Universiteit Leiden

Utility and Trustworthiness of Generative AI in Peer Review

Vuk Tomić (s3724409)

Supervisors:
Dr. Tom Heyman & Evert van Nieuwenburg

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 1, 2026

Abstract

The peer-review process plays a central role in maintaining the quality and integrity of scientific research. However, human peer review is often criticized for being time-consuming, resource-intensive, and subject to substantial variability between reviewers. Recent advances in Large Language Models have raised interest in their potential use as assistants in the reviewing process. This thesis investigates the reliability of AI-generated peer reviews by comparing evaluations produced by a Large Language Model with reviews written by human experts.

A dataset of 99 scientific papers was constructed from publicly available conference submissions and reviews obtained through OpenReview. For each paper, human review scores were collected and compared with AI-generated evaluations produced using OpenAI's GPT-4o-mini model. To assess consistency, the model reviewed each paper three independent times using identical instructions. The AI was additionally required to provide a confidence score for each evaluation. Statistical analyses included descriptive statistics, score difference analysis, linear regression, correlation analysis, and consistency measurements based on standard deviations across repeated runs.

The results demonstrate a statistically significant positive relationship between AI-generated and human-generated review scores ($R^2 = 0.384$, $p < 0.001$), indicating that the model was able to identify broad differences in paper quality in a manner that partially aligns with human judgment. At the same time, the AI system exhibited a tendency to assign systematically higher scores than human reviewers. Analysis of repeated evaluations showed relatively low score variability across runs, suggesting a reasonably high degree of internal consistency. Confidence scores were positively associated with AI review scores but showed little relationship with either agreement with human reviewers or score variability across repeated evaluations.

Overall, the findings suggest that modern Large Language Models can provide useful and consistent support for scientific peer review. However remaining differences between AI-generated and human-generated evaluations indicate that current systems should be viewed as complementary tools that assist human reviewers rather than replacements for expert human judgment.

Contents

1 Introduction

1.1 Topic

Scientific peer review is one of the fundamental mechanisms used to evaluate the quality, validity, and originality of academic research before publication. By relying on expert assessment, peer review aims to ensure that published scientific work meets accepted standards of methodology and offers a substantial contribution to the field.. In modern academia, peer review is widely used in conferences, journals, and grant evaluation processes, particularly within rapidly evolving fields such as Artificial Intelligence and Computer Science.

Despite its important role in the publication of scientific papers themselves, traditional peer review faces several well-known challenges. Human reviewers in a large number of cases have very different attitudes and opinions about the same paper, which results in inconsistent evaluation and extremely different aspects, focuses and comments for the same scientific paper [4]. Also, recently we have witnessed an increased number of submitted scientific papers, which has overloaded the human peer review process itself, which results in delays, as well as concerns about the quality and objectivity of the reviews themselves [3]. All these problems have contributed to researchers investigating whether automated systems and artificial intelligence can really contribute to the peer review process.

Recent advancements in Large Language Models (LLMs), such as GPT-based systems, have demonstrated strong capabilities in natural language understanding, summarization, reasoning, and text generation. These kinds of developments have raised interest in the possibility of using generative AI systems to support scientific reviewing tasks. Modern LLMs are capable of processing scientific papers, generating detailed textual feedback, identifying strengths and weaknesses, and assigning evaluation scores that may resemble those produced by human reviewers. //

The growing capabilities of Large Language Models have also influenced the policies of major Artificial Intelligence conferences. Recent editions of ICML have introduced specific guidelines regarding the use of generative AI during the scientific publication process. While authors are generally allowed to use LLM-based tools to assist with writing and research under certain conditions, the use of generative AI during the reviewing process itself remains highly debated and, in some cases, restricted by conference policies [2] . These discussions reflect the increasing importance of understanding how AI systems may influence scientific evaluation and whether they can be integrated into peer review workflows in a reliable and trustworthy manner. Recent studies have also begun investigating AI-assisted reviewing systems and their potential impact on review quality, reviewer workload, and evaluation consistency. [8]

While existing research has shown promising results regarding AI-generated reviews, important questions remain unanswered regarding their reliability, consistency, and practical usefulness. In particular, it remains unclear how closely AI-generated review scores align with human reviewer scores, how stable AI evaluations remain across repeated runs, and whether AI-generated feedback provides meaningful peer review.

In addition to generating answers and numerical evaluations, modern Large Language Models can also provide confidence estimates associated with their outputs. Such confidence scores are

intended to reflect the model's internal certainty regarding a prediction, judgment, or recommendation. Confidence estimates have attracted increasing research interest because they may help users assess the trustworthiness of AI-generated outputs and identify situations in which the model is less certain about its conclusions. However, previous research [10] has shown that confidence and correctness do not always coincide, making confidence calibration an important topic in the evaluation of Large Language Models.

This thesis investigates the use of generative AI for scientific peer review by comparing AI-generated reviews with human reviews obtained from scientific conference datasets. The study focuses on evaluating scoring differences, consistency, reliability, confidence, and feedback quality between AI-generated and human-generated reviews. This thesis aims to contribute to the understanding of whether generative AI can serve as a reliable support tool in academic peer review workflows.

1.2 Problem Statement

The increasing number of scientific paper submissions has intensified concerns regarding the scalability, consistency, and reliability of traditional peer review systems. Human reviewers often produce substantially different evaluations for the same paper, while time constraints and reviewer fatigue may negatively affect review quality. At the same time, recent progress in generative AI has enabled Large Language Models to generate detailed scientific feedback and numerical evaluations that resemble human reviews. [4]

Although generative AI systems show potential for assisting scientific reviewing, it remains unclear whether their evaluations are sufficiently reliable and consistent to support real-world peer review processes. In particular, there is limited understanding of how AI-generated review scores compare to human reviewer scores, how stable AI evaluations remain across repeated evaluations of the same paper, and whether the generated feedback provides useful and meaningful criticism comparable to human peer review.

The problem statement of this thesis is therefore:

How do AI-generated peer reviews compare to human peer reviews in terms of scoring consistency, reliability, and feedback quality?

Specifically, this problem statement can be broken down into 3 research questions that are going to be answered in the rest of this thesis:

- RQ1: How consistent are AI-generated reviews across multiple evaluations of the same paper?
- RQ2: How closely do AI-generated review scores align with human review scores?
- RQ3: How is average AI confidence related to average AI generated review scores?

1.3 Contribution

The key contributions of this thesis are:

- We evaluate the consistency and reliability of AI-generated reviews through repeated evaluations of the same scientific papers.
- We provide a comparative analysis between AI-generated peer reviews and human peer reviews using real scientific conference review data.
- We analyze scoring differences between AI-generated and human-generated reviews using statistical metrics and variability analysis.
- We investigate differences in confidence levels between AI-generated review runs and related scores.

1.4 Thesis overview

This bachelor thesis was carried out at the Leiden Institute of Advanced Computer Science (LIACS) under the supervision of Tom Heyman and Evert van Nieuwenburg.

The remainder of this thesis is organized as follows. Chapter 2 provides the necessary background on Large Language Models, their capabilities, confidence estimation, and their potential use in scientific peer review. Chapter 3 reviews related work on AI-assisted peer reviewing and discusses previous research on review quality, reliability, and confidence calibration. Chapter 4 describes the methodology, including the dataset construction, sampling strategy, AI review generation process, and statistical analyses used throughout the study. Chapter 5 presents and discusses the experimental results, with the consistency of AI-generated reviews, their correlation with human reviewers, and the relationship between review scores and confidence. Finally, Chapter 6 summarizes the main findings, discusses the limitations of the study, and provides recommendations for future research and Chapter 7 contains the code.

2 Background

The rapid development of Artificial Intelligence has significantly transformed the way people interact with digital information and computational systems. Among the most influential recent developments are Large Language Models, which have demonstrated remarkable capabilities in understanding, generating, and reasoning about natural language. These systems have attracted considerable attention from researchers, businesses, educators, and the general public due to their ability to perform a wide variety of language-related tasks with a high degree of fluency and flexibility.

Large Language Models are machine learning models trained on massive collections of text data obtained from books, websites, scientific articles, and other publicly available sources. Through this training process, the models learn statistical patterns in language, allowing them to predict and generate coherent text based on a given input. Modern LLMs are based primarily on the Transformer architecture introduced by Vaswani, which enabled substantial improvements in the processing of long text sequences and contextual information [5].

