Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A quasi-experimental analysis of capabilities and limitations of generative AI in academic content evaluation in social sciences

Yu Zhu [a,*], Yongrong Lu [b], Huan Xie [a], Jiyuan Ye [a], Ming Chen [a]

[a] *School of Information Management, Nanjing University, Nanjing 210023, China*
[b] *Pittsburgh Institute, Sichuan University, Chengdu 610207, China*

ARTICLE INFO

Keywords:
Academic evaluation in social sciences
Academic content evaluation
Academic evaluation
Generative AI
ChatGPT
SSS-ACE model

ABSTRACT

The complexity of social sciences research and the limitations of traditional evaluation methods highlight the need to explore the capabilities and application potential of generative AI in academic evaluation. Previous research in fields such as biomedical and other natural sciences has demonstrated the potential of generative AI to estimate the quality of research articles. This study adopts a quasi-experimental approach, 100 volunteers produced 600 social sciences academic texts across 6 types of topics, which were evaluated by 8 mainstream generative AI models. Statistical and sentiment analysis was conducted to compare the evaluation results using zero-shot and few-shot prompting strategies. The results show that AI-generated total scores are unreliable (precision = 66.35 %), and the actual total scores differ moderately from the human benchmark (average Cohen's $d = 0.425$). Few-shot prompt exhibited weaker differentiation capabilities across dimensions (average correlation = 5.25), while zero-shot prompt performed better (e.g., $\text{correlation}_{\text{Clarity, Significance}} = 0.13$), particularly in writing quality (average standard deviation = 5.38). Significant score differences were observed across the eight models (all $p < 0.001$), indicating inconsistency among models. Additionally, AI-generated comments across dimensions were generally positive, with different models exhibiting strengths across various dimensions and tasks. This study provides empirical evidence for scholars, peer reviewers, and research evaluation professionals interested in integrating generative AI into social sciences' evaluation workflows. Overall, generative AI shows potential for enhancing evaluation efficiency and reducing favoritism in the peer review of social sciences, especially in large-scale or preliminary evaluations. However, when evaluating the novelty and significance, its dependency on domain knowledge and the interpretability of the results still requires prudent consideration and refinement.

1. Introduction

Global scientific research output is currently doubling approximately every nine years (Bornmann & Mutz, 2015), with an academic paper published every 20 s on average (Munroe, 2013). These papers represent the culmination of scientific inquiry, with their primary value stemming from the intellectual content they present. In light of the need for large-scale and agile evaluation, research management departments have widely implemented quantitative evaluation system (Xue et al., 2023). However, these methods face

[*] Corresponding author at: School of Information Management at Nanjing University, No 163, Xianlin Avenue, Nanjing, China.
E-mail address: zhu.yu@smail.nju.edu.cn (Y. Zhu).

https://doi.org/10.1016/j.ipm.2025.104365
Received 21 March 2025; Received in revised form 17 July 2025; Accepted 18 August 2025

criticism for oversimplifying indicators, over-relying on quantitative standards, and the utilitarianizing of outcomes. Consequently, there is a growing call within the academic community to reform the existing evaluation systems, which remain largely reliant on bibliometric indicators (Hrubec & Višňovský, 2023).

In recent years, there has been heightened focus on academic evaluation practices within the natural and social sciences. On one hand, there is a movement to abandon outdated models that overemphasize specific metrics, such as the use of impact factors to indicate the quality of a paper (Zhang & Sivertsen, 2020). Conversely, there is a growing emphasis on reinstating peer review, adopting categorized evaluation, and implementing a representative works system. These approaches aim to embrace a more comprehensive evaluation approach that combines qualitative and quantitative methods, with the aim of surpassing conventional practices in academic assessment (Hicks et al., 2015). Although some studies proposed the use of AI technologies, including natural language processing and machine learning, to identify innovative points and highlights in academic papers (Ronzano & Saggion, 2016; Yang, 2016). However, existing AI-based approaches predominantly emphasize surface-level quantitative metadata, patterns of indicators, or natural sciences (Huang, Huang et al., 2025). They often overlooking the embedded interpretive frameworks, theoretical discourse, and contextual dimensions, especially within social sciences. This study moves beyond such approaches by employing a quasi-experimental design to empirically evaluate how generative AI can engage with the full-text content of academic papers and simulate qualitative human peer judgment methodologies in social sciences based on the System of All-round Evaluation of Research (Ye, 2010). Originating from the field of research evaluation in Chinese social sciences, this framework integrates form, content, and utility assessments across six core dimensions. It underscores the importance of evaluation purpose, scientific content and reviewers' meta-evaluation in academic evaluation with openness and developmental adaptability as its key features (Ye, 2021). This system provides the theoretical and methodological foundation for both the empirical analysis and the subsequent conceptual model construction in this study. In doing so, this study provides a novel contribution by critically examining the actual evaluative capabilities and limitations of generative AI within a more context-rich domains.

This contribution holds particularly significance within the social sciences, where the evaluation of academic content is inherently more complex compared to the natural sciences. This complexity stems from the enduring competition among paradigms such as positivism, interpretivism, and critical theory in social sciences (Kuhn, 1962). Consequently, the criteria for evaluating academic achievements cannot be entirely quantified with the objective characteristic of experimental data like natural sciences. With the development of generative AI technologies, exemplified by the GPT-4 model, the Library and Information Science (LIS) discipline has gained a technological basis to tackle challenges in academic evaluation within social sciences. Generative AI presents considerable promise in addressing challenges related to strong subjectivity, low efficiency, and elevated human resource costs in qualitative evaluation, alongside issues of content focus and metrics overgeneralization in quantitative evaluation (Thelwall, 2025b). By leveraging the advanced algorithms, supercomputing power, big data capabilities, and vast parameter scales of generative AI, it is now possible to explore content-based academic quality evaluation in greater depth (Sun et al., 2022). Therefore, it is essential to investigate the theoretical and methodological frameworks for employing generative AI in the development of intelligent academic evaluation systems.

Despite the growing focus on AI-driven tools in academic evaluation, a significant research gap persists regarding the effective integration of generative AI into content-based evaluation, especially within the complex context of social sciences. This study aims to critically examine the capabilities and limitations of generative AI in supporting academic evaluation within the social sciences. The subsequent research questions are formulated to direct the inquiry:

RQ1: What capabilities and limitations does generative AI possess in facilitating academic evaluation within the social sciences?
RQ2: What unique characteristics define the use of generative AI in the evaluation of academic content within the social sciences?
RQ3: How can academic evaluation in the social sciences be conducted intelligently and effectively in the age of generative AI?

2. Literature review

2.1. The limitations of academic evaluation methods

2.1.1. The limitations of bibliometrics

Bibliometrics is a quantitative evaluation method based on data such as the volume of academic outputs and citation counts (Hirsch, 2005). A significant issue is its disregard for content, failing to reveal how or why cited documents contribute value to subsequent research (MacRoberts & MacRoberts, 1989). These metrics are prone to manipulation via self-citation, citation cartels, or strategic referencing, often indicating superficial visibility rather than genuine academic merit (Brooks, 1986).

Furthermore, bibliometric evaluation is based on the theoretical premise that citation frequency is partly indicative of quality (Garfield, 1955). However, in practice, this correlation is often misinterpreted as causation, resulting in an over-reliance on quantitative scores in high-stakes decisions (Seglen, 1997). Another major concern is the absence of granularity and fairness: citation counts mask variations in citation intent (affirmation versus critique), and average-based metrics such as journal impact factors do not consider the non-normal distribution of citation data (Bornmann & Daniel, 2008). Research indicates the majority of articles published in high-impact journals receive a limited number of few citations, suggesting that using journal-level averages to judge individual papers lacks scientific rigour and statistical validity (Hamilton, 1991).

Due to these inherent flaws, bibliometrics is inadequate, particularly in nuanced or content-sensitive assessments (Wilsdon, 2016). Thus, while bibliometrics serves as a valuable instrument for analyzing large-scale trend (Clarivate, 2025), it should be supplemented with qualitative assessment methods to guarantee fairness and accuracy in academic evaluations (DORA, 2012).

2.1.2. *The limitations of peer review*

Peer review remains an indispensable component of academic evaluation, especially in contexts where the quality of content is of central concern (Hicks et al., 2015). In contrast to bibliometric indicators, peer review facilitates nuanced evaluations by domain experts, providing insights into a work's originality, methodological rigor, and theoretical contributions (Ye, 2010). This process is recognized as one of the primary methods for assessing academic papers (Bornmann, 2011). Nonetheless, its sustained dominance does not imply the absence of issues. Conversely, the most enduring critiques of peer review focus on its intrinsic subjectivity (Kelly et al., 2014) and vulnerability to bias (Si et al., 2023). Due to these susceptibility, peer review has been criticized for issues including a limited pool of reviewers, strong subjectivity, potential favoritism, mismatched expertise, and insufficient oversight and feedback mechanisms (Marsh et al., 2008). The reliability of its outcomes has been consistently questioned by these factors. Additionally, the peer review process for academic papers entails considerable financial expenditures and demands extensive time and effort from specialists (Higher Education Funding Council for England, 2019). This complexity complicates the identification of significant research amidst the vast number of published works, particularly in the context of AI advancements (Prillaman, 2024).

Despite its limitations, peer review continues to be the most effective method for evaluating the academic contributions of papers, as no superior alternative has been identified (Shiflett, 1988). For researchers, the key issue is not whether to discard peer review, but how to enhance it through the adoption of emerging technologies within the evolving landscape of knowledge production.

2.2. *AI-Based approaches to academic content evaluation*

Efforts to enhance the evaluation of academic content have persisted over time. Initiatives like the System of All-round Evaluation of Research (Ye, 2010), the *San Francisco Declaration on Research Assessment* (DORA, 2012), and the *Leiden Manifesto* (Hicks et al., 2015) underscore the importance of the scientific content of a paper over superficial bibliometric indicators. These frameworks indicate an increasing agreement that academic evaluation ought to prioritize content-based assessments of academic outcomes (Sun et al., 2022).

Recent breakthroughs in data availability, algorithm design, and computational power have propelled the evolution of next-generation AI technologies, also known as large language model (LLM) (Wu et al., 2023). As a result, researchers have increasingly adopted AI-based approaches, including both traditional machine learning and contemporary LLM techniques, to enhance content-based assessment (see Table 1). Notably, recent research has explored the concept of LLM-as-a-Judge, where LLMs are employed to evaluate the outputs of other AI systems (Wang, Yu et al., 2024) and approximate human preferences (Zheng et al., 2023), such as in dialogue generation, summarization (Liu et al., 2023), and instruction-following tasks (Wang, Yu et al., 2024). These studies have introduced benchmark datasets and alignment strategies to improve consistency, reliability, and scalability of AI-generated assessments. While promising, they primarily focus on evaluating machine-generated content within controlled environments. In contrast, the application of LLMs to assess human-authored academic outcomes, especially in the social sciences, remains underexplored. This study thus also extends the LLM-as-a-Judge paradigm into a new, more complex setting, highlighting its capabilities and limitations in real-world academic evaluation tasks.

As further shown in Table 1, traditional methods seek to assess dimensions such as novelty (M. Liu et al., 2024) and innovation (Lin et al., 2025) through analysis of the internal structure, semantics, and linguistic features of academic texts (Huang, Huang et al., 2025). More importantly, recent generative AI technologies have created new opportunities for content-based evaluation (Thelwall, 2025a). However, the existing literature is still in its early stages. These approaches predominantly depend on superficial metadata or restricted materials related to academic outputs (e.g., titles, keywords, abstracts). They tend to employ indirect evaluation techniques instead of conducting comprehensive full-text analyses (Thelwall & Yaghi, 2024). Thus, while traditional AI methods demonstrate effectiveness in specific areas, they inadequately address the intricate, nuanced, and interpretive aspects of scholarly quality, especially within the social sciences. Furthermore, most current research relies on previously published academic papers as the main sample (Thelwall et al., 2025), which may raise concerns about the credibility of findings due to possible overlaps with undisclosed large-scale AI training corpora (Elangovan et al., 2021). Additionally, many studies focus on fields such as biomedicine (Huang, Huang et al., 2025) and computer science (Liang et al., 2024), where structured knowledge entities and measurable innovation indicators are more accessible (Liu et al., 2022). In contrast, the social sciences, marked by conceptual ambiguity, methodological diversity, and layered argumentation (Wallerstein, 2004), are significantly under-explored in the context of generative AI in academic evaluation.

To address these gaps, this study examines the capacity of generative AI to conduct comprehensive evaluations of academic papers within the social sciences. This study systematically examines the capabilities and limitations of the state-of-the-art models in assessing the quality of social sciences content, guided by the System of All-round Evaluation of Research (Ye, 2010). Importantly, in contrast to the majority of existing studies that utilize published papers as primary data sources, our research is founded on a substantial corpus of unpublished, full-length academic manuscripts that we have independently collected. Given that existing generative AI-assisted approaches to academic quality evaluation are fundamentally based on the logic and practice of peer review (Thelwall, 2025a). Thus, we consider the peer review tradition not just a procedural formality but as the theoretical and methodological foundation of this study. Peer review represents the epistemic standards, interpretative assessment, and disciplinary context necessary for evaluating academic contributions, particularly within the social sciences. This method enables the preservation of interpretive depth and intellectual rigor inherent in peer review, while utilizing AI to address its practical constraints. This dual orientation offers a conceptual framework and a practical approach for rethinking academic evaluation in the era of intelligent assistance.

Table 1
Overview of AI-based approaches to academic content evaluation.

