*applied sciences*

MDPI

*Article*

# Applied with Caution: Extreme-Scenario Testing Reveals Significant Risks in Using LLMs for Humanities and Social Sciences Paper Evaluation

**Hua Liu [1], Ling Dai [2,*] and Haozhe Jiang [3,*]**

[1] Institute of Higher Education, Taizhou University, Taizhou 225300, China; 661008@tzu.edu.cn
[2] Faculty of Education, East China Normal University, Shanghai 200062, China
[3] College of Education, Zhejiang University, Hangzhou 310058, China
* Correspondence: perfectdaring@outlook.com (L.D.); haozhejiang@zju.edu.cn (H.J.); Tel.: +86-15952689203 (L.D.); +86-13812952896 (H.J.)

**Abstract**

The deployment of large language models (LLMs) in academic paper evaluation is increasingly widespread, yet their trustworthiness remains debated; to expose fundamental flaws often masked under conventional testing, this study employed extreme-scenario testing to systematically probe the lower performance boundaries of LLMs in assessing the scientific validity and logical coherence of papers from the humanities and social sciences (HSS). Through a highly credible quasi-experiment, 40 high-quality Chinese papers from philosophy, sociology, education, and psychology were selected, for which domain experts created versions with implanted "scientific flaws" and "logical flaws". Three representative LLMs (GPT-4, DeepSeek, and Doubao) were evaluated against a baseline of 24 doctoral candidates, following a protocol progressing from 'broad' to 'targeted' prompts. Key findings reveal poor evaluation consistency, with significantly low intra-rater and inter-rater reliability for the LLMs, and limited flaw detection capability, as all models failed to distinguish between original and flawed papers under broad prompts, unlike human evaluators; although targeted prompts improved detection, LLM performance remained substantially inferior, particularly in tasks requiring deep empirical insight and logical reasoning. The study proposes that LLMs operate on a fundamentally different "task decomposition-semantic understanding" mechanism, relying on limited text extraction and shallow semantic comparison rather than the human process of "worldscape reconstruction → meaning construction and critique", resulting in a critical inability to assess argumentative plausibility and logical coherence. It concludes that current LLMs possess fundamental limitations in evaluations requiring depth and critical thinking, are not reliable independent evaluators, and that over-trusting them carries substantial risks, necessitating rational human-AI collaborative frameworks, enhanced model adaptation through downstream alignment techniques like prompt engineering and fine-tuning, and improvements in general capabilities such as logical reasoning.

**Keywords:** large language models (LLMs); academic paper evaluation; humanities and social sciences (HSS); AI trustworthiness; human-AI collaboration

check for updates

Received: 25 August 2025
Revised: 1 October 2025
Accepted: 1 October 2025
Published: 3 October 2025

**Citation:** Liu, H.; Dai, L.; Jiang, H. Applied with Caution: Extreme-Scenario Testing Reveals Significant Risks in Using LLMs for Humanities and Social Sciences Paper Evaluation. *Appl. Sci.* **2025**, *15*, 10696. https://doi.org/10.3390/app151910696

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Large Language Models (LLMs) are increasingly deployed in academic paper evaluation contexts, leveraging their sophisticated natural language understanding and gen-

eration capabilities across diverse assessment scenarios, from scholarly peer review to dissertation and coursework evaluation [1]. Evidence from recent empirical investigations underscores this accelerating trend: systematic analysis of peer review practices reveals that 6.5% to 16.9% of academic reviews demonstrate substantial LLM influence across multiple scholarly venues [2,3], while examination of recent peer review data indicates that at least 15.8% of evaluative reports were generated with LLM assistance [4,5]. These findings collectively signal a paradigm shift toward systematic integration of LLMs within academic evaluation frameworks.

LLMs present distinctive advantages over both traditional human evaluation and earlier Automated Essay Scoring (AES) systems [6]. These computational evaluators offer adaptive, customizable assessments without requiring domain-specific pre-training; dramatically reduce temporal costs while enhancing evaluation efficiency; deliver scalable formative feedback to facilitate iterative manuscript refinement; and potentially mitigate subjective biases inherent in human judgment, thereby promoting evaluative objectivity and fairness [7].

Nevertheless, inherent limitations of LLMs—including hallucination, inconsistency, systematic biases, sycophantic behavior, and limited interpretability—raise fundamental questions regarding their evaluative accuracy, reliability, and validity, particularly in the context of HSS paper assessment, which relies heavily on contextual nuance, theoretical reasoning, and interpretive depth [8]. The trustworthiness of LLMs as academic evaluators remains contentious. Recent empirical investigations attempting to characterize LLM evaluation capabilities have yielded contradictory findings: while some studies report substantial reliability and validity, others expose critical vulnerabilities and warn of significant risks [9]. These ostensibly conflicting conclusions stem largely from variations in research design and methodological approaches. When the evaluation task involves judging whether a paper adheres to specific rules (e.g., grammar, syntax, formatting conventions, or research norms), LLMs often excel [10]. Conversely, their performance is often unsatisfactory when the task requires a deep understanding and critical assessment of the scholarly content. In everyday scenarios, with appropriate prompt guidance, LLMs can generally provide passable evaluations, albeit with issues such as generic responses and a lack of depth. However, these performances may obscure fundamental flaws in their evaluative capabilities [11]. A limited number of meticulously designed studies have attempted to expose their true abilities through extreme-scenario testing.

Building upon critical analysis of existing methodological frameworks, this study employs extreme-scenario testing to probe the lower performance boundaries of LLM evaluation capabilities, specifically within HSS contexts. Through systematic exploration of capability limits, we derive evidence-based human-AI collaboration strategies and technical optimization pathways to mitigate risks associated with inappropriate deployment of LLMs in academic evaluation contexts [12,13].

This paper makes the following contributions:

- We designed a highly credible comparative experiment focused on HSS papers, which not only effectively probes the lower bounds of LLM evaluation through extreme scenarios but also addresses the challenge of measuring evaluation competence due to the highly subjective nature of HSS paper assessment.
- The study exposes fundamental limitations in LLMs' paper evaluation capabilities, particularly in the domains of scientific and logical evaluation, which are typically masked under naturalistic conditions.
- We provide a novel interpretation of the distinctive mechanisms underlying LLM-based paper evaluation, which differ fundamentally from those of human evaluators

and help explain their shortcomings in capturing fact-based logical reasoning and theoretical coherence in HSS writing.

## 2. Related Work

### 2.1. Evaluation Capabilities of LLMs in Academic Paper Assessment

Empirical investigations have yet to converge on a consensus regarding LLM evaluation capabilities, a divergence attributable both to task-dependent performance variability and methodological heterogeneity in capability assessment approaches.

#### 2.1.1. Task-Dependent Variability in LLM Evaluation Performance

Evaluation capabilities demonstrate systematic variation across assessment dimensions. LLMs exhibit robust performance in rule-based evaluation tasks, demonstrating accuracy, reliability, and efficiency when assessing manuscript compliance with formatting standards, grammatical conventions, and stylistic guidelines. However, their performance deteriorates markedly in domains requiring sophisticated comprehension. Bouziane et al. documented ChatGPT's inferior performance relative to human evaluators in assessing thematic coherence [3]. Steiss et al., comparing ChatGPT and human evaluators across 200 manuscripts on multiple dimensions of formative feedback—including criterion-based assessment, clarity, accuracy, emphasis on critical elements, and maintenance of supportive tone—found ChatGPT's feedback quality inferior to human evaluators across all dimensions except criterion-based assessment, with effect sizes ranging from small to moderate [4].

Performance further varies according to the manuscripts' academic rigor. When evaluating scholarly manuscripts characterized by high academic sophistication and originality, LLMs manifest critical limitations, including insufficient domain-specific expertise that precludes contextualizing feedback within nuanced disciplinary frameworks [5] and inadequate capacity to evaluate scientific contributions within broader disciplinary contexts or determine publication merit [12]. Multiple investigations have documented deficiencies in LLM academic evaluation capabilities, including tendencies to accept authorial claims without rigorous verification, producing superficial textual summaries rather than critical assessments [13]; inability to detect deep theoretical flaws, identify missing metrics, or recognize exaggerated conclusions [14]; and potential misinterpretation of low-quality manuscripts as methodologically sound and credible [15]. Conversely, LLMs demonstrate superior performance when evaluating less academically demanding texts. Research indicates ChatGPT can match expert mentors in analyzing reflective essays across dimensions of depth, analytical rigor, logical clarity, originality, reflective feedback quality, and collaborative engagement [16].

#### 2.1.2. Methodological Influences on Observed LLM Capabilities

Testing paradigms substantially influence observed LLM evaluation performance. Under conventional testing conditions, LLMs often demonstrate favorable performance characteristics. Recent large-scale empirical investigation revealed substantial overlap between GPT-4-generated and human-generated reviews, with 57.4% of users rating GPT-4 feedback as helpful or very helpful, and 82.4% considering it at least comparable to some human reviewer feedback [1]. However, performance under conventional conditions inadequately reflects genuine evaluation capabilities. LLMs' architectural mechanisms—predicting subsequent token probabilities, supervised fine-tuning, and reinforcement learning from human feedback—enable generation of text conforming to user expectations without necessarily producing accurate or meaningful content. This manifests as superficially compelling but factually incorrect commentary [17] and generic, surface-level rather than targeted feedback [18].

