

AI-Assisted Peer Review: Comparing Generative AI and Human Evaluations

Kasey Miller
millerkc@uncw.edu

Zealia Svecharnik
zs8464@uncw.edu

Kevin Matthews
matthewskd@uncw.edu

Geoff Stoker
stokerg@uncw.edu

Congdon School
University of North Carolina Wilmington
Wilmington, NC 28403 USA

Abstract

Rapid advances in Generative AI offer new opportunities to streamline academic peer review, yet questions remain regarding whether AI systems can match human evaluative judgment. Through a Business Process Reengineering (BPR) lens, we evaluate three leading GenAI models (ChatGPT, Claude, and Gemini) using a standardized rubric across an exploratory sample of 11 Information Systems conference papers (N=11) with historical human-reviewer scores (2021–2025). AI-generated evaluations exhibited lower scoring variance than human reviews, while repeated-run reliability varied across models and evaluation criteria. AI evaluations also diverged substantially from human outcomes, showing virtually no overlap in top-ranked papers. AI models exhibited score compression, relative severity toward methodology, and relative leniency toward writing style, while producing paper rankings that differed substantially from those of human reviewers. These findings suggest that current LLMs should not be viewed as replacements for expert human judgment, but may be useful as upstream tools to support structural and mechanical evaluation in a reengineered hybrid review workflow.

Keywords: Artificial Intelligence, Generative AI, Academic Assessment, Peer Review, Information Systems Education, Decision Support Systems.

AI-Assisted Peer Review: Comparing Generative AI and Human Evaluations

Kasey Miller, Zealia Svecharnik, Kevin Matthews and Geoff Stoker

1. INTRODUCTION

Academic peer review, the foundational mechanism through which new research is validated, is increasingly strained by growing submission volumes and a limited supply of qualified reviewers. Manuscript submissions grew 6.1% annually from 2013 to 2018 (Publons, 2018), with recent journal reports citing a 33% year-over-year increase (Crowe, 2026). The global peer-review workload is now estimated to exceed 100 million hours annually (Aczel et al., 2021). Meanwhile, conference submissions have surged—ICLR grew from 960 submissions in 2018 to nearly 5,000 in 2023 (Liang et al., 2024b), and ISCAP submissions increased 51% between 2018 and 2025.

Recent advances in LLMs offer promising capabilities for scholarly review, including rapid manuscript evaluation, consistent rubric application, and reduced reviewer fatigue. While prior studies have examined AI-generated reviews in large journal and conference datasets, few have evaluated Information Systems (IS) education conferences or analyzed the consistency of repeated AI evaluations on identical manuscripts. Consequently, how different AI systems compare against human reviewers and one another under identical rubrics remains understudied.

To address this gap, this study examines 11 ISCAP conference submissions (2021–2025) with original human-reviewer scores. ChatGPT, Claude, and Gemini independently evaluated each paper using the same rubric. We compare AI and human scores and examine differences among the AI systems through the following research questions:

- RQ1: To what extent do AI-generated review scores correlate with human-reviewer scores?
- RQ2: Does AI demonstrate different scoring variances than humans?
- RQ3: How stable are AI-generated evaluations across repeated review sessions?

The paper is organized as follows: Section 2 reviews the current literature; Section 3 describes the methodology; Section 4 presents the results; Section 5 provides discussion; and Section 6 concludes and outlines future work.

2. LITERATURE REVIEW

The increasing availability of GenAI has prompted researchers to examine how it can augment knowledge-intensive organizational processes traditionally performed by humans. One such process is academic peer review, where rising manuscript submission rates, reviewer fatigue, and extended review cycles have challenged the scalability of traditional methods. While much of the emerging literature has focused on the capabilities and limitations of AI-assisted manuscript evaluation, less attention has been devoted to how AI can be integrated into the broader peer-review workflow and to whether AI-generated evaluations demonstrate the consistency and reliability necessary for operational adoption.

Business Process Reengineering (BPR) offers a useful theoretical lens for examining this question, as it emphasizes the fundamental redesign of organizational processes to achieve substantial improvements in productivity, quality, service, and cycle time (Hammer, 1990). While BPR predates modern artificial intelligence, it is applied here as a general process-redesign framework to examine how AI capabilities can be incorporated into the peer-review workflow to optimize task allocation between human and automated raters. Prior research has suggested that BPR techniques can be used to model the effects of AI-assisted automation, cognitive information biases, and machine learning on productivity and cycle-time improvement (Hammer, 1990; Housel & Bell, 2001; Miller & Mun, 2023). From this perspective, the relevant question is not whether AI can replace human reviewers, but whether it can provide reliable, consistent evaluations that justify its inclusion in the scholarly review process.

2.1 AI Use in the Peer Review Process

A growing body of literature has begun examining AI's role in scholarly evaluation (Checco et al., 2021). Liang et al. (2024a) used a maximum-likelihood framework to analyze major AI conferences (ICLR, NeurIPS, CoRL, EMNLP), estimating that 6.5%–16.9% of review sentences were substantially AI-modified. Higher AI involvement correlated with late submissions, missing scholarly references, and standardized feedback. These findings confirm that AI use in peer review is widespread, raising legitimate concerns about feedback diversity and reviewer independence, as well as potential efficiency gains.

2.2 Quality of AI-Generated Review Compared to Human Review

Assessing the quality of AI-generated feedback reveals critical content asymmetries. Comparing GPT-4 against human reviewers across 3,096 Nature and 1,709 ICLR papers, Liang et al. (2024b) found moderate overlap (30.8%–39.23%). A shuffling experiment confirmed paper specificity, as overlap dropped to near-zero when GPT-4 reviews were paired with human reviews from different manuscripts. However, GPT-4 commented on research implications 7.27 times more often than humans while raising novelty concerns 10.69 times less often. Joachim et al. (2025) confirmed this novelty gap, finding only weak correlations between AI originality scores and human rejection decisions across four models (GPT-4o, Claude 3.5, Gemini 1.5, GPT-o1) reviewing 22 oral and maxillofacial surgery manuscripts. Human reviewers rejected 68.2% of submissions, compared with 0%–9.1% for AI models. All models demonstrated high temporal stability but only moderate agreement with human decisions (kappa 0.38–0.46). Similarly, Biswas et al. (2023) noted that ChatGPT feedback lacked the citation-level specificity and depth typical of expert reviewers. Nevertheless, Liang et al. (2024b) reported that in a survey of 308 researchers, over half rated GPT-4's feedback as helpful, and over 80% considered it more useful than at least some human reviews. Taken together, these studies reveal an unresolved tension: AI-generated reviews offer meaningful overlap with human feedback yet diverge on the dimensions human reviewers weigh most critically—novelty, depth, and citation specificity.

2.3 Reviewer Agreement and Reliability

An important consideration in evaluating AI-generated reviews is the inherent variability of human peer review. Prior research has demonstrated that reviewer agreement is often lower than expected, particularly when evaluating novel ideas, emerging methods, or interdisciplinary contributions. Bornmann et al. (2010) conducted a meta-analysis of journal peer-review inter-rater reliability (IRR), and the results confirm that IRR is limited. Across over 19,000 manuscripts, the mean Cohen's Kappa was just 0.17, meaning reviewers agreed only 17% more than chance would predict. Differences in reviewers' expertise, methodological preferences, and interpretations of evaluation criteria lead to divergent assessments (Cole et al., 1981; Baxt et al., 1998). If human reviewers do not fully agree with one another, AI-human disagreement should not automatically be interpreted as evidence of AI failure. Instead, comparisons between human and AI reviewers should be considered within the broader context of reviewer variability and reliability. Consequently, measures such as inter-rater reliability, variance, and consistency are critical for understanding human and AI evaluation processes. This perspective is relevant to the present study because it examines not only agreement between humans and AI systems but also consistency within and across AI systems over multiple reviews.

2.4 Ethical and Sociological Considerations

These findings must be evaluated alongside broader ethical and institutional concerns. Hosseini and Horbach (2023) analyzed AI use across five reviewer dimensions, highlighting opacity, confidentiality risks, and poor reproducibility — for example, submitting identical prompts three weeks apart yielded markedly different outputs. Unattributed AI use also threatens the credibility of peer review and reviewers' skill development. In response, Elsevier (2024) permitted AI for manuscript drafting with mandatory disclosure but explicitly prohibited its use in review and editorial decision-making, citing the risk of incorrect or biased conclusions, while Farber (2025) advocated a hybrid approach that pairs human domain expertise with AI efficiency. These considerations highlight an important distinction:

rather than treating agreement with human reviewers as proof of AI accuracy, empirical evaluations should examine observable properties of AI assessments — consistency across repeated scoring, cross-model agreement, score variance, and rank ordering. Collectively, the literature suggests a potential tiered model in which AI supports screening and structural evaluation while expert human judgment remains central to evaluating novelty, rigor, and merit, as well as to publication decisions.

2.5 Research Gap

Despite this growing evidence, three gaps remain. Little research addresses AI review behavior within IS education conferences. Few studies compare multiple contemporary AI systems against the same human-reviewed manuscript corpus using a shared rubric — Joachim et al. (2025) is a notable exception, though their kappa-based analysis focused on binary accept/reject decisions on a medical manuscript sample. And while temporal stability has been examined at the model level, cross-model agreement within and across repeated review sessions remains largely unexamined. The present study begins to address these gaps by comparing repeated evaluations from ChatGPT, Claude, and Gemini against human-reviewer scores on ISCAP conference submissions, assessing agreement, variance, ranking behavior, relative scoring differences, and reliability across eight review criteria to inform whether AI-generated evaluations demonstrate the consistency and alignment needed to support AI-assisted peer review.

3. METHODOLOGY

3.1 Human Reviewer Dataset

The dataset consists of a convenience sample of 11 papers submitted to the ISCAP conference (formerly EDSIGCON/CONISAR) between 2021 and 2025, along with their original human-reviewer scores. All 11 manuscripts were accepted for publication and share at least one common co-author. Because ISCAP does not archive historical submissions or reviewer scores, this sample reflects all historical reviewer data cataloged by the authoring team. Given the small sample size ($N=11$), the restricted score range inherent in accepted papers, and the single-conference context, this study is explicitly structured as an exploratory benchmark rather than as a basis for generalized conclusions about AI review capabilities across broader academic domains.

