CITY UNIVERSITY OF HONG KONG
香港城市大學

Enhancing The Performance of Knowledge Workers Through Large Language Models: A Case Study of Academic Paper Peer Review in Scholarly Services

應用大型語言模型提升知識型工作者的績效——以科研學術服務論文評審為例

Submitted to
College of Business
商學院
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Business Administration
工商管理博士學位

by

Liu Guoxing
刘国兴

October 2025
二零二五年十月

摘 要

隨著全球學術研究產出持續增長，學術期刊和會議論文的投稿量呈指數級攀升，傳統的人工同行評審系統已難以應對日益繁重和複雜的評審需求。人力資源瓶頸、評審週期延長、評審標準不一以及主觀性偏差等問題日益突出，嚴重影響著研究成果的及時傳播與學術共同體的健康發展。在這一背景下，人工智慧（AI），特別是大型語言模型（Large Language Models, LLMs）以其強大的自然語言處理和推理能力，為重塑學術評審流程帶來了新的機遇和挑戰。

本論文聚焦於 LLM 驅動的 AI 工具在學術論文同行評審流程中的應用效果，系統考察 AI 輔助對評審效率、準確性、品質、認知負荷以及用戶體驗的影響，為構建更高效、更科學、更可持續的學術評審生態提供理論依據和實踐路徑。研究採用了兩項實證實驗，分別探討 AI 工具的獨立效應及其與人類評審員協作的機制，並基於大樣本的定量與定性數據，全面揭示人機協同模式下的優勢、局限與未來發展方向。

實驗一採用隨機對照實驗設計（RCT），將參與者隨機分為實驗組（AI 輔助）和對照組（人工獨立評審），以便模擬真實的期刊評審環境。評審任務涵蓋內容豐富性、論證邏輯、表達與寫作、創新性與重要性、整體評價五個核心維度，所有參評材料均由相關領域的專家遴選，並以專家共識評分為參照基準。通過對評審效率、評分偏差、認知負荷及信任感等多項指標的定量分析，實驗一發現 LLM 輔助能顯著提升評審流程的整體效率，實驗組在初篩與格式化等機械性任務上平均用時大幅縮短。此外，AI 輔助在“內容豐富性”和“整體評價”等客觀性、可量化維度的評分準確性表現尤為突出，其結果與專家評分高度一致，顯著降低了評審員之間的評分分歧，提高了評分一致性。然而，在“創新性與重要性”和“論證邏輯”這些需要更深層次專業判斷與批判性思維的維度，LLM 的輔助效果相對有限。數據分析顯示，儘管 AI 工具可以在一定程度上降低評分偏差，但在這些主觀性強的維度上，人

工專家的判斷仍然不可或缺。此外， AI 輔助評審初期可能略微增加認知負荷，但隨著用戶適應度的提高，評審員對 AI 輸出的信任逐步增強，認知負擔亦趨於回落。

為進一步探究 AI 與人類評審員的協作潛力與策略優化，實驗二在 AiScholar 平臺上通過設定多組不同人機協作與反思機制的實驗組，對比了對照組（無 AI 輔助）與三類 AI 輔助模式的差異：僅在 AI 弱項維度觸發反思（定向反思組）、僅在 AI 強項維度反思（優勢反思組）、全維度多輪反思（全面反思組）。評審流程包括初評—AI 評分提示—反思—終評，並綜合採集評分改進、評審時長、主觀負荷、用戶滿意度等多元數據。結果表明，在 AI 能力較弱的主觀複雜維度（如創新性、論證邏輯）引入定向反思干預，可以實現與全面反思近似的評分準確性提升，而認知和時間成本顯著低於全面反思組。與此相比，僅在 AI 優勢維度反思則未能有效改善評審結果，反而有可能引發注意力分散和額外負擔。進一步的統計檢驗和回歸分析證實，合理的任務分配和流程設計（如“定向反思”）可在保證評審品質的同時優化資源配置，實現人機協作效能的最優平衡。開放式問卷和用戶回饋亦顯示，AI 建議的結構化和可解釋性對提升評審員信任度、激發批判性反思具有重要作用，評審員在反思干預環節更願意結合 AI 與自身判斷，主動修正初步直覺偏差。

通過整合兩項實驗的結果，論文進一步梳理了 AI 與人類專家在評審流程中的優勢互補特性及其交互機制。AI 擅長快速處理大批量、結構化的數據分析和初步篩選，顯著緩解了評審員的機械性工作負擔，並在高一致性、低主觀性維度上表現出色；而人類評審員則能針對 AI 短板的主觀複雜任務施展深度分析與創造性判斷，有效避免“錨定效應”及 AI 潛在的系統性偏見。論文還系統探討了雙過程理論（System 1/2）在 AI 人機協作評審中的理論價值：AI 建議往往激發系統 1 的快速直覺決策，而反思環節則喚起系統 2 的深度推理，從而實現評審品質與效率的動態平衡。通過反思機制精准啟動系統 2，最大限度發揮 AI 與人工的協同優勢。