The popularity of Large Language Models increased dramatically following the public release of ChatGPT by OpenAI in late 2022. Millions of users gained access to a conversational AI system capable of answering questions, generating text, summarizing documents, assisting with programming tasks, and engaging in complex discussions [1]. Since then, the adoption of LLM-based systems has expanded rapidly across many domains.

One of the key reasons for the success of modern LLMs is their versatility. Unlike traditional machine learning systems that are typically designed for a single task, Large Language Models can perform a wide range of activities using the same underlying architecture. Some applications are text summarization, translation, question answering, code generation, information extraction, content creation, and decision support. Recently researchers have begun investigating whether LLMs can also contribute to scientific workflows, including literature review, manuscript evaluation, and peer review support.

The capabilities of modern LLMs have improved significantly [1] with the introduction of increasingly larger and more sophisticated models such as GPT-5.5, GPT-5 Pro, Claude, Gemini, and other state-of-the-art systems [1]. These models demonstrate strong performance on various reasoning, comprehension, and knowledge-intensive tasks. As a result, they have become valuable tools for assisting human experts in tasks that traditionally required extensive manual effort.

Despite these advances, important limitations remain. Large Language Models do not truly understand information in the same way humans do and may occasionally generate inaccurate, misleading, or inconsistent outputs. Furthermore, their responses can vary across repeated evaluations of the same input due to the probabilistic nature of text generation. These characteristics are particularly relevant when considering the use of LLMs in scientific peer review, where reliability, consistency, and trustworthiness are essential requirements. [12]

3 Related Work

Peer review has long been studied as a central mechanism for maintaining scientific quality, but it has also been criticized for inconsistency, subjectivity, and scalability problems. With the recent development of large language models, researchers have started to investigate whether AI systems can assist in scientific reviewing by generating feedback, identifying weaknesses, and assigning evaluation scores. This chapter reviews the most relevant work related to human peer review reliability, AI-assisted scientific feedback, and the use of large language models in academic evaluation.

A major motivation for this thesis comes from the known limitations of traditional peer review. Human reviewers often disagree with one another, even when evaluating the same paper. This disagreement can occur because reviewers focus on different aspects of a paper, interpret quality differently, or apply scoring criteria inconsistently. As a result, the final evaluation of a paper may depend not only on the quality of the work itself, but also on the particular reviewers assigned to it. This makes reliability and consistency important dimensions when evaluating any reviewing system, whether human or AI-based.

Recent work has explored the ability of large language models to provide feedback on research papers. Liang et al. conducted a large-scale empirical analysis using GPT-4 to generate feedback on scientific manuscripts. Their study compared GPT-4 feedback with human peer reviews across Nature family journals and ICLR papers. They found that the overlap between GPT-4 comments and human reviewer comments was comparable to the overlap between two human reviewers, suggesting that LLM-generated feedback can identify many of the same issues raised by human experts. In a user study, more than half of researchers found GPT-4 feedback helpful or very helpful, although the authors also emphasized important limitations of AI-generated feedback, such as generating overly generic comments, missing technical or methodological issues, and lacking the domain expertise and critical judgment of human reviewers.. [13]

This work is particularly relevant because it directly investigates whether large language models can provide useful scientific feedback. However, the focus of that study is mainly on feedback overlap and perceived usefulness, while this thesis focuses more specifically on scoring differences, consistency across repeated AI evaluations, confidence levels, and comparison with human review scores. In this sense, the present thesis builds on the idea that LLMs may be useful in peer review, but examines their reliability from a different empirical perspective.

Other recent research has focused on creating larger datasets of AI-generated reviews. For example, the Gen-Review dataset contains a large collection of LLM-written reviews linked to human reviews and ICLR submissions from multiple years. This dataset was designed to study questions such as bias in LLM reviews, detectability of AI-generated reviews, instruction-following ability, and alignment between LLM ratings and paper outcomes. [7] This line of work shows that the analysis of AI-generated peer reviews is becoming an important research area, especially as AI systems become more capable of producing realistic academic evaluations.

A further concern is the detectability and transparency of AI-generated reviews. Some recent studies investigate whether peer reviews written by LLMs can be distinguished from human reviews. [6] This is relevant because if AI-generated reviews become difficult to detect, conferences

and journals may face challenges in maintaining transparency about how reviews are produced. However, the focus of this thesis is not on detecting AI-written reviews, but on evaluating how their scores, consistency, confidence, and feedback quality compare to human reviews. [11]

Existing literature suggests that large language models can generate useful scientific feedback and may support parts of the peer review process. At the same time, the literature also shows that AI-generated reviews must be evaluated carefully, especially regarding reliability, scoring behavior, and overconfidence [12] [7]. This thesis contributes to this research direction by empirically comparing AI-generated reviews with human peer reviews on a selected set of scientific papers, focusing on score alignment, repeated-run consistency and confidence.

4 Methodology

This chapter describes the methodology used to compare AI-generated peer reviews with human peer reviews. The methodology focuses on the collection and preparation of scientific review data, the generation of AI-based evaluations using a Large Language Model, and the statistical analysis used to compare the resulting reviews. The chapter also describes the experimental setup, review generation process, and evaluation metrics used throughout the thesis.

4.1 Human Peer Review Dataset

The dataset used in this thesis was obtained from the OpenReview platform, which is widely used for peer review management in major Artificial Intelligence and Machine Learning conferences. More specifically, the data originates from the International Conference on Learning Representations (ICLR), where submissions, reviews, reviewer ratings, and review discussions are publicly accessible for many conference years.

The choice of OpenReview and ICLR data was motivated by several factors. First, ICLR is one of the leading conferences in Artificial Intelligence and Machine Learning, meaning that the review data originates from real scientific reviewing performed by expert reviewers in highly competitive academic settings. Second, the OpenReview platform provides structured metadata and publicly accessible review information, making it suitable for reproducible scientific analysis. Finally, the dataset contains both quantitative and qualitative review information, enabling a broad comparison between AI-generated and human-generated evaluations.

The collected dataset contains:

- Paper metadata.
- Reviewer scores.
- Confidence scores.
- Textual reviewer feedback.
- Links to the associated paper PDFs.
- Additional information such as date, authors and so on.

The numerical review scores assigned by human reviewers represent the primary evaluation signal used throughout the thesis. These scores typically reflect the reviewer's overall assessment of the paper quality, originality, significance, and technical correctness.

The data collection process involved extracting review information from OpenReview records and converting the data into a structured tabular format suitable for further analysis. Since the OpenReview data contains nested structures and multiple review entries per paper, additional preprocessing steps were required to correctly associate reviewer evaluations with their corresponding submissions.

More specifically, the steps taken to get this dataset are:

1. Collecting paper metadata and review information from OpenReview.

2. Matching reviews to their corresponding submissions.

3. Extracting numerical ratings and reviewer confidence values.

4. Constructing structured tables for statistical evaluation.

4.2 Sampling Strategy

The complete OpenReview dataset contains a very large number of papers and reviews. Since generating AI reviews for full scientific papers requires substantial computational resources and API usage costs, it was not feasible to evaluate the entire dataset. As a result, a sampling strategy was required in order to construct a representative but computationally manageable subset of papers.

The sampling procedure used in this thesis was based on stratified sampling. The main goal of this strategy was to ensure that the final dataset included papers with varying levels of perceived quality rather than concentrating only on highly rated or poorly rated submissions. We devided the sampled papers into three categories based on average human score. This is important because comparing AI and human reviewing behavior across different quality levels provides a more balanced and informative evaluation of AI-generated peer reviews.

The sampling process began by calculating the average human review score for each paper. Since most papers in the dataset were reviewed by multiple reviewers, the average score provided a more stable estimate of overall reviewer evaluation than individual reviewer ratings alone.

After average scores were computed, papers were divided into score-based groups. Samples were then selected from different score ranges in order to preserve variation in paper quality. After average scores were computed, papers were divided into score-based groups using the average human review score. The original ICLR review process uses a recommendation scale from 1 to 10, where lower scores indicate stronger rejection recommendations and higher scores indicate stronger acceptance recommendations. To ensure representation of papers with different quality levels, papers with an average score of 3 or below were classified as low-scoring, papers with scores between 3 and 6 as medium-scoring, and papers with scores above 6 as high-scoring. Samples were then selected from each group to preserve variation in paper quality and create a proportinatelyt balanced dataset for comparison. This ensured that the final dataset included:

1. Highly rated papers.

2. Moderately rated paper.

3. Lower rated papers.

An additional consideration during sampling involved conference year selection. Papers submitted before 2022 were prioritized. The primary motivation for this decision was to ensure that the original human reviews were produced without assistance from modern Large Language Models, such as ChatGPT and similar systems, which were not publicly available at the time. As a result, the human review scores used in this study are highly likely to represent purely human evaluations. This helps improve the validity of the comparison between AI-generated and human-generated

peer reviews by reducing the possibility that the human reviewers themselves relied on AI-assisted reviewing tools.