Year	Paper ID	# Human Reviewers	# Co-Authors
2021	1-1	3	6
2021	1-2	2	5
2021	1-3	3	3
2021	1-4	4	7
2022	2-1	3	2
2022	2-2	3	4
2022	2-3	3	2
2023	3-1	3	2
2024	4-1	3	2
2025	5-1	3	4
2025	5-2	3	4

Table 1. Basic data for the 11 papers. Note # of reviewers for papers 1-2 and 1-4.

3.2 AI Review Generation

All AI reviews were generated using three popular general-purpose GenAI tools:

- OpenAI’s ChatGPT 5.5 Instant (OpenAI, 2026)
- Anthropic’s Claude Sonnet 4.6 Low (Anthropic, 2026)
- Google DeepMind’s Gemini 3.5 Flash (Google DeepMind, 2026)

We used the same rubric as the human reviewers (Figure 1), applied identical instructions and prompts across tools, and constrained results to numeric scores from 1 to 5 across eight categories. To generate each review, we opened a temporary/incognito chat window, attached the manuscript, attached a file containing the rubric in Figure 1, and provided the following prompt:

Using the guidance provided by Allen Lee in "Reviewing a Manuscript for Publication" (<https://allenslee.com/referee.htm>), please evaluate the attached manuscript and provide rankings on the attached scale for the ISCAP Conference (<https://iscap.us/>).

Evaluation By Category

[Scale: 1. N/A; 2. Reject; 3. Resubmit w/Changes; 4. Accept w/Minor; 5. Accept as Submitted]

Category	Your Ranking
Importance to IS/CIS Educators	☐ 1. N/A ☐ 2 ☐ 3 ☐ 4 ☐ 5
Originality of Approach	☐ 1. N/A ☐ 2 ☐ 3 ☐ 4 ☐ 5
Clarity of Subject/Paper	☐ 1. N/A ☐ 2 ☐ 3 ☐ 4 ☐ 5
Literature Review	☐ 1. N/A ☐ 2 ☐ 3 ☐ 4 ☐ 5
Justification of Methodology	☐ 1. N/A ☐ 2 ☐ 3 ☐ 4 ☐ 5
Summary and Conclusion	☐ 1. N/A ☐ 2 ☐ 3 ☐ 4 ☐ 5
Writing Style	☐ 1. N/A ☐ 2 ☐ 3 ☐ 4 ☐ 5

Overall recommendations for the conference and its proceedings

(Scale: 1. Not Appropriate for our Audience; 2. Reject; 3. Resubmit w/Changes; 4. Recommend w/Minor; 5. Highly Recommended)

Conference Recommendation
 ☐ 1
 ☐ 2
 ☐ 3
 ☐ 4
 ☐ 5

Figure 1: Rubric used to rate the 8 criteria for each paper

Evaluations were conducted from June to July 2026, with each evaluation initiated in a new temporary or incognito chat session. Each paper was uploaded alongside the rubric (Figure 1) and assigned a 1–5 numeric rating across the eight criteria, following Allen Lee's referee guidance. Originally submitted, anonymized manuscript versions were evaluated, matching the versions seen by human reviewers; both AI and human reviewers therefore evaluated the same manuscript version under the same blinding conditions, eliminating manuscript version or blinding as a source of discrepancy between conditions. Each paper was evaluated three times per model, producing 99 total synthetic reviews and 792 individual scores. To enable direct comparison with human reviewer sets, evaluations were aggregated both across platforms per run (e.g., Run 1 across ChatGPT, Claude, and Gemini) and within platforms across runs (e.g., ChatGPT Runs 1–3).

3.3 Measures and Aggregation

Eight evaluation criteria were scored by both human reviewers and AI systems using the conference's five-category ordinal rating scale (Tables 2–3).

Importance to IS/CIS Educators	Originality of Approach	Clarity of Subject/Paper
Literature Review	Justification of Methodology	Summary and Conclusion
Writing Style	Conference Recommendation	

Table 2. Evaluation criteria provided to paper reviewers.

Ordinal Score	First 7 Criteria	Conference Recommendation only
1	N/A	Not Appropriate for our Audience
2	Reject	Reject
3	Resubmit w/Changes	Resubmit w/Changes
4	Accept w/Minor	Recommend w/Minor
5	Accept as Submitted	Highly Recommended

Table 3. ISCAP ordinal rating scale guidance provided to the raters (see Figure 1).

Because a score of 1 indicates 'N/A' rather than a quality rating below 'Reject' for seven criteria, these ratings are strictly ordinal. Human raters assigned four 'N/A' scores, whereas AI raters selected none. To mirror conference protocol, these four entries were retained as numerical 1s; a sensitivity check that treated them as missing values confirmed that their inclusion slightly increased human variance without altering the direction of correlations or the primary score-compression findings.

Once completed, analyses were conducted across eight review groupings:

- Human reviewers (Human)
- AI Run 1 (AI1) – the first run from ChatGPT, Claude, and Gemini each
- AI Run 2 (AI2)
- AI Run 3 (AI3)
- ChatGPT runs 1-3 (ChatGPT3)
- Claude runs 1-3 (Claude3)
- Gemini runs 1-3 (Gemini3)
- All nine AI reviews combined (AI-All)

This multi-tiered aggregation scheme supported three analytic perspectives: cross-model agreement during a shared evaluation cycle (AI1–AI3), model-specific stability across repeated runs (ChatGPT3, Claude3, Gemini3), and overall collective AI behavior (AI-All).

3.4 Analysis Approach

We examined agreement, variance, and consistency via visual and quantitative comparisons, including rankings, heat maps, box plots, and histograms.

- RQ1: Human-AI Score Agreement. Analyzed through individual paper score correlations and ranking comparisons between humans and AI tools.
- RQ2: Variance and Scoring Behavior. Examined through distributional comparisons, score variance, and mean AI-human scoring differences.
- RQ3: Stability of AI Scores Over Repeated Runs. Assessed using quantitative comparisons over time and between AI systems. Inter-rater reliability was evaluated using Intraclass Correlation Coefficients (ICCs) based on Shrout and Fleiss (1979). Depending on rater distribution and measurement objective, we calculated single-measure variants -- One-Way Random Effects ICC(1,1), Two-Way Random Effects for Absolute Agreement ICC(2,1), and Two-Way Mixed Effects for Consistency ICC(3,1) -- and their corresponding average-measure counterparts based on k raters: ICC(1,k), ICC(2,k), and ICC(3,k) (McGraw & Wong, 1996). The one-way model was applied to the human panel because rater counts varied by paper (2, 3, or 4 reviewers) and reviewer identities were not matched across papers, whereas the two-way models were applied to the AI systems because every paper received a complete, identically structured set of model/run ratings.

Statistical calculations were conducted using multiple GenAI tools and web-based statistical calculators. Krippendorff's alpha was calculated using the K-Alpha Calculator (Marzi et al., 2024), with an ordinal level of measurement specified. For the primary analyses, N/A ratings were retained as numerical values of 1 to mirror the original conference scoring protocol; the sensitivity analysis treated these values as missing.

4. RESULTS

Because the 11 papers constitute a small, convenience-based sample, these statistical analyses are primarily exploratory.

4.1 Overview of Sample and Descriptive Statistics

The dataset comprises 11 papers reviewed by 33 human reviewers (3 per paper, except papers 1-2 and 1-4, which had 2 and 4 reviewers, respectively) across eight rubric criteria. Each AI tool reviewed each paper three times, producing 99 AI-generated reviews (Appendix A).

Table 4 summarizes the overall mean and standard deviation, while Table 5 presents category-level

breakdowns. Human reviewers demonstrated greater score variance across papers and criteria than AI systems. In this sample, ChatGPT showed descriptively ~25% less variability than humans, while Claude and Gemini showed descriptively ~34% less variability; these are sample differences rather than statistically tested effects. Across criteria, human standard deviations ranged from 0.68 to 1.19, compared to AI standard deviations generally ranging from 0.39 to 0.74 (excluding ChatGPT's 0.00 SD for Summary and Conclusion).

Human	ChatGPT	Claude	Gemini
4.14 +/- 0.93	3.82 +/- 0.70	3.90 +/- 0.61	3.90 +/- 0.61

Table 4. Mean and standard deviation for assigned scores by reviewer type.

Category	Human	ChatGPT	Claude	Gemini
Conference Recommendation	4.33 +/- 0.68	3.58 +/- 0.49	3.82 +/- 0.52	3.64 +/- 0.59
Importance to IS/CIS Educators	4.21 +/- 0.98	4.58 +/- 0.55	4.39 +/- 0.49	4.12 +/- 0.59
Originality of Approach	4.15 +/- 0.70	3.64 +/- 0.54	3.73 +/- 0.45	3.36 +/- 0.69
Clarity of Subject/Paper	4.27 +/- 0.79	4.18 +/- 0.46	3.94 +/- 0.65	4.24 +/- 0.43
Literature Review	4.09 +/- 1.19	3.48 +/- 0.74	3.76 +/- 0.70	3.48 +/- 0.74
Justification of Methodology	3.97 +/- 1.06	2.88 +/- 0.41	3.48 +/- 0.66	3.15 +/- 0.74
Summary and Conclusion	3.97 +/- 0.83	4.00 +/- 0.00	3.97 +/- 0.39	3.88 +/- 0.48
Writing Style	4.15 +/- 0.99	4.21 +/- 0.41	4.12 +/- 0.48	4.64 +/- 0.48

Table 5. Mean and standard deviation for assigned scores by category and reviewer type.

Figure 2 illustrates score frequency distributions. Humans utilized the full 1–5 range, whereas AI systems selected no 'N/A' ratings and assigned far fewer 5s, exhibiting a compressed score distribution.

