducibility, and ethics; minor issues cover clarity, typography, and formatting. All items cite $J(a)$ with page references.

PRR as a Repeated Game

Let $R = \{r_1, \ldots, r_k\}$ be potential rejection reasons. For rounds $t \in \{1, 2, 3\}$, estimate

$$p_t(r_i) \triangleq \Pr(r_i \text{ holds at round } t \mid b_t, G_t, E), \qquad i = 1, \ldots, k, \tag{13}$$

with b_{t+1} and G_{t+1} updated from round-t feedback. Aggregate terminal rejection mass and implied acceptance:

$$q \triangleq \sum_{i=1}^{k} p_3(r_i), \tag{14}$$

$$\hat{a} \triangleq 1 - q. \tag{15}$$

The PRR table reports $p_t(r_i)$ with quotes and pages for each i, t.

Double-Blind Simulation

Instantiate two independent reviewers R_1, R_2 with distinct priors and thresholds

$$b_1 \neq b_2, \qquad (\theta_1^{(j)}, \theta_2^{(j)}, \lambda^{(j)}) \text{ for } j \in \{1, 2\}. \tag{16}$$

Enforce non-communication via epistemic constraints

$$K_1 \neg K_2(c) \ \wedge \ K_2 \neg K_1(c), \qquad \forall c \in C, \tag{17}$$

where K_j is the knowledge modality for reviewer j. Each produces δ_j. The editor aggregates by

$$\delta^* \triangleq \mathcal{A}(\delta_1, \delta_2, S(M), r_f)\,, \tag{18}$$

where $\mathcal{A}$ is a deterministic policy (e.g., disagreement $\Rightarrow$ *Revise*, otherwise majority).

Dynamic Epistemic Rules

Let events be

$$\text{upload}(M):\ b \leftarrow b_0(M), \quad \text{reset}:\ b \leftarrow b_{\text{vac}}, \quad \text{retask}(\tau'):\ \tau \leftarrow \tau'. \tag{19}$$

Each event re-initializes state consistent with (3)–(12).

Field-specific Verification

Attach checklist modules by domain $d \in \{\text{STEM}, \text{HUM}, \text{SOC}\}$:

$$\varphi_d:\ M \mapsto \{\text{tests, thresholds, evidence requirements}\}, \tag{20}$$

which parameterize $\text{MethVal}(M)$ in (10) and populate majors/minors in (11).

Procedural Policy and Guard Enforcement

If $F(\text{no\_manuscript})$ or $F(\text{off\_topic})$ holds, halt and request the missing input. If two attempts fail to satisfy $O(\text{ground\_to\_pages})$, emit guardrail refusal. Formally,

$$\neg\text{Admissible} \ \Rightarrow \ \delta(M, \tau) = \textit{Refuse-with-instructions}. \tag{21}$$

xPeerd Implementation

xPeerd is an research-grade peer review simulation system from KNOWDYN (UK). It applies the framework and model presented here using the computational workflow illustrated in Figure 1. The *Manuscript* class embodies the triple $M = \langle C, E, P \rangle$ in (1), with review tasks τ instantiated as in (2). The *Belief* class and the *Update* operation implement the epistemic dynamics of (3), anchored by the grounding constraint (4) and guarded by the deontic obligations and forbiddances in (5). The construction of the Dung graph $G = (A, R)$ in (6), the acceptance set $\mathrm{Ext}_\sigma(G)$, and the credibility weights $w(a)$ in (7) are captured by the *GraphG* and *AcceptRule* nodes. Justifications $J(a)$ also appear as a dedicated function. The *FraudRisk* and *Score* functions reflect the definitions of r_f (8) and $S(M)$ (9), with $W_{\min}$ from (10) constraining admissible outcomes. The *Decision* class, linked through the *DecisionRule* operation, corresponds to the decision function $\delta(M, \tau)$ and the policy thresholds in (11)–(12). Finally, the PRR branch reproduces the three–round repeated game of (13)–(15), and the DBReviewSim branch represents the double–blind simulation with distinct priors and thresholds (16)–(18), aggregated by the editor. Thus the diagram provides a graphical counterpart to the mathematical model, with structural classes, functional mappings, and procedural rules aligned to their respective equations.