Extreme-scenario testing more effectively exposes LLM limitations. Recent studies employing systematic extreme testing to reveal critical vulnerabilities in LLM academic review processes. By substituting author affiliations with prestigious researchers, universities, or companies, they exposed inherent institutional bias, with average positive ratings increasing from 36.8% to 41.2%, 40.8%, and 41.6%, respectively. Through adversarial manipulation—injecting machine-readable but humanly invisible text emphasizing manuscript strengths—they demonstrated LLMs' susceptibility to authorial manipulation, with 90% consistency between generated reviews and injected content, while human reviewer consistency decreased from 53% to 16% [19].

The design of prompts also significantly influences the manifestation of LLMs' evaluation capabilities. Zhong et al. employed extensive prompts—which included the ICLR review guidelines, randomly selected human-written reviews for both accepted and rejected papers, and a template for the ICLR 2024 review format—to generate LLM-based reviews [20]. Separately, Wong et al. distilled the review criteria of a leading cell biology journal into key dimensions: originality, accuracy, conceptual advance, timeliness, and significance [21]. These defined criteria were then used to prompt an LLM to evaluate a given PubMed ID paper on a three-star rating system, culminating in an overall assessment. Such meticulously crafted prompts facilitate the LLM's understanding of domain-specific evaluation requirements and conventions, thereby enabling improved performance.

### *2.2. Factors Influencing LLM Paper Evaluation Capabilities*

Deficiencies in LLM paper evaluation capabilities can be traced to inherent limitations, including inconsistency, hallucination, and constrained logical reasoning capacities.

#### 2.2.1. Inconsistency

Inconsistency manifests when semantically equivalent inputs yield semantically non-equivalent outputs [22], generating user confusion and eroding trust. Contributing factors encompass stochastic sampling procedures, model updates, platform-internal operations, and hardware specifications. Additionally, conflicting or ambiguous information within training corpora may exacerbate inconsistency. This phenomenon intensifies markedly when evaluating lengthy documents such as academic papers. Multiple investigations document that LLMs generate unstable and inconsistent assessments when repeatedly reviewing identical manuscripts, exhibiting both score fluctuations and commentary variations [23,24].

#### 2.2.2. Hallucination

Hallucination refers to LLM generation of content that appears confident yet remains meaningless or unfaithful to source material [25]. Huang et al. categorize hallucinations into factual and faithfulness dimensions, with factual hallucinations further subdivided into factual contradictions and factual fabrications [26]. Factual contradictions involve outputs containing real-world information that exhibits internal contradictions. Factual fabrications encompass outputs containing facts unverifiable against established knowledge, further classified into unverifiable hallucinations and overclaim hallucinations [27]. The former represents statements entirely nonexistent or unverifiable through existing sources—for instance, when queried about "major environmental impacts of Eiffel Tower construction," an LLM erroneously claimed "construction led to extinction of Parisian tigers." The latter involves claims lacking universal validity due to subjective bias, such as asserting "Eiffel Tower construction is widely recognized as the event triggering the global green building movement." Such subjective assertions prove particularly challenging to identify and eliminate [28].

Hallucination exerts dual negative impacts on LLM evaluation capabilities. First, LLMs may fabricate information when generating commentary, presenting it alongside

spurious or irrelevant supporting references—representing active confabulation. Second, hallucination impairs LLMs' capacity to accurately discern factual veracity and argumentative validity within manuscripts. Quantitative assessments have revealed accuracy rates of merely 19.3% and 15.3% for evaluating manuscript soundness and providing supplementary content suggestions, respectively. Evaluating manuscript soundness requires assessing statement correctness, argumentative validity, and evidential completeness; providing supplementary suggestions demands integration of domain knowledge with critical manuscript comprehension [29]. Both deficiencies correlate strongly with entrenched hallucination tendencies. Furthermore, LLMs may assign higher scores to incomplete manuscripts compared to complete versions—another consequence of hallucination.

#### 2.2.3. Limited Logical Reasoning Capabilities

Research demonstrates that LLMs excel in reasoning tasks with well-defined problems, explicit knowledge frameworks, and decomposable reasoning steps. However, they lag behind human performance in scenarios requiring rigorous logic or novel combinations. Mounting evidence indicates LLMs can provide superficially plausible yet ultimately incorrect or invalid justifications. Careful evaluations of chain-of-thought explanations revealed they typically fail to accurately reflect LLMs' underlying reasoning processes. When controlled biased features are introduced (e.g., consistently placing correct answers in option A), LLMs fail to acknowledge dependence on these features in their chain-of-thought [30]. Studies have demonstrated that ChatGPT can reach correct mathematical theorem conclusions through flawed or invalid logical steps [31].

Capability analyses of critical logical reasoning tasks (reading comprehension and natural language inference) further highlight LLM reasoning limitations. Beaulieu et al. found ChatGPT and GPT-4 performance deteriorated significantly on novel datasets requiring logical reasoning, despite relatively strong performance on existing benchmarks, suggesting current success may rely on exploiting dataset-specific patterns rather than robust human-like reasoning [32]. Additionally, LLMs exploit superficial spurious patterns rather than meaningful logic in reasoning tasks, such as heavy reliance on lexical overlap between premises and hypotheses in NLI benchmarks [33].

Lacking rigorous logical reasoning capabilities in novel combinatorial scenarios inevitably impacts LLMs' comprehension and evaluation of manuscript logical rigor, though empirical data in this domain remains scarce.

Furthermore, extreme-scenario testing, relative to conventional paradigms, facilitates isolation from sycophantic text generation influences, exposing genuine LLM evaluation capabilities. Future investigations should prioritize probing lower performance boundaries under extreme conditions, providing empirical foundations for judicious LLM deployment and risk mitigation. Moreover, research on LLM inconsistency, hallucination, and logical limitations indicates critical directions for capability assessment—evaluative consistency, scientific validity assessment, and logical coherence evaluation—representing core competencies in academic evaluation.

## 3. Research Methodology

This study employs extreme-scenario testing to probe LLMs' evaluative consistency and their capacity to assess scientific validity and logical coherence in academic manuscripts.

### *3.1. Quasi-Experimental Design*

This study employed a quasi-experimental design with three manuscript variants (original, scientifically flawed, and logically flawed versions) as the independent variable. The dependent variables comprised evaluation outcomes from both LLMs and human

evaluators, including numerical scores and written commentary. Prompting strategies and LLM versions served as control variables to ensure experimental consistency.

#### 3.1.1. Test Data Preparation

Domain experts from philosophy, sociology, education, and psychology (doctoral degree holders and professors in their respective fields with $\geq$3 high-quality publications within the past three years) selected 10 high-quality papers from each discipline (comprehensive rating $\geq$ 8.0 from two domain experts per discipline) from CSSCI-indexed journals published within the past two years. This yielded 40 manuscripts (10 from each of the four disciplines) spanning empirical and theoretical research paradigms. Following standardized protocols (detailed criteria omitted for brevity), original manuscripts underwent targeted modifications (single-point additions, deletions, repositioning, or substitutions) to introduce substantial scientific or logical flaws. Scientific flaws encompassed: methodological inadequacies, data misinterpretation, inappropriate result interpretation, and unsound arguments or exposition. Logical flaws included: poor semantic coherence between passages and compromised logical continuity across sections. Relevant examples can be found in Appendix A.

#### 3.1.2. Selection of Large Language Models

To ensure the comprehensiveness and representativeness of our findings, we selected three distinct LLMs as evaluators: GPT-4, DeepSeek, and Doubao. This selection was strategic and based on the following rationale. GPT-4 (OpenAI) was included as a benchmark model due to its widely recognized state-of-the-art performance in general NLP tasks. DeepSeek (DeepSeek AI) was chosen as a leading, powerful open-source alternative, known for its competitive capabilities often cited in comparative studies. Finally, Doubao (ByteDance) was selected to represent a category of highly accessible and influential LLMs within the Chinese digital ecosystem. Its inclusion is crucial for our objective to assess the capabilities of models that are readily available to a broad Chinese-speaking audience, thereby addressing the practical implications of LLM usage in evaluating Chinese HSS papers. This tripartite selection—covering a global frontier model, a top-tier open-source model, and a dominant domestic model—ensures a diverse and pragmatic analysis.

#### 3.1.3. Testing Protocol

The testing protocol comprised three sequential phases designed to systematically evaluate LLM performance under varying conditions. In the first phase, three representative LLMs (Doubao-1.6, DeepSeek-V3, and GPT-4.1) evaluated all three manuscript versions using standardized broad prompts that specified evaluation dimensions of scientific validity, logical coherence, and formal compliance with corresponding score allocations. Each manuscript version underwent duplicate evaluation to enable assessment of the models' capacity to detect deliberately introduced scientific and logical flaws.

The second phase examined whether targeted guidance could enhance flaw detection capabilities. Targeted prompts directing specific attention to methodological rigor (e.g., "Please re-evaluate with particular attention to methodological rigor") or logical continuity (e.g., "Please re-evaluate focusing on logical continuity between sections") were deployed to test whether such explicit guidance improved detection and evaluation performance (Broad prompts and targeted prompts see Appendix B).