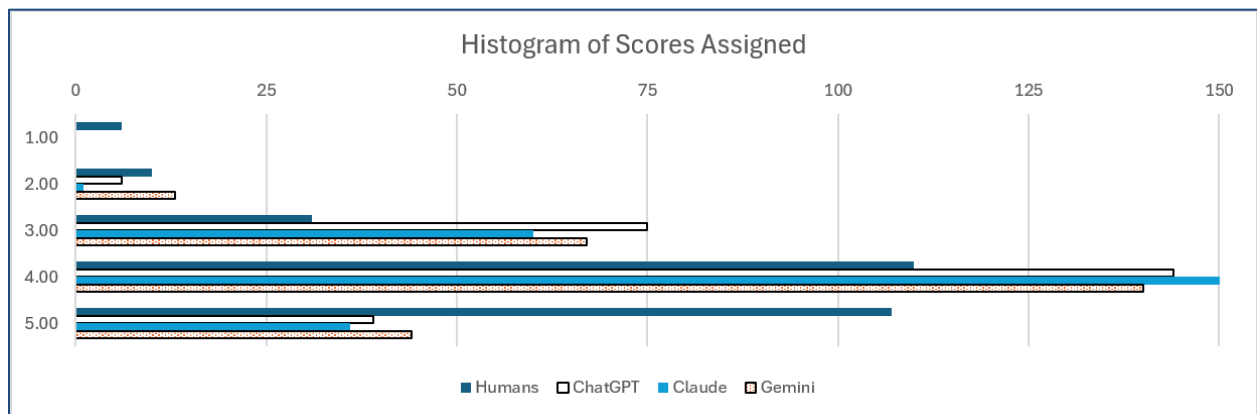


Figure 2: Histogram of the 264 scores assigned by each of the various reviewers.

Table 6 presents mean paper scores as a heat map (complete heat maps are in Appendices B, D, and E).

4.2 RQ1: Human-AI Score Agreement

RQ1: To what extent do AI-generated review scores correlate with human-reviewer scores?

Overall, human-AI rating agreement was weak to negative across most settings, indicating limited alignment between human and AI scores. Table 7 presents criterion-level associations. The strongest positive association was observed for Importance to IS/CIS Educators ($r = 0.753$, ChatGPT). The strongest inverse association was observed for Claude on Writing Style ($\rho = -0.778$), with a similarly strong inverse Pearson association for Clarity of Subject/Paper ($r = -0.688$).

Human - mean											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	4.00	4.50	4.33	4.75	3.67	5.00	4.33	4.67	4.67	3.67	4.00
Importance to IS/CIS Educators	3.33	3.00	4.67	4.75	3.33	4.33	4.67	4.67	4.33	4.00	4.67
Originality of Approach	3.33	5.00	4.67	4.50	3.67	4.00	4.00	4.00	4.00	4.00	4.67
Clarity of Subject/Paper	4.00	5.00	4.33	4.75	3.00	5.00	4.33	4.67	4.67	3.67	3.67
Literature Review	4.33	4.50	4.33	5.00	2.00	5.00	3.00	4.67	4.67	3.33	4.00
Justification of Methodology	3.67	4.50	4.00	4.75	2.67	4.00	3.00	5.00	4.67	3.33	4.00
Summary and Conclusion	4.00	5.00	4.33	4.50	2.67	4.00	4.00	4.33	4.67	3.00	3.33
Writing Style	4.00	5.00	4.67	4.50	3.00	5.00	4.33	4.33	3.33	3.33	4.33
ChatGPT - mean											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	4.00	3.00	3.00	3.33	3.67	3.67	3.67	3.33	3.67	4.00	4.00
Importance to IS/CIS Educators	4.33	3.67	4.67	5.00	4.00	4.67	4.67	5.00	5.00	5.00	4.33
Originality of Approach	3.00	3.00	4.33	3.67	3.00	3.67	3.67	3.67	4.00	4.00	4.00
Clarity of Subject/Paper	4.33	4.00	3.67	4.00	4.67	4.33	4.00	4.67	4.00	4.00	4.33
Literature Review	3.33	2.67	3.00	4.33	3.00	4.33	3.00	3.00	3.67	3.67	4.33
Justification of Methodology	3.00	3.00	2.33	2.67	3.00	2.67	2.67	3.00	3.33	3.00	3.00
Summary and Conclusion	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Writing Style	4.33	4.00	4.00	4.00	4.67	4.33	4.00	4.67	4.00	4.00	4.33
Claude - mean											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	3.67	3.33	4.00	3.67	3.67	4.33	3.67	4.00	3.67	4.00	4.00
Importance to IS/CIS Educators	4.33	4.00	4.33	4.67	4.67	5.00	4.00	4.67	4.00	4.00	4.67
Originality of Approach	3.33	3.00	4.00	3.33	4.00	4.00	3.67	4.00	3.67	4.00	4.00
Clarity of Subject/Paper	4.33	3.67	3.00	3.67	4.00	4.33	4.00	4.00	4.00	4.00	4.33
Literature Review	3.67	2.67	3.67	3.67	4.00	4.33	3.33	3.67	4.33	4.67	3.33
Justification of Methodology	3.33	3.00	3.00	3.00	3.33	3.33	3.33	4.67	3.67	4.33	3.33
Summary and Conclusion	3.67	4.00	4.00	4.00	3.67	4.33	4.33	4.00	4.00	4.00	3.67
Writing Style	4.33	3.67	4.00	3.67	4.33	4.33	4.33	4.00	4.00	4.00	4.67
Gemini - mean											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	3.67	2.67	4.00	3.33	4.00	3.33	3.67	3.33	3.67	4.33	4.00
Importance to IS/CIS Educators	4.00	3.00	4.33	4.33	4.33	4.00	4.00	4.00	4.00	4.67	4.67
Originality of Approach	3.00	2.33	4.00	3.33	3.33	3.00	3.00	3.00	3.33	4.33	4.33
Clarity of Subject/Paper	4.33	4.00	4.00	4.00	4.00	4.00	4.00	4.67	4.33	4.67	4.67
Literature Review	3.33	2.00	3.33	3.67	3.33	4.00	2.67	3.67	4.00	4.00	4.33
Justification of Methodology	3.00	2.67	3.00	2.33	3.33	2.33	3.00	3.67	4.00	4.00	3.33
Summary and Conclusion	4.00	3.33	4.00	3.33	4.33	3.67	4.00	4.33	3.67	4.00	4.00
Writing Style	4.67	4.00	5.00	4.33	5.00	4.67	4.67	4.67	4.33	5.00	4.67

Table 6. Mean score heat map by paper across categories for each rater type. Note: Color gradients in Table 6 and Appendices B and D apply a relative 3-color scale (Green = 5.00/High, Yellow = 3.50–4.00/Moderate, Red = 1.00–2.50/Low) to visually emphasize score distributions and highlight divergence across evaluation criteria and rater groups.

Category	ChatGPT		Claude		Gemini	
	Pearson r	Spearman rho	Pearson r	Spearman rho	Pearson r	Spearman rho
Conference Recommendation	-0.497	-0.526	-0.509	-0.361	-0.238	-0.146
Importance to IS/CIS Educators	0.753	0.571	0.137	0.276	0.570	0.421
Originality of Approach	0.322	0.384	0.330	0.272	0.071	0.161
Clarity of Subject/Paper	-0.340	-0.246	-0.688	-0.637	-0.044	0.051
Literature Review	0.380	0.377	-0.095	-0.028	-0.214	-0.076
Justification of Methodology	0.127	0.106	-0.044	0.053	0.054	0.211
Summary and Conclusion	n/a	n/a	-0.168	-0.091	-0.224	-0.307
Writing Style	-0.212	-0.242	-0.642	-0.778	-0.131	0.068

Table 7. Measures of association between humans and the different AI tools. Note: Given the small sample size ($N=11$), the correlation estimates are highly uncertain. For example, the strongest observed Pearson correlation ($r = 0.753$) has an approximate 95% CI of $[0.29, 0.93]$, while a correlation of $r = 0.10$ has an approximate 95% CI of $[-0.53, 0.66]$. Accordingly, the correlations in Table 7 should be interpreted as exploratory estimates rather than precise measures of association.

At the paper level, human and AI rank ordering showed limited overlap (Table 8). There was no overlap with human raters' #1 or #2 papers, and only partial overlap with #3. The humans' #2 paper was ranked last by all AI groups. Conversely, AI groups showed notable mutual overlap on papers 5-1 and 5-2; paper 5-1 ranked in the top 3 across all AI groupings but second-to-last among human reviewers. Overall paper rankings in Table 8 were calculated as the unweighted mean across all eight rubric criteria for each reviewer group, treating each criterion as equally important; no criterion-specific weighting was applied.

Paper	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
1-1	9	4	8	9	7	7	7	7
1-2	2	11	11	11	11	11	11	11
1-3	5	7	6	7	8	10	9	5
1-4	1	10	9	8	9	6	10	8
2-1	11	6	4	6	6	8	5	4
2-2	4	9	2	3	4	3	1	9
2-3	8	8	10	10	10	9	8	10
3-1	3	5	1	5	3	5	3	3
4-1	6	3	7	4	5	4	6	6
5-1	10	1	3	1	1	2	2	1
5-2	7	2	5	2	2	1	4	2

Table 8: Comparing top-ranked papers among reviewer groups (Section 3.3).

Examination of the two largest human–AI ranking reversals suggests that different mechanisms contributed to the discrepancies. For Paper 5-1, which ranked near the top for AI but near the bottom for humans, the human ratings showed substantial inter-reviewer variation, with one reviewer assigning markedly lower scores than the others. In contrast, Paper 1-2 received consistently high human scores on most criteria but substantially lower AI scores for Originality of Approach, Literature Review, and Justification of Methodology. Thus, the largest ranking reversals appear to reflect both human inter-reviewer variability and criterion-specific differences between human and AI evaluations.

4.3 RQ2: Variance and Scoring Behavior

RQ2: Does AI demonstrate different scoring variances than humans?