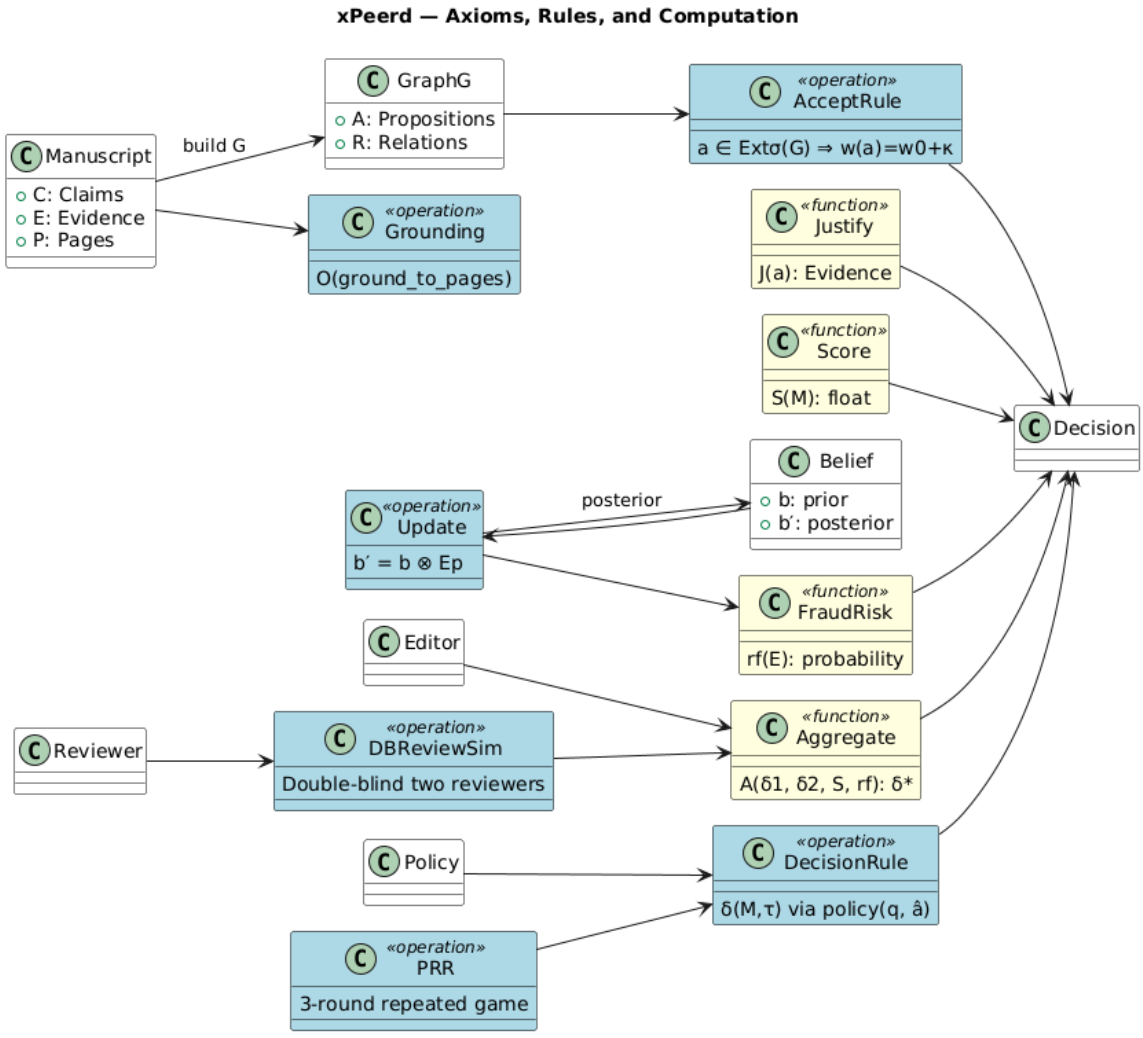

Figure 1: UML diagram of xPeerd axioms and computation flow. Structural classes (white), functions (yellow), and operations (blue) are distinguished, and arrows indicate the decision workflow.

Evaluating Peer-Review Simulations Produced by xPeerd

The evaluation methodology reported here was applied to a sample of 500 peer–review simulation cases (n=500) was conducted on xPeerd between February and May 2025. The xPeerd users who participated in the study agreed to share their anonymized peer-review reports for the purpose of this study. The evaluation workflow was implemented using an open Python pipeline, released under open access at this Github repository. The code provides reproducible data parsing, classification, analytics, and visualization routines that investigate the performance of xPeerd by subject matter and stable review type.

Subject Matter Classification

The first step in the evaluation applies semantic classification to each peer-review report. Using the *All Science Journal Classification* (ASJC) system, each simulation text is embedded with a sentence transformer model and compared to a set of descriptive ASJC labels. The highest–scoring label is assigned as the primary subject area, and the top–3 alternatives with similarity z–scores are retained for uncertainty quantification. The resulting distribution shows how the 500 cases map onto disciplinary categories such as Medicine, Computer Science, and Engineering.

Compliance Metrics

The second step measures adherence to the framework's rules. Two indicators are computed by review type: (i) *decision coverage*, the fraction of cases yielding a valid decision, and (ii) *anchoring compliance*, the proportion of issues that cite page references above the minimum rule threshold. These metrics are visualized as paired bar charts, enabling direct comparison of coverage versus anchoring across the five task types (`/HCReview`, `/DAReview`, `/ConfReview`, `/PRR`, `/DBReviewSim`).

Decision Composition and Workload

Next, decisions are decomposed into the three categorical outcomes (`Reject`, `Revise`, `Accept`). Shares are reported by review type, together with total case counts n. Workload is summarized as the distribution of issue counts per case, visualized with violin plots. A companion analysis examines the relation between workload and evidence anchoring using scatter plots and hexbin density maps.

Task-Specific Analyses

Two specialized procedures are then evaluated:

PRR. For `/PRR` cases, the terminal rejection mass q is estimated after three rounds, and the complementary acceptance probability $\hat{a}$ is reported. Histograms of q and bar charts of the most frequent rejection reasons provide insight into iterative critique dynamics.

DBReviewSim. For `/DBReviewSim` cases, the two reviewer decisions δ_1 and δ_2 are compared. Agreement versus disagreement counts are tabulated and shown in heatmaps, and the editor aggregation policy $\mathcal{A}$ is applied to generate the final decision δ^*.

Uncertainty in Classification

Finally, the semantic classification results are evaluated for robustness. Top–1 assignment confidence is computed from softmax–normalized similarity scores, with simulations below a threshold (e.g., $p < 0.25$) flagged as uncertain. Confidence distributions are plotted as histograms and boxplots stratified by ASJC area, and bar charts summarize the most common second–best labels. This layer highlights which disciplinary assignments are stable and which remain ambiguous.

Results

The initial corpus consisted of $n = 500$ peer-review simulation records. From this set, $n = 352$ reports were retained for the final analysis. A total of 148 records were excluded based on pre-defined criteria, primarily targeting simulations where user input did not conform to the system's operational requirements. These exclusions comprised cases of misfired prompts, where users provided commands unrelated to a specific review task, and freestyle prompts, where users simulated bespoke aspects of peer-review in conversational interaction rather than using one of the stable review prompts analysed here. This filtering step ensures that the analyzed dataset consists only of valid, completed review simulations as defined by the system's intended function, thereby preventing the inclusion of incomplete or irrelevant interactions from skewing the performance metrics. Each review report was semantically classified into subject domains using the All Science Journal Classification (ASJC) supergroups.

Figure 2(a) shows the distribution of review counts across supergroups. The majority of reports fell into *Physical Sciences* ($n = 109$) and *Health Sciences* ($n = 113$), followed by *Humanities* ($n = 70$) and *Social Sciences* ($n = 46$), while *Life Sciences* were sparsely represented ($n = 14$). No reports defaulted to the "Multidisciplinary" fallback category, indicating robust coverage of domain-specific classification.