本研究還對倫理、隱私和數據安全等核心問題進行了嚴格規範和制度設

計。所有實驗流程均經過知情同意與倫理審批，所有採集數據去標識化、加密存儲，僅限學術分析使用。AI 建議在評審介面匿名呈現，杜絕因身份資訊導致的任何偏見，確保實驗結果的科學性和公正性。同時，明確告知 AI 工具僅為輔助支持，所有最終判斷權均歸屬評審員自身，嚴格防止自動化偏見及 AI 過度依賴。

本研究的實踐意義在於為學術出版領域的智能化轉型和評審體系優化提供了直接可操作的經驗和決策參考。首先，研究結果為期刊主編和出版機構提供了數據驅動的證據，表明採用 LLM 驅動的 AI 工具不僅可以顯著提升評審效率，還能優化評審資源配置，縮短出版週期。其次，定向反思機制的設計為評審平臺開發者和管理者提供了流程改進的範式，幫助實現精准干預、合理分工，避免過度依賴 AI 或評審員單一判斷，從而提升評審品質和一致性。再次，研究提出的協作策略和交互流程（如“先人工初評-再 AI 提示-後反思終評”）可作為構建智能評審平臺的核心邏輯，為未來 AI 與人類協作的學術評審系統開發提供理論支撐和實踐範本。此外，主觀負荷與滿意度測評為學術共同體提供了判斷 AI 介入合理性與適用邊界的定量依據，有助於制定合理的評審標準和品質監控機制。從更廣泛的角度來看，本研究經驗可推廣至醫療、法律、工程、金融等知識密集型行業，助力建設高效、可控、可信賴的智能化知識工作流程，推動相關領域數位化、智能化轉型升級。

在理論貢獻方面，本論文以豐富的實驗數據和嚴密的統計分析驗證了 AI 賦能同行評審的價值，明確界定了 AI 在不同評審維度的優勢與不足，並提出了“人機互補—定向反思”模式的優化策略，豐富了人機協作決策與認知理論的實證內涵。實踐層面，本研究為學術期刊、科研管理和出版行業提供了切實可行的 AI 評審流程設計建議，如針對 AI 短板的維度精准激發人工批判性反思、在客觀性強的流程環節充分發揮 AI 自動化優勢等，為推動學術評審系統的智能升級和管理創新奠定了堅實基礎。

總的來說，本論文通過多維度、多環節的實驗設計與數據分析，全面揭示了 LLM 驅動的 AI 工具在提升學術評審效率、品質和用戶體驗方面的巨大潛力，厘清了 AI 與人類專家優勢互補、動態協作的最佳實踐路徑，並為後

續學術評審智能化、標準化發展提供了理論支撐和實踐範式。隨著 LLM 技術的持續進步和學科領域應用的不斷深化，AI 與人類專家高效協作的知識工作模式有望在更廣泛的科研場景中推廣應用，推動學術出版與創新生態的高質量發展。

關鍵字：人工智慧，學術論文評審，AI 輔助，人機協作，評審效率，評審準確性，創新性，方法論，GPT-4

Abstract

As global scholarly research output continues to grow, the number of submissions to academic journals and conference proceedings is increasing at an exponential rate. Traditional manual peer review systems are finding it increasingly difficult to cope with the escalating volume and complexity of review demands. Issues such as resource bottlenecks, prolonged review cycles, inconsistencies in evaluation standards, and subjective bias have become increasingly prominent, severely impacting the timely dissemination of research findings and the healthy development of the academic community. Against this backdrop, artificial intelligence (AI)—and especially large language models (LLMs), with their powerful natural language processing and reasoning capabilities—offer both new opportunities and challenges for reshaping the scholarly review process.

This thesis focuses on the application and effectiveness of LLM-driven AI tools in the peer review process for academic manuscripts. It systematically examines the impact of AI-assisted review on efficiency, accuracy, quality, cognitive load, and user experience, with the aim of providing both theoretical foundations and practical pathways for constructing a more efficient, scientific, and sustainable academic review ecosystem. Two empirical experiments were conducted to investigate both the independent effects of AI tools and the mechanisms of their collaboration with human reviewers. Based on large-sample quantitative and qualitative data, this study comprehensively reveals the advantages, limitations, and future directions of human-AI collaborative models.