The third phase established human baseline performance through evaluation by 24 doctoral candidates from philosophy, sociology, education, and psychology disciplines. These evaluators, each possessing at least one high-quality publication within the preceding three years, assessed the manuscript versions following the same evaluation framework. Each evaluator assessed 10 distinct papers to ensure each paper received two independent

evaluations. Meanwhile, the versions of the papers were randomly distributed and kept confidential from the evaluators to prevent contamination effects.

### 3.2. Statistical Analysis

The statistical analysis employed multiple complementary approaches to examine evaluation consistency and flaw detection capabilities. For intra-rater reliability assessment, two-way mixed-effects models (ICC(3,1)) were applied to evaluate the consistency of each LLM's duplicate ratings across all 120 manuscripts, providing insight into the stability of individual model performance.

Inter-rater reliability was examined through two distinct approaches. Two-way mixed-effects models (ICC(3,k)) assessed consistency among mean scores across the three LLMs, while two-way random-effects models (ICC(2,k)) evaluated agreement between each LLM's mean scores and human evaluator means. This dual approach enabled comprehensive understanding of both within-group and between-group consistency patterns.

Version comparison analysis utilized paired-samples *t*-tests to examine mean score differences between original and modified manuscripts for both LLM and human evaluators. Detection rates were calculated as the proportion of correctly identified flaws, with Cohen's Kappa coefficients quantifying the level of agreement between LLMs and human evaluators in flaw detection performance [34].

The impact of prompting strategies was assessed through paired-samples *t*-tests comparing scores under broad versus targeted prompting conditions. Changes in detection rates corresponding to different prompting approaches were systematically documented to quantify the effect of explicit guidance on evaluation performance.

Heterogeneity in detection capabilities across different flaw types was examined using Kruskal–Wallis tests for multi-category comparisons and Mann–Whitney U tests for binary comparisons. These non-parametric approaches were selected to accommodate potential violations of normality assumptions and to provide robust insights into differential detection patterns across flaw categories.

### 3.3. Structured Content Analysis

Researchers coded LLM commentary regarding detection of predetermined flaws using a three-tier classification: "Yes" (accurate identification and analysis of the flaws, or provision of targeted improvement suggestions), "Partial" (incomplete coverage of the flaws, or identification without precise analysis), "No" (failure to identify the flaws, or provision of positive evaluation despite the flaws).

### 3.4. Case Analysis

Representative cases of flaw detection (both successful and unsuccessful instances where commentary addressed the modified sections) underwent in-depth qualitative analysis to elucidate mechanisms underlying accurate versus erroneous evaluations.

## 4. Findings

### 4.1. Evaluation Consistency

Intra-rater reliability reflects the consistency of a specific rater's evaluation criteria over time, indicating the stability of the scoring results. This study adopts the criteria proposed by Koo and Li for interpreting the Intraclass Correlation Coefficient (ICC): ICC < 0.50 indicates poor consistency and is unacceptable [35]; $0.50 \leq$ ICC < 0.75 indicates moderate consistency and is tolerable; $0.75 \leq$ ICC < 0.90 indicates good consistency and is acceptable; ICC $\geq 0.90$ indicates excellent consistency, which is highly desirable. Analysis of the intra-rater reliability for all 120 papers (under unified, broad prompts) revealed that the two scoring

attempts by the three large language models (LLMs) all exhibited poor consistency (with ICC values below 0.75). This suggests substantial random error in the models' scoring, casting doubt on the reliability of their evaluation results (in contrast, the human expert showed good consistency between two scoring attempts).

Inter-rater reliability measures the uniformity of scoring standards applied by different raters. This study refers to the evaluation standards established by Fleiss [36] and Cicchetti [37]: ICC < 0.40 indicates poor consistency and is unacceptable; 0.40 ≤ ICC < 0.60 indicates fair/moderate consistency and is tolerable; 0.60 ≤ ICC < 0.75 indicates good consistency and is acceptable; 0.75 ≤ ICC < 0.90 indicates very good consistency, which is highly desirable; ICC ≥ 0.90 indicates excellent consistency. The relatively lower thresholds of these criteria account for the greater difficulty in controlling differences and subjectivity among individuals. Analysis of the inter-rater reliability between the LLMs themselves and between the LLMs and the human expert showed that only the consistency between DeepSeek and the human expert reached a fair level. The consistency between Doubao, GPT-4 and the human expert, as well as the consistency among the three LLMs themselves, was poor. This indicates a significant discrepancy in the practical application of scoring standards between large language models and human evaluators (see Table 1 for details).

**Table 1.** Score consistency test results.

| Reliability Type | Evaluators | Metric | ICC | 95% CI | *p*-Value |
|---|---|---|---|---|---|
| Intra-rater Reliability | DeepSeek | ICC(3, 1) | 0.219 | [0.042, 0.280] | 0.008 |
| | Doubao | ICC(3, 1) | 0.461 | [0.309, 0.591] | 0.000 |
| | GPT-4 | ICC(3, 1) | 0.192 | [0.013, 0.359] | 0.018 |
| | Human Experts | ICC(2, k) | 0.785 | [0.651, 0.852] | 0.000 |
| Inter-rater Reliability | deepseek; Doubao; GPT-4 | ICC(3, k) | 0.394 | [0.188, 0.556] | 0.000 |
| | deepseek; Human Experts | ICC(2, k) | 0.525 | [0.262, 0.687] | 0.000 |
| | Doubao; Human Experts | ICC(2, k) | 0.100 | [−0.145, 0.311] | 0.187 |
| | GPT-4; Human Experts | ICC(2, k) | 0.100 | [−0.215, 0.343] | 0.250 |

### *4.2. Detection and Evaluation of Scientific Flaws*

#### 4.2.1. Score Differences Between Versions Under Broad Prompts

Paired-samples *t*-tests comparing evaluator scores under broad prompting conditions revealed distinct patterns (Table 2). All three LLMs showed no significant differences between their scores for original manuscripts (mean of two evaluations, hereafter) and scientifically flawed versions. In contrast, human evaluators demonstrated significant score differences between versions, with original manuscripts receiving significantly higher mean scores than scientifically flawed versions. These differences exhibited not only statistical significance but also substantial practical significance, as evidenced by large effect sizes.

**Table 2.** Results of Paired *t*-test on scores between the original version and the version with scientific issues.

| Pair | Version | Mean | N | SD | t | *p* | r | Cohen's d |
|---|---|---|---|---|---|---|---|---|
| Pair 1 (DeepSeek) | Original | 88.85 | 40 | 2.551 | 0.723 | 0.474 | 0.193 | 0.114 |
| | Scientific flaws | 88.48 | 40 | 2.526 | | | | |
| Pair 2 (Doubao) | Original | 88.89 | 40 | 1.866 | −0.270 | 0.789 | 0.695 | 0.043 |
| | Scientific flaws | 89.95 | 40 | 1.884 | | | | |
| Pair 3 (GPT-4) | Original | 88.36 | 40 | 2.279 | −1.285 | 0.206 | 0.100 | 0.203 |
| | Scientific flaws | 89.93 | 40 | 1.922 | | | | |
| Pair 4 (Human Experts) | Original | 89.83 | 40 | 2.427 | 7.343 | 0.000 | 0.541 | 1.161 |
| | Scientific flaws | 85.89 | 40 | 4.023 | | | | |

#### 4.2.2. Score Differences and Detection Rates: Broad vs. Targeted Prompts

Paired-samples *t*-tests comparing LLM scores for scientifically flawed versions under broad versus targeted prompting conditions revealed differential patterns. DeepSeek

and Doubao demonstrated significant score reductions under targeted prompts, with DeepSeek exhibiting a Cohen's d of 1.620—indicating very large effects both statistically and practically—while Doubao showed a moderate effect size (Cohen's d = 0.646). GPT-4 scores remained stable across prompting conditions (Table 3).

**Table 3.** Results of paired *t*-test on scores of the version with scientific issues under broad vs. targeted prompts.

| Pair | Prompt Type | Mean | N | SD | t | *p* | r | Cohen's d |
|---|---|---|---|---|---|---|---|---|
| Pair 1 (DeepSeek) | Broad | 88.48 | 40 | 2.526 | 10.243 | 0.000 | 0.520 | 1.620 |
| | Targeted | 82.78 | 40 | 4.098 | | | | |
| Pair 2 (Doubao) | Broad | 89.95 | 40 | 1.884 | 4.083 | 0.000 | 0.453 | 0.646 |
| | Targeted | 88.23 | 40 | 2.931 | | | | |
| Pair 3 (GPT-4) | Broad | 88.94 | 40 | 1.922 | 1.401 | 0.169 | 0.479 | 0.221 |
| | Targeted | 88.20 | 40 | 3.791 | | | | |

Regarding scientific flaw detection rates (including partial detection), targeted prompts yielded substantial improvements: 500% for DeepSeek, 200% for Doubao, and 200% for GPT-4.