To assess the reliability of classification, we computed confidence scores $\hat{p}_i \in [0, 1]$ for each report, where $\hat{p}_i$ denotes the posterior probability assigned to the selected supergroup. Figure 2(b) displays the histogram of these scores. The distribution was centered around $\mu = 0.45$ with a spread of $\sigma \approx 0.12$, and the critical acceptance threshold was defined at

$$\tau = 0.20\,.$$

All retained reports exceeded this threshold, with the bulk of classifications concentrated in the range $[0.30, 0.60]$. This indicates that the semantic assignment of supergroups was statistically stable across the analytic set.

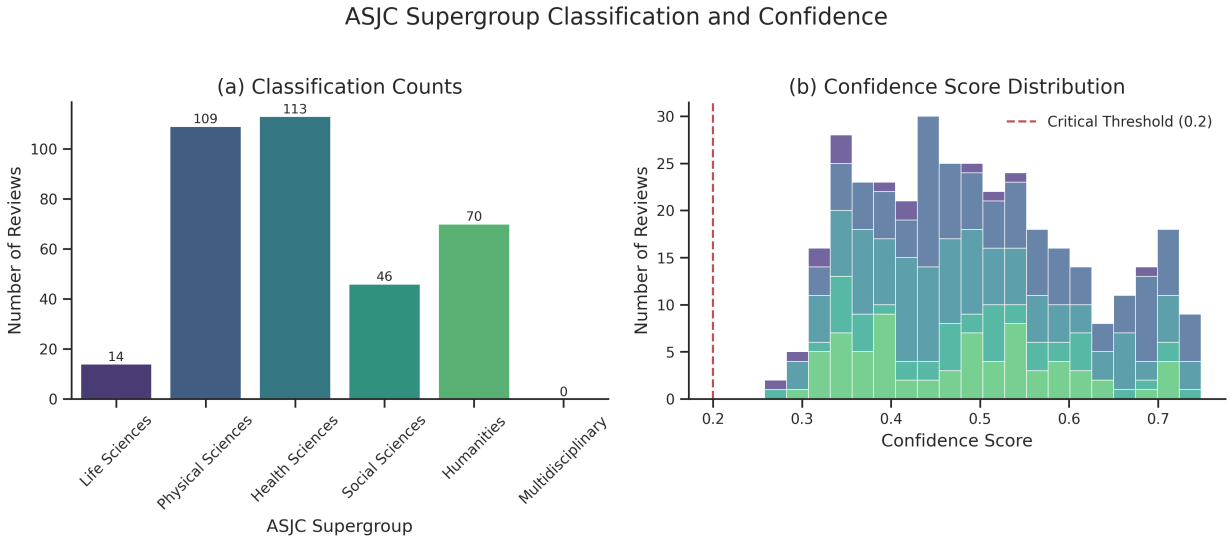

Figure 2: ASJC Supergroup Classification and Confidence. (a) Classification counts of $n = 352$ valid reports across the ASJC supergroups. (b) Distribution of classification confidence scores $\hat{p}_i$ with a critical threshold $\tau = 0.20$ indicated by the dashed line.

Figure 3 summarizes the distribution of simulated editorial decisions across the ASJC supergroups for the analytic set of $n = 352$ reports. Each decision outcome was categorized into three discrete states:

$$D \in \{\text{Reject}, \text{Revise}, \text{Accept}\}\,.$$

Figure 3 indicates the relative frequency of each decision within supergroups. *Revise* decisions dominated across all domains, with probabilities exceeding $p(D = \text{Revise} \mid g) > 0.5$ for every group g. The likelihood

of outright *Reject* was highest in the *Life Sciences* and *Health Sciences*, with observed proportions $p(D = \text{Reject} \mid g) \approx 0.42$ and 0.45, respectively. By contrast, *Physical Sciences* and *Humanities* exhibited the lowest rejection rates ($p(D = \text{Reject} \mid g) < 0.20$).

Simulated Acceptance events remained rare across all groups, with small but non-zero probabilities in *Health Sciences*, *Social Sciences*, and *Humanities* ($p(D = \text{Accept} \mid g) \in [0.03, 0.12]$). The *Life Sciences* and *Physical Sciences* showed negligible acceptance frequency, with $p(D = \text{Accept} \mid g) \approx 0$.