AI-human scoring differences were calculated by subtracting the human mean score from the AI mean score for each criterion (Figure 3 and Table 9). Positive values indicate relative AI leniency and negative values indicate relative AI severity.

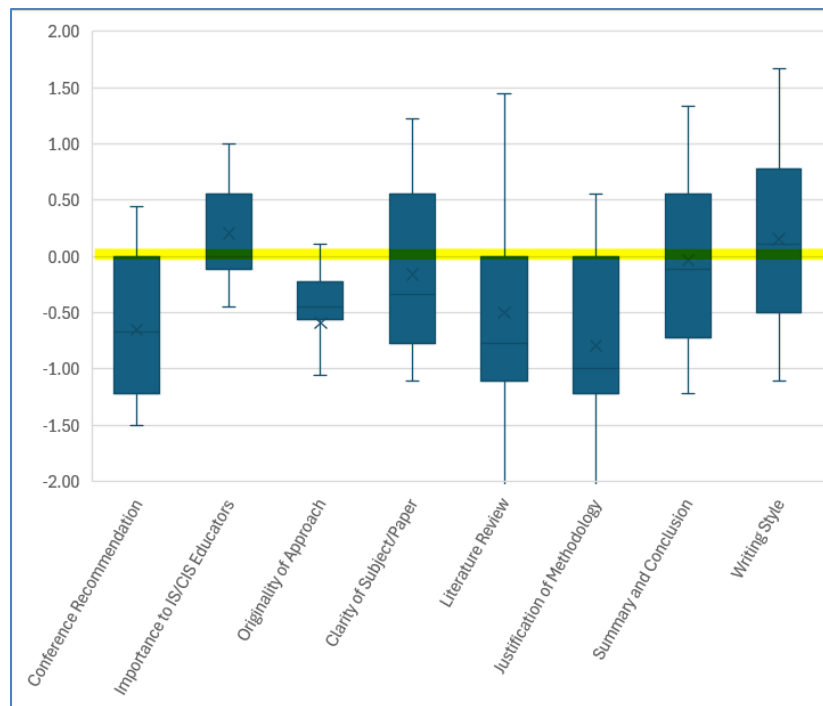


Figure 3: Box plots of the difference between the mean score for all 9 AI runs and the mean score for all human ratings across the 11 papers.

AI Severity	AI Leniency	Even
Conference Recommendation	Importance to IS/CIS Educators	Clarity of Subject/Paper
Originality of Approach	Writing Style	Summary and Conclusion
Literature Review		
Justification of Methodology		

Table 9. Distribution of review categories by AI severity, leniency, or evenness.

Broken out by model (Figure 4 and Table 10), Claude displayed higher leniency for Importance to IS/CIS Educators, Gemini was most severe on Originality of Approach, and ChatGPT exhibited the strongest severity toward Justification of Methodology (-1.08). Gemini showed notable leniency on Writing Style (+0.42).

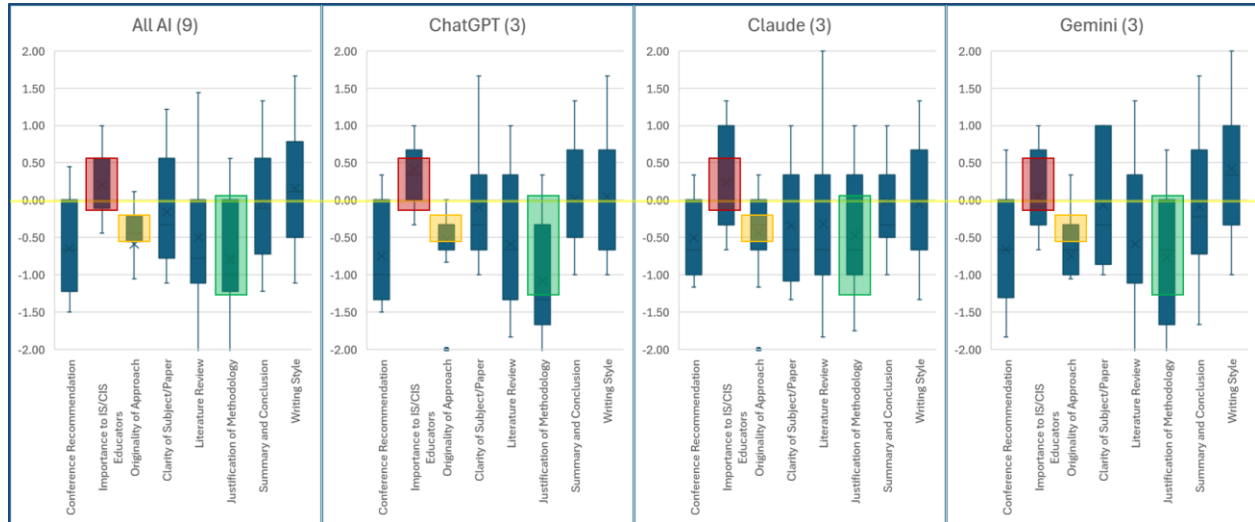


Figure 4: Box plots of the mean difference broken out by specific AI tools. Note: The first panel is a smaller version of Figure 3.

Category	All AI	ChatGPT	Claude	Gemini
Conference Recommendation	-0.65	-0.75	-0.51	-0.66
Importance to IS/CIS Educators	0.20	0.42	0.23	0.04
Originality of Approach	-0.59	-0.53	-0.44	-0.74
Clarity of Subject/Paper	-0.16	-0.10	-0.34	-0.07
Literature Review	-0.50	-0.59	-0.32	-0.59
Justification of Methodology	-0.79	-1.08	-0.48	-0.77
Summary and Conclusion	-0.04	0.02	-0.02	-0.09
Writing Style	0.16	0.05	-0.05	0.42

Table 10: Mean AI-human scoring differences by AI tool; negative values (red) indicate relative AI severity and positive values (green) indicate relative AI leniency.

4.4 RQ3: Stability of AI Scores Over Repeated Runs

RQ3: How stable are AI-generated evaluations across repeated review sessions?

Krippendorff's Alpha and ICC results (Appendix C) provide two complementary perspectives on AI reliability: stability of each individual model across repeated evaluations and agreement among different AI models.

Within-model stability. To evaluate the stability of each AI model across its three repeated evaluations, the three runs were treated as a fixed set of raters and evaluated using two-way mixed-effects consistency ICCs, ICC(3,1) and ICC(3,3). Results for ChatGPT, Claude, and Gemini varied substantially across criteria (Appendix C), indicating that repeated evaluations by the same model were more stable for some rubric dimensions than others. Originality of Approach showed relatively strong

repeated-run stability across the AI models, while several other criteria showed considerably weaker stability. These results indicate that repeated evaluations from the same model are not fully deterministic.

Cross-model agreement. A separate analysis examined agreement among ChatGPT, Claude, and Gemini within each evaluation run. Because this analysis concerns absolute agreement among different models, two-way random-effects ICCs, ICC(2,1) and ICC(2,3), were used. Across Runs 1 and 2, ICC(2,3) values of 0.653 and 0.592 indicated moderate cross-model agreement. In Run 3, however, ICC(2,3) fell to -0.818. The corresponding consistency estimate, ICC(3,3), was -1.080. Negative sample ICC estimates can occur when within-paper disagreement exceeds the estimated between-paper variance. Thus, the negative estimates observed in Run 3 indicate very poor reliability among the three models for this sample and should not be interpreted as evidence that the models systematically rated papers in opposite directions.

Given the small sample of 11 papers, these reliability estimates should be interpreted as exploratory. Confidence intervals around the estimates are expected to be wide.

4.5 Results Summary

Taken together, the results suggest:

1. Human and AI review scores showed weak to negative overall agreement.
2. AI scores exhibited lower variance than human scores, while repeated-run reliability varied across models and evaluation criteria.
3. AI scoring displayed a relative severity pattern for four of the eight categories.
4. Rank-ordered comparisons showed minimal overlap between AI and human judgments for top papers.

5. DISCUSSION

AI systems exhibited lower variance in scores than human reviewers but reached different conclusions about paper quality. Human reviewers exhibited modest agreement across criteria, reflecting the inherent variability of scholarly peer review. AI repeated-run reliability varied across models and evaluation criteria, while agreement with human reviewers remained weak. This distinction between reliability and agreement is critical: an evaluator can be internally consistent while applying evaluation standards that differ substantially from human raters.

5.1 Human-AI Agreement

While generative AI models exhibited lower variance in scores than human reviewers, they did not replicate human evaluative outcomes. Weak or negative human-AI correlations were driven by divergent paper rankings (e.g., papers 5-1 and 5-2 were top-ranked by AI but low-ranked by humans). One possible explanation for this divergence is that human reviewers may place greater weight on domain novelty, methodological rigor, and contextual significance, whereas AI models may respond more strongly to structural and presentation attributes. Because the present study analyzes numerical scores rather than reviewer reasoning, this interpretation requires further validation. Furthermore, because human inter-rater reliability was low across several categories, AI-human disagreement partly reflects the inherently subjective nature of human peer review.

5.2 Variance, Score Compression, and Relative Severity/Leniency

Across nearly every analysis, AI systems exhibited lower score variability than human reviewers. By selecting no 'N/A' ratings and assigning fewer perfect scores, AI models demonstrated score compression, concentrating evaluations toward the middle of the scale. While this compression reduces scoring variability, it may also limit AI's ability to distinguish exceptional papers. Additionally, AI scoring showed a consistent pattern of relative severity toward Conference Recommendation, Originality, Literature Review, and Methodology, and relative leniency toward Importance to Educators and Writing Style, when measured against the human reviewer baseline. This pattern may reflect differences in how AI and human reviewers weight observable structural characteristics and domain-specific

considerations, although the present study does not directly measure reviewer reasoning.