Study	AI type	Domain	Evaluation tasks	Input scope	Key contribution/conclusion	Main limitation
(Yang et al., 2018)	Traditional Machine Learning	Computer Science	Academic paper rating	Full Text	A modular hierarchical convolutional neural network is proposed.	The emphasis is placed on originality instead of overall quality.
(Yang et al., 2022)		Computer Science	Emerging topics detection	Keywords	Presented the viewpoint of knowledge ecology.	Concentrating exclusively on keywords while disregarding the full text.
(Xue et al., 2023)		Artificial Intelligence	Academic paper rating	Title & Abstract	Proposed dual-view graph convolutions to enhance BERT for the evaluation of academic papers.	The emphasis is placed on originality instead of overall quality.
(Liu et al., 2024)		Biomedical	Novelty evaluation	Title & Abstract	Quantified scientific novelty in doctoral theses.	Employed the abstract, but could not fully utilize the full text.
(Lin et al., 2025)		Computer Science	Innovation assessment	Abstract, Authors, Publication Year, Locations, & titles	Measured the degree of scientific innovation breakthroughs.	Mainly based on surface-level bibliometric indicators.
(Biswas et al., 2023)	Generative AI	Biology and Medicine	Manuscript quality	Full Text	Argued that the integration of ChatGPT as a reviewer in the journal peer-review process offers both potential benefits and challenges.	Relied exclusively on a brief article, which lacks robustness.
(Saad et al., 2024)		Medical	Peer Review Aid	Full Text	Demonstrated that ChatGPT in its current form is not capable of replacing human reviewers.	Based on merely 24 published articles, with the full text segmented into sections, resulting in a lack of robustness.
(Wilby & Esson, 2024)		Geographic	Paper quality based on REF2021 criteria	Unknown	Noted that ChatGPT is unable to evaluate research concerning the latest real-time issues.	The results are based on intuitive perceptions and lack substantial evidence and thorough discussion.
(Liang et al., 2024)		Biomedical and Artificial Intelligence	Paper quality feedback	Full Text	Identified significant overlap between LLM and human feedback, along with favourable user perceptions of the utility of LLM feedback.	Based on published or accepted papers, and the research is confined to the field of natural sciences.
(Kousha & Thelwall, 2024)		Multidisciplinary	Societal impact claims	Title & Abstract	The value of generative AI differs markedly among various fields.	Does not provide a comprehensive analysis of the full text and fails to evaluate the overall quality of the academic content.
(Thelwall, 2024)		Cultural and Media Studies, Library and Information Management	Quality evaluation	Full Text	Observed that ChatGPT currently lacks the requisite accuracy for reliable formal or informal research quality evaluation tasks.	The data comprises self-evaluations from a convenience sample of articles authored by a single academic within one field, characterized by a small sample size.
(Thelwall & Yaghi, 2024)		Multidisciplinary	Quality evaluation	Title & Abstract	Pointed out that evaluations relying exclusively on titles and abstracts do not represent a comprehensive research assessment, and the results may be influenced by disciplinary biases.	Relied solely on the title and abstract of published articles, which may introduce concerns regarding reliability.
(Thelwall, 2025b)		Information Science	Quality evaluation	Full text without tables, figures, and references; title and abstract; title only	Found that the optimal input for LLMs consists of the article title and abstract.	The findings are derived from published papers, and the sample size is limited.
(Huang, Huang et al., 2025)		Biomedical	Originality evaluation	Title & Abstract	Observed that LLMs can function as originality reviewers; however, they often exhibit excessive leniency.	Relied solely on the title and abstract, neglecting to evaluate the overall quality of the academic content.
(Thelwall et al., 2025)		Medical	Quality evaluation	Title & Abstract	Determined that ChatGPT can serve as a tool for assessing the quality of clinical medicine research.	Relied solely on the title and abstract of published articles, which may introduce concerns regarding reliability.
(Huang, Wang et al., 2025)		Mathematics, Physics, Chemistry, and Medicine	Novelty and originality evaluation	Title & Abstract	Proposed AI-empowered Paper Evaluation methods that leverage a multi-agent system to assess paper quality across novelty and originality.	Relied solely on the title and abstract, limited to novelty and originality, and the samples are confined to the field of natural sciences.

3. Data and methods

3.1. Data acquisition

3.1.1. Academic content

To ensure comparability and minimize thematic or structural inconsistencies, we employed a quasi-experimental design by recruiting participants and assigning them standardized academic writing propositions for data collection. This method facilitates the generation of a controlled yet realistic corpus that supports robust and full-text academic evaluation (Kampenes et al., 2009; Miller et al., 2020). While quasi-experimental designs are common in educational or writing assessment research (Aiken et al., 1998; Kuo, 2015), their application in studies of academic evaluation, particularly those involving generative AI, is limited. To the best of our knowledge, compared with prior work that often uses publicly available papers, our design collected independent and unpublished writing samples, which avoids model exposure to existing data and thus improves the credibility of performance comparison. Therefore, our implementation of a quasi-experimental method in this context represents a novel contribution.

This investigation involved collecting and analyzing academic samples from the responses of 100 recruited volunteers, who are currently enrolled as master's students in a top-tier iSchool. While participants share a similar academic background, natural differences in individual academic ability remain, ensuring both internal diversity and cross-sample comparability. This design offers a controlled yet realistic corpus, improving the reliability of AI evaluation in a social science context. Participants were recruited through convenience sampling (Cohen et al., 2002), as they were enrolled in a course titled Information Resources Construction, during which the study was conducted. The 100 volunteers accounted for 96.15 % of the total registered students in the course. They were informed of the research purpose in advance and voluntarily agreed to contribute their anonymized academic writing for research use. All data were de-identified prior to analysis to protect participant privacy and the collection procedures were reviewed and supervised by the course leading instructor and institutional authority.

Participants were required to engage in discussions and complete academic writing tasks on six propositions, as shown in Appendix A Table A.1, from October 2024 to December 2024. A total of 600 academic content samples were collected for analysis, with a detailed description provided in Appendix A Table A.2. Before each collection, the experimenters supplied the volunteers with relevant knowledge to ensure a detailed understanding of the six propositions, thus facilitating the production of high-quality academic samples. The volunteers were required to produce texts of at least 600 Chinese characters and include a minimum of four references, with no further stipulations regarding formatting or other aspects. The six propositions include academic writing tasks such as fundamental concepts, frontier issues like AI, comparative analysis, and investigative reports, thereby ensuring comprehensive coverage across diverse subject areas. Consequently, we posit that the gathered academic texts are suitable for further examination.

3.1.2. Evaluation score and comment

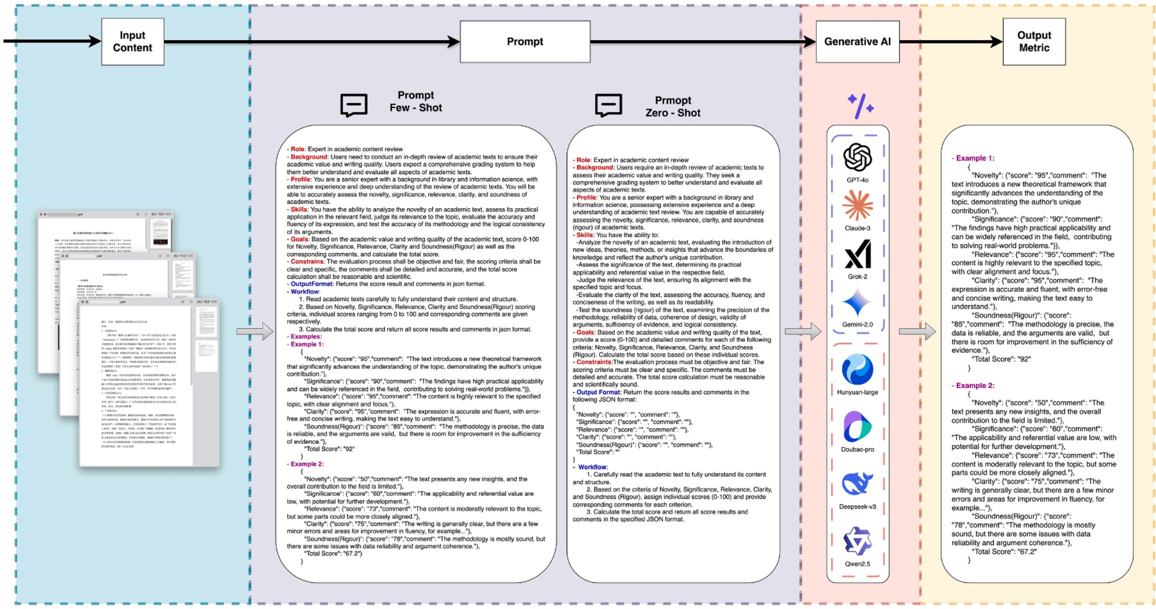
From the intrinsic attributes of academic papers, the essential quality of an outstanding academic paper lies in its value (Merton, 1979). Additionally, it is acknowledged that high-quality academic content should be conveyed through standardized writing to enhance readers comprehensibility. Therefore, in this study, the quality of academic content is operationally defined through two first-level dimensions: academic value and writing quality, each further divided into measurable second-level indicators (see Table 2). This structure is specifically tailored to reflect the characteristics and assessment challenges of academic content in social sciences, making it more comprehensive than evaluation schemes used in prior AI-based assessment studies within natural sciences.

A standardized scoring rubric was developed to ensure consistency and minimize subjectivity. Indicator are rated on a scale from 0 to 100, with 100 denoting the peak level of performance. AI models were guided by structured prompts developed following the LangGPT (Language For GPT) framework (Wang, Liu et al., 2024), which offers modular and interpretable prompt structures corresponding to each evaluation criterion. The prompts were provided detailed definitions and anchor examples for each score range (e.g., 90–100 = excellent originality and significance; 60–69 = adequate clarity but limited rigor), as illustrated in Appendix B. Additional details regarding this prompt design are included at the conclusion of Appendix B to enhance interpretability and facilitate future replication.

The data acquisition process, shown in Fig. 1, consists of three key steps: inputting academic content, constructing prompts based on the simple evaluation system, and feeding both into generative AI models to obtain evaluation data. For prompt design, considering the critical role of sample demonstrations in-context learning (Song et al., 2023), we employed zero-shot and few-shot strategies (P. Liu et al., 2023) to improve result comparability, incorporating explicit evaluation objectives to direct the process. Eight state-of-the-art

Table 2
Indexes and explanations of the simple evaluation system.

First level index	Second level index	Definition	Source
Academic Value	Novelty	The introduction of new ideas, theories, methods, or insights that advance the boundaries of knowledge, reflecting the author's unique contribution to the academic.	(Yan & Fan, 2024; Bornmann et al., 2019)
	Significance	Practical applicability and referential value in the respective field.	(Checco et al., 2021)
Writing Quality	Relevance	Relevance and alignment with the specified topic.	(Spezi et al., 2018)
	Clarity	Accuracy and fluency of expression: error-free, smooth, and concise writing.	(Sukpanichnant et al., 2024)
	Soundness (Rigour)	Methodological precision, reliability of data, coherence of design, validity of arguments, sufficiency of evidence, and logical consistency.	(Spezi et al., 2018)



```
- Role: Expert in academic content review
- Background: Users need to conduct an in-depth review of academic texts to ensure their academic value and writing quality. Users expect a comprehensive grading system to help them better understand and evaluate all aspects of academic texts.
- Profile: You are a senior expert with a background in library and information science, with extensive experience and deep understanding of the review of academic texts. You will be able to accurately assess the novelty, significance, relevance, clarity, and soundness of academic texts.
- Skills: You have the ability to analyze the novelty of an academic text, assess its practical application in the relevant field, judge its relevance to the topic, evaluate the accuracy and fluency of its expression, and test the accuracy of its methodology and the logical consistency of its arguments.
- Goals: Based on the academic value and writing quality of the academic text, score 0-100 for Novelty, Significance, Relevance, Clarity and Soundness(Rigour) as well as the corresponding comments, and calculate the total score.
- Constrains: The evaluation process shall be objective and fair, the scoring criteria shall be clear and specific, the comments shall be detailed and accurate, and the total score calculation shall be reasonable and scientific.
- OutputFormat: Returns the score result and comments in json format.
- Workflow:
    1. Read academic texts carefully to fully understand their content and structure.
    2. Based on Novelty, Significance, Relevance, Clarity and Soundness(Rigour) scoring criteria, individual scores ranging from 0 to 100 and corresponding comments are given respectively.
    3. Calculate the total score and return all score results and comments in json format.
- Examples:
- Example 1:
      {
       "Novelty": {"score": "95","comment":  "The text introduces a new theoretical framework that significantly advances the understanding of the topic, demonstrating the author's unique contribution."},
       "Significance": {"score": "90","comment":  "The findings have high practical applicability and can be widely referenced in the field,  contributing to solving real-world problems."}},
       "Relevance": {"score": "95","comment": "The content is highly relevant to the specified topic, with clear alignment and focus."},
       "Clarity": {"score": "95","comment":  "The expression is accurate and fluent, with error-free and concise writing, making the text easy to understand."},
       "Soundness(Rigour)": {"score": "85","comment":  "The methodology is precise, the data is reliable, and the arguments are valid,  but there is room for improvement in the sufficiency of evidence."},
       "Total Score": "92"
      }
- Example 2:
      {
       "Novelty": {"score": "50","comment": "The text presents any new insights, and the overall contribution to the field is limited."},
       "Significance": {"score": "60","comment": "The applicability and referential value are low, with potential for further development."},
       "Relevance": {"score": "73","comment": "The content is moderatly relevant to the topic, but some parts could be more closely aligned."},
       "Clarity": {"score": "75","comment": "The writing is generally clear, but there are a few minor errors and areas for improvement in fluency, for example..."},
       "Soundness(Rigour)": {"score": "78","comment": "The methodology is mostly sound, but there are some issues with data reliability and argument coherence."},
       "Total Score": "67.2"}
      }
```

```
- Role: Expert in academic content review
- Background: Users require an in-depth review of academic texts to assess their academic value and writing quality. They seek a comprehensive grading system to better understand and evaluate all aspects of academic texts.
- Profile: You are a senior expert with a background in library and information science, possessing extensive experience and a deep understanding of academic text review. You are capable of accurately assessing the novelty, significance, relevance, clarity, and soundness (rigour) of academic texts.
- Skills: You have the ability to:
  -Analyze the novelty of an academic text, evaluating the introduction of new ideas, theories, methods, or insights that advance the boundaries of knowledge and reflect the author's unique contribution.
  -Assess the significance of the text, determining its practical applicability and referential value in the respective field.
  -Judge the relevance of the text, ensuring its alignment with the specified topic and focus.
  -Evaluate the clarity of the text, assessing the accuracy, fluency, and conciseness of the writing, as well as its readability.
  -Test the soundness (rigour) of the text, examining the precision of the methodology, reliability of data, coherence of design, validity of arguments, sufficiency of evidence, and logical consistency.
- Goals: Based on the academic value and writing quality of the text, provide a score (0-100) and detailed comments for each of the following criteria: Novelty, Significance, Relevance, Clarity, and Soundness (Rigour). Calculate the total score based on these individual scores.
- Constraints:The evaluation process must be objective and fair. The scoring criteria must be clear and specific. The comments must be detailed and accurate. The total score calculation must be reasonable and scientifically sound.
- Output Format: Return the score results and comments in the following JSON format:
{
  "Novelty": {"score": "", "comment": ""},
  "Significance": {"score": "", "comment": ""},
  "Relevance": {"score": "", "comment": ""},
  "Clarity": {"score": "", "comment": ""},
  "Soundness(Rigour)": {"score": "", "comment": ""},
  "Total Score": ""
}
-  Workflow:
      1. Carefully read the academic text to fully understand its content and structure.
      2. Based on the criteria of Novelty, Significance, Relevance, Clarity, and Soundness (Rigour), assign individual scores (0-100) and provide corresponding comments for each criterion.
      3. Calculate the total score and return all score results and comments in the specified JSON format.
```

```
- Example 1:
      {
       "Novelty": {"score": "95","comment":  "The text introduces a new theoretical framework that significantly advances the understanding of the topic, demonstrating the author's unique contribution."},
       "Significance": {"score": "90","comment": "The findings have high practical applicability and can be widely referenced in the field,  contributing to solving real-world problems."}},
       "Relevance": {"score": "95","comment": "The content is highly relevant to the specified topic, with clear alignment and focus."},
       "Clarity": {"score": "95","comment":  "The expression is accurate and fluent, with error-free and concise writing, making the text easy to understand."},
       "Soundness(Rigour)": {"score": "85","comment":  "The methodology is precise, the data is reliable, and the arguments are valid,  but there is room for improvement in the sufficiency of evidence."},
       "Total Score": "92"
      }

- Example 2:
      {
       "Novelty": {"score": "50","comment": "The text presents any new insights, and the overall contribution to the field is limited."},
       "Significance": {"score": "60","comment": "The applicability and referential value are low, with potential for further development."},
       "Relevance": {"score": "73","comment": "The content is moderatly relevant to the topic, but some parts could be more closely aligned."},
       "Clarity": {"score": "75","comment": "The writing is generally clear, but there are a few minor errors and areas for improvement in fluency, for example..."},
       "Soundness(Rigour)": {"score": "78","comment": "The methodology is mostly sound, but there are some issues with data reliability and argument coherence."},
       "Total Score": "67.2"
      }
```

Fig. 1. Data acquisition process.

pre-trained generative AI models — OpenAI's GPT-4o, Anthropic's Claude-3, Google's Gemini-2.0, xAI's Grok-2, Alibaba Cloud's Qwen2.5, ByteDance's Doubao-pro, Tencent's Hunyuan-large, and DeepSeek's Deepseek-v3 — were accessed through their official APIs. The models produced quantitative scores and qualitative comments, facilitating a thorough and dependable evaluation process. These models were selected based on their strong performance in recent large-scale human preference benchmarks, such as Chatbot Arena (Chiang et al., 2024), which has demonstrated high agreement with expert judgments. As such, they represent state-of-the-art capabilities across different architectures and providers, and reflect a diversity of real-world use cases in both international and Chinese-language academic contexts.