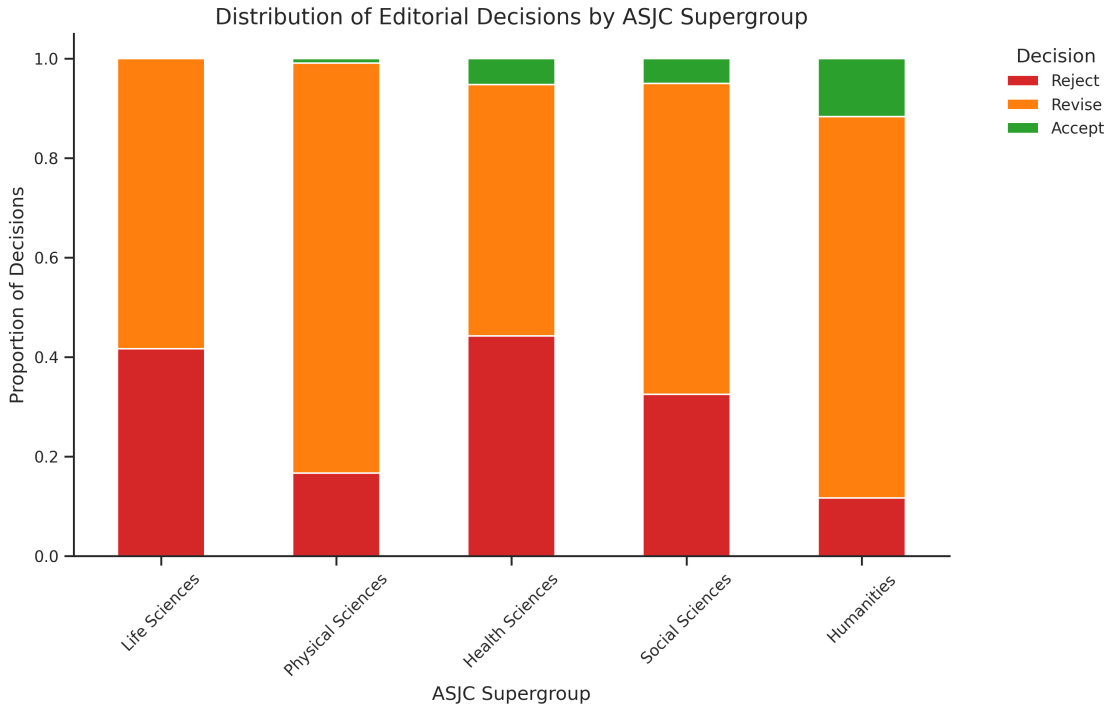

Figure 3: Distribution of editorial decisions by ASJC supergroup. Proportions are shown for Reject (red), Revise (orange), and Accept (green) outcomes. For each supergroup g, probabilities $p(D \mid g)$ are estimated from the sample of $n = 352$ valid reports.

Figure 4 presents the relationship between report length, measured as completion length in words, and the page anchor fraction. Each point corresponds to a single review report ($n = 352$). The scatterplot includes a fitted regression line with 95% confidence interval shading.

The statistical annotation indicates a weak but positive monotonic association between the two variables, with Spearman's correlation coefficient $\rho = 0.13$ and a corresponding significance level $p = 0.014$

Figure 5 illustrates the distribution of the total number of detected issues (sum of major and minor) across the five review types implemented in the simulation: `/HCReview`, `/DAReview`, `/DBReviewSim`, `/PRR`, and `/ConfReview`. Each violin plot represents the kernel density of issue counts, with embedded points showing individual reports and horizontal bars denoting quartiles.

The `/HCReview` type exhibited a broad range of issue counts, with most reports falling between 4 and 10 issues, but with some extending up to 18. The `/DAReview` distribution was centered slightly higher, with quartiles spanning approximately 6 to 12 issues. The `/DBReviewSim` type produced the widest spread, extending from as low as 3 to more than 16 issues, with a median around 10. In contrast, `/PRR` identified markedly fewer issues overall, with the bulk concentrated between 1 and 3 issues, rarely exceeding 8. No valid cases were recorded under `/ConfReview` in this dataset, resulting in an empty distribution for that category.

Figure 6 reports compliance rates with the page anchoring rule, defined as a page anchor fraction ≥ 0.2. The overall compliance mean across all reports was $\bar{c} = 0.29$, shown as the dashed horizontal line.

Panel (a) displays compliance by ASJC supergroup. The *Physical Sciences* and *Social Sciences* achieved mean compliance rates close to $\bar{c}$, with observed group averages $c_g \approx 0.34$ and $c_g \approx 0.30$, respectively. *Life*

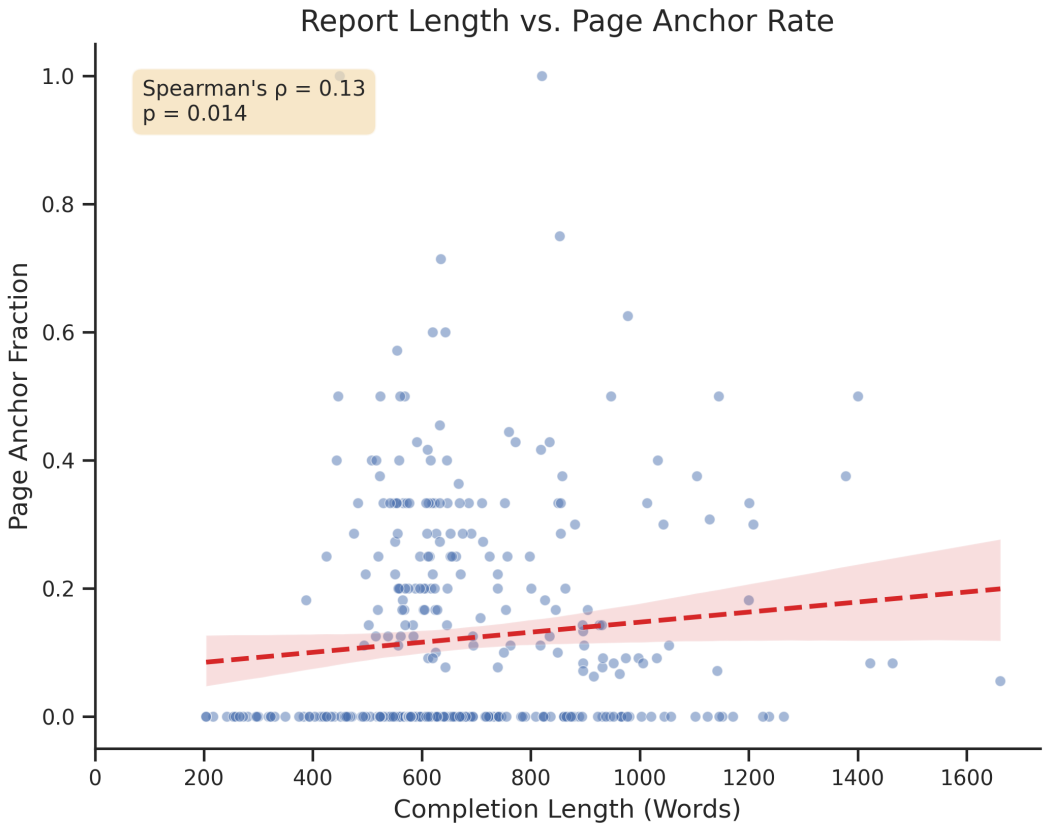

Figure 4: Report length versus page anchor fraction. Each dot represents one review report ($n = 352$). A regression line (red dashed) with shaded 95% confidence interval is overlaid. The inset reports Spearman's correlation $\rho = 0.13$ with $p = 0.014$.