5.3 Stability of AI Evaluations

Repeated evaluations indicate that AI-generated reviews are non-deterministic, with stability varying across models and evaluation criteria. Scoring variations across criteria, such as Clarity of Subject/Paper, demonstrate that LLMs can reach different conclusions on identical manuscripts under identical instructions. The mixed reliability results, including poor cross-model agreement in one evaluation run, reinforce the need for caution when considering AI-generated evaluations as substitutes for human reviewers.

5.4 Implications for AI-Assisted Peer Review

Through a BPR lens, AI provides value by enabling a structural redesign of the peer-review workflow to optimize task allocation. The empirical findings suggest a potential division of labor within a reengineered review process. The observed scoring patterns indicate that AI may be useful in supporting the evaluation of more directly observable structural and mechanical dimensions, while substantive judgments concerning novelty, methodological rigor, and publication merit remain the appropriate responsibility of expert human reviewers. This proposed allocation should be evaluated directly in future research.

5.5 Discussion Summary

In summary, contemporary LLMs cannot substitute for human reviewers because they exhibit score compression, display a systematic severity pattern relative to the human baseline, and produce paper rankings that diverge from human reviewer rankings. However, their lower scoring variability and ability to apply a common evaluation rubric suggest a potential supporting role for mechanical evaluation tasks within a reengineered peer-review process.

5.6 Limitations

This study has several limitations that should be considered when interpreting the results.

- **Sample Constraints:** The sample is limited to 11 accepted ISCAP papers sharing at least one co-author, serving as an exploratory benchmark rather than a basis for generalized conclusions. The authors are currently working with ISCAP conference organizers to obtain a substantially larger sample (likely 100+ papers, pending author permission) from this year's submissions, including rejected manuscripts, to address this limitation in future work.
- **Absence of Rejections:** With no rejected papers available, the study could not evaluate AI systems' ability to identify unacceptable submissions.
- **Human Benchmark Variability:** Modest agreement among human reviewers indicates that AI-human divergence partly reflects underlying human disagreement.
- **Measurement Scale:** Because a score of 1 denotes "N/A" rather than a quality judgment below "Reject" for seven of the eight criteria, these ratings are ordinal rather than true Likert-interval data. A sensitivity check confirmed this did not alter the study's primary conclusions, but means, correlations, and reliability coefficients calculated on this scale should be interpreted accordingly.
- **Reviewer Expertise:** Human reviewer expertise and domain familiarity were not controlled or measured, and may vary across the original review panels in ways that also contribute to human-AI disagreement.
- **Prior Model Exposure:** Because all sample papers were previously accepted and are potentially indexed online, current AI models may have encountered these manuscripts (or citing works) during pretraining, which could influence scores in ways not present when evaluating truly unseen submissions.
- **Temporal Snapshot:** Evaluated AI model versions represent a specific snapshot in time (mid-2026); future architectures may display different scoring behaviors.
- **Quantitative Focus:** The study focused on numerical scores without qualitative text analysis of reviewer comments.

Despite these limitations, the study provides a rare comparison of three leading AI systems against

actual evaluations by human reviewers. The use of multiple AI platforms, repeated review sessions, and several complementary measures of agreement, severity/leniency, and reliability provides an initial empirical foundation for understanding how generative AI may augment the scholarly review process in the future.

6. CONCLUSIONS & FUTURE WORK

Applying a Business Process Reengineering (BPR) framework, this study evaluated Generative AI as a workflow-enabling technology rather than a replacement for human judgment. AI systems exhibited lower variance in scores than human reviewers (RQ2), while repeated-run reliability varied across models and evaluation criteria (RQ3). AI evaluations also diverged substantially from human outcomes and top-paper rankings (RQ1), exhibiting score compression, relative methodological severity, and relative writing-style leniency. Furthermore, variation across repeated runs demonstrates that current LLM evaluations are not fully deterministic.

Collectively, these results suggest a potential role for GenAI within a hybrid review architecture, including upstream support for structural, mechanical, and formatting evaluation while retaining human authority over research innovation, methodological rigor, and publication decisions. The effectiveness of such a division of labor requires further empirical validation.

6.1 Future Work

Future research should expand this exploratory benchmark in three directions. First, the authors are working with ISCAP conference organizers to obtain this year's submitted manuscripts and associated reviewer scores, pending author permission, which would provide a substantially larger sample (likely 100+ papers) than the present study's 11. Critically, this dataset would include both rejected and accepted submissions, allowing a direct test of whether AI systems can meaningfully distinguish paper quality across the full acceptance-range distribution rather than only among already accepted work. Second, repeating this evaluation with updated or fine-tuned model versions would indicate whether the observed score compression and weak human-AI agreement are stable properties of current LLMs or artifacts of the specific models and prompting approach used here. Third, incorporating qualitative analysis of AI-generated review text, alongside the numeric scores examined in this study, would clarify whether AI reasoning aligns with human reviewers' stated rationale even when final scores diverge.

7. REFERENCES

- Aczel, B., Szász, B., & Holcombe, A. (2021). A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, 6(1). <https://doi.org/10.1186/s41073-021-00118-2>.
- Anthropic. (2026, February 17). Introducing Claude Sonnet 4.6. <https://www.anthropic.com/news/claude-sonnet-4-6>
- Baxt, W. G., Waeckerle, J. F., Berlin, J. A., & Callahan, M. L. (1998). Who reviews the reviewers? Feasibility of using a template to reduce reviewer bias. *Annals of Emergency Medicine*, 32(5), 570-578. [https://doi.org/10.1016/S0196-0644\(98\)70034-3](https://doi.org/10.1016/S0196-0644(98)70034-3)
- 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., Mutz, R., & Daniel, H.D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLOS ONE*, 5(12). <https://doi.org/10.1371/journal.pone.0014331>
- Checchio, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, 8(25). <https://doi.org/10.1057/s41599-020-00703-8>

- Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer review. *Science*, 214(4523), 881-886. <https://doi.org/10.1126/science.7302566>
- Crowe, I. (2026, June 23). 33% surge in scholarly submissions strains US publishing. *Academic Jobs*. <https://www.academicjobs.com/us/higher-education-news/33percent-surge-in-scholarly-submissions-strains-us-publishing-or-academicjobs-24546>
- Elsevier. (2024). Generative AI policies for journals. Elsevier. <https://www.elsevier.com/about/policies/publishing-ethics/generative-ai-policies-for-journals>
- Farber, S. (2025). Comparing human and AI expertise in the academic peer review process: towards a hybrid approach. *Higher Education Research & Development*, 44(4), 871-885. <https://doi.org/10.1080/07294360.2024.2445575>
- Google DeepMind. (2026). Gemini 3.5 Flash. <https://deepmind.google/models/gemini/flash/>
- Hammer, M. (1990). Reengineering work: Don't automate, obliterate. *Harvard Business Review*, 1-8.
- 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(4). <https://doi.org/10.1186/s41073-023-00133-5>
- Housel, T., & Bell, A. (2001). *Measuring and managing knowledge*. McGraw-Hill/Irwin.
- Joachim, M. V., Dodson, T. B., & Laviv, A. (2025). How artificial intelligence differs from humans in peer review. *Journal of Oral and Maxillofacial Surgery*, 83, 1040-1050. <https://doi.org/10.1016/j.joms.2025.03.015>
- Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024a). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *Proceedings of the 41st International Conference on Machine Learning*, 235, 29575-29620. <https://dl.acm.org/doi/10.5555/3692070.3693262>
- 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. (2024b). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, 1(8). <https://doi.org/10.1056/AIoa2400196>
- Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha Calculator—Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12, 102545. <https://doi.org/10.1016/j.mex.2023.102545>
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1), 30. <https://doi.org/10.1037/1082-989X.1.1.30>
- Miller, K., & Mun, J. (2023). Cyber technologies, machine learning, additive manufacturing, and cloud in the box to enable optimized maintenance processes in extreme conditions. In A. Moallem (Ed.), *HCI for Cybersecurity, Privacy and Trust* (pp. 672-684). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35822-7_43
- OpenAI. (2026, May 5). GPT-5.5 Instant: Smarter, clearer, and more personalized. <https://openai.com/index/gpt-5-5-instant/>
- Publons. (2018). Global state of peer review 2018. Clarivate Analytics. <https://doi.org/10.14322/publons.GSPR2018>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2), 420. <https://doi.org/10.1037/0033-2909.86.2.420>

Appendix A – All Scores by Reviewer Type

Humans	5-2	5-1	4-1	3-1	2-3	2-2	2-1	1-4	1-3	1-2	1-1
Conference Recommendation	4 4 4	5 4 2	4 5 5	5 5 4	5 4 4	5 5 5	4 4 3	5 5 4 5	4 5	4 5 4	4 4 4
Importance to IS/CIS Educators	5 5 4	5 5 2	4 5 4	5 5 4	5 4 5	4 5 4	3 4 3	4 5 5 5	4 5	5 1 5	3 4 3
Originality of Approach	4 5 5	5 4 3	3 4 5	4 4 4	4 4 4	3 5 4	4 4 3	4 5 5 4	4 5	5 5 5	3 3 4
Clarity of Subject/Paper	3 4 4	5 3 3	4 5 5	5 5 4	5 4 4	5 5 5	3 4 2	5 4 5 5	4 4	5 5 5	4 4 4
Literature Review	5 4 3	3 5 2	4 5 5	5 5 4	4 4 1	5 5 5	3 1 2	5 5 5 5	3 5	5 4 5	4 4 5
Justification of Methodology	4 4 4	4 4 2	4 5 5	5 5 5	4 4 1	3 5 4	3 4 1	4 5 5 5	3 4	5 4 5	3 4 4
Summary and Conclusion	3 4 3	4 3 2	4 5 5	4 5 4	4 4 4	3 5 4	3 3 2	4 4 5 5	4 4	5 5 5	4 4 4
Writing Style	4 4 5	4 4 2	4 5 1	5 5 3	5 4 4	5 5 5	3 4 2	5 4 4 5	4 5	5 5 5	4 4 4

Figure A-1: All human-reviewer scores; note highlighted cells call attention to the papers that have 4 and 2 human reviews versus the standard 3.