Another important aspect of this is that we wanted to make sure that all scientific papers were written exclusively by humans without any help from AI, which means that we only had to use papers before 2022, so our dataset consists of scientific papers submitted from 2017 to 2022.

The final sampled dataset was constructed using a separate sampling script implemented in RStudio. After the stratified sampling procedure was completed, 100 papers were selected from the three predefined quality groups (low, medium, and high scoring papers) using random sampling within each group. The selected papers were then automatically downloaded from OpenReview and stored locally on the researcher's computer. These locally stored PDF files were subsequently used as input for the AI review generation pipeline.

Only one paper was unavailable to be downloaded within our selected papers, so it was not successfully retrieved for further analysis. As a result, the final analysis was performed on 99 scientific papers. The resulting dataset preserved representation across all three score groups, ensuring that papers with low, medium, and high human review scores were included in the evaluation. This sample size was selected as a compromise between:

- Statistical diversity.
- Experimental feasibility.
- Computational limitations.
- and API cost constraints.

In addition to enabling statistical analysis, the selected sample size also allowed repeated AI evaluations of the same paper, which was necessary for studying review consistency and variability across multiple model runs.

4.3 AI Review Generation

After the human review data had been collected and processed, the next stage of the methodology focused on generating AI-based peer reviews for the sampled scientific papers. The primary objective of this phase was to create reviews that could be systematically compared with human peer reviews in terms of scoring behavior, confidence estimation, consistency across repeated evaluations, and the overall quality. In order to achieve this, a controlled review-generation pipeline was designed using a Large Language Model capable of processing scientific documents and producing structured academic evaluations.

The AI-generated reviews were produced using the OpenAI API and a GPT-based Large Language Model, more specifically, 4o-mini model. GPT-4o-mini was selected because it provides strong performance on reasoning, summarization, text analysis, and feedback generation tasks while remaining computationally efficient. This was particularly important given that each paper was evaluated three times, resulting in several hundred review generations throughout the experiment. Compared to larger frontier models, GPT-4o-mini offers significantly lower API costs and faster processing times, making large-scale evaluation feasible within the practical resource constraints of

our project. In addition the model supports long-context inputs, allowing complete scientific papers to be processed while maintaining a consistent and reproducible review generation framework across all experiments. These capabilities make Large Language Models particularly suitable for scientific reviewing tasks, where the model must interpret technical material, identify strengths and weaknesses, and provide coherent evaluative feedback. Since the objective of the thesis was not simply to generate text, but to evaluate the reliability of AI reviewing behavior, it was important that the review generation process remain consistent, reproducible, and sufficiently structured across all evaluated papers.

The review generation pipeline consisted of several sequential stages. First, the PDF file corresponding to each scientific paper was obtained from the sampled dataset. The textual content of the paper was then extracted and prepared for input into the language model. Since scientific papers often contain large amounts of technical information, including equations, tables, references, and experimental sections, preprocessing was necessary in order to ensure that the extracted text remained sufficiently readable and coherent for the model. The extracted paper text was then combined with a structured review prompt specifically designed to simulate the behavior of a scientific conference reviewer. The complete prompt used to instruct the language model is provided in Figure 1.

```
analysis_prompt <- paste(
  "You are acting as a careful academic peer reviewer.",
  "Read the attached scientific paper PDF and produce a structured review.",
  "Base your judgment only on the information available in the paper.",
  "If important information is missing or unclear, state that explicitly.",
  "",
  "Return ONLY valid JSON.",
  "",
  "Use exactly the following structure:",
  "{",
  '  "paper_title": "string",',
  '  "short_summary": "string",',
  '  "main_claims": ["string", "string"],',
  '  "strengths": ["string", "string"],',
  '  "weaknesses": ["string", "string"],',
  '  "methodology_summary": "string",',
  '  "suggested_review_score_1_to_100": "integer",',
  '  "confidence_in_score_1_to_5": "integer",',
  '  "score_justification": "string",',
  '  "r_code_for_followup_analysis": "string"',
  "}",
  "",
  "Guidelines:",
  "- strengths should capture the main positive contributions of the paper",
  "- weaknesses should capture problems in reasoning, evaluation, clarity, novelty, or methodology",
  "- confidence_in_score_1_to_5 means confidence in the assigned review score",
  "- suggested_review_score_1_to_100 should be an integer on a peer-review style scale",
  "- for suggested_review_score_1_to_100 take into account the entire range",
  "- for suggested_review_score_1_to_100 consider that one third of papers should have score bellow 35, other third above 63, and other third of papers
should have score between 35 and 65",
  "- do not include markdown",
  "- do not include explanations outside the JSON",
  sep = "\n"
)
```

Figure 1: Alaysis Prompt

For each evaluation, the model assigned a numerical review score on a scale from 1 to 100, where higher values indicated a more favorable assessment of the paper. The use of a 100-point scale was intended to provide greater granularity and allow the model to express smaller differences between papers. For comparison with the human review scores available in the OpenReview dataset, the AI-generated scores were subsequently transformed to a 1–10 scale by dividing the generated score by ten.

In addition to the review score, the model was instructed to provide a confidence score on a scale from 1 to 5. This confidence value represented the model's self-assessed certainty regarding the

assigned review score, with higher values indicating greater confidence in the evaluation. Confidence scores were collected for all review runs and were later used to investigate the relationship between model confidence, assigned scores, and agreement with human reviewers.

In addition to assigning a numerical score, the model was instructed to generate detailed written feedback explaining the reasoning behind the evaluation. The generated reviews therefore included both quantitative and qualitative components, similar to real conference peer reviews.

A key aspect of the review generation process involved maintaining prompt consistency across all experiments. The exact same prompt structure and evaluation instructions were used for all papers in order to ensure that the resulting differences between generated reviews originated from model behavior and paper content rather than from variations in prompting style. API calls were separate per paper. This was important because prompt wording can significantly influence the behavior of Large Language Models. By standardizing the review prompt, the experimental setup aimed to preserve fairness and comparability across all generated evaluations.

Another important methodological consideration concerned the probabilistic nature of Large Language Models. Unlike deterministic algorithms, modern generative AI systems do not necessarily produce identical outputs when evaluating the same input multiple times. Even under identical prompting conditions, slight variations may occur due to probabilistic token sampling during text generation. This characteristic is highly relevant for peer review research because consistency and reproducibility are critical properties of reliable reviewing systems. If an AI system assigns substantially different scores to the same paper across repeated evaluations, this raises important questions regarding the stability and reliability of AI-assisted reviewing.

For this reason, each scientific paper in the sampled dataset was evaluated multiple times rather than only once, more specifically 3 times. Repeated evaluation allowed the thesis to examine the consistency of AI-generated reviews and measure the degree of variability present across different runs of the same paper. This repeated-run methodology formed one of the central components of the thesis because it enabled quantitative analysis of AI review stability.

In order to preserve realistic reviewing behavior while avoiding excessive randomness, the temperature parameter was set to 0.7 throughout all experiments. This value provided a balance between output consistency and natural variation, allowing the thesis to examine AI review stability under conditions that resemble real-world use. Keeping the temperature fixed across all evaluations ensured that observed differences between runs originated from the model's probabilistic behavior rather than changes in experimental settings.

For every paper, three AI-generated reviews were produced using the same evaluation pipeline, the same prompt structure, and the same review criteria. The generated outputs were then processed and transformed into structured data suitable for statistical analysis. From these repeated evaluations, several important metrics were generated, including:

- AI-generated score for each run.
- Average AI-generated review score.
- Standard deviation of AI-generated scores.

- AI-generated confidence score for each run.
- Average AI-generated confidence score.
- Difference in average AI-generated review scores - Human scores, general and absolute.
- In which group based on average human score does the paper belong to.

An important difference between the AI-generated and human-generated reviews concerns the number of evaluations available for each paper. While every paper in this study received exactly three AI-generated reviews, the number of human reviewers varied across papers in the OpenReview dataset. Some papers received fewer reviews, while others received more, reflecting the original conference reviewing process.