#### 4.2.3. Comparative Analysis of Flaw Detection: LLMs vs. Human Evaluators

Analysis of optimal performance across all trials (including targeted prompt conditions) revealed that human evaluators substantially outperformed all LLMs in detecting scientific flaws, categorized as complete detection, partial detection, or non-detection (Figure 1). Kappa consistency analysis yielded values of 0.186 ($p$ = 0.022) for DeepSeek, 0.120 ($p$ = 0.090) for Doubao, and 0.117 ($p$ = 0.057) for GPT-4 when compared with human evaluators, indicating only DeepSeek achieved statistically significant, albeit slight, agreement with human detection patterns.

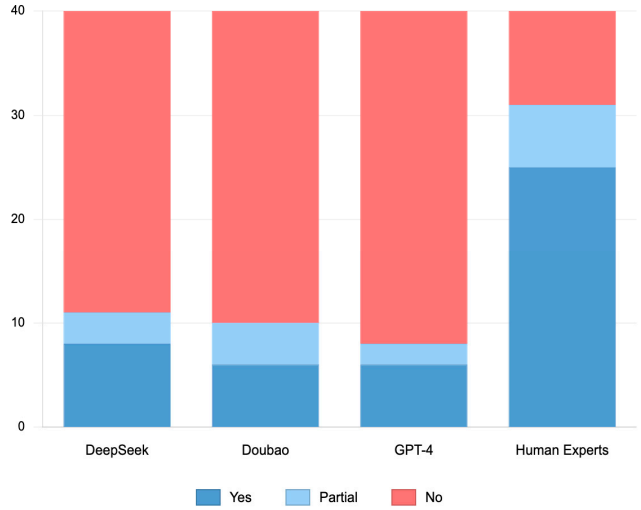



**Figure 1.** Results of Scientific Flaw Detection by Various Evaluators.

#### 4.2.4. Heterogeneity in LLM Detection of Scientific Flaws

The type of scientific flaws significantly influenced detection rates (Figure 2). Kruskal–Wallis test revealed significant differences across the four flaws categories (H(3) = 11.490, $p$ = 0.009). Post hoc analysis indicated that detection rates for "methodological inadequacies" were significantly higher than those for "unsound arguments and exposition" ($p$ = 0.001).

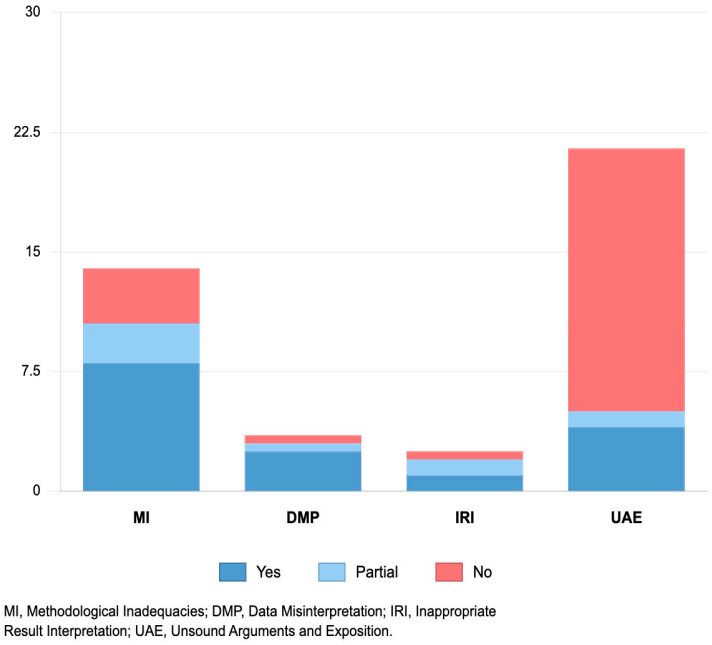



**Figure 2.** Frequency of Different Scientific Flaw Type Detection by LLMs.

Qualitative analysis of LLM commentary revealed competent evaluation of empirical methodology and data interpretation. LLMs accurately identified methodological issues such as uncontrolled confounding variables in model construction and common method bias risks in survey research, while recommending enhanced robustness checks for observational studies and corrections for sample self-selection bias. They precisely recognized borderline statistical significance (e.g., $p = 0.06$). However, LLMs demonstrated limited capacity for context-specific methodological evaluation. For instance, all three models failed to detect the omission of city-level generalized difference-in-differences robustness tests in the scientifically flawed version of "Effects of 'Double Reduction' Policy on Household Extracurricular Tutoring Participation and Expenditure." This oversight likely occurred because the manuscript already included PSM-DID robustness checks and placebo tests. Yet the specific research context, characterized by heterogeneous policy implementation intensity across cities, necessitated additional city-level generalized DID robustness testing—a judgment requiring deep contextual understanding beyond LLM capabilities.

Analysis further revealed substantial deficiencies in LLMs' capacity to evaluate argumentative soundness. Among the four successful detections, LLMs either flagged viewpoints as controversial (e.g., "learning modes: from active knowledge construction to passive comprehension")—actually reflecting inconsistency with training data—or identified internal textual contradictions—essentially exercising text parsing rather than critical evaluation. While some researchers attribute these limitations to inadequate assessment capabilities for complex, frontier topics and niche disciplines, the issue may be more fundamental: LLMs may lack genuine epistemic capabilities, rendering them unable to evaluate validity beyond atomic facts and consensual viewpoints.

Two supplementary tests explored this hypothesis. First, when presented with isolated viewpoints containing obvious errors alongside original versions, LLMs identified errors in modified versions but also critiqued original viewpoints, providing detailed analyses of simplifications and ambiguities. This hypercritical, microscopic analysis, while appearing comprehensive, proves impractical for manuscript evaluation—exhaustive scrutiny of every claim would preclude holistic assessment, as individual arguments exist within broader contextual frameworks permitting reasonable simplification and authorial perspective. Second, when presented with contextually embedded viewpoints without obvious errors,

LLMs favored modified over original versions. This preference stemmed not from superior validity but from greater semantic proximity between modified viewpoints and extracted context, whereas original viewpoints embodied deeper critical engagement with empirical reality. This pattern reveals LLMs' reliance on contextual coherence rather than experiential reflection in evaluating argumentative validity.

### *4.3. Detection and Evaluation of Logical Flaws*

#### 4.3.1. Score Differences Between Versions Under Broad Prompts

Paired-samples *t*-tests examining scores under broad prompting conditions demonstrated that all three LLMs showed no significant differences between original manuscripts (mean of two evaluations, hereafter) and their logically flawed version. Conversely, human evaluators exhibited significant differences, with original manuscripts receiving significantly higher mean scores than their logically flawed versions, demonstrating very large effect sizes both statistically and practically (Table 4).

**Table 4.** Results of paired *t*-test on scores between the original version and the version with logical Issues.

| Pair | Version | Mean | N | SD | t | *p* | r | Cohen's d |
|---|---|---|---|---|---|---|---|---|
| Pair 1 (DeepSeek) | Original | 88.85 | 40 | 2.551 | 1.491 | 0.144 | 0.433 | 0.236 |
| | Logical flaws | 88.14 | 40 | 3.066 | | | | |
| Pair 2 (Doubao) | Original | 88.89 | 40 | 1.866 | −1.032 | 0.309 | 0.563 | 0.163 |
| | Logical flaws | 90.15 | 40 | 1.511 | | | | |
| Pair 3 (GPT-4) | Original | 88.36 | 40 | 2.279 | −0.373 | 0.712 | 0.113 | 0.059 |
| | Logical flaws | 88.54 | 40 | 2.182 | | | | |
| Pair 4 (Human Experts) | Original | 89.83 | 40 | 2.427 | 9.896 | 0.000 | 0.345 | 1.565 |
| | Logical flaws | 83.83 | 40 | 3.922 | | | | |

#### 4.3.2. Score Differences and Detection Rates: Broad vs. Targeted Prompts

Paired-samples *t*-tests comparing scores for logically flawed version under broad versus targeted prompting conditions revealed divergent patterns. DeepSeek and GPT-4 showed no significant score reductions following targeted prompting. Conversely, Doubao demonstrated significant score increases under targeted prompts, with mean scores rising by 1.48 units (95% CI [−1.97, −0.98]) and exhibiting a large effect size (Table 5).

**Table 5.** Results of paired *t*-test on scores of the version with logical issues under broad vs. targeted prompts.

| Pair | Version | Mean | N | SD | t | *p* | r | Cohen's d |
|---|---|---|---|---|---|---|---|---|
| Pair 1 (DeepSeek) | Broad | 88.14 | 40 | 3.066 | 1.136 | 0.263 | 0.601 | 0.180 |
| | Targeted | 87.45 | 40 | 4.782 | | | | |
| Pair 2 (Doubao) | Broad | 90.15 | 40 | 1.511 | −6.042 | 0.000 | 0.594 | 0.955 |
| | Targeted | 91.63 | 40 | 1.849 | | | | |
| Pair 3 (GPT-4) | Broad | 88.54 | 40 | 2.182 | 0.856 | 0.397 | 0.296 | 0.135 |
| | Targeted | 88.00 | 40 | 4.026 | | | | |

Regarding logical flaw detection rates, targeted prompting yielded marked improvements: DeepSeek and GPT-4 progressed from zero detection to 30% and 15% detection rates, respectively, while Doubao's detection rate increased by 33%.