ChatGPT	5-2	5-1	4-1	3-1	2-3	2-2	2-1	1-4	1-3	1-2	1-1
Conference Recommendation	4 4 4	4 4 4	3 4 4	3 4 3	4 4 3	3 4 4	4 4 3	3 4 3	3 3 3	3 3 3	4 4 4
Importance to IS/CIS Educators	4 4 5	5 5 5	5 5 5	5 5 5	4 5 5	4 5 5	4 4 4	5 5 5	4 5 5	4 3 4	4 4 5
Originality of Approach	4 4 4	4 4 4	4 4 4	3 4 4	3 4 4	3 4 4	3 3 3	3 4 4	4 4 5	3 3 3	3 3 3
Clarity of Subject/Paper	5 4 4	4 4 4	4 4 4	4 5 5	4 4 4	4 5 4	4 5 5	4 4 4	3 4 4	4 4 4	4 5 4
Literature Review	3 5 5	3 4 4	3 4 4	3 3 3	3 3 3	4 4 5	3 3 3	4 5 4	3 3 3	3 2 3	3 4 3
Justification of Methodology	3 3 3	3 3 3	3 4 3	3 3 3	3 3 2	2 3 3	3 3 3	3 3 2	2 2 3	3 3 3	3 3 3
Summary and Conclusion	4 4 4	4 4 4	4 4 4	4 4 4	4 4 4	4 4 4	4 4 4	4 4 4	4 4 4	4 4 4	4 4 4
Writing Style	5 4 4	4 4 4	4 4 4	4 5 5	4 4 4	4 5 4	4 5 5	4 4 4	4 4 4	4 4 4	4 5 4
Claude	5-2	5-1	4-1	3-1	2-3	2-2	2-1	1-4	1-3	1-2	1-1
Conference Recommendation	4 4 4	4 4 4	4 3 4	4 4 4	4 3 4	4 5 4	3 5 3	4 4 3	4 4 4	3 4 3	4 4 3
Importance to IS/CIS Educators	5 4 5	4 4 4	4 4 4	5 5 4	4 4 4	5 5 5	5 5 4	5 5 4	5 4 4	4 4 4	5 4 4
Originality of Approach	4 4 4	4 4 4	4 3 4	4 4 4	4 3 4	4 4 4	4 4 4	4 3 3	4 4 4	3 3 3	4 3 3
Clarity of Subject/Paper	5 4 4	4 4 4	4 4 4	5 4 3	4 3 5	4 5 4	4 5 3	3 4 4	3 3 3	3 4 4	5 4 4
Literature Review	4 3 3	4 5 5	5 4 4	4 4 3	3 3 4	4 5 4	4 4 4	4 4 3	4 4 3	2 3 3	4 4 3
Justification of Methodology	4 3 3	4 5 4	4 3 4	5 5 4	3 3 4	3 4 3	3 4 3	3 3 3	3 3 3	3 3 3	4 3 3
Summary and Conclusion	4 4 3	4 4 4	4 4 4	4 4 4	4 4 5	4 5 4	4 4 3	4 4 4	4 4 4	4 4 4	4 3 4
Writing Style	5 4 5	4 4 4	4 4 4	4 4 4	4 4 5	4 5 4	4 5 4	4 3 4	4 4 4	3 4 4	5 4 4
Gemini	5-2	5-1	4-1	3-1	2-3	2-2	2-1	1-4	1-3	1-2	1-1
Conference Recommendation	4 4 4	5 4 4	4 3 4	3 4 3	4 4 3	3 3 4	4 4 4	3 3 4	4 4 4	3 3 2	4 3 4
Importance to IS/CIS Educators	4 5 5	5 4 5	4 4 4	4 4 4	4 4 4	4 4 4	4 4 5	4 4 5	4 5 4	4 3 2	4 4 4
Originality of Approach	4 4 5	4 4 5	4 3 3	3 3 3	3 3 3	3 3 3	3 3 4	3 3 4	4 4 4	3 2 2	3 3 3
Clarity of Subject/Paper	5 4 5	5 4 5	5 4 4	4 5 5	4 4 4	4 4 4	4 4 4	4 4 4	4 4 4	4 4 4	4 4 5
Literature Review	4 5 4	4 4 4	4 4 4	3 4 4	3 3 2	4 4 4	4 3 3	3 4 4	3 4 3	2 2 2	3 3 4
Justification of Methodology	3 3 4	5 3 4	5 3 4	4 4 3	3 3 3	2 2 3	3 3 4	2 2 3	3 3 3	3 2 3	3 3 3
Summary and Conclusion	4 4 4	4 4 4	4 4 3	4 5 4	4 4 4	4 3 4	4 5 4	3 3 4	4 4 4	4 3 3	4 4 4
Writing Style	5 4 5	5 5 5	5 4 4	4 5 5	5 4 5	4 5 5	5 5 5	4 4 5	5 5 5	4 4 4	5 5 4

Figure A-2: All AI-reviewer scores; each paper was reviewed 3 times by each AI.

Appendix B – Mean Category Scores by Paper by Reviewer Groups

Human - mean											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	4.00	4.50	4.33	4.75	3.67	5.00	4.33	4.67	4.67	3.67	4.00
Importance to IS/CIS Educators	3.33	3.00	4.67	4.75	3.33	4.33	4.67	4.67	4.33	4.00	4.67
Originality of Approach	3.33	5.00	4.67	4.50	3.67	4.00	4.00	4.00	4.00	4.00	4.67
Clarity of Subject/Paper	4.00	5.00	4.33	4.75	3.00	5.00	4.33	4.67	4.67	3.67	3.67
Literature Review	4.33	4.50	4.33	5.00	2.00	5.00	3.00	4.67	4.67	3.33	4.00
Justification of Methodology	3.67	4.50	4.00	4.75	2.67	4.00	3.00	5.00	4.67	3.33	4.00
Summary and Conclusion	4.00	5.00	4.33	4.50	2.67	4.00	4.00	4.33	4.67	3.00	3.33
Writing Style	4.00	5.00	4.67	4.50	3.00	5.00	4.33	4.33	3.33	3.33	4.33
ChatGPT - mean											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	4.00	3.00	3.00	3.33	3.67	3.67	3.67	3.33	3.67	4.00	4.00
Importance to IS/CIS Educators	4.33	3.67	4.67	5.00	4.00	4.67	4.67	5.00	5.00	5.00	4.33
Originality of Approach	3.00	3.00	4.33	3.67	3.00	3.67	3.67	3.67	4.00	4.00	4.00
Clarity of Subject/Paper	4.33	4.00	3.67	4.00	4.67	4.33	4.00	4.67	4.00	4.00	4.33
Literature Review	3.33	2.67	3.00	4.33	3.00	4.33	3.00	3.00	3.67	3.67	4.33
Justification of Methodology	3.00	3.00	2.33	2.67	3.00	2.67	2.67	3.00	3.33	3.00	3.00
Summary and Conclusion	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Writing Style	4.33	4.00	4.00	4.00	4.67	4.33	4.00	4.67	4.00	4.00	4.33
Claude - mean											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	3.67	3.33	4.00	3.67	3.67	4.33	3.67	4.00	3.67	4.00	4.00
Importance to IS/CIS Educators	4.33	4.00	4.33	4.67	4.67	5.00	4.00	4.67	4.00	4.00	4.67
Originality of Approach	3.33	3.00	4.00	3.33	4.00	4.00	3.67	4.00	3.67	4.00	4.00
Clarity of Subject/Paper	4.33	3.67	3.00	3.67	4.00	4.33	4.00	4.00	4.00	4.00	4.33
Literature Review	3.67	2.67	3.67	3.67	4.00	4.33	3.33	3.67	4.33	4.67	3.33
Justification of Methodology	3.33	3.00	3.00	3.00	3.33	3.33	3.33	4.67	3.67	4.33	3.33
Summary and Conclusion	3.67	4.00	4.00	4.00	3.67	4.33	4.33	4.00	4.00	4.00	3.67
Writing Style	4.33	3.67	4.00	3.67	4.33	4.33	4.33	4.00	4.00	4.00	4.67
Gemini - mean											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	3.67	2.67	4.00	3.33	4.00	3.33	3.67	3.33	3.67	4.33	4.00
Importance to IS/CIS Educators	4.00	3.00	4.33	4.33	4.33	4.00	4.00	4.00	4.00	4.67	4.67
Originality of Approach	3.00	2.33	4.00	3.33	3.33	3.00	3.00	3.00	3.33	4.33	4.33
Clarity of Subject/Paper	4.33	4.00	4.00	4.00	4.00	4.00	4.00	4.67	4.33	4.67	4.67
Literature Review	3.33	2.00	3.33	3.67	3.33	4.00	2.67	3.67	4.00	4.00	4.33
Justification of Methodology	3.00	2.67	3.00	2.33	3.33	2.33	3.00	3.67	4.00	4.00	3.33
Summary and Conclusion	4.00	3.33	4.00	3.33	4.33	3.67	4.00	4.33	3.67	4.00	4.00
Writing Style	4.67	4.00	5.00	4.33	5.00	4.67	4.67	4.67	4.33	5.00	4.67
All AI - mean											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	3.78	3.00	3.67	3.44	3.78	3.78	3.67	3.56	3.67	4.11	4.00
Importance to IS/CIS Educators	4.22	3.56	4.44	4.67	4.33	4.56	4.22	4.56	4.33	4.56	4.56
Originality of Approach	3.11	2.78	4.11	3.44	3.44	3.56	3.44	3.56	3.67	4.11	4.11
Clarity of Subject/Paper	4.33	3.89	3.56	3.89	4.22	4.22	4.00	4.44	4.11	4.22	4.44
Literature Review	3.44	2.44	3.33	3.89	3.44	4.22	3.00	3.44	4.00	4.11	4.00
Justification of Methodology	3.11	2.89	2.78	2.67	3.22	2.78	3.00	3.78	3.67	3.78	3.22
Summary and Conclusion	3.89	3.78	4.00	3.78	4.00	4.00	4.11	4.11	3.89	4.00	3.89
Writing Style	4.44	3.89	4.33	4.00	4.67	4.44	4.33	4.44	4.11	4.33	4.56
AI - Run 1											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	4.00	3.00	3.67	3.67	3.67	3.33	4.00	3.33	3.67	4.33	4.00
Importance to IS/CIS Educators	4.33	4.00	4.33	4.33	4.33	4.33	4.00	4.67	4.33	4.67	4.33
Originality of Approach	3.33	3.00	4.00	4.00	3.33	3.33	3.33	3.33	4.00	4.00	4.00
Clarity of Subject/Paper	4.33	3.67	3.33	3.33	4.00	4.00	4.00	4.33	4.33	4.33	5.00
Literature Review	3.33	2.33	3.33	3.33	3.67	4.00	3.00	3.33	4.00	3.67	3.67
Justification of Methodology	3.33	3.00	2.67	2.67	3.00	2.33	3.00	4.00	4.00	4.00	3.33
Summary and Conclusion	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Writing Style	4.67	3.67	4.33	4.33	4.33	4.00	4.33	4.00	4.33	4.33	5.00
AI - Run 2											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	3.67	3.00	3.67	3.67	4.33	4.00	3.67	4.00	3.33	4.00	4.00
Importance to IS/CIS Educators	4.00	3.38	4.67	4.67	4.33	4.67	4.33	4.67	4.33	4.33	4.33
Originality of Approach	3.00	2.67	4.00	3.33	3.33	3.67	3.67	3.67	3.33	4.00	4.00
Clarity of Subject/Paper	4.33	4.00	3.67	4.00	4.67	4.67	4.33	4.67	4.00	4.00	4.00
Literature Review	3.67	2.52	3.67	4.33	3.33	4.33	3.00	3.67	4.00	4.33	4.33
Justification of Methodology	3.00	2.81	2.67	2.67	3.33	3.00	3.33	4.00	3.33	3.67	3.00
Summary and Conclusion	3.67	3.67	4.00	3.67	4.33	4.00	4.33	4.33	4.00	4.00	4.00
Writing Style	4.67	4.00	4.33	3.67	5.00	5.00	4.67	4.67	4.00	4.33	4.00
AI - Run 3											
Category	1-1	1-2	1-3	1-4	2-1	2-2	2-3	3-1	4-1	5-1	5-2
Conference Recommendation	3.67	3.00	3.67	3.33	3.33	4.00	3.33	3.33	4.00	4.00	4.00
Importance to IS/CIS Educators	4.33	3.67	4.33	4.67	4.33	4.67	4.33	4.33	4.33	4.67	5.00
Originality of Approach	3.00	2.67	4.33	3.67	3.67	3.67	3.33	3.67	3.67	4.33	4.33
Clarity of Subject/Paper	4.33	4.00	3.67	4.00	4.00	4.00	3.67	4.33	4.00	4.33	4.33
Literature Review	3.33	2.67	3.00	3.67	3.33	4.33	3.00	3.33	4.00	4.33	4.00
Justification of Methodology	3.00	2.67	3.00	2.67	3.33	3.00	2.67	3.33	3.67	3.67	3.33
Summary and Conclusion	4.00	3.67	4.00	4.00	3.67	4.00	4.00	4.00	3.67	4.00	3.67
Writing Style	4.00	4.00	4.33	4.33	4.67	4.33	4.00	4.67	4.00	4.33	4.67