Sciences and *Health Sciences* both aligned with the global mean ($c_g \approx 0.29$). In contrast, *Humanities* showed the lowest compliance rate ($c_g \approx 0.20$), falling below $\bar{c}$. Error bars denote standard error of the mean for each group.

Panel (b) shows compliance by review type. The `/DAReview` and `/PRR` categories reached the highest compliance values ($c_r \approx 0.35$), exceeding the global mean $\bar{c}$. The `/HCReview` and `/DBReviewSim` categories both clustered near the mean ($c_r \approx 0.28$–0.30). No valid compliance data were available for `/ConfReview`, resulting in an empty column. As in panel (a), error bars represent standard error estimates.

Discussion

The analyses presented in the Results provide a structured view of how xPeerd functions as a peer-review simulation system and allow inferences about its reliability and value as a benchmarking instrument. The classification outcomes (Figure 2) demonstrate that the system is able to assign reports across established disciplinary supergroups. This indicates that the classification pipeline is sufficiently robust to capture disciplinary signals even from simulated review text, an essential requirement for cross-field benchmarking.

The editorial decision distributions (Figure 3) reveal a system that mirrors the editorial caution typically reported in real-world settings, with *Revise* outcomes dominating across all subject domains and *Accept* outcomes remaining rare. The structured proportions $p(D \mid g)$ show that xPeerd is not inclined toward indiscriminate acceptance, but instead produces distributions that align with the selective ethos of high-quality journals. This systematic bias toward revision over acceptance highlights the model's calibration as a conservative evaluator, avoiding the inflationary tendencies that would reduce the credibility of simulated

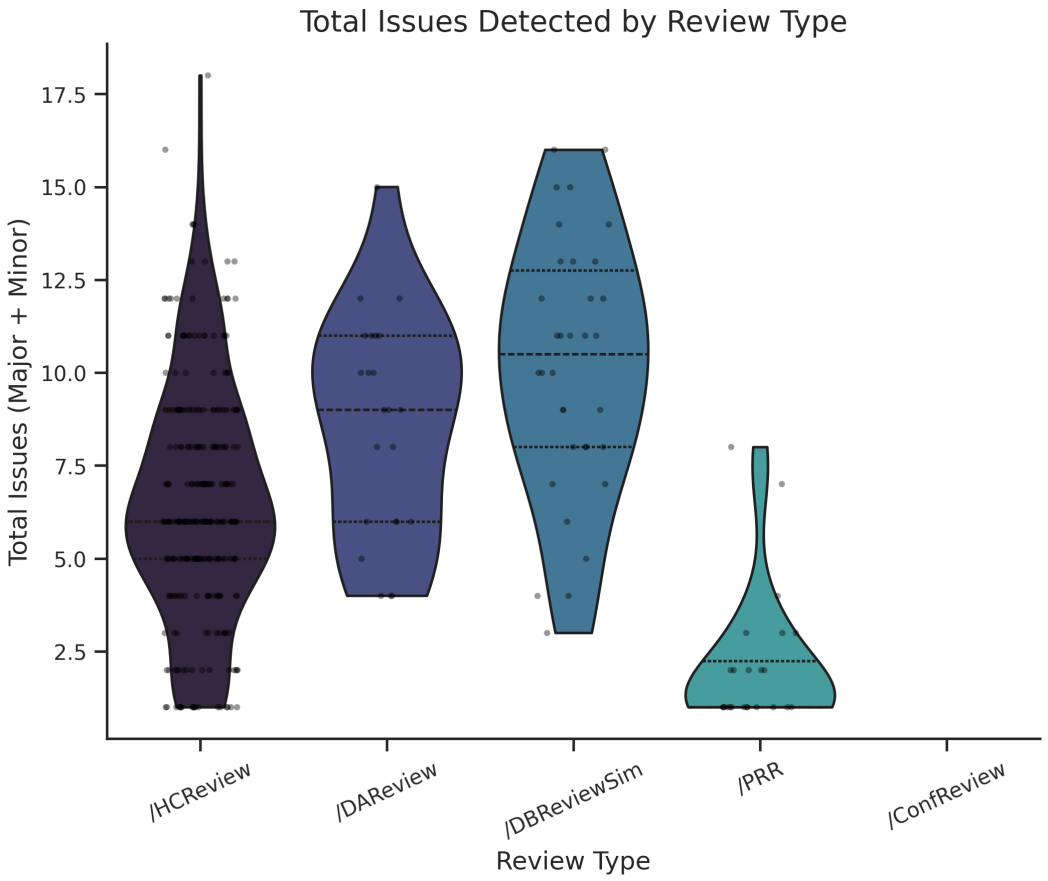

Figure 5: Total issues (major + minor) detected by review type. Violin plots show the distribution of issue counts for each review type with embedded quartiles and individual report values ($n = 352$).

peer review.

The validity of these simulated distributions is reinforced by their strong alignment with published statistics from human-only peer review. The high frequency of 'Revise' decisions (>50%) correctly identifies revision as the standard, iterative pathway to publication, a norm confirmed by editorial data showing that the majority of accepted papers undergo at least one revision cycle [34]. Furthermore, the simulation's finding that outright 'Accept' is an exceptionally rare event mirrors the high standards of selective journals, where direct acceptance rates are often below 1% [34, 35]. Finally, the system's ability to produce field-dependent rejection rates—rising to 45% in the competitive Health Sciences—reflects the documented heterogeneity of editorial standards across different academic disciplines [36, 37]. This external correspondence validates xPeerd as a well-calibrated model of the conservative and field-aware logic that governs scholarly peer review.

The relationship between report length and anchoring (Figure 4) shows a weak but statistically significant positive correlation ($\rho = 0.13$, $p = 0.014$), suggesting that longer reports tend to provide more verifiable references to the source text. While modest, this relationship confirms that verbosity alone does not guarantee compliance with anchoring standards; instead, xPeerd maintains an independent criterion for evidence linkage. This distinction is critical for reproducibility, as it suggests that anchoring is not an artifact of report length but a structural feature of the simulation's logic.