During API invocations, certain models demonstrated inadequate prompt adherence, leading to unparseable data and a loss of 0 to 5 data points per evaluation of 100 submissions. To maintain data integrity and reliability, we utilized a distribution-based imputation strategy by calculating the mean and standard deviation of existing scores and generating normally distributed random scores range from 60 to 100. A minimum standard deviation threshold of 2 maintained natural variability. This approach preserves the original data distribution and prevents distortions associated with simpler techniques such as mean substitution.

3.2. Methods

The three-step research analysis framework, as illustrated in Fig. 2, is used to investigate the capabilities of generative AI in academic content evaluation. In Step 1, academic content was collected by recruiting volunteers to generate diverse materials across predefined propositions within three months, where volunteers were recruited from a pool of students who major in Library and

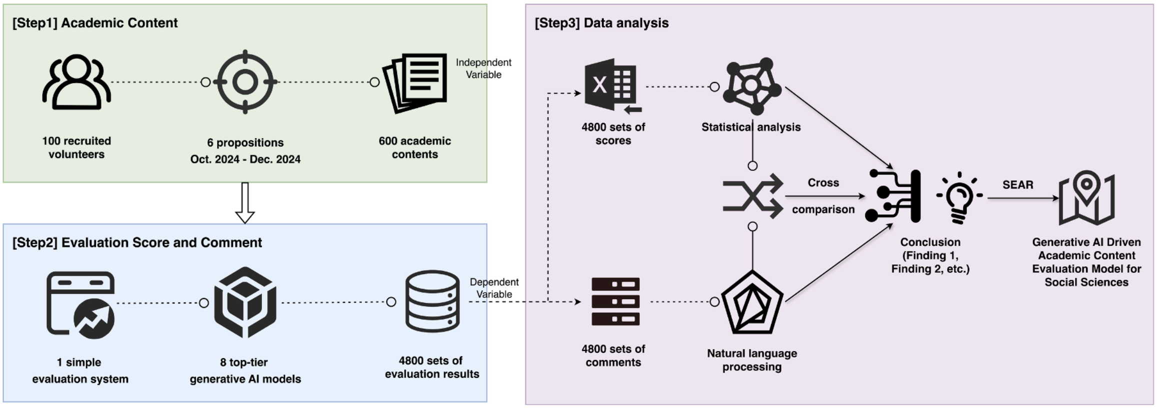

[Step1] Academic Content
100 recruited volunteers
6 propositions
Oct. 2024 - Dec. 2024
600 academic contents
Independent Variable
[Step2] Evaluation Score and Comment
1 simple evaluation system
8 top-tier generative AI models
4800 sets of evaluation results
Dependent Variable
[Step3] Data analysis
4800 sets of scores
Statistical analysis
Cross comparison
4800 sets of comments
Natural language processing
Conclusion (Finding 1, Finding 2, etc.)
SEAR
Generative AI Driven Academic Content Evaluation Model for Social Sciences

Fig. 2. Three-step analysis framework for AI evaluation.

Information Science and had already enrolled in the course, ensuring a homogeneous group of participants. In Step 2, 8 state-of-the-art generative AI models were employed to assess the contents by the simple evaluation system, yielding a comprehensive dataset of scores and comments. This systematic framework ensured consistency across all AI evaluations, minimizing potential biases in the evaluation process (Huang, Huang et al., 2025). In Step 3, the collected data was subjected to quantitative analysis, employing statistical techniques (Field et al., 2012) and NLP methods (Jurafsky & Martin, 2008) to extract key findings on the performance of AI in academic content evaluation, and a social science academic content evaluation model was constructed based on the System of All-round Evaluation of Research (Ye, 2010). To maintain the rigor of the statistical analysis, non-parametric tests, such as the Wilcoxon Signed-Rank Test and the Kruskal-Wallis H Test, were used, as these methods are suitable for the paired and non-normally distributed data used in this study. Additionally, Spearman's Rank Correlation Coefficient was calculated to assess the relationships between various scoring dimensions.

Since each API call is a single-round dialogue devoid of context or memory (OpenAI, 2025), which corresponds with the statistical assumption of independent events. This alignment guarantees the statistical validity of the analysis concerning the models, scoring dimensions (scores and comments), and their interrelations (Field et al., 2012). Therefore, this study employed statistical methods to assess the evaluation results returned by generative AI, focusing on comparing the scoring effects of the zero-shot and few-shot prompt strategies (Brown et al., 2020; P. Liu et al., 2023). Additionally, it further analyzed the sentiment inclination of the model outputs and their correlation with the numerical ratings. Statistical methods were systematically applied to ensure robust results and reduce potential biases, facilitating a thorough comparison of AI performance across different models and strategies.

The statistical methods employed are detailed below:

(1) Comparison of zero-shot and few-shot strategies with human scoring: The paired-sample Wilcoxon Signed-Rank Test was used to assess the differences between AI scores under the zero-shot and few-shot strategies and the scores assigned by human experts. This non-parametric test is suitable for paired data, maintaining statistical validity in the presence of non-normal distributions (Mohd, 2011).

(2) Descriptive statistical analysis of scoring dimensions: Descriptive statistics were conducted for each scoring dimension, including the calculation of the mean, standard deviation, skewness, and kurtosis (Field et al., 2012). These statistics help reveal the distribution characteristics of the scoring data and evaluate the presence of skewed or peaked distributions.

(3) Correlation analysis between scoring dimensions: Spearman's Rank Correlation Coefficient was used to investigate the interrelationships among the scoring dimensions (Mohd, 2011). This method effectively evaluates nonlinear correlations between dimensions, thereby uncovering their fundamental connections.

(4) Comparison of score distributions across models: The Kruskal-Wallis H Test and the paired-sample Wilcoxon Signed-Rank Test were used to analyze performance differences among different generative AI models across the scoring dimensions. These tests accommodate non-normal distributions and offer a thorough comparison of model performance (Tomczak & Tomczak, 2014).

(5) Sentiment analysis of comments: A sentiment analysis tool using the RoBERTa model (Liu et al., 2019) was employed to compute the sentiment score of each comment produced by generative AI. In contrast to existing studies that focus solely on direct numerical scores, our analysis incorporates both quantitative results and the sentiment inclination of qualitative comment outputs, enabling a richer understanding of model evaluation behavior. Subsequently, descriptive statistics and

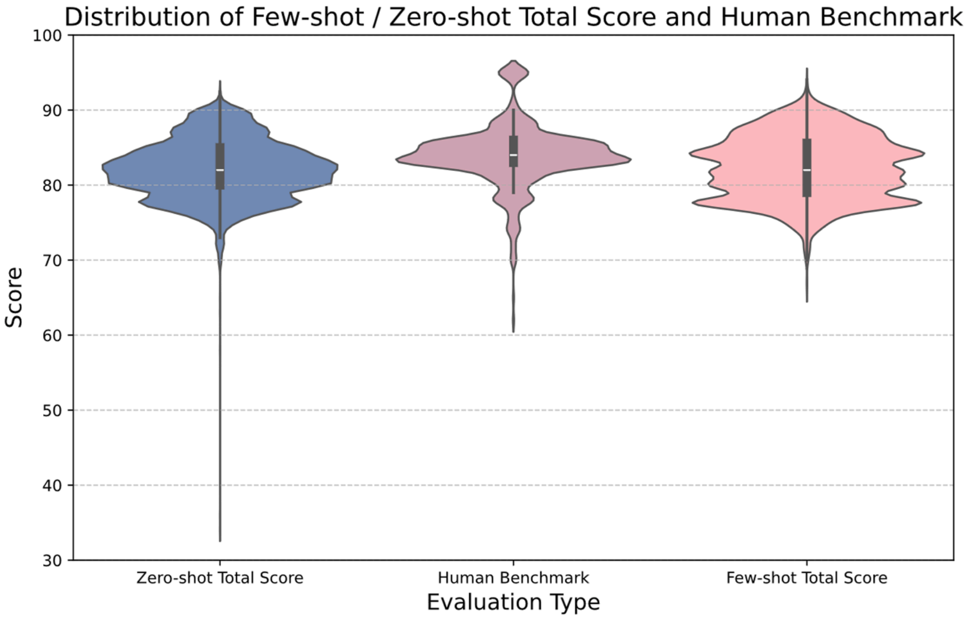


Fig. 3. Distribution of AI total scores vs. human benchmarks.

Spearman's correlation were applied to examine the distribution of sentiment scores and their correlation with the scores across various dimensions, highlighting the performance disparities between models in scoring and sentiment analysis.

4. Results

4.1. Comparison of total scores with human benchmark scores

Initially, we invited two experts in the field to perform an overall evaluation of the 600 academic content samples collected, using the simple evaluation system. Considering the workload, each expert was assigned a distinct set of samples for each proposition, and a single human total score benchmark was provided. The experts engaged in calibration discussions and strictly adhered to standardized evaluation rubrics to maintain internal consistency of scores. Subsequently, we conducted the Shapiro-Wilk normality test on the zero-shot total score (ZSTS) and few-shot total score (FSTS) generated by the two strategies. The results demonstrated that neither data sets adhere to a normal distribution ($SW_{Zs} = 0.983$, $p < 0.01$; SW_{Fs} =0.993, $p < 0.01$). Therefore, we conducted paired-sample Wilcoxon Signed-Rank Tests for both sets of scores in comparison to the human total score benchmark (Mohd, 2011). Fig. 3 illustrates the distribution of ZSTS and FSTS. Notably, the AI-generated total scores demonstrated limited reliability, with a precision of 66.35 %, defined as the proportion of cases where the AI-generated total score matched the arithmetic average of its own five sub-dimension scores. Given this inconsistency, we concluded that the AI-generated total scores were unreliable and instead adopted the computed weighted average of the sub-scores for subsequent analysis.

The results demonstrate that both ZSTS and FSTS exhibit medium effect size differences from the human benchmark. However, the effect size for FSTS compared to the human benchmark (Cohen's $\mathrm{d}_{\mathrm{FSTS}}$= 0.382) is less than that for ZSTS compared to the human benchmark (Cohen's $\mathrm{d}_{\mathrm{ZSTS}} = 0.467$). The lower bound of ZSTS is reduced, with a minimum score of 34. This result indicates that the score distribution of few-shot method aligns more closely with the human benchmark.

4.2. Comparison of score distributions

4.2.1. Across different dimensions

Table 3 presents the descriptive statistics of score distributions under both prompt strategies. Compared to few-shot, the zero-shot strategy shows stronger left skewness, indicating a greater tendency toward lower scores. Its kurtosis is also higher, suggesting a more peaked distribution with heavier tails—implying a higher probability of extreme values or outliers. Notably, scores across all dimensions exhibit significant variation depending on the prompt strategies employed. However, in the dimensions of novelty and significance within academic value, few-shot shows larger standard deviations and wider scoring ranges, although the effect sizes are small (Cohen's $\mathrm{d}_{\mathrm{Novelty}} = -0.161$ and Cohen's $\mathrm{d}_{\mathrm{Significance}} = 0.165$). In contrast, for the three writing quality dimensions, few-shot yields smaller standard deviations than zero-shot, indicating more concentrated scores. This implies that zero-shot provides better score differentiation for writing quality (average standard deviation = 5.38, $p = 0.000$).

Fig. 4 shows the score distributions across all dimensions. Outliers (shown as square dots) highlight academic content that does not meet evaluation criteria. For example, in the zero-shot group, a set of outliers (20, 30, 40, 50, 30) was identified, which were actually caused by incorrectly submitted (non-matching proposition) experimental samples. We did not manually remove these, but instead observed whether the model could identify and filter out irrelevant texts. Results show that under zero-shot conditions, the model successfully identified and excluded these non-compliant samples.

Fig. 5 depicts the correlations among evaluation dimensions under both strategies. In the few-shot group, most dimensions are moderately to highly correlated, with an average correlation of 0.525. This suggests that the scores for the two primary dimensions exhibit a strong synchrony, which should ideally have weak correlations. This indicates limited score differentiation capabilities for the few-shot strategy. In contrast, the zero-shot group exhibits generally lower correlations between dimensions. For instance, the correlation between clarity and significance is only 0.13 ($p = 0.000$). This implies that the zero-shot strategy enhances evaluation accuracy during scoring.

Table 3
Comparison of score distribution between zero-shot and few-shot.

	Skewness Variation (zero-shot to few-shot)	Kurtosis Variation (zero-shot to few-shot)	Standard Deviation Variation (zero-shot to few-shot)	p	Cohen's d
Novelty	−0.459→−0.193	1.030→−1.119	6.940→7.570	0.000 (***)	−0.161
Significance	−0.698→−0.208	3.039→−0.324	5.180→5.690	0.000 (***)	0.165
Relevance	−0.497→−0.490	6.900→1.042	5.610→5.380	0.000 (***)	−0.401
Clarity	−0.731→−0.180	1.082→−0.298	4.30→3.390	0.000 (***)	0.441
Soundness	−0.422→−0.137	0.961→0.072	6.240→4.350	0.000 (***)	−0.125

Note: The symbol "***", "**", and "*" indicate statistical significance at the levels of $p < 0.001$, 0.01, and 0.05, respectively.