Because of this variability, individual human review scores were not selectively included or excluded when calculating the overall results. Instead, all available human review scores associated with a paper were used when computing average human scores and measures of score variability. This approach preserves the original evaluation process used by the conference and avoids introducing additional selection bias that could arise from manually removing or prioritizing particular reviewer evaluations.

The use of repeated evaluations also allowed the thesis to compare AI reviewer variability with the natural variability already present among human reviewers. Since human reviewers frequently disagree with one another when evaluating the same paper, measuring AI variability provides important context regarding whether AI-generated reviews are more stable, equally variable, or potentially less reliable than human evaluations.

In addition to numerical scoring, the generated textual feedback itself was preserved and stored for further analysis, but it was not used for the purposes of this research paper. The textual feedback generated by the model includes strength and weaknesses, summaries of the paper and suggestions for improvement.

Another important aspect of the AI review generation process involved generation settings and model configuration. Since Large Language Models are probabilistic systems, repeated evaluations of the same paper may produce variations in scores and textual feedback even under identical prompting conditions. In this thesis, the temperature parameter was set to 0.7, the default temperature for the model which we used, providing a balance between output consistency and natural variation in generated reviews. This configuration reduced excessive randomness while still allowing the model to generate realistic and non-deterministic reviewer behavior. Preserving some degree of variability was important because one of the central objectives of the thesis was to analyze the consistency and reliability of AI-generated peer reviews across multiple evaluations of the same paper.

In a small number of cases the model did not return a complete evaluation, resulting in missing review scores or confidence values for individual runs. Consequently, a limited number of derived variables, such as average scores and standard deviations, could not be calculated for those papers. These missing observations were retained as missing values and excluded from analyses requiring the unavailable variables.

4.4 Data Analysis

After AI-generated reviews and human review data had been collected, a series of statistical analyses were performed to answer the research questions introduced in Chapter 1. The analysis focused on measuring score agreement between AI-generated and human-generated reviews, evaluating the consistency of repeated AI evaluations, and examining the relationship between review confidence and scoring behavior.

For each paper, the three AI-generated review scores were aggregated by calculating the average AI score. The average score was used as the primary AI evaluation and compared against the average human review score calculated from all available human reviewers. In addition, the standard deviation of the three AI-generated scores was calculated as a measure of review consistency. Lower standard deviation values indicate that the model assigned similar scores across repeated evaluations, while higher values indicate greater variability.

To evaluate review consistency - RQ1, pairwise correlations between the three AI review runs were calculated. In addition, the standard deviation of AI-generated scores was analyzed across all papers in order to measure stablity of the model's evaluations. These analyses provide insight into whether our model produces similar judgments when evaluating the same paper multiple times.

To investigate the relationship between AI-generated and human-generated evaluations - RQ2, correlation analysis and linear regression were performed using the average AI score and average human score for each paper. Scatter plots and regression lines were used to visualize the strength of the relationship and identify potential disagreements between the two evaluation methods.

For confidence analysis - RQ3, the average confidence score generated by the model was compared with both the assigned review scores and the observed differences between AI-generated and human-generated evaluations. Correlation and regression analyses were used to investigate whether higher confidence was associated with higher review scores and whether more confident evaluations exhibited stronger agreement with human reviewers. This analysis provides insight into whether the model expresses greater confidence when assigning particularly high or low scores and whether confidence serves as meaningful indicator of agreement with human judgment.

5 Results

5.1 Descriptive Statistics

The final dataset consisted of 99 scientific papers successfully retrieved from OpenReview and processed by the AI review generation pipeline. For each paper, three independent AI-generated reviews were produced and compared with the corresponding human review scores obtained from OpenReview.

Figure 2 summarizes the main descriptive statistics. The average human review score across all papers was 5.22 (SD = 1.45), while the average AI-generated score was 6.22 (SD = 0.86). On average, the AI model assigned scores approximately one point higher than human reviewers. The mean difference between AI and human scores was 1.00, while the mean absolute difference was 1.21.

variable	mean	sd
Average Human Score	5.22	1.45134907634729
Average AI Score	6.22104377104377	0.857567969508881
Human Score SD	1.0868068659061	0.752448865933413
AI Score SD	0.375025433810306	0.280007619279281
Average AI Confidence	3.80808080808081	0.272890420520836
AI Confidence SD	0.214735946069596	0.283832456259691
AI - Human Difference	1.02575757575758	1.12843648160264
Absolute AI - Human Difference	1.20589225589226	0.931272972235836

Figure 2: Descriptive statistics of the main study variables

The average AI confidence score was 3.81 on the five-point confidence scale (SD = 0.27), indicating that the model generally reported relatively high confidence in its evaluations.

Taken together, the descriptive statistics suggest that the AI model tended to evaluate papers more positively than human reviewers while exhibiting lower variability in its scores.

5.1.1 AI Review Scores Consistency

Figure 3 illustrates the scores assigned by the AI model across the three repeated evaluations for each paper. Most score trajectories remain relatively stable across runs, with only modest fluctuations between evaluations. Although several papers exhibit larger score changes, the majority of papers receive scores within approximately one point across all three runs. This visual evidence supports the conclusion that the model demonstrates relatively high internal consistency when repeatedly evaluating the same scientific paper.

Figure 4 presents the distribution of the standard deviation of AI-generated review scores across the three evaluation runs. The distribution is concentrated at relatively low values, with most papers exhibiting standard deviations below 0.6 points. Only a small number of papers show

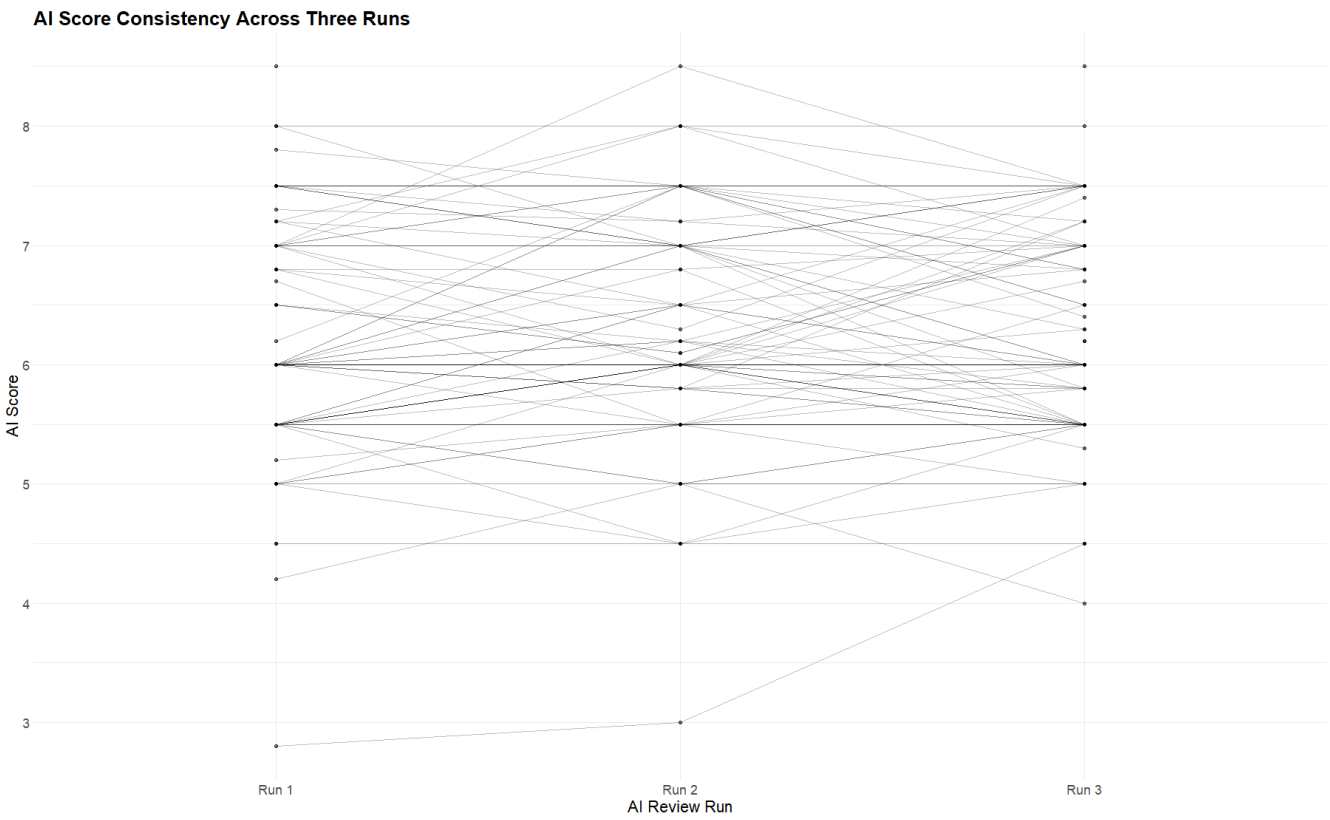

Figure 3: AI Score Consistency Across Three Runs

substantially larger variability. These results indicate that repeated evaluations of the same paper generally produce similar scores, suggesting that the model's scoring behavior is reasonably stable despite the probabilistic nature of LLM generation.