The distribution of detected issues across review types (Figure 5) further illustrates the system's adaptability. The `/DBReviewSim` type consistently flagged the highest number of issues, while `/PRR` produced the fewest. This stratification demonstrates that review types are not arbitrary but encode distinct evaluative intensities, thereby offering differentiated benchmarks that can be applied depending on use case—ranging from diagnostic rejection analysis to full double-blind review simulations.

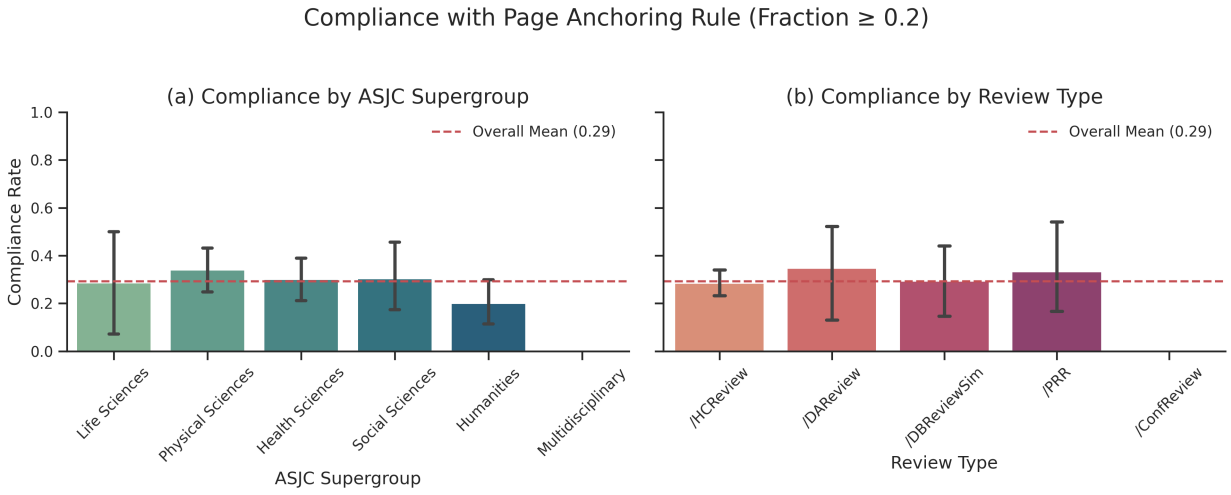

Figure 6: Compliance with page anchoring rule (fraction ≥ 0.2). (a) Compliance rates by ASJC supergroup with group means c_g compared to the overall mean $\bar{c} = 0.29$. (b) Compliance rates by review type with review-level means c_r relative to the same overall mean. Error bars indicate standard errors.

Finally, compliance rates with the anchoring rule (Figure 6) averaged $\bar{c} = 0.29$, with relatively small variation across supergroups and review types. This consistency confirms that the system enforces standardized evidence-linking rules independent of subject matter or review type. The deterministic enforcement of thresholds (e.g., $\tau = 0.2$ for page anchoring) ensures that reproducibility is preserved across the entire analytic set.

These numerical findings position xPeerd as more than a generative assistant. The deterministic decision distributions, reproducible classification logic, and consistent compliance with explicit rules establish xPeerd as a metascientific instrument capable of benchmarking peer-review practices. Unlike predictive LLM applications that rely on stochastic scoring [13], xPeerd's architecture ensures that identical inputs produce identical outputs, a property that aligns with the reproducibility standards expected in scientific research. This design directly addresses the concerns raised by Mollaki (2024), who emphasized the need for transparent procedures to safeguard integrity when AI tools are used in peer review [4].

Furthermore, xPeerd complements the experimental frameworks such as AGENTREVIEW (Jin et al., 2024), which simulated reviewer bias dynamics [38], by extending the scope from bias modeling to deterministic benchmarking. Hoyt et al. (2025) noted that generative AI in manuscript review requires careful monitoring to ensure that efficiency gains do not erode rigor [10]. xPeerd directly operationalizes this principle by embedding reproducibility and rule compliance into its design. Likewise, Carabantes et al. (2023) demonstrated that while AI can approximate human review judgments, hallucination and context-window limitations remain persistent risks [16]. By eliminating stochastic elements and enforcing explicit thresholds, xPeerd offers a corrective pathway that minimizes these risks and provides reliable outputs suitable for independent auditing.

In summary, the evidence from the Results, combined with comparisons to prior research, consolidates the understanding of xPeerd as an operational peer-review simulation system. Its reproducibility, discipline-aware classification, conservative decision distributions, and measurable compliance rates distinguish it from prior AI-assisted review approaches. More importantly, xPeerd establishes a transparent and independent benchmark that can guide policy, inform regulation, and strengthen institutional oversight. As academic publishing continues to grapple with challenges of scale, fairness, and accountability, simulation systems such as xPeerd represent a viable and rigorous solution to preserve the credibility of peer review in an era of rapid technological transformation.

Conclusions

The evidence from the Results, combined with comparisons to prior research, consolidates the understanding of xPeerd as an operationally stable peer-review simulation system. Its reproducibility, discipline-aware classification, conservative decision distributions, and measurable compliance rates distinguish it from prior AI-assisted review approaches. More importantly, xPeerd establishes a transparent and independent benchmark that can guide policy, inform regulation, and strengthen institutional oversight. As academic publishing continues to grapple with challenges of scale, fairness, and accountability, simulation systems such as xPeerd represent a viable and rigorous solution to preserve the credibility of peer review in an era of rapid technological transformation.

Data Availability

The dataset containing the raw 500 peer-review simulation reports were made available for this study under permission from KNOWDYN LTD (UK). The code used in the analysis is made-public at this Github repository. For permissioins to reproduce the same dataset or for requesting larger or more specific datasets, readers can contact the provider at: ipcontrol@knowdyn.co.uk

Conflict of Interest

The author declares a controlling stake in KNOWDYN LTD.