#### 4.3.3. Comparative Analysis of Logical Flaw Detection: LLMs vs. Human Evaluators

Analysis of optimal performance across all trials (including targeted prompt conditions) demonstrated human evaluators' substantial superiority in detecting logical flaws, categorized as complete detection, partial detection, or non-detection (Figure 3). Kappa consistency analysis yielded values of 0.016 ($p = 0.559$) for DeepSeek, 0.009 ($p = 0.714$) for

Doubao, and 0.008 ($p$ = 0.737) for GPT-4 when compared with human evaluators, indicating no statistically significant agreement between LLMs and human evaluators in logical flaw detection.

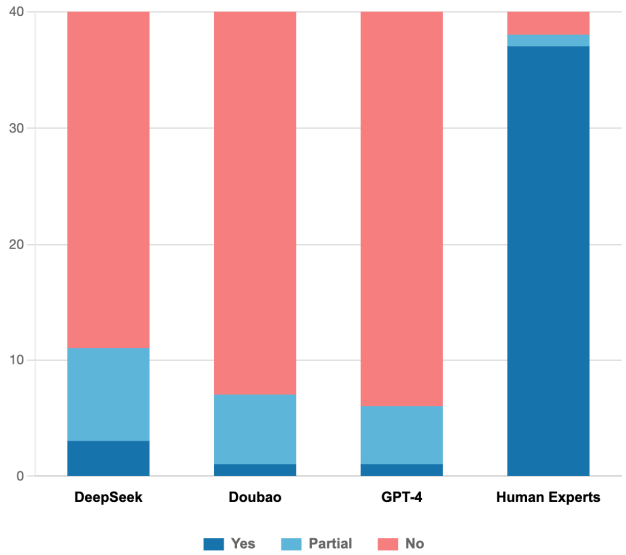



**Figure 3.** Results of Logical Flaw Detection by Various Evaluators.

#### 4.3.4. Heterogeneity in LLM Detection of Logical Flaws

Types of logical flaws showed no significant impact on detection rates (Figure 4), with Mann–Whitney U test revealing no significant differences between groups (U = 131.5, $p$ = 0.577).

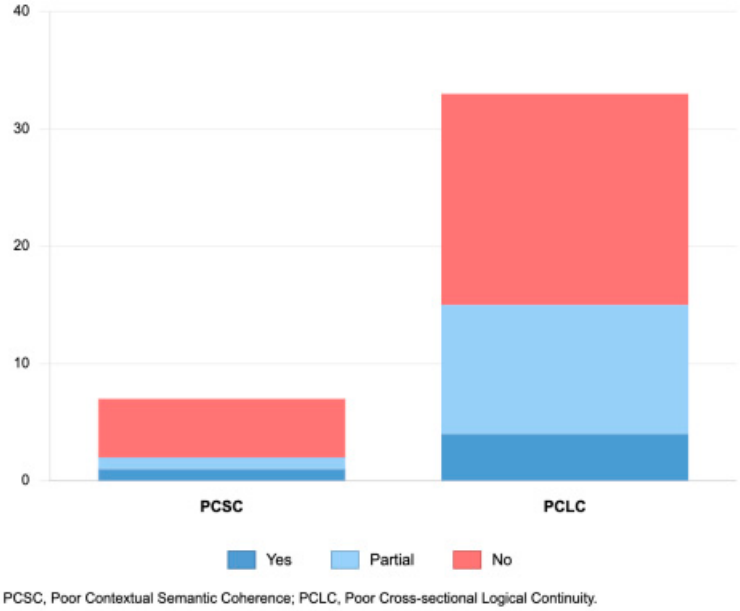



**Figure 4.** Frequency of Different Logical Flaw Type Detection by LLMs.

Qualitative analysis of LLM commentary revealed partial but unstable capacity to detect global logical discontinuities, typically requiring targeted prompting. Consider a study on the "Double Reduction" policy's impact on household tutoring expenditure: the logically flawed version replaced all original policy recommendations with three implementation strategies disconnected from empirical findings. Under broad prompts, all three models provided positive assessments of logical coherence, exemplified by: "From policy background ('Double Reduction' objectives) → literature gap (lacking categorical assessment and causal inference) → research design (DID model + heterogeneity analysis) → conclusions (policy

effects and group differences) → recommendations (holistic governance framework), the logical chain remains complete." Only under targeted prompting did GPT-4 identify the discontinuity—"While conclusions and policy recommendations propose holistic governance approaches, specific implementation pathways insufficiently integrate with prior data analysis, lacking continuous causal chain demonstration"; DeepSeek partially detected flaws—"Recommendations could more tightly integrate heterogeneity findings, such as enhanced regulatory measures for high-income households"; while Doubao maintained its positive assessment.

However, even with targeted prompts, all three models failed to detect numerous instances of conclusion-recommendation disconnection. This likely stems from high semantic complexity and similarity between conclusions and recommendations, extended inferential chains from findings to implications, and implicit value orientations. For instance, in a study on internet dependency's well-being effects under digital economy development, one finding highlighted urban-rural heterogeneity—"Urban residents with higher education levels tend to utilize internet for information capital accumulation, thereby enhancing well-being." The original version proposed "enhancing residents' digital capabilities and literacy," while the logically flawed version exclusively recommended network infrastructure development, entirely failing to address the finding. None of the models detected this logical rupture.

When logical discontinuities were localized rather than global—such as single recommendation points disconnected from findings, or individual measures misaligned with identified problems—LLM detection became even more challenging. For example, in "Three Forms of Modern Family Education," the original version proposed and elaborated "protecting life-form family education" (secondary heading), corresponding to the prior section "life-form family education shows marked contraction in modern society" (tertiary heading). The modified version instead proposed "guiding life-form family education," elaborating through "active intervention in family life ethical education" and "strengthening dynamic supervision of inappropriate educational actions"—clearly disconnected from the preceding context. All three models failed to detect this discontinuity, likely due to its localized nature (point-to-point disconnection) and the paper's lack of structured empirical research templates.

Comparatively, LLMs demonstrated greater sensitivity to semantic coherence between passages. Some logical continuity issues were detected through semantic association rather than argumentative logic. For instance, in a modified version of "Understanding Zhuangzi's 'Joy of Fish' from Jin Yuelin's Perspective," deletion of passages discussing anthropocentric interpretation difficulties and epistemological deficiencies created logical rupture. Doubao identified abrupt transitions between "anthropocentric view" and "authentic consciousness view" sections, recommending transitional sentences—detecting flaws through semantic association rather than argumentative logic. However, LLM detection of semantic coherence issues remained unstable. When semantic disconnection occurred within narrow scopes, such as between claims and supporting arguments, LLMs struggled with detection. For example, in a modified version of "Gains and Losses of Sociological Field Research Characterized by 'Meaning Inquiry,'" where supporting examples for two sub-arguments in section four were swapped, all three models failed to capture the misalignment between claims and evidence.

## 5. Discussion

### *5.1. Reliability of Scoring and Prompt Optimization*

This study revealed notably low internal consistency in the scores assigned by the models across multiple evaluations. This issue stems fundamentally from the inherent

stochasticity of the LLM sampling mechanism [25]. Consequently, enhancing the stability of LLM-based paper evaluation constitutes a complex systemic challenge, for which prompt optimization is a crucial measure. Our findings indicate that low scoring consistency is associated with insufficiently specific prompts. For instance, when evaluating a theoretical paper, an LLM might alternately apply paradigms suited for empirical research or theoretical study, leading to significant score discrepancies [27]. Furthermore, even when prompts define evaluation dimensions and score distributions, LLMs tend to refine these criteria randomly [6]. When these refined criteria include indicators of "deficiency" or factors unfavorable to the paper, the result is an undeservedly lower score. Therefore, providing LLMs with targeted, detailed evaluation standards, along with exemplars from specific academic domains, is anticipated to mitigate scoring inconsistency to a certain extent.

Research also demonstrates that targeted prompts moderately improved detection rates, reflecting LLMs' evaluation mechanisms. For lengthy texts, LLMs initially extract key sections (abstracts, headings) based on evaluation dimensions, then conduct semantic-level interpretation and assessment [26]. Only under specific guidance do they perform detailed analysis of individual sections from particular perspectives. When prompted to attend to methodological rigor, LLMs carefully examine specific methodological applications. When directed toward logical continuity, they comprehensively examine inter-sectional connections [13]. However, targeted prompts may induce hypercritical evaluation, flagging issues that human readers would not consider problematic.

### *5.2. Deficiencies in Evaluating Scientific Rigor and Logical Coherence*

The study found that the three large models (LLMs) did not exhibit a significant difference in their scores between the original papers and the versions containing scientific or logical flaws. Furthermore, the proportion of comments in which the LLMs accurately identified these errors was substantially lower than that of human scholars. Heterogeneity analysis and qualitative examination revealed that the detection of scientific errors was highly dependent on the type of error. For flaws that could be checked against explicit rules, such as violations of empirical research methodology, the LLMs were capable of accurate identification. However, for problems that transcended clear rules and required concrete, profound empirical insight, the LLMs often failed to make correct judgments. In assessing logical coherence, the LLMs primarily relied on structural features (e.g., section headings) and superficial semantic cues rather than the deep logical argumentation of the paper. Papers in the HSS heavily employ fact-based logical reasoning—such as inductive, abductive, and practical reasoning—which premises on objective facts, universal principles, or accepted common knowledge to derive conclusions through rigorous logical relationships, emphasizing factual evidence and causal linkages over subjective conjecture or formal symbolic manipulation. The LLMs' lack of embodied cognition, and consequently a unified world schema grounded in such cognition [38], underlies their incompetence in evaluating the logic of HSS papers.