Figure 5 compares the distributions of AI-generated scores across the three evaluation runs. The boxplots show nearly identical medians, interquartile ranges, and overall score distributions for all runs. This similarity indicates that no systematic drift occurred between runs and further supports the conclusion that the model produced consistent evaluations across repeated assessments of the same papers.

To provide a baseline for interpreting AI review consistency, agreement between individual human reviewers was also examined. Figure 6 presents the pairwise Pearson correlations between the review scores assigned by different human reviewers. Moderate positive correlations were observed between Reviewers 1–3, ranging from 0.47 to 0.56, indicating a moderate level of agreement. Reviewer 4 exhibited considerably weaker agreement with the remaining reviewers, with correlations ranging from 0.08 to 0.25 . These findings illustrate that substantial variability also exists among human reviewers, providing important context for interpreting the consistency of AI-generated evaluations.

5.1.2 Relationship Between AI and Human Review Scores

To address RQ2, the relationship between average human review scores and average AI-generated review scores was examined using correlation and linear regression analysis. Figure 7 presents the

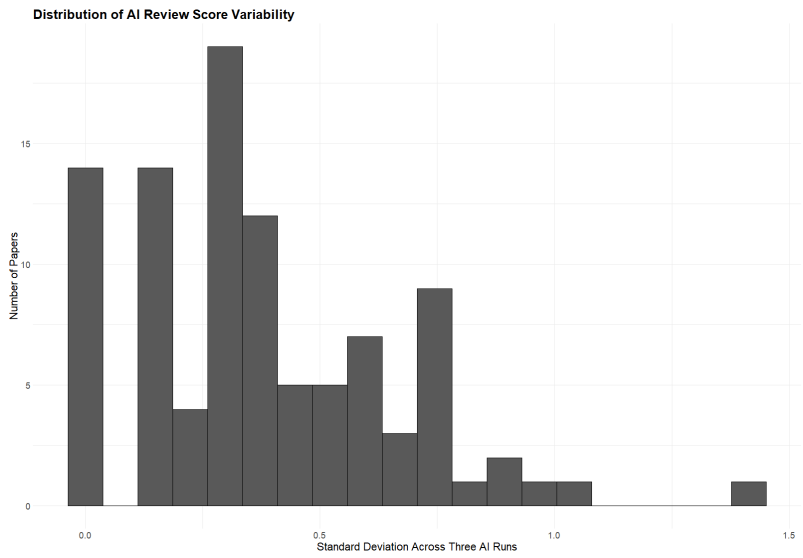

Figure 4: Distribution of AI Review Score Variability

scatter plot together with the fitted regression line.

A clear positive relationship can be observed between human and AI evaluations. Papers receiving higher scores from human reviewers generally also received higher scores from the AI model. Although individual disagreements are present, the overall trend suggests that the AI model was capable of distinguishing between lower- and higher-quality papers in a manner broadly consistent with human reviewers.

Linear regression analysis confirmed this relationship. The estimated regression coefficient for average AI score was 0.370 1.039 (SE = 0.134, $p < 0.001$), indicating a statistically significant positive association between human and AI evaluations. The fitted regression model explained approximately 38.4% of the variance in AI-generated scores ($R^2 = 0.384$).

The regression equation can be expressed as:

$$HumanScore = -1.27 + 1.04 \times AIScore$$

This result indicates that increases in human review scores were associated with higher AI-generated scores, supporting the existence of substantial agreement between the two evaluation approaches.

5.1.3 Regression Diagnostics

Figure 8 presents the residual plot of the fitted regression model. The residuals are distributed around zero without a clearly visible systematic pattern, suggesting that the linear regression model provides a reasonable representation of the relationship between human and AI review scores.

While several papers exhibit relatively large residual values, no strong evidence of non-linearity is visible. The residual analysis therefore supports the use of a linear model for examining the relationship between AI-generated and human-generated evaluations.

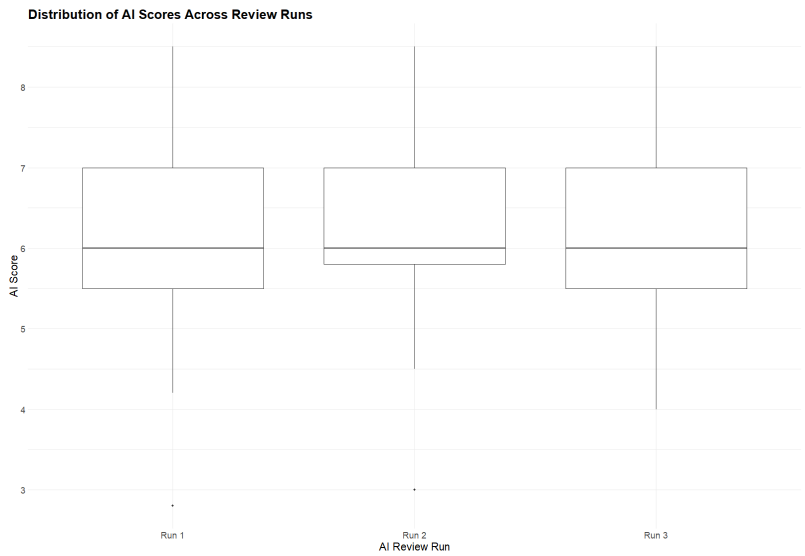

Figure 5: Distribution of AI Scores Across Review Runs

5.1.4 Distribution of Score Differences

To further examine disagreement between AI-generated and human-generated evaluations, score differences were calculated by subtracting the average human score from the average AI score.

Figure 9 illustrates the distribution of these differences. The distribution is centered above zero, indicating that the AI model generally assigned higher scores than human reviewers. This observation is consistent with the descriptive statistics, which showed an average positive difference of approximately 1.00 points.

Although most score differences were relatively moderate, several papers displayed substantially larger deviations, indicating cases where AI-generated evaluations differed considerably from human reviewer judgments.

5.1.5 Distribution of Human and AI Scores

Figures 10 and 11 present the distributions of average human review scores and average AI-generated review scores respectively.

The human review scores span a relatively wide range of values, reflecting the inclusion of papers from low-, medium-, and high-scoring quality groups.

The distribution of AI-generated scores follows a similar pattern but is shifted toward higher values. The concentration of AI scores around the 6–7 range further supports the observation that the model generally evaluated papers more favorably than human reviewers.

Figure 6: Pairwise Pearson correlations between individual human reviewers

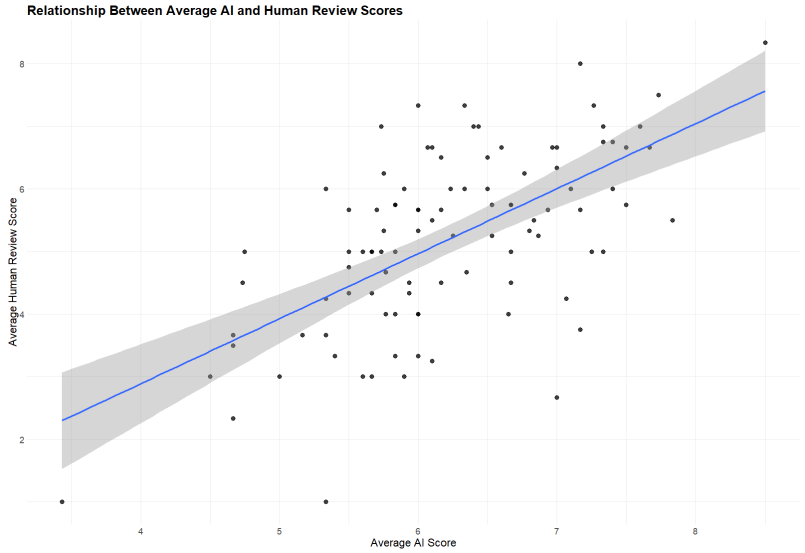

Figure 7: Relationship between average human review scores and average AI-generated review scores.