Experiment I employed a randomised controlled trial (RCT) design, randomly assigning participants to an experimental group (AI-assisted) and a control group (human-only independent review), simulating a real journal review environment. Review tasks covered five core dimensions: content richness, argument logic, expression and writing, innovation and significance, and overall evaluation. All materials for review were selected by domain experts, with expert consensus scores serving as reference standards. Through quantitative analysis of multiple indicators—review efficiency, scoring deviation, cognitive load, and trust—

Experiment I finds that LLM assistance significantly enhances the overall efficiency of the review process, with the experimental group showing marked reductions in time spent on preliminary screening and mechanical formatting tasks. Moreover, AI assistance demonstrates particularly strong scoring accuracy in objective, quantifiable dimensions such as “content richness” and “overall evaluation,” with results highly consistent with expert scores. This markedly reduces score divergence among reviewers and increases scoring consistency. However, in dimensions such as “innovation and significance” and “argument logic,” which require deeper professional judgement and critical thinking, the effect of LLM assistance is relatively limited. Data analysis indicates that while AI tools can reduce scoring deviation to some extent, human experts' judgement remains indispensable in these highly subjective dimensions. Additionally, AI-assisted review may initially increase cognitive load slightly, but as users become more familiar, their trust in AI output grows, and cognitive burden tends to decrease accordingly.

To further explore the potential for and strategies of AI–human reviewer collaboration, Experiment II was conducted on the AiScholar platform by establishing multiple experimental groups with distinct human–AI collaboration and reflection mechanisms. The study compared a control group (no AI assistance) with three types of AI-assisted models: targeted reflection (reflection only on AI-weakness dimensions), strengths-based reflection (reflection only on AI-strength dimensions), and comprehensive reflection (multi-round reflection on all dimensions). The review process consisted of initial review, AI score prompts, reflection, and final review, with comprehensive collection of data on score improvement, review duration, subjective cognitive load, and user satisfaction. The results show that introducing targeted reflection interventions on subjective and complex dimensions where AI capability is relatively weak (e.g., innovation, argument logic) can achieve accuracy improvements close to those of comprehensive reflection, but with significantly lower cognitive and time costs. In contrast, reflection only on AI-strength dimensions does not effectively improve review outcomes and may even cause distraction and unnecessary burden. Further statistical tests and regression analyses confirm that rational task allocation and

process design—such as targeted reflection—optimise resource allocation and achieve optimal human–AI collaborative efficiency while maintaining review quality. Open-ended surveys and user feedback further indicate that the structured and explainable nature of AI recommendations plays a critical role in increasing reviewers' trust and stimulating critical reflection. During the reflection intervention phase, reviewers are more willing to combine AI recommendations with their own judgement, proactively correcting initial intuitive biases.

By integrating the findings of the two experiments, this thesis further delineates the complementary advantages and interaction mechanisms between AI and human experts within the review process. AI excels at rapidly processing large-scale, structured data analysis and preliminary screening, which significantly alleviates reviewers' mechanical workload and demonstrates outstanding performance in dimensions characterised by high consistency and low subjectivity. In contrast, human reviewers are able to deploy deep analysis and creative judgement in addressing the more subjective and complex tasks where AI is less effective, thereby effectively avoiding the "anchoring effect" and potential systemic biases of AI. The thesis also systematically discusses the theoretical value of Dual Process Theory (System 1/2) in AI–human collaborative peer review: AI-generated suggestions often stimulate rapid, intuitive decision-making (System 1), while the reflection stage activates deeper reasoning (System 2), thereby achieving a dynamic balance between review quality and efficiency. By precisely activating System 2 through targeted reflection mechanisms, the collaborative strengths of both AI and human reviewers are maximised.

This study also establishes strict protocols and institutional safeguards concerning core issues of ethics, privacy, and data security. All experimental procedures underwent informed consent and ethical approval; all collected data were de-identified and stored in encrypted form for academic analysis only. AI suggestions were presented anonymously in the review interface, precluding any bias arising from identity information and ensuring the scientific validity and fairness of the results. Simultaneously, it was made explicit that AI tools serve only as auxiliary support, and all final judgements rested solely with the reviewers themselves, strictly preventing automation bias and over-reliance on AI.

The practical significance of this study lies in providing directly actionable experiences and decision-making references for the intelligent transformation and optimisation of the academic publishing field's review system. First, the research results offer data-driven evidence for journal editors and publishing organisations, demonstrating that the adoption of LLM-driven AI tools can not only significantly enhance review efficiency but also optimise resource allocation and shorten publication cycles. Second, the design of the targeted reflection mechanism presents a process improvement paradigm for review platform developers and administrators, facilitating precise interventions and rational task allocation to avoid over-reliance on either AI or individual reviewer judgement, thereby improving review quality and consistency. Third, the collaboration strategies and interactive processes proposed in this study—such as the "initial human review, followed by AI prompt, and subsequent reflection for final review"—can serve as the core logic for building intelligent review platforms, providing both theoretical foundations and practical templates for the future development of collaborative human–AI academic review systems. Moreover, the evaluation of subjective cognitive load and satisfaction offers the academic community quantitative criteria for determining the appropriateness and application boundaries of AI intervention, which is valuable for formulating reasonable review standards and quality monitoring mechanisms. More broadly, the experiences of this study are transferrable to other knowledge-intensive industries such as healthcare, law, engineering, and finance, supporting the establishment of efficient, controllable, and trustworthy intelligent knowledge workflows and driving digital and intelligent transformation in relevant fields.