The findings of this study exhibit a complementary and corroborative relationship with the conclusions of two other studies that employed the method of deliberately implanted errors. Voutsa et al. found that LLMs performed well in detecting "overclaiming," "citation issues," and "theoretical errors" but showed lower detection rates for "technical errors" and "ethical problems." This discrepancy may stem from the former having more distinct linguistic cues or referenceable standard norms, while the latter relies more heavily on contextual understanding. It should be noted that in their study, the errors implanted into the papers were primarily generated by LLMs themselves (except for ethical issues, which were manually added). This approach may have introduced a bias favoring LLM performance. Dennstädt et al. reported that LLMs identified deliberately implanted

errors in 7 out of 13 test papers, covering both mathematical and conceptual errors [39]. Importantly, the test papers used were artificially constructed, short computer science papers, and three different prompting strategies were employed: Direct Prompting, One-Shot Prompting, and Part-Based Prompting. These experimental settings could influence the difficulty of error detection.

In summary, this study focuses on the sensitivity and critical ability of LLMs regarding scientific and logical issues in HSS papers, which differs from detecting scientific-technical errors or formal reasoning flaws in natural science or engineering papers. Nevertheless, the conclusions of these two types of studies can still be mutually verifying and complementary. Furthermore, this study emphasizes probing the lower bound of LLMs' HSS paper evaluation capability. This aims to reveal potential risks and establish a baseline for improving their evaluation performance through techniques like prompt engineering and fine-tuning. Consequently, the testing intentionally avoided providing excessive guidance to the LLMs.

### *5.3. The Distinctive Evaluation Mechanism of LLMs for Scholarly Papers*

An examination of LLMs' practical performance reveals that they often automatically adopt a "task decomposition-semantic understanding" process when evaluating lengthy papers. Specifically, based on the prompts and the paper's key features, LLMs first decompose the overall task into smaller, potentially multi-layered evaluation subtasks. Each subtask encompasses specific evaluation angles, criteria, and particular sections or aspects of the paper. Subsequently, the LLM extracts limited information from the paper relevant to each subtask for relatively in-depth semantic understanding and comparison. However, the computational resources allocated to each subtask are minimal, as is the amount of textual information processed. Consequently, the scope and depth of the subsequent semantic understanding and comparison are severely constrained. Compounded by the inherent limitations of LLMs, such as their lack of embodied cognition and real-world experience [40], this leads to evaluations from each subtask that lack critical depth and comprehensiveness, particularly when the prompts are broad. This mechanistic limitation likely constitutes the fundamental reason for the low detection rates of issues related to argument plausibility and logical coherence observed in HSS paper evaluation.

Based on schema theory [41], a competent evaluator does not directly proceed by filling out a detailed evaluation form item by item. Instead, they must first undergo an internal psychological process of "worldscape reconstruction → meaning construction and critique." First, using the language of the paper as cues, the human evaluator "reconstructs" the worldscape depicted within it (including the methods of its construction). This process of "reconstruction" is grounded in the evaluator's own knowledge and experience and incorporates their reflection and creativity. The evaluator keenly discerns which parts of this worldscape are crucial and which might be problematic. Subsequent meaning construction and critique are then conducted upon this reconstructed worldscape: judging the validity of viewpoints and their exposition becomes a process of mutual constitution between the whole and its parts, with this worldscape as the backdrop; assessing logical coherence equates to examining the structural soundness of the worldscape itself. Furthermore, the acute awareness of key and potentially problematic parts formed during "reconstruction" makes the evaluation more targeted and productive.

Therefore, to mitigate the risks of over-trusting and over-relying on LLMs, it is necessary to develop human-AI collaborative evaluation [42]. Since the strength of human evaluators lies in deep textual understanding and critiquing through reflection from within the text, and the strength of LLMs lies in their vast knowledge, rich data, rapid retrieval, and comparison capabilities—excelling at objective analysis and comparison from outside

the text—combining their advantages can balance evaluation efficiency with quality, and breadth with depth, particularly in the complex domain of HSS scholarship.

### 5.4. Limitations

The conclusions of this study are derived from evaluation tests conducted by LLMs on academic papers (in Chinese) in the HSS, and thus cannot be directly generalized to their ability to evaluate papers in other disciplines. The prompts used in the study, including both broad and targeted prompts, were relatively simple. They neither provided specific evaluation criteria for particular types of papers nor included exemplars of paper evaluations. Consequently, the results reflect the lower bound of LLMs' evaluation capability rather than their best possible performance. Furthermore, the proposed explanation for the identified deficiencies—the "task decomposition-semantic understanding" evaluation mechanism—is primarily inferred from limited cases and external performance observations, and thus lacks sufficient empirical evidence.

## 6. Conclusions

In summary, this study exposes fundamental deficiencies in LLMs for paper evaluation, including low internal consistency in scoring and low detection rates for flaws in argument plausibility and logical coherence. It confirms a significant gap between LLM capabilities and human evaluators when assessing academic papers (in Chinese) within the HSS. These findings highlight the substantial risks associated with over-trusting and over-relying on LLMs—the warning that "automatically generating reviews without thoroughly reading the manuscripts will undermine the rigorous assessment process fundamental to scientific progress" is not an exaggeration [43].

To mitigate these risks, a two-pronged approach is necessary. Firstly, it is essential to fully recognize the limitations and the unique operational mechanism of LLMs in paper evaluation, and on this basis, establish rational human-AI collaborative evaluation frameworks. Secondly, efforts should focus on enhancing the LLMs' adaptation to the paper evaluation task through downstream alignment techniques such as prompt engineering, supervised fine-tuning, and reinforcement learning. Concurrently, improving the general capabilities of LLMs is crucial. This could involve integrating reasoning models to shift LLMs from statistical pattern matching towards an intelligent review paradigm closer to human critical thinking [44]. Alternatively, enhancing logical reasoning capabilities, particularly fact-based reasoning, might be achieved by incorporating reasoning tasks (e.g., chain-of-thought prompting, counterfactual generation) during pre-training, or by explicitly separating factual memory from reasoning abilities within the model (e.g., through knowledge neuron localization).

Regarding future research on LLMs' paper evaluation capability, investigating their upper limit for evaluating HSS papers could be pursued through complex prompt engineering, constructing paper evaluation agents (multi-agent systems), and model fine-tuning on high-quality paper evaluation datasets. Furthermore, to validate the proposed "task decomposition-semantic understanding" evaluation mechanism, future work could employ methods such as visualizing attention mechanisms or inserting classifiers at intermediate model layers. Another promising direction involves integrating graph-based reasoning tools, such as Graph Neural Networks (GNNs), to guide LLMs towards incorporating a human-like evaluation process of "worldscape reconstruction $\rightarrow$ meaning construction and critique".

**Author Contributions:** Conceptualization, H.L. and H.J.; methodology, H.L.; software, L.D.; validation, H.L., L.D. and H.J.; formal analysis, H.L.; investigation, H.L. and L.D.; resources, H.J.; data curation, L.D.; writing—original draft preparation, H.L.; writing—review and editing, L.D. and H.J.; visualization, L.D.; supervision, H.J.; project administration, H.J.; funding acquisition, H.L. and H.J. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Fundamental Research Funds for the Central Universities, grant number YBNLTS2025-036, and the Jiangsu Province Education Science Planning Key Project, grant number B/2-23/01/122.