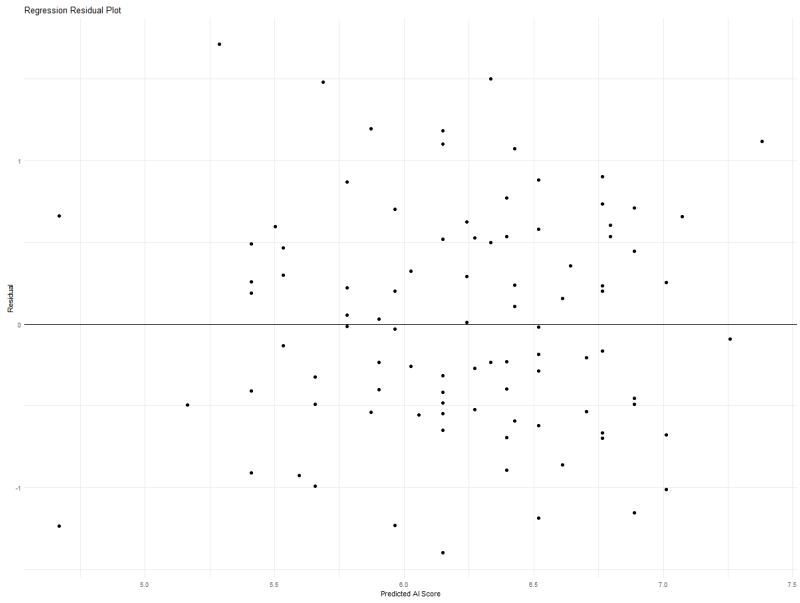

Figure 8: Residual plot for the linear regression model relating average human review scores and average AI-generated review scores.

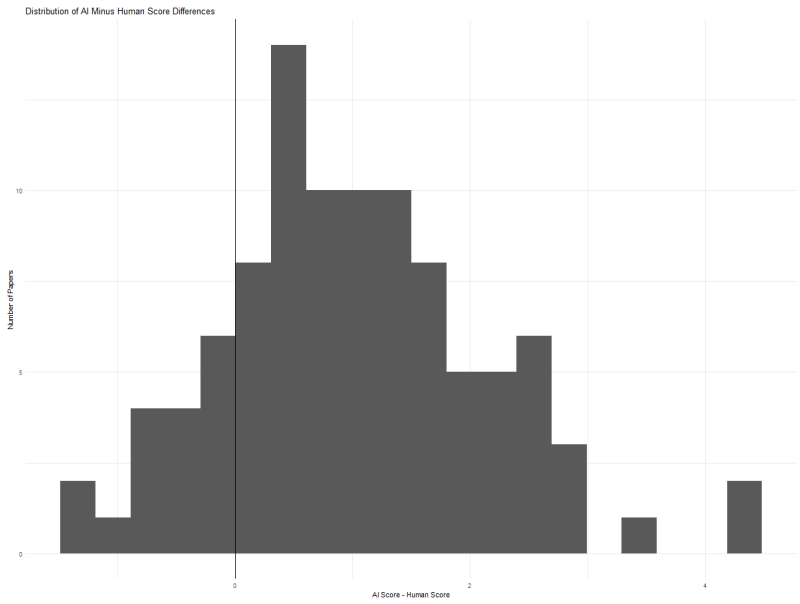

Figure 9: Distribution of AI-human score differences.

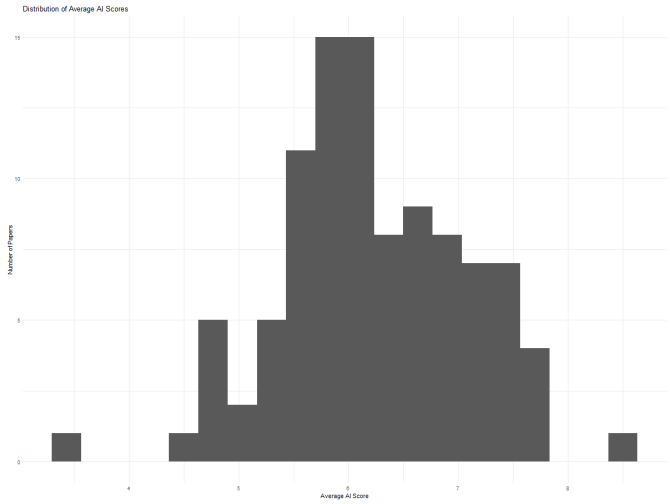

Figure 10: Distribution of average AI-generated review scores.

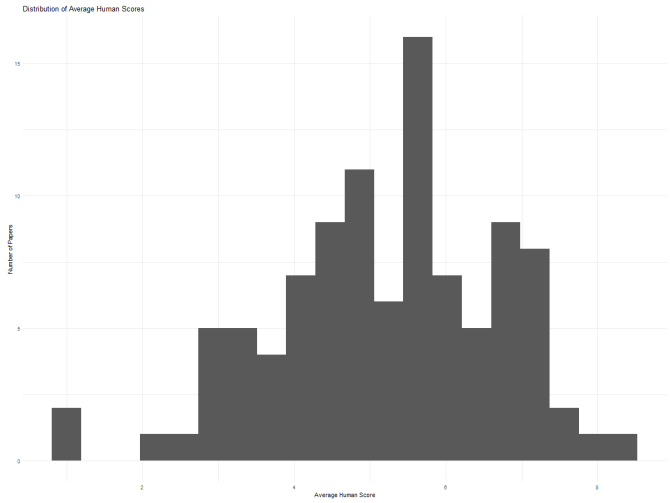

Figure 11: Distribution of average human review scores.

5.2 Comparison of score variability between human reviewers and AI

Figure 12 compares the variability of review scores assigned by human reviewers and the AI model. Human reviewers exhibited greater variability in their assigned scores than the AI across repeated review runs. The average standard deviation of human reviewer scores was 1.09 (SD = 0.75), whereas the average standard deviation of AI-generated scores was 0.38 (SD = 0.28). This indicates that the AI model produced more consistent review scores than the human reviewers.

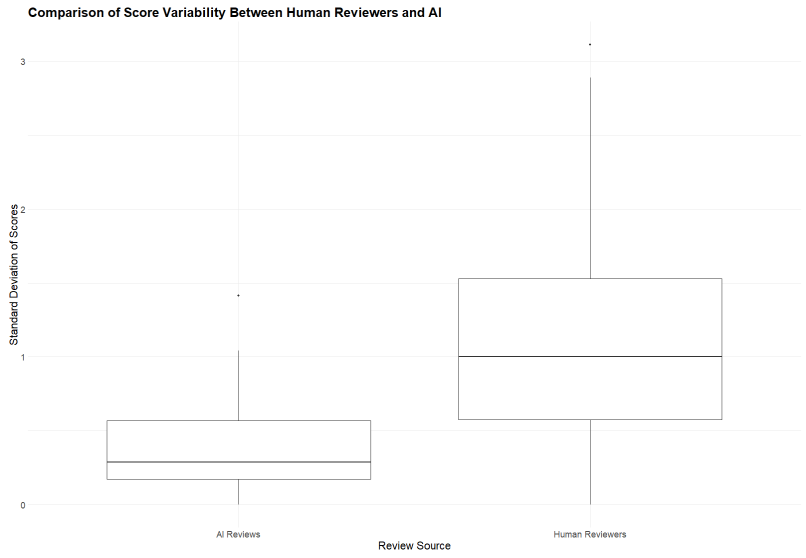

Figure 12: Comparison of the standard deviation of review scores between human reviewers and AI-generated reviews

To determine whether this difference was statistically significant, a paired-samples t-test was conducted. The analysis showed that the standard deviation of human reviewer scores was significantly higher than that of the AI-generated scores, $t(97) = 8.79$, $p < .001$. The mean difference between the two standard deviations was 0.73, with a 95% confidence interval ranging from 0.56 to 0.89. These results provide statistical evidence that the AI model generated significantly more consistent review scores than the human reviewers.

5.2.1 Absolute Score Differences

To evaluate disagreement independently of direction, absolute score differences between AI-generated and human-generated evaluations were calculated.

Figure 13 shows that most papers exhibited relatively small to moderate differences between AI and human scores. However, a limited number of papers produced substantially larger differences, with some exceeding four points. These cases represent the strongest disagreements between the AI model and human reviewers and may provide valuable insight into situations where AI-assisted reviewing is less reliable.