In terms of theoretical contributions, this thesis, through abundant experimental data and rigorous statistical analysis, validates the value of AI-empowered peer review, clearly delineates the strengths and limitations of AI across different review dimensions, and proposes the optimised "human–AI complementarity—targeted reflection" model, thus enriching the empirical content of decision-making and cognitive theory in human–AI collaboration. On the practical level, the study provides concrete and feasible design recommendations for AI review processes in academic journals, research management, and publishing

industries, such as precisely activating human critical reflection on AI-weak dimensions and fully leveraging AI automation advantages in objective workflow stages, laying a solid foundation for the intelligent upgrading and management innovation of academic review systems.

In summary, through multidimensional and multi-stage experimental design and data analysis, this thesis comprehensively reveals the great potential of LLM-driven AI tools in improving the efficiency, quality, and user experience of academic review, clarifies best practices for the complementary advantages and dynamic collaboration between AI and human experts, and provides both theoretical support and practical paradigms for the future intelligent and standardised development of academic review. With the ongoing advancement of LLM technology and its deepening application across disciplines, efficient knowledge work modes characterised by close collaboration between AI and human experts are expected to be promoted in a broader range of research contexts, thereby facilitating the high-quality development of scholarly publishing and innovation ecosystems.

Keywords: Artificial Intelligence; Academic Paper Review; AI Assistance; Human-AI Collaboration; Review Efficiency; Review Accuracy; Innovation; Methodology; GPT-4

CITY UNIVERSITY OF HONG KONG
Qualifying Panel and Examination Panel

Surname:	LIU
First Name:	Guoxing
Degree:	Doctor of Business Administration
College/Department:	College of Business

The Qualifying Panel of the above student is composed of:

Supervisor(s)

Prof. MA Jian	Department of Information Systems City University of Hong Kong
Prof. HUANG Lihua	Department of Information Management and Information Systems Fudan University

Qualifying Panel Member(s)

Prof. XU Jingjun David	Department of Information Systems City University of Hong Kong
Prof. LU Xianghua	Department of Information Management and Information Systems Fudan University

This thesis has been examined and approved by the following examiners:

Prof. SIA Choon Ling	Department of Information Systems City University of Hong Kong
Prof. MA Jian	Department of Information Systems City University of Hong Kong
Prof. ZHANG Chenghong	Department of Information Management and Information Systems Fudan University
Prof. HUANG Lihua	Department of Information Management and Information Systems Fudan University
Prof. WANG Gang	School of Management Hefei University of Technology

答 謝

時光荏苒，博士求學之路轉瞬即逝。謹以此頁，向所有在我攻讀博士學位期間給予我指導、支持與鼓勵的師長、親友及夥伴，致以最誠摯的謝意。

首先，我要由衷感謝我的指導教授們。在論文的研究與寫作過程中，您們以淵博的學識、嚴謹的治學態度與無私的奉獻精神，引領我探索學術前沿，突破研究瓶頸。您們的悉心指導與諄諄教誨，不僅為本論文的順利完成奠定了堅實基礎，更使我學會了獨立思考與科學探究的方法，這份恩情將令我終身受益。

同時，我也要感謝論文答辯委員會的各位委員。您們提出的寶貴意見與深刻洞見，為本論文的完善提供了關鍵性的指導，使研究的深度與廣度得以提升。我也要感謝香港城市大學商學院，為我提供了一個優質的學術環境與豐富的研究資源，讓我在這片學術沃土上得以專心致學，不斷成長。

本研究的順利完成，離不開所有參與實證研究的專家學者與同學們。感謝您們在百忙之中抽出寶貴時間參與實驗，您們的熱情參與和提供的寶貴數據，是本論文實證分析的基石。此外，感謝在研究過程中給予支持與協助的合作機構與朋友們，您們的幫助讓研究得以順利推進。

在此，我也要向求學路上的同儕與摯友們表達謝意。我們在學術上的交流探討與生活中的相互扶持，是我克服困難、堅持不懈的動力源泉。

最後，我要將最深的感謝獻給我的家人。是您們無條件的愛、理解與包容，成為我身後最堅實的後盾，支持我度過這段充滿挑戰與收穫的旅程。您們的付出與鼓勵，是我完成學業的最大動力。

論文的完成是一個階段的結束，也是一個新起點的開始。再次向所有關心與幫助過我的人們，致以最由衷的感謝！

目 錄

表目錄

圖目錄