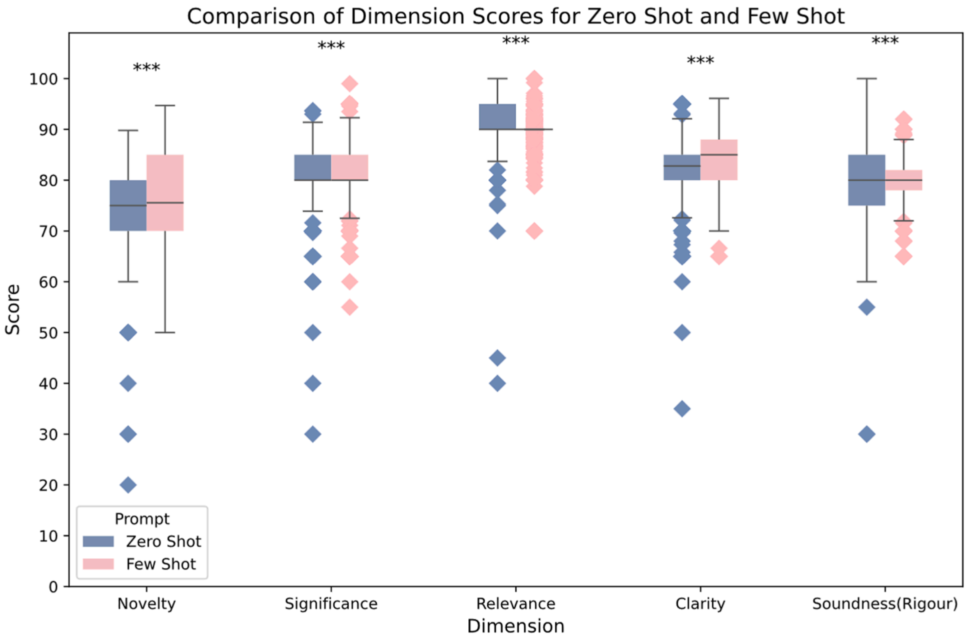


Fig. 4. Score distributions across evaluation dimensions with identifiable outliers.
Note: The symbol "***", "**", and "*" indicate statistical significance at the levels of $p < 0.001$, 0.01, and 0.05, respectively.

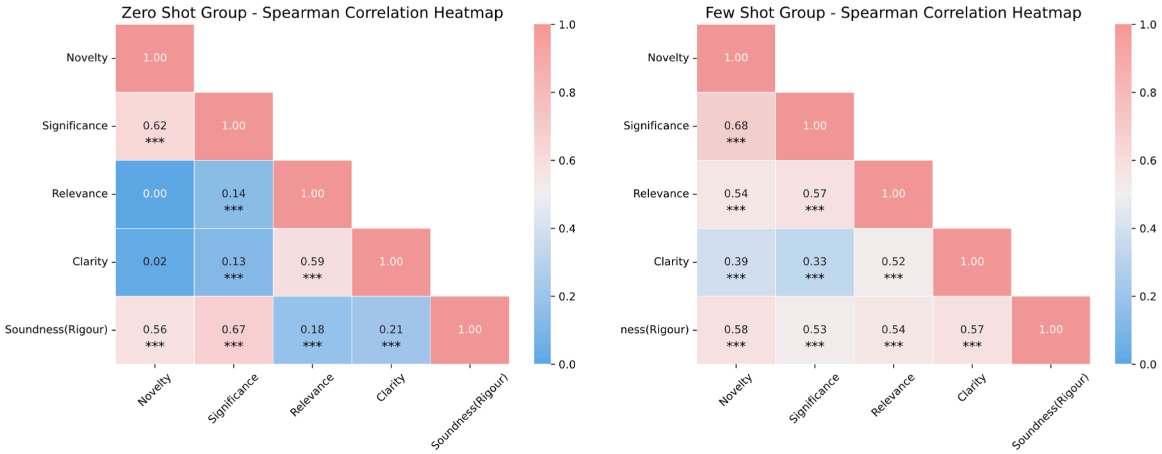


Fig. 5. Correlations between evaluation dimensions across both prompt strategies.
Note: The symbol "***", "**", and "*" indicate statistical significance at the levels of $p < 0.001$, 0.01, and 0.05, respectively.

4.2.2. *Across different models*

Fig. 6 presents the score distributions of various models across different dimensions, sorted by median scores. Following the categorization of various prompt strategies and dimensions, a Kruskal-Wallis test was conducted based on the groupings of models. The results indicated significant differences among the models overall (all $p < 0.001$).

Table 4 presents the results of the Mann-Whitney U test, which compares the mean differences between each model and the overall mean within the few-shot and zero-shot strategy groups.

Regarding score averages, both prompt strategies indicated that Qwen2.5 and Doubao-pro assigned higher scores ($\text{Mean}_{\text{Qwen2.5}} = 83.885$; $\text{Mean}_{\text{Doubao-pro}} = 84.07$), whereas Grok-2 and Gemini-2.0 assigned lower scores ($\text{Mean}_{\text{Grok-2}} = 80.315$; $\text{Mean}_{\text{Gemini-2.0}} = 80.355$). The scoring discrepancies between Deepseek-v3 and Hunyuan-large were significant under both the few-shot and zero-shot conditions ($\text{Difference}_{\text{Deepseek-v3}} = 2.55$; $\text{Difference}_{\text{Hunyuan-large}} = 2.67$). Furthermore, it is worth noting the diagnosis of outliers. As mentioned earlier, erroneous sample data was submitted. Although identified as outliers in the zero-shot condition, only Hunyuan-large (20, 30, 40, 50, 20) and Deepseek-v3 (30, 50, 70, 80, 60) appropriate scores across all dimensions, whereas other models did not recognize their irrelevance. For instance, Grok-2 in the zero-shot condition assigned a score of 100 for relevance to this erroneous

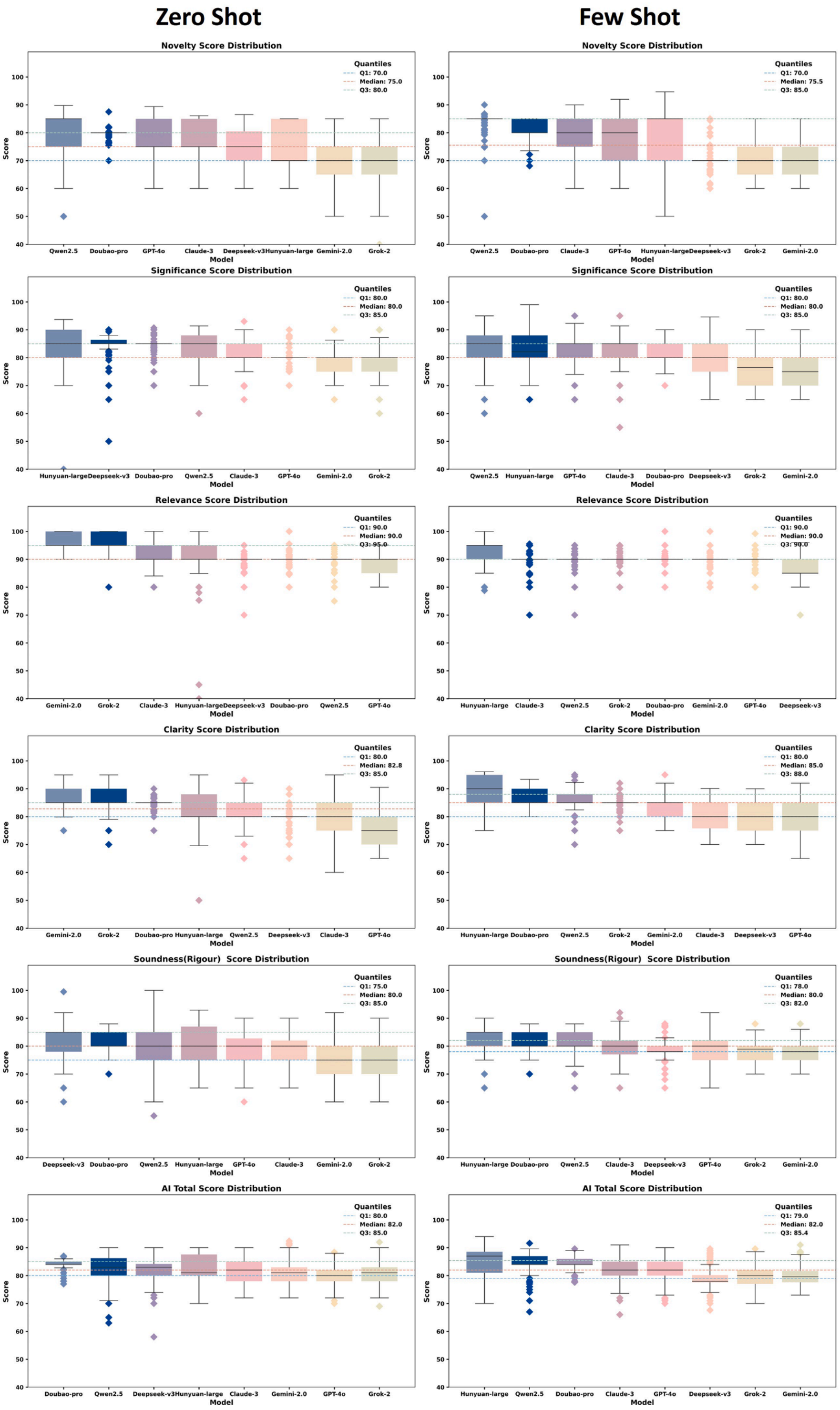


Fig. 6. Score distributions across models and dimensions.

Table 4
Significance test results of score differences across different models.

Zero-Shot					Few-Shot				
Model	Mean	U Statistic	*P* value	Result	Model	Mean	U Statistic	*P* value	Result
Grok-2	80.70	1155,041	0.000 (***)	R1	Gemini-2.0	79.67	743,543.5	0.000 (***)	R1
GPT-4o	80.85	1153,115	0.000 (***)	R1	Grok-2	79.93	785,069.5	0.000 (***)	R1
Gemini-2.0	81.04	1218,031.5	0.000 (***)	R1	Deepseek-v3	80.06	1450,355.5	0.000 (***)	R1
Claude-3	81.51	1096,868.5	0.000 (***)	R1	GPT-4o	82.03	1407,501	0.366	R2
Hunyuan-large	82.54	1481,070	0.253	R2	Claude-3	82.52	1247,297	0.145	R2
Deepseek-v3	82.61	1826,620.5	0.000 (***)	R3	Doubao-pro	84.32	1883,137.5	0.000 (***)	R3
Qwen2.5	82.76	1627,464	0.000 (***)	R3	Qwen2.5	85.01	2014,749	0.000 (***)	R3
Doubao-pro	83.82	1961,789.5	0.000 (***)	R3	Hunyuan-large	85.21	1988,347	0.000 (***)	R3
Overall	81.98				Overall	82.34			

* Note: The symbol "***", "**", and "*" indicate statistical significance at the levels of $p < 0.001$, 0.01, and 0.05, respectively. R1 indicates significantly lower value than the overall mean; R2 indicates no significant difference from the overall mean; R3 indicates significantly higher value than the overall mean.

sample. This highlights that notable performance differences exist among models, with considerable variation in their stability.

4.3. Comparison of sentiment inclination scores of comments

4.3.1. Across different dimensions

For sentiment analysis, we employed a pre-trained RoBERTa-based sentiment classification model (Liu et al., 2019) to assign each AI-generated comment a sentiment inclination score ranging from −1 (strongly negative) to +1 (strongly positive), where 0 indicates a neutral sentiment. These scores were computed using a continuous scale rather than discrete labels, and the absolute value of the score reflects the intensity of sentiment expression. The model was chosen for its robust performance in fine-grained sentiment detection in human text (Lengkeek et al., 2023).

As shown in Fig. 7, regardless of whether few-shot or zero-shot prompts are used, the comments generated by generative AI consistently exhibit a positive tone across all dimensions. Specifically, the zero-shot prompt typically receives higher sentiment scores, with the most significant differences observed in the novelty (median difference = 0.193) and soundness (median difference = 0.065) dimensions. This indicates that comments generated with the zero-shot prompt are tend to be more positively inclined.

Table 5 presents a comparison of score differences between samples receiving positive and negative comments. In both the few-shot and zero-shot groups, positive comments were associated with higher scores across all five dimensions. Notably, few-shot showed a difference of 7.28 for novelty, whereas zero-shot exhibited a difference of 5.29 for novelty 2 and 3.39 for soundness, revealing a more pronounced score disparity between positive and negative comments.

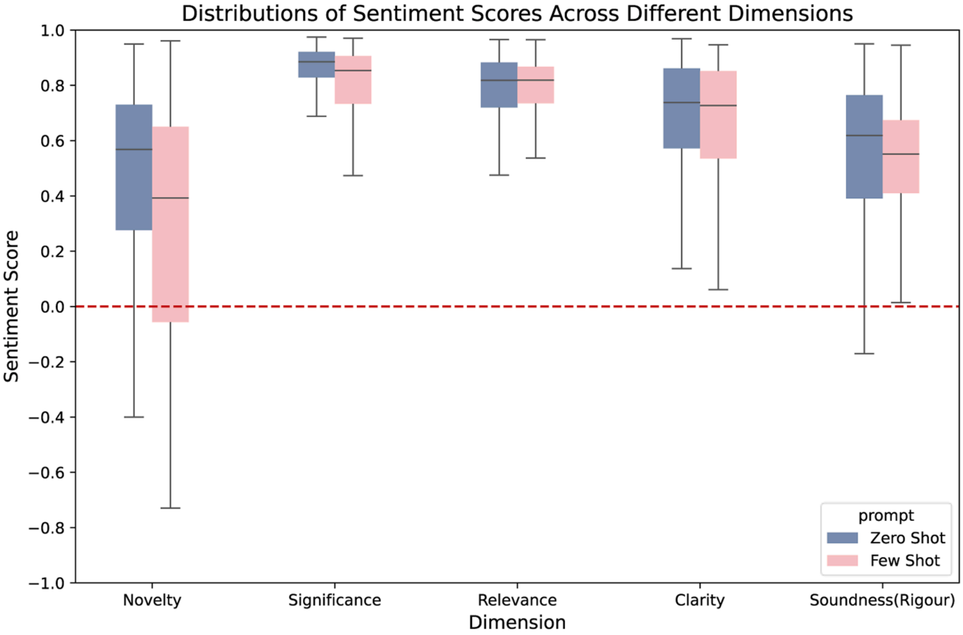
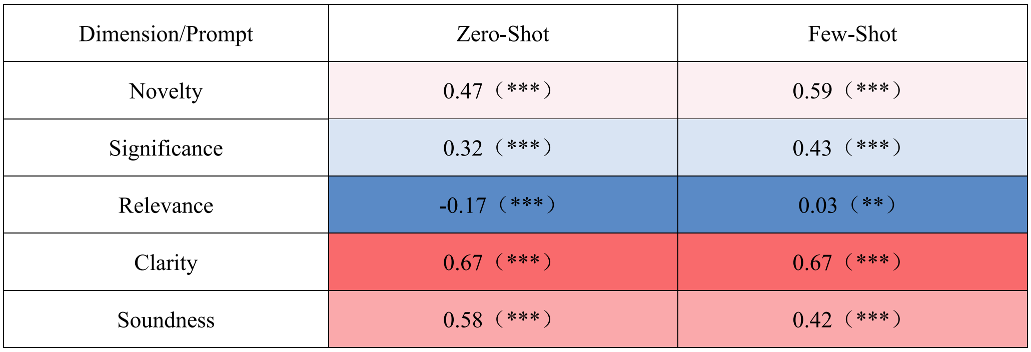


Fig. 7. Sentiment inclination of AI-generated comments across evaluation dimensions.