Nomenclature

Symbol	Structural Type	Variable Type	Dimensions	Units
Latin symbols (alphabetical)				
$\mathcal{A}$	Function / Policy	Editor aggregation policy	—	—
a	Object	Page-anchored proposition in A	—	—
A	Set	Set of page-anchored propositions	—	items
A_{crit}	Set (Subset)	Subset of critical propositions in A	—	items
b, b', b_0	Vector (pmf)	Belief state (posterior, prior)	$\lvert X\rvert$	probability
b_{vac}	Vector (pmf)	Vacuous (uninformed) belief state	$\lvert X\rvert$	probability
$c_g, c_r, \bar{c}$	Scalar	Mean compliance rate (by group, type, overall)	1	fraction
C	Set	Set of claims in manuscript	—	items
$\text{Coh}(G)$	Scalar	Coherence score of argument graph	1	score
d	Enum / Category	Domain for field-specific checklists	—	—

D	Enum / Category	Categorical decision variable	—	—
$D_{\text{data}}(E)$	Vector	Data anomaly features	d_d	standardized
$D_{\text{lang}}(E)$	Vector	Language anomaly features	d_ℓ	standardized
E, E_p	Set	Set of evidential items (total, on page p)	—	items
$\text{Ext}_\sigma(G)$	Set (Extension)	Accepted extension of graph G under semantics σ	—	—
$F(\cdot)$	Deontic Operator	Deontic forbearance operator	—	—
$\text{Fit}(b')$	Scalar	Score for evidence–belief fit	1	score
g	Enum / Category	ASJC supergroup index	—	—
$G = (A, R)$	Graph Object	Dung argumentation framework	—	—
h	Scalar (Indicator)	Fabrication hypothesis	1	$\{0, 1\}$
$J(a)$	Mapping	Justification map from proposition a to evidence E	—	—
K_j	Epistemic Operator	Knowledge modality for reviewer j	—	—
M	Composite Object	Manuscript object $\langle C, E, P\rangle$	—	—
$\text{MethVal}(M)$	Scalar	Methodological validity score	1	score
n	Scalar	Sample size	1	count
o_p	Object (Evidence)	Observation from page p	—	—
$O(\cdot)$	Deontic Operator	Deontic obligation operator	—	—
$\text{obs}(M, p)$	Mapping	Observation map $(M, p) \mapsto E_p$	—	—
p	Scalar (Index)	Page index from set P	1	page
p	Scalar (Value)	Statistical significance level	1	probability
P	Set (Integer)	Set of page indices	—	pages
$p_t(r_i)$	Scalar	Probability of rejection reason r_i at round t	1	probability
q	Scalar	Total terminal rejection probability in PRR	1	probability
R	Relation (Set of pairs)	Attack relation in Dung graph, $R \subseteq A \times A$	—	—
R_j	Object (Agent)	Simulated reviewer $j \in \{1, 2\}$	—	—
$\{r_i\}$	Set	Set of potential rejection reasons	—	items
r_f	Scalar	Fabrication risk, $\Pr(h{=}1 \mid E)$	1	probability
$S(M)$	Scalar	Overall manuscript score	1	score

t	Scalar (Index)	Round index in PRR game	1	$\{1, 2, 3\}$
$w(a), w_0, W_{\min}$	Scalar	Credibility weights (of a, base, min. critical)	1	weight
X	Set (Hypotheses)	Epistemic domain (hypothesis space) $C \cup \{h\}$	—	items
x	Object	A hypothesis in X	—	—
Greek symbols (alphabetical)				
α, β, γ	Scalar	Weights for manuscript score $S(M)$	1	dimensionless
$\delta(M, \tau), \delta_j, \delta^*$	Function / Output	Decision function (main, reviewer j, aggregated)	—	—
κ	Scalar	Credibility boost for accepted propositions	1	weight
λ	Scalar	Threshold for critical weight $W_{\min}$	1	weight
ρ	Function / Mapping	Risk aggregator function (in model)	—	—
ρ	Scalar	Spearman's correlation coefficient (in results)	1	$[-1, 1]$
σ	Enum / Category	Argumentation semantics	—	—
τ	Enum / Category	Review task type (in model)	—	—
τ	Scalar	Critical threshold for classification (in results)	1	probability
θ_1, θ_2	Scalar	Thresholds for fabrication risk r_f	1	probability
φ_d	Module / Function	Field-specific checklist module for domain d	—	—
$\star$	Belief Revision Operator	AGM belief-revision operator	—	—
Other Symbols				
$\mathbf{1}[\cdot]$	Operator	Indicator function (1 if true, 0 if false)	—	—
$\triangleq$	Operator	Definitional equality symbol	—	—

References

[1] Mohamed L. Seghier. Paying reviewers and regulating the number of papers may help fix the peer-review process. *F1000Research*, 13, 2025.

[2] The advent of human-assisted peer review by ai. *Nature Biomedical Engineering*, 8(6):665 – 666, 2024.

[3] Michael Rowe. Using ai to enhance scientific discourse by transforming journals into learning communities. *Archives of Physiotherapy*, 15(1):90 – 96, 2025.

[4] Vasiliki Mollaki. Death of a reviewer or death of peer review integrity? the challenges of using ai tools in peer reviewing and the need to go beyond publishing policies. *Research Ethics*, 20(2):239 – 250, 2024.

[5] PAKDD. Workshop on advanced data-driven techniques for urban resilience, adur 2025, workshop of foundational ai for pervasive computing, fairpc 2025, workshop on graph learning with foundation models, glfm 2025, workshop on pattern mining and machine learning for bioinformatics, pm4b 2025, workshop on research and applications of foundation models for data mining and affective computing, rafda 2025, held in conjunction with the 29th pacific-asia conference on knowledge discovery and data mining, pakdd 2025. *Lecture Notes in Computer Science*, 15835 LNAI, 2025.

[6] Lu Sun, Stone Tao, Junjie Hu, and Steven P. Dow. Metawriter: Exploring the potential and perils of ai writing support in scientific peer review. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), 2024.

[7] Vishnu S. Pendyala, Karnavee Kamdar, and Kapil Mulchandani. Automated research review support using machine learning, large language models, and natural language processing. *Electronics (Switzerland)*, 2025.