The average absolute difference across all papers was 1.21 points, indicating that while agreement between AI and human evaluations was generally positive, meaningful disagreements remained present for a number of papers.

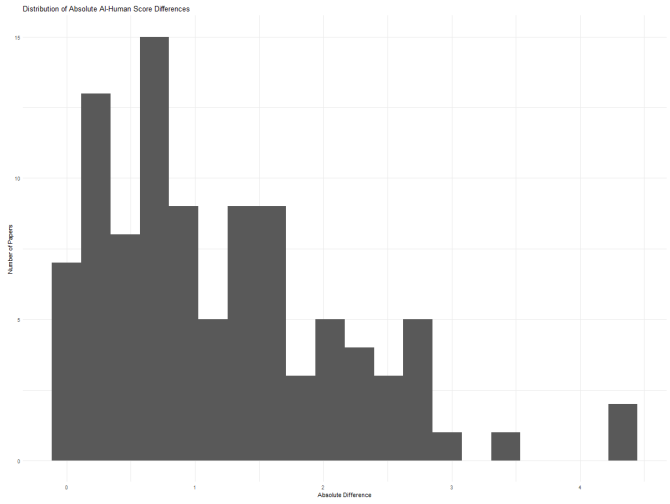

Figure 13: Distribution of absolute AI-human score differences.

5.3 Confidence Analysis

To assess the consistency of AI-generated confidence scores across repeated evaluations, the distribution of confidence scores for each of the three review runs was examined - Figure 14. The distributions were nearly identical across all runs, with confidence values concentrated almost exclusively between 3 and 4 on the 5-point confidence scale. No noticeable differences in the median or overall spread of the confidence scores were observed between the three runs. This suggests that the AI model produced a similar overall level of confidence across repeated evaluations, indicating that the confidence estimates remained stable at the distribution level.

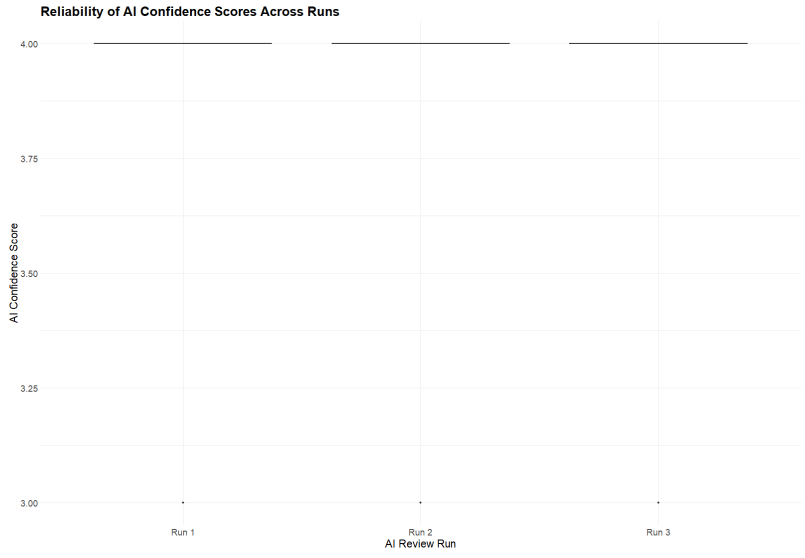

Figure 14: Distribution of AI confidence scores across the three AI review runs

To address RQ3, the relationships between AI confidence, AI-generated review scores, agreement

with human reviewers, and score variability across repeated evaluations were examined. Figure 15 presents the relationship between average AI confidence scores and absolute AI-human score differences.

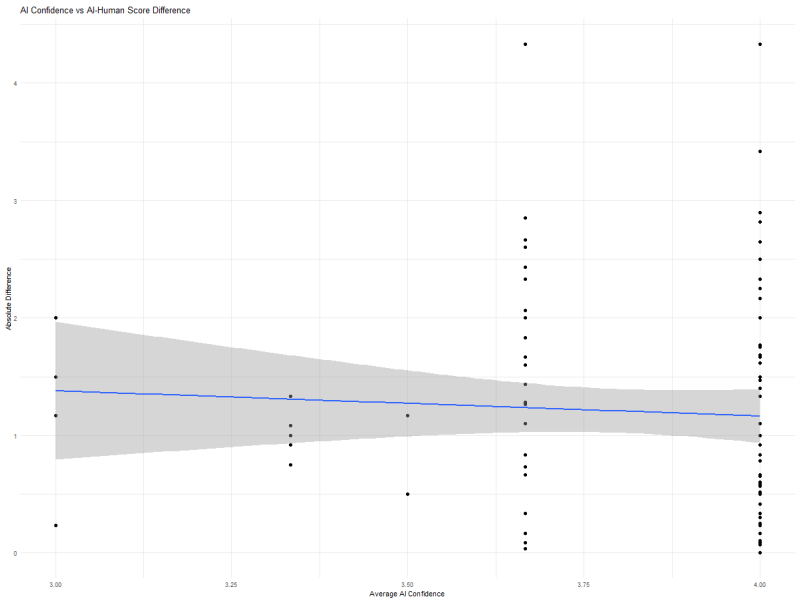

Figure 15: Relationship between average AI confidence and absolute AI-human score differences

Visual inspection of the scatter plot suggests only a weak relationship between confidence and agreement. This observation was confirmed through linear regression analysis. The regression coefficient for average AI confidence was -0.215 (SE = 0.346, p = 0.535), indicating that confidence was not a statistically significant predictor of disagreement between AI-generated and human-generated scores.

Consequently, higher confidence values were not associated with significantly smaller score differences. In other words, the model's reported confidence did not reliably indicate whether its evaluation would be closer to or further from the judgments of human reviewers.

This finding suggests that confidence scores should be interpreted as a measure of the model's internal certainty rather than as an indicator of agreement with human reviewers.

Figure 16 examines the relationship between average AI confidence and score variability across repeated evaluations. The fitted regression line is nearly flat, indicating little association between the confidence reported by the model and the consistency of its scores. Highly confident evaluations were not necessarily more stable across runs than evaluations with lower confidence. This suggests that the model's self-reported confidence is not a reliable indicator of review consistency.

Figure 17 presents the relationship between average AI-generated review scores and average AI confidence. A positive relationship can be observed, with higher review scores generally associated with higher confidence values. Papers receiving stronger evaluations from the model tend to be accompanied by greater self-reported confidence. This finding suggests that the model is more certain when assigning favorable evaluations, although additional statistical testing would be required to determine the strength and significance of this relationship.