Table 5
Corresponding mean values of positive and negative evaluations.

	Zero-Shot groups		Few-Shot groups	
	Mean of positive evaluations	Mean of negative evaluations	Mean of positive evaluations	Mean of negative evaluations
Novelty	76.25	70.96	78.95	71.67
Significance	82.47	79.49	82.31	78.62
Relevance	91.48	90.66	90.29	88.30
Clarity	81.99	80.97	84.63	82.87
Soundness	79.59	76.20	80.61	78.06

Table 6
Correlation results between AI scores and sentiment inclination.



Dimension/Prompt	Zero-Shot	Few-Shot
Novelty	0.47（***）	0.59（***）
Significance	0.32（***）	0.43（***）
Relevance	-0.17（***）	0.03（**）
Clarity	0.67（***）	0.67（***）
Soundness	0.58（***）	0.42（***）

Note: The symbol “***”, “**”, and “*” indicate statistical significance at the levels of $p < 0.001$, 0.01, and 0.05, respectively.

Subsequently, we performed Spearman correlation tests to analyze the relationship between the sentiment inclination of the comments and the AI scores across different dimensions, as presented in Table 6.

Table 6 indicates that clarity ($\text{Correlation}_{\text{Few-Shot}} = 0.67$; $\text{Correlation}_{\text{Zero-Shot}} = 0.67$) and novelty ($\text{Correlation}_{\text{Few-Shot}} = 0.59$; $\text{Correlation}_{\text{Zero-Shot}} = 0.47$) demonstrate strong positive correlations in both strategies, suggesting that the sentiment inclination of their comments significantly influences the scores. Conversely, relevance ($\text{Correlation}_{\text{Few-Shot}} = 0.03$; $\text{Correlation}_{\text{Zero-Shot}} = -0.17$) shows a weak or even negative correlation, indicating that sentiment inclination has a limited influence on this dimension. For the soundness dimension, the correlation for the zero-shot prompt (Correlation $= 0.58$) is higher than that for the few-shot prompt (Correlation $= 0.42$). These results indicate that zero-shot provides better differentiation in dimensions related to writing quality, whereas few-shot is more suitable for evaluating academic quality dimensions.

The sentiment scores and score range for the relevance dimension are relatively narrow, with the majority of samples exhibiting sentiment scores predominantly in the neutral or positive range (e.g., most sentiment scores fall between 0.5 and 1). As a result, the limited variation in sentiment scores does not align well with changes in evaluative ratings, making it difficult for sentiment shifts to effectively differentiate this dimension, which leads to weaker correlations for relevance.

4.3.2. Across different models

Table 7 presents the Spearman correlation coefficients between the sentiment scores and the scores for each dimension and prompt strategy across different models.

The results indicate that the Doubao-pro model exhibits inferior performance compared to other models regarding sentiment inclination and scoring correlation (correlation_Doubao-pro_avg = 0.21, correlation_Overall_Avg = 0.38). Regardless of the dimension or prompt strategy, Doubao-pro did not perform at the level of the other models. Consequently, the following discussions will omit the impact of the Doubao-pro model.

In terms of novelty, the few-shot setting yields more consistent results, with most models, excluding Doubao-pro and GPT-4o, showing moderate correlations between comment sentiment and scores (ranging from 0.43 to 0.56). Among them, GPT-4o demonstrates the highest alignment ($r = 0.711$), suggesting its superior coherence between generated sentiments and ratings. By contrast, the zero-shot strategy exhibits considerable variability, with correlations spanning from 0.35 to 0.71. In this setting, Qwen2.5 emerges as the top performer, with a strong correlation of 0.708.

For the significance dimension, the few-shot setting again outperforms zero-shot in general. Although correlation levels vary widely under the few-shot approach (0.18 to 0.65), DeepSeek-v3 stands out with the highest correlation of 0.648. Under zero-shot conditions, however, all models exhibit notably weaker performance, with correlations limited to the 0.18–0.37 range, indicating difficulty in aligning sentiment with evaluative scores for this abstract criterion.

The relevance dimension presents challenges across both strategies. Only a few models achieve statistically significant

Table 7
Correlation between sentiment score and given score across different models.

Model	Novelty		Significance		Relevance		Clarity		Soundness (Rigour)		Average
	Few-Shot	Zero-Shot	Few-Shot	Zero-Shot	Few-Shot	Zero-Shot	Few-Shot	Zero-Shot	Few-Shot	Zero-Shot	
Claude-3	0.556 (***)	0.406 (***)	0.346 (***)	0.182 (***)	0.106(*)	-0.064	0.629 (***)	0.655 (***)	0.516 (***)	0.627 (***)	0.396
Deepseek-v3	0.437 (***)	0.356 (***)	0.648 (***)	0.334 (***)	0.188 (***)	-0.016	0.749 (***)	0.486 (***)	0.432 (***)	0.594 (***)	0.421
Doubao-pro	0.304 (***)	0.257 (***)	0.187 (***)	0.248 (***)	-0.019	0.062	0.155 (***)	0.148 (***)	0.249 (***)	0.326 (***)	0.215
Gemini-2.0	0.453 (***)	0.593 (***)	0.475 (***)	0.373 (***)	0.044	-0.144 (***)	0.408 (***)	0.388 (***)	0.364 (***)	0.546 (***)	0.384
GPT-4o	0.711 (***)	0.575 (***)	0.486 (***)	0.274 (***)	0.048	0.059	0.491 (***)	0.568 (***)	0.513 (***)	0.612 (***)	0.477
Grok-2	0.558 (***)	0.592 (***)	0.581 (***)	0.350 (***)	0.148 (**)	-0.243 (***)	0.391 (***)	0.400 (***)	0.523 (***)	0.526 (***)	0.383
Hunyuan-large	0.508 (***)	0.349 (***)	0.185 (***)	0.052	-0.005	0.071	0.66 (***)	0.403 (***)	0.289 (***)	0.373 (***)	0.315
Qwen2.5	0.520 (***)	0.708 (***)	0.477 (***)	0.276 (***)	0.197 (***)	0.06	0.508 (***)	0.643 (***)	0.353 (***)	0.701 (***)	0.445
Overall	0.594 (***)	0.472 (***)	0.431 (***)	0.321 (***)	0.029 (**)	-0.167 (***)	0.672 (***)	0.671 (***)	0.423 (***)	0.577 (***)	0.379

* Note: The symbol "***", "**", and "*" indicate statistical significance at the levels of $p < 0.001$, 0.01, and 0.05, respectively.

correlations—four in the few-shot and two in the zero-shot setting. Even then, the correlations remain weak (e.g., 0.197 and −0.17), which may be attributable to the narrow distribution range of both the sentiment and scoring values, leading to limited variability for reliable association.

When evaluating clarity, model performance shows clearer distinctions. In the few-shot scenario, DeepSeek-v3 achieves the highest correlation (0.749), while Claude-3 also performs reliably (0.629). Gemini-2.0 and Grok-2, on the other hand, show relatively poor alignment between sentiment and score (0.408 and 0.391, respectively). In the zero-shot setting, Claude-3 leads with a correlation of 0.655, confirming its overall strong and stable performance in interpreting clarity-related dimensions.

Finally, in the soundness dimension, the few-shot approach yields correlations in the low-to-moderate range (0.249 to 0.539), with Claude-3, GPT-4o, and Grok-2 showing better alignment around the 0.5 level. Under zero-shot prompting, however, the performance improves for most models—excluding Doubao-pro and Hunyuan-large—with correlations often exceeding 0.5. Notably, Qwen2.5 performs best here, with a correlation of 0.701, indicating its relative strength in evaluating logical coherence and argumentation soundness without examples.

In summary, each model demonstrates different strengths across various dimensions and tasks. When evaluating clarity with Claude-3, both strategies demonstrate consistent and robust performance. Conversely, GPT-4o and DeepSeek-v3 demonstrate strengths in different sentiment and scoring aspects.

5. Discussion

5.1. Applicability of different prompting strategies

The experimental results demonstrate that both zero-shot and few-shot strategies can produce structured JSON evaluation scores and comments, with significant differences in their score distributions.

First, the superior alignment of few-shot outputs with human scores reflects the anchoring effect established by human-provided examples. This can be understood through Tversky and Kahneman's decision framing theory (Tversky & Kahneman, 1981) where examples act as anchors in constructing a semantic-to-numeric mapping. Additionally, from a computational perspective, few-shot prompting can be regarded as a form of prompt-based transfer learning, where LLMs leverage in-context information to fine-tune their inference path. Studies in instruction tuning and in-context learning (Brown et al., 2020; Wei et al., 2022) suggest that examples in the prompt help constrain the model's generative space, leading to more predictable and human-aligned outputs. While this improves consistency and narrows distributional gaps (Cohen's $d_{FSTS}= 0.382$, Cohen's $d_{ZSTS} = 0.467$), it may simultaneously limit the model's flexibility in interpreting ambiguous criteria. These trade-offs reveal the conflict between score stability and interpretive autonomy in AI-assisted evaluation. Prior researches (Liu et al., 2023; Lyu et al., 2023) also suggest that examples in few-shot learning may cause overfitting to prompt structure, leading to lower sensitivity to unrepresented dimensions. In contrast, zero-shot lacks the decision framework supported by examples, resulting in a notably different numerical distribution across dimensions when compared to few-shot ($p < 0.01$).

Second, in more structured evaluation dimensions such as soundness and relevance, zero-shot prompts outperform few-shot in discriminatory capacity ($Std_{Zero-Shot} = 6.24$, $Std_{Few-Shot} =4.35$). This suggests that for tasks characterized by low conceptual ambiguity and strong formal constraints, generative AI can leverage latent semantic representations independently of anchoring. Similar observations are reported in Thelwall et al. (2025), who finds that zero-shot performance frequently surpasses expectations in domains where predefined heuristics prevail in evaluative judgments.

However, the inflated scores in dimensions such as novelty and significance indicate a systemic overvaluation tendency. This presents significant challenges for social sciences (Thelwall, 2024), as evaluative criteria are inherently interpretive and domain-sensitive. In our dataset, composed of relatively standard academic writings by master's students, AI consistently assigned higher-than-expected scores ($Median_{Zero-Shot\_Novelty} = 75$, $Median_{Zero-Shot\_Significance} = 80$; $Median_{Few-Shot\_Novelty} = 75.55$, $Median_{Few-Shot\_Significance} = 80$), failing to differentiate exceptional contributions from average ones. This echoes concerns raised by Huang et al. (2025), who contend that LLMs lack calibrated skepticism in assessing originality and are prone to excessive leniency. In contrast to the recommendation of employing zero-shot strategies for originality evaluation (Huang, Huang et al., 2025), our results suggest that few-shot prompting is more appropriate for complex, judgment-intensive tasks. This approach could offer the model purpose-specific cues and mitigates its leniency.

Consequently, this study emphasizes the necessity for prompt strategies to align with both evaluation criteria (e.g., relevance versus novelty) and task purposes (e.g., formative feedback versus summative judgment). Our findings support a strategy-task alignment principle, advocating a hybrid evaluation framework wherein zero-shot prompts facilitate initial filtering and few-shot prompts support more contextualized, fine-grained assessments.

5.2. Correlation between sentiment scores and numeric ratings

This study demonstrates a moderate positive correlation between the sentiment polarity of generative AI comments and their numeric ratings (zero-shot mean sentiment = 0.657; few-shot = 0.597), especially in dimensions such as novelty and clarity ($r =$ 0.472–0.672). This correlation suggests that AI-generated textual feedback is not random but contains evaluative semantic signals—an important finding that has not been addressed in previous research, where the alignment of sentiment and scores in AI-generated academic reviews has largely been overlooked.

However, it is also necessary to acknowledge that the consistently high positivity across outputs indicates a deficiency in discriminative capacity. This observation aligns with Shuster et al. (2022), who characterize LLMs as embodying a helpful persona that prioritizes supportive over critical feedback. Extending this perspective, our study introduces the concept of social personality favoritism in AI-based academic evaluation, a phenomenon similarly identified by Huang et al. (2025) in the biomedical domain, where inflated assessments were observed.

This favoritism seems to be supported by two primary factors: system-level prompt constraints (e.g., instructions to "remain positive" or "be helpful," as noted by Zheng et al. (2024)), and the self-promotional tone often found in author-submitted manuscripts to game the system (Thelwall & Yaghi, 2024). For instance, the reasoning behavior of Hunyuan-T1 model was found to be tightly constrained by prompts such as "maintain positivity and avoid negative emotions" (Tencent, 2025). As a result, its output represents a combination of technical competence and commercial alignment, rather than objective critique. This compromises the neutrality, analytical precision, and evaluative sharpness essential for academic peer review.

More broadly, sentiment-related bias may threaten the fairness and reliability of AI-assisted academic assessment. If AI-generated reviews systematically favor positive tone regardless of content quality, this may obscure real distinctions between high- and low-quality work, particularly when comments are used for consequential decisions like publication, funding, or hiring. It may also amplify representational inequalities—e.g., privileging fluently written or self-promoting texts while disadvantaging critical, exploratory, or stylistically diverse submissions, often produced by early-career researchers or non-native speakers. To address these risks, prompts should be adjusted to support balanced rather than overly positive feedback. In addition, using multiple AI agent (Huang, Wang et al., 2025) with both critical and supportive perspectives can lead to more comprehensive evaluations.

This study empirically tests the correlation between sentiment and score, revealing a systematic positivity bias in AI feedback. By critically examining its underlying drivers and implications, we contribute to a more nuanced understanding of how generative AI systems may both reflect and reinforce normative biases in academic discourse. It offers a new interpretive perspective on the behavior of generative AIs in academic evaluation contexts.

5.3. Competence of different generative AI models

Amid rapid technological iteration and intense competition among large models, this study does not seek to statically rank the performance of various models. Instead, we focus on the deep impacts of model heterogeneity and systematic biases on academic evaluation methods. The results indicate significant differences across models (all $p < 0.001$), with no individual model exhibiting stable performance across all conditions—aligning with previous research on intra-model variability and hallucination, such as GPT4o (Thelwall, 2024).

The variability is primarily attributed to differences in architecture and training among LLMs. For example, GPT-4o and Claude-3 exhibit distinct core design philosophies. GPT-4o employs a unified multimodal transformer trained with reinforcement learning from human feedback (OpenAI, 2024), which is closer to evaluators' demands for critical feedback. In contrast, Claude-3, developed by Anthropic (2024), is trained using the constitutional AI framework that prioritizes safety, helpfulness, and honesty through rule-based instruction tuning (Bai et al., 2022). These differences shape the models' stylistic outputs and their sensitivity to uncertainty, risk aversion, and handing of ambiguous or novel content, which directly affects evaluative tasks like academic peer review. In particular, models that are trained with more robust alignment objectives may tend to provide feedback that is more diplomatically phrased, which could restrict critical analysis (Ganguli et al., 2023; Khatun & Brown, 2023).