**Institutional Review Board Statement:** Not applicable. This study did not involve humans or animals as research subjects. The research utilized only publicly available, previously published academic papers from peer-reviewed journals as test materials. The doctoral candidates who participated as evaluators served in their capacity as subject matter experts providing professional assessments, rather than as research subjects themselves. Their evaluations constituted expert judgment similar to standard academic peer review processes, which do not require ethical approval. Therefore, institutional review board approval was not required for this study.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data are not publicly available due to privacy/ethical restrictions, but are available from the corresponding author on reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AES | Automated Essay Scoring |
| AI | Artificial Intelligence |
| CSSCI | Chinese Social Sciences Citation Index |
| DID | Difference-in-Differences |
| HSS | Humanities and Social Sciences |
| GPT | Generative Pre-trained Transformer |
| ICC | Intraclass Correlation Coefficient |
| ICLR | International Conference on Learning Representation |
| LLM | Large Language Model |
| NLI | Natural Language Inference |
| PSM-DID | Propensity Score Matching Difference-in-Differences |
| STEM | Science, Technology, Engineering, and Mathematics |

# Appendix A

**Table A1.** Types and Examples of Flaws in the Test Data.

| Version | Flaw Type | Quantity | Example |
| --- | --- | --- | --- |
| Scientific Issues Version | Methodological Inadequacies | 14 | Deleted the robustness test using the city-level generalized difference-in-differences (DID) model (from "Effects of the 'Double Reduction' Policy on Household Extracurricular Tutoring Participation and Expenditure"). |
| | Data Misinterpretation | 3 | Changed the diminishing marginal effect trend derived from the data to a trend of no diminishing marginal effect (from "The 'Audience Effect' in Social Surveys: The Impact of Bystanders on the Response Performance of the Elderly"). |
| | Inappropriate Result Interpretation | 2 | Interpreted the finding in the discussion section—that the eigenvalues for positive emotion words were higher than for negative emotion words—as a positive signal (from "A Psycholinguistic Analysis of Adolescent Suicide Notes: Based on the Interpersonal Theory of Suicide"). |
| | Unsound Arguments and Exposition | 21 | In the second point, "Learning Transformation to Promote Deep Understanding in the Digital Intelligence Era," reversed the third sub-point to its opposite: "Learning Modes: From Active Knowledge Construction to Passive Comprehension," and provided corresponding arguments (from "Deep Learning in the Digital Intelligence Era: From Rote Memorization to Deep Understanding"). |
| Logical Issues Version | Poor Contextual Semantic Coherence | 7 | Swapped the case studies for the two sub-arguments in the fourth section (Argument 1: The ability to make sociological abstractions and generalizations from the interviewee's narrative to form concepts and judgments; Argument 2: The accumulation of common sense and understanding of social customs also helps to quickly grasp the sociological significance of the text) (from "Gains and Losses of Sociological Field Research Characterized by 'Meaning Inquiry'"). |
| | Poor Cross-sectional Logical Continuity | 33 | Replaced the entire recommendations section with "measures to strengthen internet infrastructure," which was unrelated to the research findings (from "The Well-being Effect and Mechanism of Residents' Internet Dependency in the Context of the Digital Economy Development"). |

## Appendix B

**Table A2.** Broad Prompts and Targeted Prompts for LLMs.

| Flaw Type | Broad Prompt | Targeted Prompt |
| --- | --- | --- |
| Methodological Inadequacies | Please evaluate the following paper from three aspects: scientific content and intellectual depth, formal compliance, and innovation and practicality. Provide a score (out of 100, with 50 for scientific content/intellectual depth, 30 for formal compliance, and 20 for innovation/practicality). | Please re-evaluate, paying special attention to methodological rigor. |
| Data Misinterpretation | | Please re-evaluate, paying special attention to the scientific validity of the data interpretation. |
| Inappropriate Result Interpretation | | Please re-evaluate, paying special attention to the reasonableness of the interpretation. |
| Unsound Arguments and Exposition | | Please re-evaluate, paying special attention to the soundness of the paper's arguments and their exposition. |
| Poor Contextual Semantic Coherence | | Please re-evaluate, paying special attention to the semantic coherence between different parts of the text. |
| Poor Cross-sectional Logical Continuity | | Please re-evaluate, paying special attention to the logical continuity between different sections of the paper. |

## Appendix C

**Table A3.** Criteria for Flaw Detection and Judgment with Application Examples.

| Flaw Type | Yes (Accurately identifies and analyzes the issue, or provides targeted suggestions for improvement) | Partial (Partially addresses the issue, or identifies the location of the issue but with imprecise analysis) | No (Fails to identify the issue, or even provides a positive evaluation) |
| --- | --- | --- | --- |
| Methodological Inadequacies | "...supplementing with quantile combination and clustering combination methods for robustness is good, but the comparative discussion of the results is slightly brief. It could be further detailed to explain the robustness and differences of the conclusions under each method." —GPT-4, Targeted Prompt (Flaw source: Three robustness tests were deleted.—From "What is the Impact of Effort on Occupational Status Attainment?...") | "Control variables: including age, gender, marriage, health, education, etc., basically cover the main influencing factors, but potential variables such as family income and occupation type were not considered. The variable selection is reasonable, but the explanatory power and robustness of some indicators could be further optimized." —DeepSeek, Targeted Prompt (Flaw source: The 'family economic status' control variable was deleted.—From "The Well-being Effect and Mechanism of Residents' Internet Dependency...") | "The research design is rigorous: Based on data from the 2018 and 2020 China Family Panel Studies (CFPS), ... it effectively controls for endogeneity issues... and verifies the reliability of the conclusions through robustness tests." —Doubao, Broad Prompt (Flaw source: The 'family economic status' control variable was deleted.—From "The Well-being Effect and Mechanism of Residents' Internet Dependency...") |

**Table A3.** *Cont.*

| | | | |
|---|---|---|---|
| Data Misinterpretation | "...but the significance boundary of some data is slightly ambiguous (e.g., $p = 0.060$ is marginal), and the critical analysis of ChatGPT's limitations could be deeper." —GPT-4, Broad Prompt (Flaw source: The *p*-value was changed to 0.060, making the result non-significant.—From "Empowering Undergraduate Research with Generative AI...") | (No corresponding case) | "The research design is reasonable, the data source is reliable, the sample size is sufficient (7139 people), and robustness and heterogeneity analyses were conducted, enhancing the credibility of the conclusions." —DeepSeek, Broad Prompt (Flaw source: The diminishing marginal effect trend derived from the data was entirely changed to a trend of no diminishing marginal effect.—From "The 'Audience Effect' in Social Surveys...") |
| Inappropriate Result Interpretation | (No corresponding case) | "Some suggestions, such as expanding the scope of technology application, lack specific implementation paths, and their practicality needs further improvement." —Doubao, Broad Prompt (Flaw source: The first point in the research outlook was changed from "focus on experimental design, validity verification..." to "focus on experimental promotion, rather than experimental design...".—From "Technology for Good: Cultivating Adolescent Psychological Resilience...") | (Fails to identify the issue, omitted) |
| Unsound Arguments and Exposition | "Classifying the 'transition from lifestyle consumption to ritualistic consumption' in 'weak refinement' as a manifestation of 'consumerism fatigue' seems slightly contradictory to the logic of 'escaping from meaninglessness' in the core concept." —Doubao, Targeted Prompt (Flaw source: The second sub-point... was changed to the opposite view...—From "Escaping 'Meaninglessness': ... 'Consumerism Fatigue'...") | "The argument for the 'male-dominated' characteristic is relatively brief and lacks comparative data with female emotional practices, which weakens the universality of some conclusions." —Doubao, Targeted Prompt (Flaw source: The subheading 'female-dominated' and its discussion were changed to 'male-dominated'...—From "Parent–Child Integration...") | (Fails to identify the issue, omitted) |
| Poor Contextual Semantic Coherence | "When discussing the 'associative thinking and literary theory' in the third part, the discussion of the ideographic nature of Chinese characters is abruptly inserted..." —DeepSeek, Targeted Prompt (Flaw source: The epilogue was completely changed to a discussion on the ideographic nature of Chinese characters, departing from the theme.—From "On the Associative Thinking in Chinese Aesthetics") | "For example, regarding the progression of 'benevolence → righteousness → propriety,' although it is presented in the text, its theoretical development and its connection to resolving the aforementioned conflict between hierarchy and equality need to be further supplemented." —GPT-4, Targeted Prompt (Flaw source: The second of the two basic justice principles...—the principle of appropriateness—was deleted...—From "The Tension Between Confucius's Concepts of Hierarchy and Equality...") | "The paper's structure is rigorous, forming a tightly linked logical loop from... 'Epilogue'... For instance, the 'shaping forms according to categories' method of character creation in the 'Epilogue' echoes the 'observing things...' method from the I Ching mentioned earlier..." —Doubao, Targeted Prompt (Flaw source: The epilogue was completely changed to a discussion on the ideographic nature of Chinese characters...—From "On the Associative Thinking in Chinese Aesthetics") |

**Table A3.** *Cont.*

| | | | |
|---|---|---|---|
| Poor Cross-sectional Logical Continuity | "Although the conclusion and policy recommendations propose a holistic governance approach, the specific implementation paths are not closely integrated with the data analysis results from the preceding sections, lacking a continuous causal chain demonstration." —GPT-4, Targeted Prompt (Flaw source: The specific content of the policy recommendations has no intrinsic logical connection to the preceding research conclusions.—From "Effects of the 'Double Reduction' Policy…") | "The recommendation section could be more tightly integrated with the heterogeneity findings, such as enhanced regulatory measures for high-income households." —DeepSeek, Targeted Prompt (Flaw source: The specific content of the policy recommendations has no intrinsic logical connection…—From "Effects of the 'Double Reduction' Policy…") | "The research conclusions provide a direct basis for policy optimization… However, the recommendation section could be more specific, for example, by proposing quantitative indicators…" —DeepSeek, Broad Prompt (Flaw source: The specific content of the policy recommendations has no intrinsic logical connection…—From "Effects of the 'Double Reduction' Policy…") |