Table B-1: Means across all rater sets for 8 criteria across the 11 papers.

Appendix C – Inter-rater Reliability Scores for Reviewer Groups

Human					
Category	Krippendorff's Alpha	ICC(1,1)	ICC(1,k ₀)		
Conference Recommendation	0.170	0.130	0.309		
Importance to IS/CIS Educators	0.029	-0.045	-0.149		
Originality of Approach	0.153	0.158	0.360		
Clarity of Subject/Paper	0.438	0.444	0.705		
Literature Review	0.319	0.414	0.679		
Justification of Methodology	0.425	0.282	0.541		
Summary and Conclusion	0.528	0.531	0.772		
Writing Style	0.233	0.159	0.362		
ChatGPT					
Category	Krippendorff's Alpha	ICC(2,1)	ICC(2,3)	ICC(3,1)	ICC(3,3)
Conference Recommendation	0.278	0.318	0.583	0.359	0.627
Importance to IS/CIS Educators	0.413	0.459	0.718	0.520	0.765
Originality of Approach	0.513	0.537	0.777	0.630	0.837
Clarity of Subject/Paper	0.156	0.189	0.412	0.206	0.438
Literature Review	0.509	0.497	0.748	0.538	0.777
Justification of Methodology	0.120	0.123	0.296	0.121	0.292
Summary and Conclusion	n/a	n/a	n/a	n/a	n/a
Writing Style	0.121	0.138	0.324	0.143	0.333
Claude					
Category	Krippendorff's Alpha	ICC(2,1)	ICC(2,3)	ICC(3,1)	ICC(3,3)
Conference Recommendation	-0.072	-0.087	-0.316	-0.089	-0.324
Importance to IS/CIS Educators	0.262	0.310	0.574	0.365	0.633
Originality of Approach	0.407	0.444	0.706	0.500	0.750
Clarity of Subject/Paper	0.033	0.000	0.000	0.000	0.000
Literature Review	0.371	0.415	0.681	0.423	0.687
Justification of Methodology	0.380	0.467	0.724	0.455	0.714
Summary and Conclusion	0.023	0.000	0.000	0.000	0.000
Writing Style	0.093	0.079	0.205	0.073	0.191
Gemini					
Category	Krippendorff's Alpha	ICC(2,1)	ICC(2,3)	ICC(3,1)	ICC(3,3)
Conference Recommendation	0.306	0.339	0.606	0.325	0.591
Importance to IS/CIS Educators	0.267	0.328	0.594	0.310	0.574
Originality of Approach	0.663	0.647	0.846	0.671	0.859
Clarity of Subject/Paper	0.200	0.219	0.457	0.226	0.467
Literature Review	0.552	0.643	0.844	0.643	0.844
Justification of Methodology	0.389	0.397	0.664	0.439	0.701
Summary and Conclusion	0.232	0.221	0.459	0.207	0.440
Writing Style	0.111	0.103	0.255	0.098	0.245
ALL AI					
Category	Krippendorff's Alpha	ICC(2,1)	ICC(2,9)	ICC(3,1)	ICC(3,9)
Conference Recommendation	0.166	0.186	0.673	0.197	0.689
Importance to IS/CIS Educators	0.116	0.185	0.672	0.227	0.725
Originality of Approach	0.401	0.427	0.870	0.492	0.897
Clarity of Subject/Paper	0.140	0.152	0.618	0.164	0.639
Literature Review	0.403	0.441	0.876	0.461	0.885
Justification of Methodology	0.259	0.281	0.779	0.348	0.827
Summary and Conclusion	-0.010	-0.010	-0.102	-0.017	-0.172
Writing Style	0.095	0.111	0.529	0.156	0.624

Table C-1: ICC values across rater sets for 8 criteria across the 11 papers.

AI - Run 1					
Category	Krippendorff's Alpha	ICC(2,1)	ICC(2,3)	ICC(3,1)	ICC(3,3)
Conference Recommendation	0.269	0.300	0.562	0.319	0.584
Importance to IS/CIS Educators	-0.143	-0.071	-0.250	-0.091	-0.333
Originality of Approach	0.289	0.375	0.643	0.545	0.783
Clarity of Subject/Paper	0.370	0.386	0.653	0.393	0.660
Literature Review	0.180	0.304	0.567	0.369	0.637
Justification of Methodology	0.333	0.355	0.623	0.423	0.687
Summary and Conclusion	0.000	0.000	0.000	0.000	0.000
Writing Style	0.201	0.293	0.554	0.414	0.680

AI - Run 2					
Category	Krippendorff's Alpha	ICC(2,1)	ICC(2,3)	ICC(3,1)	ICC(3,3)
Conference Recommendation	-0.022	0.010	0.029	0.011	0.032
Importance to IS/CIS Educators	0.056	0.174	0.387	0.188	0.409
Originality of Approach	0.327	0.391	0.659	0.479	0.734
Clarity of Subject/Paper	0.305	0.326	0.592	0.354	0.621
Literature Review	0.396	0.447	0.708	0.441	0.703
Justification of Methodology	0.034	0.133	0.315	0.169	0.379
Summary and Conclusion	-0.113	-0.159	-0.702	-0.145	-0.611
Writing Style	0.496	0.519	0.764	0.600	0.818

AI - Run 3					
Category	Krippendorff's Alpha	ICC(2,1)	ICC(2,3)	ICC(3,1)	ICC(3,3)
Conference Recommendation	0.299	0.333	0.600	0.324	0.590
Importance to IS/CIS Educators	-0.004	0.167	0.375	0.217	0.453
Originality of Approach	0.341	0.363	0.631	0.356	0.624
Clarity of Subject/Paper	-0.214	-0.176	-0.818	-0.209	-1.080
Literature Review	0.330	0.309	0.572	0.293	0.554
Justification of Methodology	0.025	0.109	0.268	0.149	0.344
Summary and Conclusion	-0.032	-0.042	-0.136	-0.040	-0.130
Writing Style	-0.016	0.091	0.231	0.133	0.316

Table C-2: ICC values across rater groupings by run for 8 criteria across the 11 papers.