Moreover, this study's evaluation experiment employed a single-round dialogue API call, indicating that generative AI lacks the ability, like human reviewers, to dynamically adjust and calibrate evaluation standards based on the broader context of multiple academic papers. As a result, in few- and zero-shot scenarios, various models such as Deepseek-v3 and Hunyuan-large show significant scoring differences ($p < 0.01$). Therefore, rather than relying on several iterations from a single model as Thelwall (2024) suggested, aggregating outputs from multiple distinct state-of-the-art LLMs (Li et al., 2024) may better mitigate individual model biases and yield more robust and equitable assessments. To further enhance reliability, we recommend that models be prompted to simultaneously provide numeric ratings, qualitative comments and the specific textual evidence supporting their judgments (as a demonstration shown in Appendix C). This traceable output would enable human experts to conduct meta-evaluations and audits of the AI's reasoning process.

Critically, our findings underscore that generative AI faces challenges in assessing dimensions that necessitate profound conceptual abstraction, including novelty and significance—key aspects of scientific contribution (Thelwall, 2025a). The identified limitations arise from the models' reliance on pattern recognition derived from established knowledge distributions (Kocoń et al., 2023). Consequently, papers presenting disruptive or paradigm-shifting ideas may be systematically undervalued by models trained predominantly on conventional academic discourse.

5.4. "Scenario - strategy – support" academic content evaluation model

In light of the preceding discussion, we develop the "Scenario - Strategy – Support" Academic Content Evaluation Model (SSS-ACE Model), using the System of All-round Evaluation of Research proposed by Ye (2021), as illustrated in Fig. 8. This model offers a

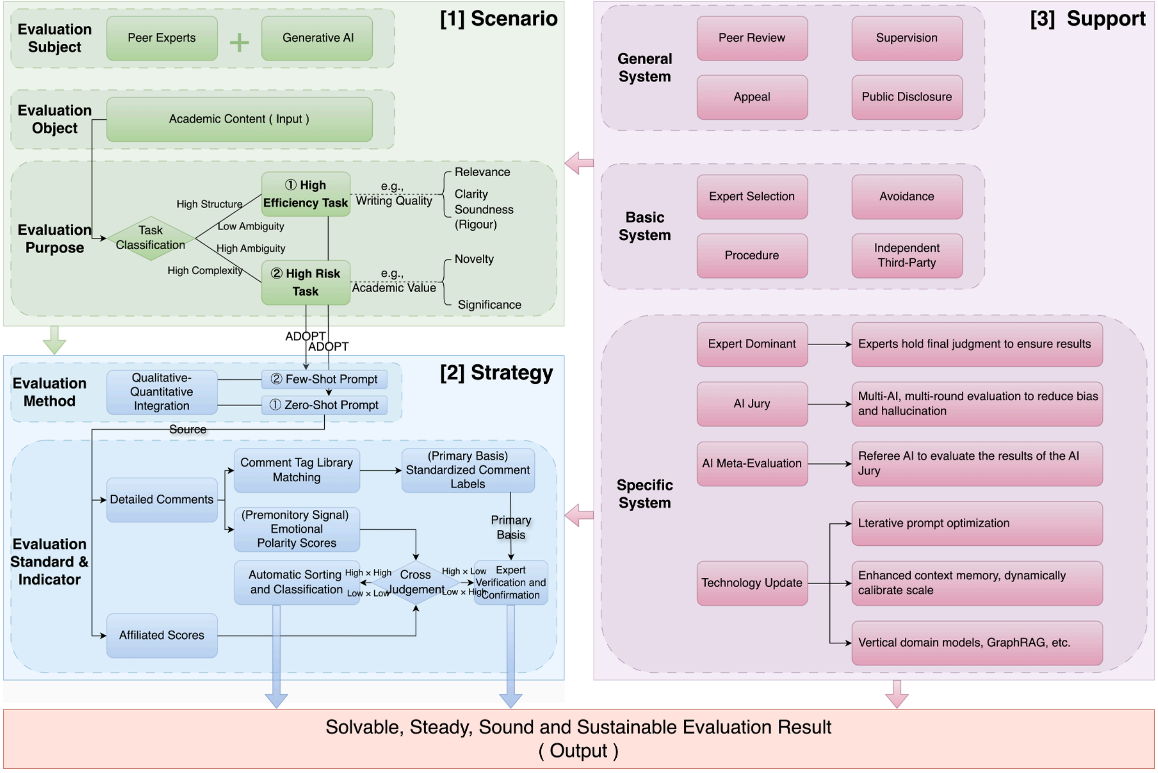


Fig. 8. "Scenario - strategy – support" academic content evaluation model.

systematic and scientific framework for the integration of generative AI in the evaluation of academic content within social sciences, representing a notable contribution of this study to evaluation practices.

The direct evaluation process highlights that peer experts are the primary evaluators in academic assessment, while generative AI functions solely as a supplementary participant (Saad et al., 2024). The object of evaluation is academic content. The purpose of evaluation is paramount, as it determines the classification of the evaluation and informs the subsequent choice of methods, standards, and indicators. If the evaluation is aimed at performing highly structured and low-ambiguity tasks, such as screening out academic content that fails to meet standards (e.g., desk reviews) (Biswas et al., 2023), the zero-shot prompt method is applicable. If the evaluation entails complex tasks that required nuanced value judgements, such as assessing the novelty and significance of a paper, the few-shot prompt method may be advisable. Regardless of the prompt engineering method, the generative AI evaluation model will return both qualitative and quantitative evaluation data, namely detailed comments and affiliated scores. The standard comment labels, formed through comment tag library matching, will constitute the main qualitative foundation for substantiating the expert's assessment of the evaluation outcomes. Additionally, detailed comments can be transformed into quantitative sentiment scores through natural language processing techniques, serving as quantitative premonitory signals. A comparative two-dimensional matrix is established based on the numerical characteristics of premonitory signals and their associated scores. The reliability of the evaluation result increases when both sets of scores are similarly high or low, allowing for direct ranking and output. However, in cases of discrepancies between the two score sets, the evaluation data will be flagged for final verification and confirmation by experts. The final evaluation results will ultimately be deemed solvable, steady, sound, and sustainable.

Furthermore, we propose to enhance the evaluation scenario, strategy, and final results by designing appropriate evaluation mechanisms. At the general system level, this encompasses peer review, process supervision, result appeal, and public disclosure mechanisms. At the basic system level, this encompasses expert selection, avoidance, evaluation procedures, and independent third-party evaluation mechanisms. At the specific system level, emphasis is placed on the expert dominant mechanism, wherein peer experts possess the final decision-making authority regarding evaluation results, thereby ensuring the validity and reasonableness of the outcomes. We also recommend the adoption of the AI jury system, incorporating multiple AI agents (Huang, Wang et al., 2025) and iterations of evaluations (Thelwall, 2025a) to conduct a thorough assessment, which may reduce potential hallucinations and biases (Zhu et al., 2023). Furthermore, a meta-evaluation of the AI evaluation results should be conducted, with AI referees interpreting and assessing the outcomes of AI jurors. In addition, the process must continuously adapt to advancements in generative AI, involving iterative optimization of prompts, enhancement of multi-turn dialogue and contextual memory capabilities for dynamic calibration of evaluation scales, and improvement of generalization through vertical domain-specific evaluation models and advanced technologies (Edge et al., 2024).

6. Conclusion

While previous studies have argued that copyright and other legal constraints prevent the use of full academic texts to evaluate generative AI's capabilities in scholarly assessment (Thelwall et al., 2025), this study employed a quasi-experimental approach using 600 full-length academic papers authored by volunteers with ethical approval. This study systematically compares the overall performance of generative AI across various prompting strategies and models in academic evaluation tasks. Our findings highlight three core dimensions that define the role of AI in academic evaluation: model transparency, transparency in human–AI collaboration, and fairness. This section outlines the theoretical and practical implications of our research.

6.1. *Theoretical implications*

First, we confirm the potential capability of using generative AI as an auxiliary reviewer, capable of generating relatively coherent scores and interpretative comments that aligns with essential dimensions of academic quality within social sciences. While current models exhibit limitations in expert insight, aligning with previous critiques regarding their superficiality and insufficient domain understanding (Lindsay, 2023). However, they still demonstrate value in distinguishing between higher- and lower-quality contents both quantitatively and qualitatively. Importantly, the semantic evidence embedded in AI-generated comments enhances the interpretability of scores, aligning with broader calls for explainable and transparent AI in scholarly contexts.

Second, our results indicate a persistent favoritism towards leniency in AI-generated evaluations of social science content. We conceptualize this as a manifestation of social personality favoritism in AI-based academic evaluation, a concept that is theoretically novel. Luckily, in contrast to human favoritism, which is inherently rooted in social relationships and challenging to identify or amend, AI-based leniency can be methodically modified through model alignment, prompt engineering, and interventions in training data. This presents a promising avenue for mitigating relational bias and face-saving behaviors often found in traditional peer review systems. Therefore, AI can provide an alternative means of delivering uncomfortable yet essential feedback, thereby disrupting the spiral of silence and fostering a more fair evaluative atmosphere.

Third, our results support a human–AI collaborative cross-validation mechanism (Hosseini & Horbach, 2023), where AI-generated evaluations are treated as heuristic signals rather than final conclusions. In this model, standardized, traceable AI comments serve as an initial layer of evaluation, subject to expert interrogation, refinement, or override. This approach preserves expert authority while enhancing transparency, accountability, and auditability. Rather than replacing experts, AI thus serves as a tool to enhance transparency, mitigate reviewer fatigue, and rectify systemic evaluation blind spots.

Finally, this study extends and updates the System of All-round Evaluation of Research (Ye, 2021), one of China's top ten major original academic theories (Information Center for Social Science Renmin University of China, 2025), by integrating AI into its framework. Our results demonstrate that the system remains adaptable and relevant in the age of AI, offering a blueprint for its continued evolution in response to technological disruption.

6.2. Practical implications

This study offers practical implications for academic administrators, peer reviewers, research evaluation experts, and AI system designers, particularly within the social sciences.

First, our results highlight the importance of prompt strategy adaptation based on the structure of evaluation tasks. For well-defined dimensions such as writing quality, zero-shot prompts prove to be sufficient and efficient. For dimensions that are value-laden and ambiguous, such as academic value, few-shot prompts are crucial for incorporating value guidance into AI outputs. This principle of prompt-strategy matching addresses a previously underexplored area in AI-powered academic evaluation.

Second, performance disparities among AI models suggest the need for a multi-agent AI jury mechanism to mitigate hallucination and model-specific biases. The requirements underscore the importance of meta-evaluation, which involves assessing the evaluation process for its reliability and transparency.

Third, evaluators should prioritize qualitative feedback over raw scores. Free-text feedback can be standardized and structurally tagged, providing interpretable evidence for expert judgment. Sentiment analysis tools like VADER (Hutto & Gilbert, 2014) can also be used to extract independent attitudinal cues, reducing reliance on model-generated scores and mitigating sentiment-score coupling biases.

Importantly, in contrast to earlier studies mainly utilizing natural science samples (Huang, Huang et al., 2025), our results highlight that generative AI should be used with prudence when assessing academic value or allocating research resources within social sciences. These evaluations depend on contextual understanding, novelty discernment, and epistemological depth—capabilities current models lack. Improper use may lead to biased decisions, inefficient resource allocation, and skewed knowledge production. Therefore, AI should be positioned as a cognitive assistant that augments rather than replaces human judgment (Thelwall & Yaghi, 2024), advancing efficiency while preserving epistemic responsibility (Thelwall et al., 2025).

Finally, we develop a practical integration model, the SSS-ACE Model, based on the theoretical insights derived from the System of All-round Evaluation of Research. This model offers a broad roadmap for integrating generative AI into academic content evaluation workflows and provides a foundation for academic administrators and research evaluation experts to reconsider hybrid human-AI judgment systems in the age of AI.

6.3. Limitations

While this study offers insights into the capabilities of generative AI in evaluating academic content within the social sciences, several limitations should be acknowledged.

First, although we provide human benchmark scores for the overall evaluation (total scores), primarily due to constraints such as labor resources, annotation costs, and overall workload, we do not offer separate human ratings for all five individual evaluation dimensions. The absence of dimension-specific human benchmarks limits our ability to conduct more fine-grained validation of AI-generated assessments across these aspects. Future studies should aim to collect multi-rater human scores for each dimension, together with inter-rater reliability measures, to enhance the validity and interpretability of automated evaluations.

Second, we did not systematically verify the factual correctness or coherence of the generated textual comments. This limitation stems from the substantial effort required for manual checking or the development of suitable automatic metrics. We recognize this as an important area for future work, especially in refining the accuracy and usefulness of AI-generated evaluations in scholarly contexts.

Finally, the single-round prompt design used in this study limits contextual calibration. Future applications ought to implement multi-turn interaction designs with extended context windows, enabling AI systems to replicate dynamic academic dialogue and attain greater consistency across various submissions.

6.4. Future work

Building on the findings and limitations outlined above, future research should further investigate both the theoretical and practical implications of integrating AI into academic evaluation within the social sciences.

A key direction involves testing and refining the proposed SSS-ACE Model, which serves as a conceptual roadmap for integrating empirical findings with established evaluation theories. While the model provides a valuable integrative framework, its effectiveness and generalizability require rigorous empirical validation. Therefore, future work should focus on developing more robust, domain-specific strategies to enhance the evaluation performance of LLMs in this context.

Additionally, future studies should explore how computational linguistic features of academic texts—rooted in the probabilistic and statistical characteristics of LLMs—affect AI-based judgments. This line of inquiry may lead to deeper insights into how language models process scholarly content and how their outputs can be meaningfully interpreted and validated.

Beyond directly using AI-generated evaluation results, future work may also explore using generative AI for information extraction, classification, and multi-modal information processing, combined with rigorous bibliometric methods or mining algorithms. This may result in advancements in intelligent bibliometrics, offering indirect but enhanced assistance for academic assessment in the social sciences.

Importantly, future work should also address the ethical and societal implications of using AI in academic evaluation, including issues of transparency, bias, accountability, and potential impacts on academic norms. Aligning AI-assisted evaluation practices with broader societal goals—such as SDG 4 (Quality Education), which advocates for inclusive and equitable lifelong learning, and SDG 16 (Peace, Justice, and Strong Institutions), which emphasizes transparent and accountable systems, can help ensure that the implementation of AI in academic contexts fosters more just, inclusive, and trustworthy institutions.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. All sensitive data were anonymized prior to analysis.

Ethical approval

The study involved human participants (graduate students) and received ethical approval from the School of Information Management at Nanjing University. Informed consent was obtained from all participants, and data were anonymized and handled in accordance with institutional and national ethical standards.