## References


1. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Xie, X. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45. [CrossRef]
2. Checco, A.; Bracciale, L.; Loreti, P.; Pinfield, S.; Bianchi, G. AI-assisted peer review. *Humanit. Soc. Sci. Commun.* **2021**, *8*, 1–11. [CrossRef]
3. Bouziane, K.; Bouziane, A. AI versus human effectiveness in essay evaluation. *Discov. Educ.* **2024**, *3*, 201. [CrossRef]
4. Steiss, J.; Tate, T.; Graham, S.; Cruz, J.; Hebert, M.; Wang, J.; Moon, Y.; Tseng, W.; Warschauer, M.; Olson, C.B. Comparing the quality of human and ChatGPT feedback of students' writing. *Learn. Instr.* **2024**, *91*, 101894. [CrossRef]
5. Biswas, S.S. AI-assisted academia: Navigating the nuances of peer review with ChatGPT 4. *J. Pediatr. Pharmacol. Ther.* **2024**, *29*, 441–445. [CrossRef]
6. Emirtekin, E. Large Language Model-Powered Automated Assessment: A Systematic Review. *Appl. Sci.* **2025**, *15*, 5683. [CrossRef]
7. Teubner, T.; Flath, C.M.; Weinhardt, C.; Van Der Aalst, W.; Hinz, O. Welcome to the era of chatgpt et al. the prospects of large language models. *Bus. Inf. Syst. Eng.* **2023**, *65*, 95–101. [CrossRef]
8. Hosseini, M.; Horbach, S.P.J.M. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Res. Integr. Peer Rev.* **2023**, *8*, 4. [CrossRef]
9. Park, S.H. Use of generative artificial intelligence, including large language models such as ChatGPT, in scientific publications: Policies of KJR and prominent authorities. *Korean J. Radiol.* **2023**, *24*, 715–718. [CrossRef] [PubMed]
10. Grippaudo, F.R.; Jeri, M.; Pezzella, M.; Orlando, M.; Ribuffo, D. Assessing the informational value of large language models responses in aesthetic surgery: A comparative analysis with expert opinions. *Aesthetic Plast. Surg.* **2025**, *49*, 1–7.
11. Hancock, P.A. Avoiding adverse autonomous agent actions. *Hum. Comput. Interact.* **2022**, *37*, 211–236. [CrossRef]
12. Kocak, B.; Onur, M.R.; Park, S.H.; Baltzer, P.; Dietzel, M. Ensuring peer review integrity in the era of large language models: A critical stocktaking of challenges, red flags, and recommendations. *Eur. J. Radiol. Artif. Intell.* **2025**, *2*, 100018. [CrossRef]
13. Qureshi, R.; Shaughnessy, D.; Gill, K.A.; Robinson, K.A.; Li, T.; Agai, E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Syst. Rev.* **2023**, *12*, 72. [CrossRef]
14. Saad, A.; Jenko, N.; Ariyaratne, S.; Birch, N.; Iyengar, K.P.; Davies, A.M.; Vaishya, R.; Botchu, R. Exploring the potential of ChatGPT in the peer review process: An observational study. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2024**, *18*, 102946. [CrossRef]
15. Mishra, T.; Sutanto, E.; Rossanti, R.; Pant, N.; Ashraf, A.; Raut, A.; Zeeshan, B. Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers. *Sci. Rep.* **2024**, *14*, 31672. [CrossRef]
16. Drori, I.; Te'eni, D. Human-in-the-Loop AI reviewing: Feasibility, opportunities, and risks. *J. Assoc. Inf. Syst.* **2024**, *25*, 98–109. [CrossRef]
17. Gawlik-Kobylińska, M. Harnessing Artificial Intelligence for Enhanced Scientific Collaboration: Insights from Students and Educational Implications. *Educ. Sci.* **2024**, *14*, 1132. [CrossRef]
18. Yavuz, F.; Çelik, Ö.; Yavaş Çelik, G. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *Br. J. Educ. Technol.* **2025**, *56*, 150–166. [CrossRef]
19. Ebadi, S.; Nejadghanbar, H.; Salman, A.R.; Khosravi, H. Exploring the impact of generative AI on peer review: Insights from journal reviewers. *J. Acad. Ethics* **2025**, *23*, 1–15. [CrossRef]
20. Zhong, J.; Xing, Y.; Hu, Y.; Lu, J.; Yang, J.; Zhang, G.; Yao, W. The policies on the use of large language models in radiological journals are lacking: A meta-research study. *Insights Into Imaging* **2024**, *15*, 186. [CrossRef]
21. Wong, R. Role of generative artificial intelligence in publishing. What is acceptable, what is not. *J. ExtraCorporeal Technol.* **2023**, *55*, 103–104. [CrossRef]
22. Elazar, Y.; Kassner, N.; Ravfogel, S.; Ravichander, A.; Hovy, E.; Schütze, H.; Goldberg, Y. Measuring and improving consistency in pretrained language models. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 1012–1031. [CrossRef]
23. Biswas, S.; Dobaria, D.; Cohen, H.L. ChatGPT and the future of journal reviews: A feasibility study. *Yale J. Biol. Med.* **2023**, *96*, 415. [CrossRef]
24. Thelwall, M. Can ChatGPT evaluate research quality? *J. Data Inf. Sci.* **2024**, *9*, 1–21. [CrossRef]
25. Rebuffel, C.; Roberti, M.; Soulier, L.; Scoutheeten, G.; Cancelliere, R.; Gallinari, P. Controlling hallucinations at word level in data-to-text generation. *Data Min. Knowl. Discov.* **2022**, *36*, 318–354. [CrossRef]
26. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Liu, T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–55. [CrossRef]
27. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
28. Cheng, K.; Sun, Z.; Liu, X.; Wu, H.; Li, C. Generative artificial intelligence is infiltrating peer review process. *Crit. Care* **2024**, *28*, 149. [CrossRef]
29. Havlík, V. Meaning and understanding in large language models. *Synthese* **2024**, *205*, 9. [CrossRef]

30. Spreitzer, C.; Straser, O.; Zehetmeier, S.; Maaß, K. Mathematical modelling abilities of artificial intelligence tools: The case of ChatGPT. *Educ. Sci.* **2024**, *14*, 698. [CrossRef]
31. Beaulieu-Jones, B.R.; Berrigan, M.T.; Shah, S.; Marwaha, J.S.; Lai, S.L.; Brat, G.A. Evaluating capabilities of large language models: Performance of GPT-4 on surgical knowledge assessments. *Surgery* **2024**, *175*, 936–942. [CrossRef]
32. Kroupin, I.G.; Carey, S.E. You cannot find what you are not looking for: Population differences in relational reasoning are sometimes differences in inductive biases alone. *Cognition* **2022**, *222*, 105007. [CrossRef]
33. Voutsa, M.C.; Tsapatsoulis, N.; Djouvas, C. Biased by Design? Evaluating Bias and Behavioral Diversity in LLM Annotation of Real-World and Synthetic Hotel Reviews. *AI* **2025**, *6*, 178. [CrossRef]
34. Lunde, Å.; Heggen, K.; Strand, R. Knowledge and power: Exploring unproductive interplay between quantitative and qualitative researchers. *J. Mix. Methods Res.* **2013**, *7*, 197–210. [CrossRef]
35. Koo, T.K.; Li, M.Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [CrossRef]
36. Fleiss, J.L. Balanced incomplete block designs for inter-rater reliability studies. *Appl. Psychol. Meas.* **1981**, *5*, 105–112. [CrossRef]
37. Cicchetti, D.V. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol. Assess.* **1994**, *6*, 284. [CrossRef]
38. Mitchell, M. AI's challenge of understanding the world. *Science* **2023**, *382*, eadm8175. [CrossRef]
39. Dennstädt, F.; Zink, J.; Putora, P.M.; Hastings, J.; Cihoric, N. Title and abstract screening for literature reviews using large language models: An exploratory study in the biomedical domain. *Syst. Rev.* **2024**, *13*, 158. [CrossRef]
40. Jiang, H.; Zhu, D.; Chugh, R.; Turnbull, D.; Jin, W. Virtual reality and augmented reality-supported K-12 STEM learning: Trends, advantages and challenges. *Educ. Inf. Technol.* **2025**, *30*, 12827–12863. [CrossRef]
41. Rest, J.; Narvaez, D.; Bebeau, M.; Thoma, S. A neo-Kohlbergian approach: The DIT and schema theory. *Educ. Psychol. Rev.* **1999**, *11*, 291–324. [CrossRef]
42. Zhu, Y.; Jiang, H.; Chugh, R. Empowering STEM teachers' professional learning through GenAI: The roles of task-technology fit, cognitive appraisal, and emotions. *Teach. Teach. Educ.* **2025**, *167*, 105204. [CrossRef]
43. Suleiman, A.; Von Wedel, D.; Munoz-Acuna, R.; Redaelli, S.; Santarisi, A.; Seibold, E.L.; Ratajczak, N.; Kato, S.; Said, N.; Sundar, E.; et al. Assessing ChatGPT's ability to emulate human reviewers in scientific research: A descriptive and qualitative approach. *Comput. Methods Programs Biomed.* **2024**, *254*, 108313. [CrossRef]
44. Zhang, W.; Cui, Y.; Zhang, K.; Wang, Y.; Zhu, Q.; Li, L.; Liu, T. A static and dynamic attention framework for multi turn dialogue generation. *ACM Trans. Inf. Syst.* **2023**, *41*, 1–30. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.