Appendix D – Mean Scores Grouped by Paper

Paper 1-1								
Category	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
Conference Recommendation	4.00	4.00	3.67	3.67	3.78	4.00	3.67	3.67
Importance to IS/CIS Educators	3.33	4.33	4.00	4.33	4.22	4.33	4.33	4.00
Originality of Approach	3.33	3.33	3.00	3.00	3.11	3.00	3.33	3.00
Clarity of Subject/Paper	4.00	4.33	4.33	4.33	4.33	4.33	4.33	4.33
Literature Review	4.33	3.33	3.67	3.33	3.44	3.33	3.67	3.33
Justification of Methodology	3.67	3.33	3.00	3.00	3.11	3.00	3.33	3.00
Summary and Conclusion	4.00	4.00	3.67	4.00	3.89	4.00	3.67	4.00
Writing Style	4.00	4.67	4.67	4.00	4.44	4.33	4.33	4.67
Mean of Means	3.83	3.92	3.75	3.71	3.79	3.79	3.83	3.75
Paper 1-2								
Category	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
Conference Recommendation	4.50	3.00	3.00	3.00	3.00	3.00	3.33	2.67
Importance to IS/CIS Educators	3.00	4.00	3.38	3.67	3.56	3.67	4.00	3.00
Originality of Approach	5.00	3.00	2.67	2.67	2.78	3.00	3.00	2.33
Clarity of Subject/Paper	5.00	3.67	4.00	4.00	3.89	4.00	3.67	4.00
Literature Review	4.50	2.33	2.52	2.67	2.44	2.67	2.67	2.00
Justification of Methodology	4.50	3.00	2.81	2.67	2.89	3.00	3.00	2.67
Summary and Conclusion	5.00	4.00	3.67	3.67	3.78	4.00	4.00	3.33
Writing Style	5.00	3.67	4.00	4.00	3.89	4.00	3.67	4.00
Mean of Means	4.56	3.33	3.26	3.29	3.28	3.42	3.42	3.00
Paper 1-3								
Category	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
Conference Recommendation	4.33	3.67	3.67	3.67	3.67	3.00	4.00	4.00
Importance to IS/CIS Educators	4.67	4.33	4.67	4.33	4.44	4.67	4.33	4.33
Originality of Approach	4.67	4.00	4.00	4.33	4.11	4.33	4.00	4.00
Clarity of Subject/Paper	4.33	3.33	3.67	3.67	3.56	3.67	3.00	4.00
Literature Review	4.33	3.33	3.67	3.00	3.33	3.00	3.67	3.33
Justification of Methodology	4.00	2.67	2.67	3.00	2.78	2.33	3.00	3.00
Summary and Conclusion	4.33	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Writing Style	4.67	4.33	4.33	4.33	4.33	4.00	4.00	5.00
Mean of Means	4.42	3.71	3.83	3.79	3.78	3.63	3.75	3.96
Paper 1-4								
Category	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
Conference Recommendation	4.75	3.33	3.67	3.33	3.44	3.33	3.67	3.44
Importance to IS/CIS Educators	4.75	4.67	4.67	4.67	4.67	5.00	4.67	4.67
Originality of Approach	4.50	3.33	3.33	3.67	3.44	3.67	3.33	3.44
Clarity of Subject/Paper	4.75	3.67	4.00	4.00	3.89	4.00	3.67	3.89
Literature Review	5.00	3.67	4.33	3.67	3.89	4.33	3.67	3.89
Justification of Methodology	4.75	2.67	2.67	2.67	2.67	2.67	3.00	2.67
Summary and Conclusion	4.50	3.67	3.67	4.00	3.78	4.00	4.00	3.78
Writing Style	4.50	4.00	3.67	4.33	4.00	4.00	3.67	4.00
Mean of Means	4.69	3.63	3.75	3.79	3.72	3.88	3.71	3.72
Paper 2-1								
Category	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
Conference Recommendation	3.67	3.67	4.33	3.33	3.78	3.67	3.67	4.00
Importance to IS/CIS Educators	3.33	4.33	4.33	4.33	4.33	4.00	4.67	4.33
Originality of Approach	3.67	3.33	3.33	3.67	3.44	3.00	4.00	3.33
Clarity of Subject/Paper	3.00	4.00	4.67	4.00	4.22	4.67	4.00	4.00
Literature Review	2.00	3.67	3.33	3.33	3.44	3.00	4.00	3.33
Justification of Methodology	2.67	3.00	3.33	3.33	3.22	3.00	3.33	3.33
Summary and Conclusion	2.67	4.00	4.33	3.67	4.00	4.00	3.67	4.33
Writing Style	3.00	4.33	5.00	4.67	4.67	4.67	4.33	5.00
Mean of Means	3.00	3.79	4.08	3.79	3.89	3.75	3.96	3.96
Paper 2-2								
Category	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
Conference Recommendation	5.00	3.33	4.00	4.00	3.78	3.67	4.33	3.33
Importance to IS/CIS Educators	4.33	4.33	4.67	4.67	4.56	4.67	5.00	4.00
Originality of Approach	4.00	3.33	3.67	3.67	3.56	3.67	4.00	3.00
Clarity of Subject/Paper	5.00	4.00	4.67	4.00	4.22	4.33	4.33	4.00
Literature Review	5.00	4.00	4.33	4.33	4.22	4.33	4.33	4.00
Justification of Methodology	4.00	2.33	3.00	3.00	2.78	2.67	3.33	2.33
Summary and Conclusion	4.00	4.00	4.00	4.00	4.00	4.00	4.33	3.67
Writing Style	5.00	4.00	5.00	4.33	4.44	4.33	4.33	4.67
Mean of Means	4.54	3.67	4.17	4.00	3.94	3.96	4.25	3.63

Table D-1: Mean category score heat maps by rater groups for papers 1-1 through 2-2.

Paper 2-3								
Category	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
Conference Recommendation	4.33	4.00	3.67	3.33	3.67	3.67	3.67	3.67
Importance to IS/CIS Educators	4.67	4.00	4.33	4.33	4.22	4.67	4.00	4.00
Originality of Approach	4.00	3.33	3.67	3.33	3.44	3.67	3.67	3.00
Clarity of Subject/Paper	4.33	4.00	4.33	3.67	4.00	4.00	4.00	4.00
Literature Review	3.00	3.00	3.00	3.00	3.00	3.00	3.33	2.67
Justification of Methodology	3.00	3.00	3.33	2.67	3.00	2.67	3.33	3.00
Summary and Conclusion	4.00	4.00	4.33	4.00	4.11	4.00	4.33	4.00
Writing Style	4.33	4.33	4.67	4.00	4.33	4.00	4.33	4.67
Mean of Means	3.96	3.71	3.92	3.54	3.72	3.71	3.83	3.63
Paper 3-1								
Category	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
Conference Recommendation	4.67	3.33	4.00	3.33	3.56	3.33	4.00	3.56
Importance to IS/CIS Educators	4.67	4.67	4.67	4.33	4.56	5.00	4.67	4.56
Originality of Approach	4.00	3.33	3.67	3.67	3.56	3.67	4.00	3.56
Clarity of Subject/Paper	4.67	4.33	4.67	4.33	4.44	4.67	4.00	4.44
Literature Review	4.67	3.33	3.67	3.33	3.44	3.00	3.67	3.44
Justification of Methodology	5.00	4.00	4.00	3.33	3.78	3.00	4.67	3.78
Summary and Conclusion	4.33	4.00	4.33	4.00	4.11	4.00	4.00	4.11
Writing Style	4.33	4.00	4.67	4.67	4.44	4.67	4.00	4.44
Mean of Means	4.54	3.88	4.21	3.88	3.99	3.92	4.13	3.99
Paper 4-1								
Category	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
Conference Recommendation	4.67	3.67	3.33	4.00	3.67	3.67	3.67	3.67
Importance to IS/CIS Educators	4.33	4.33	4.33	4.33	4.33	5.00	4.00	4.00
Originality of Approach	4.00	4.00	3.33	3.67	3.67	4.00	3.67	3.33
Clarity of Subject/Paper	4.67	4.33	4.00	4.00	4.11	4.00	4.00	4.33
Literature Review	4.67	4.00	4.00	4.00	4.00	3.67	4.33	4.00
Justification of Methodology	4.67	4.00	3.33	3.67	3.67	3.33	3.67	4.00
Summary and Conclusion	4.67	4.00	4.00	3.67	3.89	4.00	4.00	3.67
Writing Style	3.33	4.33	4.00	4.00	4.11	4.00	4.00	4.33
Mean of Means	4.38	4.08	3.79	3.92	3.93	3.96	3.92	3.92
Paper 5-1								
Category	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
Conference Recommendation	3.67	4.33	4.00	4.00	4.11	4.00	4.00	4.33
Importance to IS/CIS Educators	4.00	4.67	4.33	4.67	4.56	5.00	4.00	4.67
Originality of Approach	4.00	4.00	4.00	4.33	4.11	4.00	4.00	4.33
Clarity of Subject/Paper	3.67	4.33	4.00	4.33	4.22	4.00	4.00	4.67
Literature Review	3.33	3.67	4.33	4.33	4.11	3.67	4.67	4.00
Justification of Methodology	3.33	4.00	3.67	3.67	3.78	3.00	4.33	4.00
Summary and Conclusion	3.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Writing Style	3.33	4.33	4.33	4.33	4.33	4.00	4.00	5.00
Mean of Means	3.54	4.17	4.08	4.21	4.15	3.96	4.13	4.38
Paper 5-2								
Category	Human	AI1	AI2	AI3	AI-All	ChatGPT3	Claude3	Gemini3
Conference Recommendation	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00
Importance to IS/CIS Educators	4.67	4.33	4.33	5.00	4.56	4.33	4.67	4.67
Originality of Approach	4.67	4.00	4.00	4.33	4.11	4.00	4.00	4.33
Clarity of Subject/Paper	3.67	5.00	4.00	4.33	4.44	4.33	4.33	4.67
Literature Review	4.00	3.67	4.33	4.00	4.00	4.33	3.33	4.33
Justification of Methodology	4.00	3.33	3.00	3.33	3.22	3.00	3.33	3.33
Summary and Conclusion	3.33	4.00	4.00	3.67	3.89	4.00	3.67	4.00
Writing Style	4.33	5.00	4.00	4.67	4.56	4.33	4.67	4.67
Mean of Means	4.08	4.17	3.96	4.17	4.10	4.04	4.00	4.25

Table D-2: Mean category score heat maps by rater groups for papers 2-3 through 5-2.

Appendix E – Mean of Means by Paper