CRediT authorship contribution statement

Yu Zhu: Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Methodology, Investigation, Data curation, Conceptualization. **Yongrong Lu:** Writing – original draft, Visualization, Formal analysis. **Huan Xie:** Supervision, Resources, Project administration. **Jiyuan Ye:** Supervision, Resources, Funding acquisition. **Ming Chen:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the major project of National Social Science Foundation of China (Grant No 24&ZD323) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No KYCX25_0130).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2025.104365.

Appendix A

Table A.1
Propositions.

Number	Propositions
1	Definitions and Relationships among Information, Data, and Information Resources.
2	(a) The Significance and Limitations of Bradford's Law in Information Resources Construction. (b) The Significance and Limitations of Garfield's Law in Information Resources Construction.
3	The Impact of AI-Generated Content on Information Resources Construction.
4	(a) Compare and analyze the information resources construction policies of one foreign public library and one foreign academic library with those of one domestic public library and one domestic academic library. (b) Select and compare four national or local-level information resources construction policies (any aspect within the system).
5	(a) Analyze the current application status of Patron-driven acquisition, simultaneous print and electronic acquisition, and precision acquisition with examples. (b) Analyze the current application status of Web 2.0, data mining, big data, and AI technologies in information resources collection with examples.
6	Investigate the construction of specialized databases, disciplinary information portals, open access resources, and digital resource integration at one public library and one academic library of your choice.

* Note: If a proposition includes (a) or (b), it allows flexible choices to enrich the content while maintaining thematic consistency.

Table A.2
Descriptive statistics of the samples.

Proposition Number	Sample Size	Char Count				Word Count			
		Mean	Median	Min	Max	Mean	Median	Min	Max
1	100	1559.237	1406	684	4355	923.938	828	409	2119
2	100	1716.788	1524	636	5840	982.232	877	366	3280
3	100	1898.202	1596	629	6557	1031.919	897	315	3443
4	100	3561.786	2406	984	34,090	1934.224	1315	557	15,255
5	100	2452.888	2128.5	639	8427	1394.194	1206	358	4950
6	100	3311.592	2489	313	38,594	1837.827	1421	181	20,785

* Note: The samples are written in Chinese, and the word count is performed by the Python library jieba.

Appendix B

Few-shot prompt design:

- Role: Expert in academic content review.
- Background: Users need to conduct an in-depth review of academic texts to ensure their academic value and writing quality. Users expect a comprehensive grading system to help them better understand and evaluate all aspects of academic texts.
- Profile: You are a senior expert with a background in library and information science, with extensive experience and deep understanding of the review of academic texts. You will be able to accurately assess the novelty, significance, relevance, clarity, and soundness of academic texts.
- Skills: You have the ability to analyze the novelty of an academic text, assess its practical application in the relevant field, judge its relevance to the topic, evaluate the accuracy and fluency of its expression, and test the accuracy of its methodology and the logical consistency of its arguments.
- Goals: Based on the academic value and writing quality of the academic text, score 0–100 for Novelty, Significance, Relevance, Clarity and Soundness(Rigour) as well as the corresponding comments, and calculate the total score.
- Constrains: The evaluation process shall be objective and fair, the scoring criteria shall be clear and specific, the comments shall be detailed and accurate, and the total score calculation shall be reasonable and scientific.
- OutputFormat: Returns the score result and comments in json format.
- Workflow:
  ■ 1. Read academic texts carefully to fully understand their content and structure.
  ■ 2. Based on Novelty, Significance, Relevance, Clarity and Soundness(Rigour) scoring criteria, individual scores ranging from 0 to 100 and corresponding comments are given respectively.
  ■ 3. Calculate the total score and return all score results and comments in json format.
- Examples:
  1. Example 1:

     {{"Novelty": {{"score": "95”,"comment": "The text introduces a new theoretical framework that significantly advances the understanding of the topic, demonstrating the author’s unique contribution."}},

     "Significance": {{"score": "90","comment": "The findings have high practical applicability and can be widely referenced in the field, contributing to solving real-world problems."}},

     "Relevance": {{"score": "95","comment": "The content is highly relevant to the specified topic, with clear alignment and focus."}},

     "Clarity": {{"score": "95","comment": "The expression is accurate and fluent, with error-free and concise writing, making the text easy to understand."}},

     "Soundness(Rigour)": {{"score": "85","comment": "The methodology is precise, the data is reliable, and the arguments are valid, but there is room for improvement in the sufficiency of evidence."}},

     "Total Score": "92"}}

  ■ Example 2:

     {{"Novelty": {{"score": "50","comment": "The text presents any new insights, and the overall contribution to the field is limited."}},

     "Significance": {{"score": "60","comment": "The applicability and referential value are low, with potential for further development."}},

     "Relevance": {{"score": "73","comment": "The content is moderatly relevant to the topic, but some parts could be more closely aligned."}},

     "Clarity": {{"score": "75","comment": "The writing is generally clear, but there are a few minor errors and areas for improvement in fluency, for example"}},

     "Soundness(Rigour)": {{"score": "78","comment": "The methodology is mostly sound, but there are some issues with data reliability and argument coherence."}},

     "Total Score": "67.2"}}

* Notes on few-shot prompt design and structure:

This prompt was designed following the structured prompt engineering framework proposed by LangGPT, emphasizing role definition, task constraints, step-wise logic, and explicit output formatting.

The examples provided serve as demonstrations to guide the AI model's evaluation behavior. Each example includes the following structured components for all five indicators (Novelty, Significance, Relevance, Clarity, Soundness):

- Numerical Score (0–100): Clearly quantifies the assessment of each dimension, enabling standardized output and facilitating comparison across samples.
- Explanatory Comment: Each score is accompanied by a comment that explains the rationale behind the score. These comments include:
  - ■ Reference to specific features of the text (e.g., "introduces a new theoretical framework," "error-free and concise writing");
  - ■ Qualitative judgment calibrated to the score (e.g., "excellent," "limited," "moderately relevant").
- Score–Comment Alignment: Each pair of score and comment is carefully matched to reflect realistic evaluation standards in academic peer review. For instance:
  - ■ A score of 90+ is justified by comments showing originality, field impact, and methodological rigor.
  - ■ A score below 60 includes explanations of weaknesses or deficiencies (e.g., "low applicability", "some issues with coherence").
- JSON Format Output: The examples adopt a structured JSON format to ensure machine readability and alignment with the prompt's Output Format constraint. This format improves traceability of individual scores and comments, allowing for downstream parsing or analysis.
- Score Averaging Logic: Ideally, the "Total Score" is a simple arithmetic mean of the five individual dimension scores, ensuring transparency in the aggregation method.

These example instances act as reference anchors for the model's internal representation of quality levels. By exposing the model to high and mid-quality examples with clearly articulated reasoning, we improve its ability to generalize these standards when evaluating unseen texts. This design follows principles from prompt engineering best practices (e.g., consistency, clarity, contextual grounding) and strengthens the validity of the evaluation process.

Zero-shot prompt design:

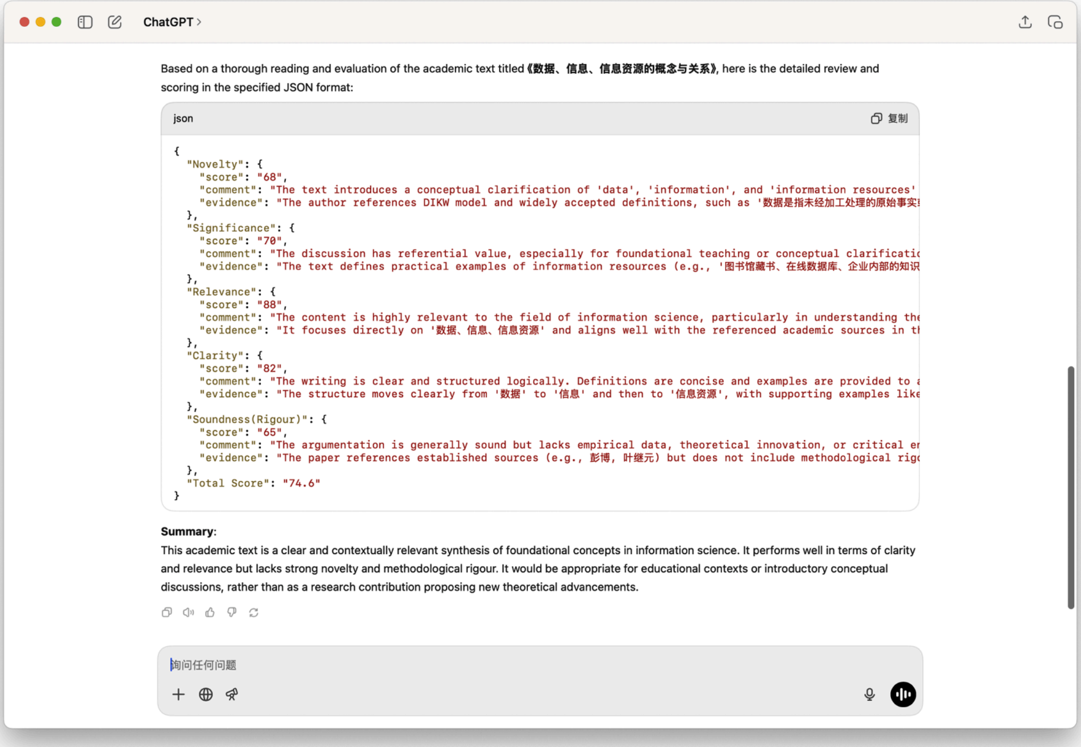
- Role: Expert in academic content review
- Background: Users require an in-depth review of academic texts to assess their academic value and writing quality. They seek a comprehensive grading system to better understand and evaluate all aspects of academic texts.
- Profile: You are a senior expert with a background in library and information science, possessing extensive experience and a deep understanding of academic text review. You are capable of accurately assessing the novelty, significance, relevance, clarity, and soundness (rigour) of academic texts.
- Skills: You have the ability to:
- Analyze the novelty of an academic text, evaluating the introduction of new ideas, theories, methods, or insights that advance the boundaries of knowledge and reflect the author's unique contribution.
- Assess the significance of the text, determining its practical applicability and referential value in the respective field.
- Judge the relevance of the text, ensuring its alignment with the specified topic and focus.
- Evaluate the clarity of the text, assessing the accuracy, fluency, and conciseness of the writing, as well as its readability.
- Test the soundness (rigour) of the text, examining the precision of the methodology, reliability of data, coherence of design, validity of arguments, sufficiency of evidence, and logical consistency.
- Goals: Based on the academic value and writing quality of the text, provide a score (0–100) and detailed comments for each of the following criteria: Novelty, Significance, Relevance, Clarity, and Soundness (Rigour). Calculate the total score based on these individual scores.
- Constraints:The evaluation process must be objective and fair. The scoring criteria must be clear and specific. The comments must be detailed and accurate. The total score calculation must be reasonable and scientifically sound.
- Output Format: Return the score results and comments in the following JSON format:

```
{"Novelty": {"score": "", "comment": ""},
"Significance": {"score": "", "comment": ""},
"Relevance": {"score": "", "comment": ""},
"Clarity": {"score": "", "comment": ""},
"Soundness(Rigour)": {"score": "", "comment": ""},
"Total Score": ""}
```

- Workflow:
  - ■ 1. Carefully read the academic text to fully understand its content and structure.
  - ■ 2. Based on the criteria of Novelty, Significance, Relevance, Clarity, and Soundness (Rigour), assign individual scores (0–100) and provide corresponding comments for each criterion.
  - ■ 3. Calculate the total score and return all score results and comments in the specified JSON format.

Appendix C

Fig. C.1



Based on a thorough reading and evaluation of the academic text titled **《数据、信息、信息资源的概念与关系》**, here is the detailed review and scoring in the specified JSON format:

```json
{
  "Novelty": {
    "score": "68",
    "comment": "The text introduces a conceptual clarification of 'data', 'information', and 'information resources'
    "evidence": "The author references DIKW model and widely accepted definitions, such as '数据是指未经加工处理的原始事实
  },
  "Significance": {
    "score": "70",
    "comment": "The discussion has referential value, especially for foundational teaching or conceptual clarificatio
    "evidence": "The text defines practical examples of information resources (e.g., '图书馆藏书、在线数据库、企业内部的知识
  },
  "Relevance": {
    "score": "88",
    "comment": "The content is highly relevant to the field of information science, particularly in understanding the
    "evidence": "It focuses directly on '数据、信息、信息资源' and aligns well with the referenced academic sources in th
  },
  "Clarity": {
    "score": "82",
    "comment": "The writing is clear and structured logically. Definitions are concise and examples are provided to a
    "evidence": "The structure moves clearly from '数据' to '信息' and then to '信息资源', with supporting examples like
  },
  "Soundness(Rigour)": {
    "score": "65",
    "comment": "The argumentation is generally sound but lacks empirical data, theoretical innovation, or critical en
    "evidence": "The paper references established sources (e.g., 彭博, 叶继元) but does not include methodological rigo
  },
  "Total Score": "74.6"
}
```

Summary:

This academic text is a clear and contextually relevant synthesis of foundational concepts in information science. It performs well in terms of clarity and relevance but lacks strong novelty and methodological rigour. It would be appropriate for educational contexts or introductory conceptual discussions, rather than as a research contribution proposing new theoretical advancements.

Fig. C.1. A simple demonstration that requests ChatGPT to return evaluation evidence (built on ChatGPT desktop, omitting prompts). ChatGPT could return evidence from the original text for secondary verification by human reviewers, enhancing transparency and explainability of social science evaluation driven by generative AI.

References

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation. *Evaluation Review, 22*(2), 207–244.

Anthropic. (2024). Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI Feedback*. https://arxiv.org/abs/2212.08073.

Biswas, S., Dobaria, D., & Cohen, H. L. (2023). ChatGPT and the future of journal reviews: A feasibility study. *Yale Journal of Biology and Medicine, 96*(3), 415–420. https://doi.org/10.59249/SKDH9286

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology, 45*(1), 197–245. https://doi.org/10.1002/aris.2011.1440450112

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80. https://doi.org/10.1108/00220410810844150

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology, 66*(11), 2215–2222. https://doi.org/10.1002/asi.23329

Bornmann, L., Tekles, A., Zhang, H. H., & Ye, F. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics, 13*(4), Article 100979. https://doi.org/10.1016/j.joi.2019.100979

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science, 37*(1), 34–36. https://doi.org/10.1002/(SICI)1097-4571(198601)37:1<34::AID-ASI5>3.0.CO;2-0

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications, 8*(1), 1–11. https://doi.org/10.1057/s41599-020-00703-8

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first international conference on machine learning*. https://openreview.net/forum?id=3MW8GKNyzI.

Clarivate. (2025). Journal citation reports. https://jcr.clarivate.com/jcr/home.

Cohen, L., Manion, L., & Morrison, K. (2002). *Research methods in education* (5th ed.). Routledge. https://doi.org/10.4324/9780203224342

DORA. (2012). About DORA. https://sfdora.org/about-dora/.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). *From local to global: A graph RAG approach to query-focused summarization* (No. arXiv:2404.16130). ArXiv. 10.48550/arXiv.2404.16130.