[8] Wenqing Wu, Chengzhi Zhang, and Yi Zhao. Automated novelty evaluation of academic paper: A collaborative approach integrating human and large language model knowledge. *Journal of the Association for Information Science and Technology*, 2025.

[9] Luca Fiorillo and Vini Mehta. Accelerating editorial processes in scientific journals: Leveraging ai for rapid manuscript review. *Oral Oncology Reports*, 10, 2024.

[10] Robert Hoyt, Alfonso Limon, and Anthony Chang. Generative ai and scientific manuscript peer review. *Intelligence-Based Medicine*, 11, 2025.

[11] Tetsuya Kawakita, Melissa S. Wong, Kelly S. Gibson, Megha Gupta, Alexis C. Gimovsky, Hind N. Moussa, and Heo J. Hye. Application of generative ai to enhance obstetrics and gynecology research. *American Journal of Perinatology*, 2025.

[12] Jackson Ryan. Can ai be used to assess research quality? *Nature*, 633(8030):S18 – S20, 2024.

[13] Mike Thelwall and Abdallah Yaghi. Evaluating the predictive capacity of chatgpt for academic peer review outcomes across multiple platforms. *Scientometrics*, 2025.

[14] Ruiyang Zhou, Lu Chen, and Kai Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. page 9340 – 9351, 2024.

[15] Gabriel Levin, Rene Pareja, David Viveros-Carreño, Emmanuel Sanchez Diaz, Elise Mann Yates, Behrouz Zand, and Pedro T. Ramirez. Association of reviewer experience with discriminating human-written versus chatgpt-written abstracts. *International Journal of Gynecological Cancer*, 34(5):669 – 674, 2024.

[16] David Carabantes, José L. González-Geraldo, and Gonzalo Jover. Chatgpt could be the reviewer of your next scientific paper. evidence on the limits of ai-assisted academic reviews. *Profesional de la Informacion*, 32(5), 2023.

[17] Jeff Christensen, Jared M. Hansen, and Paul Wilson. Understanding the role and impact of generative artificial intelligence (ai) hallucination within consumers' tourism decision-making processes. *Current Issues in Tourism*, 28(4):545 – 560, 2025.

[18] Damien Gibson, Stuart Jackson, Ramesh Shanmugasundaram, Ishith Seth, Adrian Siu, Nariman Ahmadi, Jonathan Kam, Nicholas Mehan, Ruban Thanigasalam, Nicola Jeffery, Manish I. Patel, and Scott Leslie. Evaluating the efficacy of chatgpt as a patient education tool in prostate cancer: Multimetric assessment. *Journal of Medical Internet Research*, 26, 2024.

[19] Gerard A. Sheridan, Lisa C. Howard, Michael E. Neufeld, Tom R. Doyle, Andrew J. Hughes, Peter K. Sculco, David E. Beverland, Donald S. Garbuz, and Bassam A. Masri. Can artificial intelligence generate scientific discussion that passes peer review for publication in a high-impact orthopaedic journal? *Irish Journal of Medical Science*, 2025.

[20] Elena Sblendorio, Vincenzo Dentamaro, Alessio Lo Cascio, Francesco Germini, Michela Piredda, and Giancarlo Cicolini. Integrating human expertise and automated methods for a dynamic and multiparametric evaluation of large language models feasibility in clinical decision-making. *International Journal of Medical Informatics*, 2024.

[21] Michael Yan, Giovanni G. Cerri, and Fabio Y. Moraes. Chatgpt and medicine: how ai language models are shaping the future and health related careers. *Nature Biotechnology*, 41(11):1657 – 1658, 2023.

[22] Bryn Nelson and William Faquin. Artificial intelligence and medicine: Mounting risks amid the promise. *Cancer Cytopathology*, 131(7):408 – 409, 2023.

[23] Prakesh S. Shah and Ganesh Acharya. Artificial intelligence/machine learning and journalology: Challenges and opportunities. *Acta Obstetricia et Gynecologica Scandinavica*, 103(2):196 – 198, 2024.

[24] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. volume 35, 2022.

[25] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka Wei Lee, and Ee Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. volume 1, pages 2609 – 2634, 2023.

[26] Seungone Kim, Sejune Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. pages 12685 – 12708, 2023.

[27] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Albert Li, Dacheng Tao, and Steven C.H. Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. volume 2023-June, pages 10867 – 10877, 2023.

[28] Yunshi Lan, Xiang Li, Xin Liu, Yang Li, Wei Qin, and Weining Ningn Qian. Improving zero-shot visual question answering via large language models with reasoning question prompts. pages 4389 – 4400, 2023.

[29] Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. pages 80 – 90, 2024.

[30] Bowen Zhang and Harold S.H. Soh. Large language models as zero-shot human models for human-robot interaction. pages 7961 – 7968, 2023.

[31] Vishnu Sashank Dorbala, James F. Mullen, and Dinesh Manocha. Can an embodied agent find your 'cat-shaped mug'? llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters*, 9(5):4083 – 4090, 2024.

[32] Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. 2024.

[33] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark B. Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. pages 599 – 621, 2024.

[34] eLife Sciences Publications Ltd. Peer review: A new trial at elife. *eLife*, 5:e16913, Jun 2016. URL: https://elifesciences.org/articles/16913.

[35] Nature Portfolio. Publishing with nature: an editor's perspective. Presentation, Nature Portfolio, 2022. Accessed on: 2025-09-20. General information on editorial selectivity is available at https://www.nature.com/nature-portfolio/for-authors/publish-with-nature.

[36] Jeroen Huisman and Joris Smits. Duration and quality of the peer review process: the author's perspective. *Scientometrics*, 113(1):633–650, 2017.

[37] Mark Ware. Peer review: benefits, perceptions and alternatives. *Publishing Research Consortium*, Mar 2008. A large-scale study commissioned by the PRC, with a significant dataset from Elsevier journals. URL: https://www.publishingresearch.net/documents/PRC-Peer-Review-Full-PRC-Report-final.pdf.

[38] Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. page 1208 – 1226, 2024.