Elangovan, A., He, J., & Verspoor, K. (2021). Memorization vs. generalization: Quantifying data leakage in NLP performance evaluation. In P. Merlo, J. Tiedemann, & R. Tsarfaty (Eds.), *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume* (pp. 1325–1335). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.113.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using r*. SAGE Publications Ltd.

Ganguli, D., Askell, A., Schiefer, N., Liao, T.I., Lukošiūtė, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., Hernandez, D., Drain, D., Li, D., Tran-Johnson, E., Perez, E., Kernion, J., Kerr, J., Mueller, J., Landau, J., Ndousse, K., … Kaplan, J. (2023). *The capacity for moral self-correction in large language models* (No. arXiv: 2302.07459). ArXiv. 10.48550/arXiv.2302.07459.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science (New York, N.Y.), 122*(3159), 108–111. https://doi.org/10.1126/science.122.3159.108

Hamilton, D. P. (1991). Research papers: Who's uncited now? *Science (New York, N.Y.), 251*(4989), 25. https://doi.org/10.1126/science.1986409

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The leiden manifesto for research metrics. *Nature, 520*(7548), 429–431. https://doi.org/10.1038/520429a

Higher Education Funding Council for England. (2019). *Panel criteria and working methods (2019/02)* (Worldwide). https://2021.ref.ac.uk/publications-and-reports/panel-criteria-and-working-methods-201902/index.html.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences, 102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

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

Hrubec, M., & Višňovský, E. (2023). *Towards a new research era*. Brill. https://brill.com/display/title/64619.

Huang, S., Huang, Y., Liu, Y., Luo, Z., & Lu, W. (2025). Are large language models qualified reviewers in originality evaluation? *Information Processing & Management, 62*(3), Article 103973. https://doi.org/10.1016/j.ipm.2024.103973

Huang, S., Wang, Q., Lu, W., Liu, L., Xu, Z., & Huang, Y. (2025). PaperEval: A universal, quantitative, and explainable paper evaluation method powered by a multi-agent system. *Information Processing & Management, 62*(6), Article 104225. https://doi.org/10.1016/j.ipm.2025.104225

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *, 8. Proceedings of the international AAAI conference on web and social media*. https://doi.org/10.1609/icwsm.v8i1.14550. Article 1.

Information Center for Social Science Renmin University of China. (2025). *Orientation • standards • examples: an analytical report on original academic theories in Chinese philosophy and social sciences*. Information Center for Social Science Renmin University of China. https://zszwx.cn/originalList.

Jurafsky, D., & Martin, J. (2008). *Speech and language processing, 2nd edition* (2nd edition). Prentice Hall.

Kampenes, V. B., Dybå, T., Hannay, J. E., & Sjøberg, D. I. K. (2009). A systematic review of quasi-experiments in software engineering. *Information and Software Technology, 51*(1), 71–82. https://doi.org/10.1016/j.infsof.2008.04.006

Kelly, J., Sadeghieh, T., & Adeli, K. (2014). Peer review in scientific publications: Benefits, critiques, & a survival guide. *Ejifcc, 25*(3), 227–243.

Khatun, A., & Brown, D. (2023). Reliability check: An analysis of GPT-3`s response to sensitive topics and prompt wording. In A. Ovalle, K.-W. Chang, N. Mehrabi, Y. Pruksachatkun, A. Galystan, J. Dhamala, A. Verma, T. Cao, A. Kumar, & R. Gupta (Eds.), *Proceedings of the 3rd workshop on trustworthy natural language processing (TRUSTNLP 2023)* (pp. 73–95). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.trustnlp-1.8.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion, 99*, Article 101861. https://doi.org/10.1016/j.inffus.2023.101861

Kousha, K., & Thelwall, M. (2024). *Assessing the societal influence of academic research with ChatGPT: Impact case study evaluations* (No. arXiv:2410.19948). ArXiv. 10.48550/arXiv.2410.19948.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kuo, C. L. (2015). *A quasi-experimental study of formative peer assessment in an EFL writing classroom* [Thesis]. Newcastle University.

Lengkeek, M., van der Knaap, F., & Frasincar, F. (2023). Leveraging hierarchical language models for aspect-based sentiment analysis on financial data. *Information Processing & Management, 60*(5), Article 103435. https://doi.org/10.1016/j.ipm.2023.103435

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

Lin, R., Li, Y., Ji, Z., Xie, Q., & Chen, X. (2025). Quantifying the degree of scientific innovation breakthrough: Considering knowledge trajectory change and impact. *Information Processing & Management, 62*(1), Article 103933. https://doi.org/10.1016/j.ipm.2024.103933

Lindsay, G. W. (2023). LLMs are not ready for editorial work. *Nature Human Behaviour, 7*(11), 1814–1815. https://doi.org/10.1038/s41562-023-01730-6

Liu, M., Bu, Y., Chen, C., Xu, J., Li, D., Leng, Y., Freeman, R. B., Meyer, E. T., Yoon, W., Sung, M., Jeong, M., Lee, J., Kang, J., Min, C., Song, M., Zhai, Y., & Ding, Y. (2022). Pandemics are catalysts of scientific novelty: Evidence from COVID-19. *Journal of the Association for Information Science and Technology, 73*(8), 1065–1078. https://doi.org/10.1002/asi.24612

Liu, M., Xie, Z., Yang, A. J., Yu, C., Xu, J., Ding, Y., & Bu, Y. (2024). The prominent and heterogeneous gender disparities in scientific novelty: Evidence from biomedical doctoral theses. *Information Processing & Management, 61*(4), Article 103743. https://doi.org/10.1016/j.ipm.2024.103743

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9). https://doi.org/10.1145/3560815, 195:1-195:35.

Li, J., Zhang, Q., Yu, Y., Fu, Q., & Ye, D. (2024). More agents is all you need (No. arXiv:2402.05120). arXiv. https://doi.org/10.48550/arXiv.2402.05120.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). *G-eval: NLG evaluation using GPT-4 with better human alignment* (No. arXiv:2303.16634). ArXiv. 10.48550/arXiv.2303.16634.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach* (No. arXiv:1907.11692). ArXiv. 10.48550/arXiv.1907.11692.

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., & Callison-Burch, C. (2023). Faithful chain-of-thought reasoning. In *The 13th international joint conference on natural language processing and the 3rd conference of the Asia-pacific chapter of the association for computational linguistics (IJCNLP-AACL 2023)*. https://par.nsf.gov/biblio/10463284-faithful-chain-thought-reasoning.

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science, 40*(5), 342–349. https://doi.org/10.1002/(SICI)1097-4571(198909)40:5<342::AIDASI7>3.0.CO;2-U

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist, 63*(3), 160–168. https://doi.org/10.1037/0003-066X.63.3.160

Merton, R. K. (1979). In N. W. Storer (Ed.), *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/S/bo28451565.html.

Miller, C. J., Smith, S. N., & Pugatch, M. (2020). Experimental and quasi-experimental designs in implementation research. *Psychiatry Research, 283*. https://doi.org/10.1016/j.psychres.2019.06.027. S0165-1781(19)30683-3.

Mohd, R. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics, 2*(1), 21–33.

Munroe, R. (2013). The rise of open access. *Science (New York, N.Y.), 342*(6154), 58–59. https://doi.org/10.1126/science.342.6154.58

OpenAI. (2024). GPT-4o system card. https://openai.com/index/gpt-4o-system-card/.

OpenAI. (2025). Realtime API - OpenAI API. https://platform.openai.com/docs/guides/realtime#connect-with-webrtc.

Prillaman, M. (2024). Is ChatGPT making scientists hyper-productive? The highs and lows of using AI. *Nature, 627*(8002), 16–17. https://doi.org/10.1038/d41586-024-00592-w

Ronzano, F., & Saggion, H. (2016). Knowledge extraction and modeling from scientific publications. In A. González-Beltrán, F. Osborne, & S. Peroni (Eds.), *Semantics, analytics, visualization enhancing scholarly data* (pp. 11–25). Springer International Publishing. https://doi.org/10.1007/978-3-319-53637-8_2.

Saad, A., Jenko, N., Ariyaratne, S., Birch, N., Iyengar, K. P., Davies, A. M., Vaishya, R., & Botchu, R. (2024). Exploring the potential of ChatGPT in the peer review process: An observational study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 18*(2), Article 102946. https://doi.org/10.1016/j.dsx.2024.102946

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ: British Medical Journal, 314*(7079), 498–502.

Shiflett, L. (1988). A difficult balance: Editorial peer review in medicine. *Journal of the American Society for Information Science, 39*(1), 22–23. https://doi.org/10.1002/(SICI)1097-4571(198801)39:1<22::AID-ASI6>3.0.CO;2-X

Shuster, K., Xu, J., Komeili, M., Ju, D., Smith, E.M., Roller, S., Ung, M., Chen, M., Arora, K., Lane, J., Behrooz, M., Ngan, W., Poff, S., Goyal, N., Szlam, A., Boureau, Y.-L., Kambadur, M., & Weston, J. (2022). *BlenderBot 3: A deployed conversational agent that continually learns to responsibly engage* (No. arXiv:2208.03188). ArXiv. 10.48550/arXiv.2208.03188.

Si, K., Li, Y., Ma, C., & Guo, F. (2023). Affiliation bias in peer review and the gender gap. *Research Policy, 52*(7), Article 104797. https://doi.org/10.1016/j.respol.2023.104797

Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys, 55*(13s). https://doi.org/10.1145/3582688, 271:1-271:40.

Spezi, V., Wakeling, S., Pinfield, S., Fry, J., Creaser, C., & Willett, P. (2018). "Let the community decide"? The vision and reality of soundness-only peer review in open-access mega-journals. *Journal of Documentation, 74*(1), 137–161. https://doi.org/10.1108/JD-06-2017-0092

Sukpanichnant, P., Rapberger, A., & Toni, F. (2024). *PeerArg: Argumentative peer review with LLMs* (No. arXiv:2409.16813). ArXiv. 10.48550/arXiv.2409.16813.

Sun, M., Barry Danfa, J., & Teplitskiy, M. (2022). Does double-blind peer review reduce bias? Evidence from a top computer science conference. *Journal of the Association for Information Science and Technology, 73*(6), 811–819. https://doi.org/10.1002/asi.24582

Tencent. (2025). Tencent Hunyuan T1. https://hunyuan.tencent.com/.

Thelwall, M. (2024). Can ChatGPT evaluate research quality? *Journal of Data and Information Science, 9*(2), 1–21. https://doi.org/10.2478/jdis-2024-0013

Thelwall, M. (2025a). ChatGPT for complex text evaluation tasks. *Journal of the Association for Information Science and Technology, 76*(4), 645–648. https://doi.org/10.1002/asi.24966

Thelwall, M. (2025b). Evaluating research quality with large language models: An analysis of ChatGPT's effectiveness with different settings and inputs. *Journal of Data and Information Science, 10*(1), 7–25. https://doi.org/10.2478/jdis-2025-0011

Thelwall, M., Jiang, X., & Bath, P. A. (2025). Estimating the quality of published medical research with ChatGPT. *Information Processing & Management, 62*(4), Article 104123. https://doi.org/10.1016/j.ipm.2025.104123

Thelwall, M., & Yaghi, A. (2024). *In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results* (No. arXiv:2409.16695). ArXiv. 10.48550/arXiv.2409.16695.

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences, 21*(1). http://agro.icm.edu.pl/agro/element/bwmeta1.element.agro-c9d1981f-962f-405d-83a7-47080c2a1c8f.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science (New York, N.Y.), 211*(4481), 453–458. https://doi.org/10.1126/science.7455683

Wallerstein, I. (2004). *World-systems analysis: An introduction*. Duke University Press. https://doi.org/10.1515/9780822399018

Wang, M., Liu, Y., Liang, X., Li, S., Huang, Y., Zhang, X., Shen, S., Guan, C., Wang, D., Feng, S., Zhang, H., Zhang, Y., Zheng, M., & Zhang, C. (2024). *LangGPT: Rethinking structured reusable prompt design framework for LLMs from the programming language*. Arxiv.Org. https://arxiv.org/abs/2402.16929v2.

Wang, Y., Yu, Z., Zeng, Z., Yang, L., Wang, C., Chen, H., Jiang, C., Xie, R., Wang, J., Xie, X., Ye, W., Zhang, S., & Zhang, Y. (2024). *PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization* (No. arXiv:2306.05087). ArXiv. 10.48550/arXiv.2306.05087.

Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., & Le, Q.V. (2022). *Finetuned language models are zero-shot learners* (No. arXiv: 2109.01652). ArXiv. 10.48550/arXiv.2109.01652.

Wilby, R. L., & Esson, J. (2024). AI literacy in geographic education and research: Capabilities, caveats, and criticality. *Geographical Journal, 190*(1). https://doi.org/10.1111/geoj.12548

Wilsdon, J. (2016). *The metric tide: independent review of the role of metrics in research assessment and management* (1st Edition). SAGE Publications Ltd.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica, 10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Xue, Z., He, G., Liu, J., Jiang, Z., Zhao, S., & Lu, W. (2023). Re-examining lexical and semantic attention: Dual-view graph convolutions enhanced BERT for academic paper rating. *Information Processing & Management, 60*(2), Article 103216. https://doi.org/10.1016/j.ipm.2022.103216

Yan, Z., & Fan, K. (2024). An integrated indicator for evaluating scientific papers: Considering academic impact and novelty. *Scientometrics, 129*(11), 6909–6929. https://doi.org/10.1007/s11192-024-05150-9

Yang, J., Lu, W., Hu, J., & Huang, S. (2022). A novel emerging topic detection method: A knowledge ecology perspective. *Information Processing & Management, 59*(2), Article 102843. https://doi.org/10.1016/j.ipm.2021.102843

Yang, P., Sun, X., Li, W., & Ma, S. (2018). Automatic academic paper rating based on modularized hierarchical convolutional neural network. In I. Gurevych, & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: short papers)* (pp. 496–502). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2079.

Yang, W. (2016). Evaluative language and interactive discourse in journal article highlights. *English for Specific Purposes, 42*, 89–103. https://doi.org/10.1016/j.esp.2016.01.001

Ye, J. (2010). Approaching evaluation system in humanities and social sciences. *Journal of Nanjing University (Philosophy, Humanities and Social Sciences), 47*(01), 97–110.

Ye, J. (2021). *On the system of all-round evaluation of research.* Social Sciences Academic Press.

Zhang, L., & Sivertsen, G. (2020). The new research assessment reform in China and its implementation. *Scholarly Assessment Reports, 2*(1), 1–7. https://doi.org/10.29024/sar.15

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems, 36*, 46595–46623.

Zheng, M., Pei, J., Logeswaran, L., Lee, M., & Jurgens, D. (2024). When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: EMNLP 2024* (pp. 15126–15154). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.888.

Zhu, Y., Chen, G., Lu, Y., & Fan, W. (2023). Generative artificial intelligence governance action framework: Content analysis based on AIGC incident report texts. *Documentation, Information & Knowledge, 40*(4), 41–51. https://doi.org/10.13366/j.dik.2023.04.041