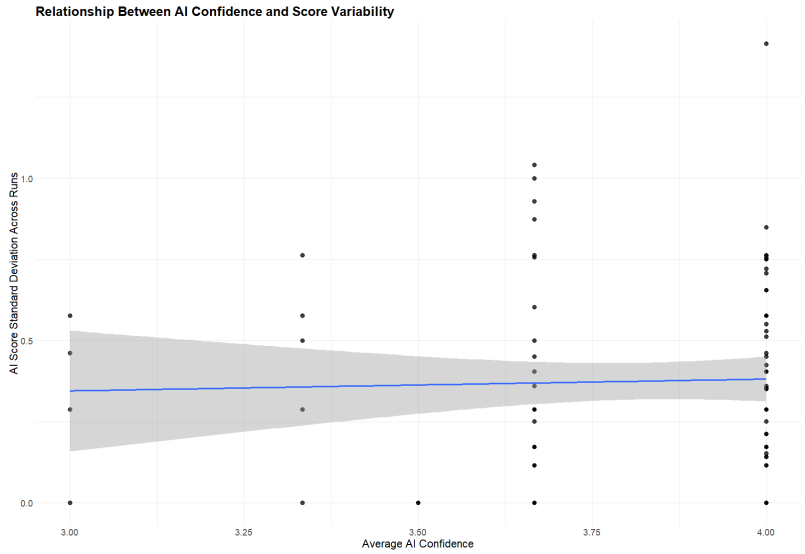

Figure 16: Relationship Between AI Confidence and Score Variability

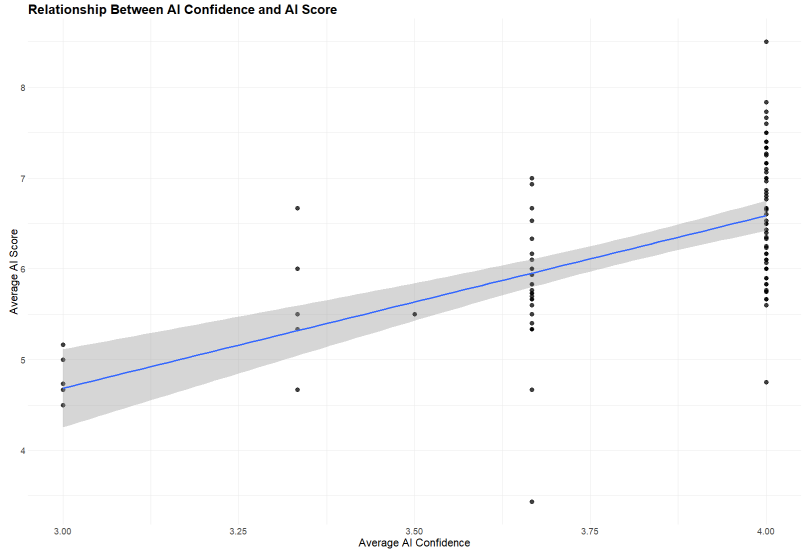

Figure 17: Relationship Between AI Score and AI Confidence

6 Conclusions and Further Research

6.1 Conclusion

This thesis investigated the potential of Large Language Models to support the scientific peer review process by comparing AI-generated reviews with human reviewer evaluations. The study focused on three main aspects: the consistency of AI-generated review scores across repeated evaluations, the agreement between AI-generated and human-generated review scores, and the relationship between AI confidence and agreement with human reviewers.

To answer these research questions, a dataset of 99 scientific papers from machine learning conferences was constructed. Papers were selected through stratified random sampling based on historical human review scores and conference year. Each paper was evaluated three times using the same LLM under identical review instructions. The resulting AI-generated scores and confidence values were then compared with corresponding human reviewer evaluations.

The results indicate that the AI model demonstrated a relatively high degree of internal consistency. Repeated evaluations of the same paper generally produced similar scores, and the observed standard deviations across runs were low. Furthermore, the variability of AI-generated review scores was significantly lower than the variability observed among human reviewers, indicating that the AI model produced more consistent numerical evaluations. Correlation analysis of the human reviewers also revealed only moderate agreement between individual reviewers, highlighting the inherent variability of the human peer-review process. Together, these findings suggest that the AI model generated stable evaluations despite the probabilistic nature of modern language models.

A statistically significant strong relationship was observed between AI-generated and human-generated review scores. Correlation and regression analyses showed that papers receiving higher scores from human reviewers also tended to receive higher scores from the AI model. However, agreement was far from perfect. While the AI model was capable of capturing general differences in paper quality, substantial disagreements remained for a number of individual papers. Nevertheless, the observed variability among human reviewers indicates that complete agreement is uncommon even within the traditional peer-review process, providing important context when interpreting disagreements between AI-generated and human-generated evaluations.

An important finding of this study was the systematic tendency of the AI model to assign higher scores than human reviewers. The average AI score exceeded the average human score by approximately one point. This pattern was visible across the score distributions and score difference analyses, suggesting that the model evaluated papers more favorably than human reviewers on average.

The confidence analysis revealed only weak relationship between the confidence expressed by the model and its agreement with human reviewers. Although higher confidence was associated with slightly smaller score differences, the relationship was limited. This finding suggests that model confidence should not be interpreted as a reliable indicator of correctness or agreement with human judgment.

The findings of this thesis are largely consistent with the literature we discussed. Liang et al. reported that LLM-generated feedback often overlaps with feedback provided by human reviewers and can identify many of the same strengths and weaknesses in scientific papers. The positive correlation observed between AI-generated and human-generated review scores in this study supports the idea that modern Large Language Models are capable of capturing important

aspects of paper quality that are also recognized by human reviewers. Furthermore, the relatively low variability observed across repeated evaluations suggests that LLMs can provide reasonably stable assessments of scientific papers.

At the same time, the results also highlight several limitations that have been emphasized in previous research. While the AI-generated scores were positively associated with human scores, substantial disagreements remained for individual papers, and the model demonstrated a tendency to assign systematically higher scores than human reviewers. In addition, the confidence analysis showed that higher confidence was not necessarily associated with stronger agreement with human evaluations. These findings support concerns raised in the literature regarding the reliability, calibration, and practical use of AI-generated peer reviews.

Overall our research suggest that modern LLMs can identify broad differences in scientific paper quality and can generate relatively consistent review scores. However, the observed disagreements with human reviewers and the tendency toward more favorable evaluations indicate that current AI systems should be viewed as support tools rather than replacements for human peer reviewers. The results therefore support the idea that AI can assist the peer review process while still requiring human oversight and final decision-making.

6.2 Further Research

This study provides insights into the reliability of AI-generated peer reviews, but several opportunities for future research remain.

First, future studies could expand the dataset beyond machine learning conferences and include papers from a wider range of scientific disciplines. Different fields often use different evaluation criteria and reviewing cultures, which may influence the level of agreement between AI-generated and human-generated reviews.

Second, future research could investigate additional Large Language Models and compare their reviewing performance. Since this thesis focused on a single model, it remains unclear whether the observed patterns generalize to other commercial or open-source LLMs. Comparative studies could help identify which model characteristics contribute most to reliable reviewing behavior.

Third, future work could examine the textual content of reviews in greater detail. While this thesis primarily focused on numerical review scores and confidence values, peer review quality also depends on factors such as feedback usefulness, specificity, constructiveness, and the ability to identify methodological weaknesses. Automated and human evaluations of review text could provide a more comprehensive understanding of AI reviewing capabilities.

Another direction of future research could also be studying different review settings and prompting strategies. Variations in review instructions, evaluation criteria, or review formats may influence both score consistency and agreement with human reviewers. Understanding these effects could help optimize AI-assisted reviewing systems for practical use. [9]

Future research could also investigate potential sources of bias in AI-generated reviews. Although bias was not examined directly in this thesis, concerns remain regarding how AI systems may respond to factors unrelated to scientific quality, such as author affiliation, institution prestige, geographic location, or research topic. Controlled experiments could provide valuable evidence regarding the fairness of AI-assisted reviewing.

Finally, future studies could explore hybrid reviewing approaches in which AI-generated reviews

are combined with human evaluations. Rather than replacing human reviewers, AI systems may prove most valuable as decision-support tools that provide preliminary assessments, identify potential weaknesses, or assist reviewers in evaluating large numbers of submissions. Understanding how humans and AI systems can most effectively collaborate remains an important area for future investigation.

Large Language Models continue to improve and their role in scientific peer review is likely to expand. Continued research will therefore be essential to ensure that AI-assisted reviewing systems remain reliable, transparent, and beneficial to the scientific community.

7 Code

The code developed for this thesis, including the scripts used for data collection, AI review generation, data preprocessing, and statistical analysis, is publicly available in the following GitHub repository:
GitHub Repossitory

References

[1] Gpt-4 technical report. *https://arxiv.org/abs/2303.08774*, 2023.

[2] Icml 2026 call for papers. *https://icml.cc/Conferences/2026/CallForPapers*, 2026.

[3] Flaminio Squazzoni Ana Marušić and Francisco Grimaldo. Publishing: Journals could share peer-review data. *Proceedings of the National Academy of Sciences*, 2017.

[4] Min Zhang Andrew Tomkins and William D. Heavlin. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 2017.

[5] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. Attention is all you need. *https://arxiv.org/abs/1706.03762*, 2017.

[6] Weixin Liang, Getahun Asfaw Tadesse, Daniel Ho, et al. Monitoring ai-modified content at scale: A case study on peer reviews. *Proceedings of the National Academy of Sciences*, 2024.

[7] Kathrin Grosse Pavel Laskov Emil Lupu Vera Rimmer Philine Widmer Luca Demetrio, Giovanni Apruzzese. Gen-review: A dataset and large-scale study of ai-generated and human-authored peer reviews. *https://openreview.net/forum?id=sBjjA7f94p*, 2025.

[8] Jake Silberg Animesh Garg Nanyun Peng Fei Sha Rose Yu Carl Vondrick James Zou Nitya Thakkar, Mert Yuksekgonul. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025. *https://arxiv.org/abs/2504.09737*, 2025.

[9] Chayapatr Achiwaranguprok Pattie Maes Pat Pataranutaporn, Nattavudh Powdthavee. Can ai solve the peer review crisis? a large scale cross model experiment of llms' performance and biases in evaluating over 1000 economics papers. *https://arxiv.org/abs/2502.00070*, 2025.

[10] Amanda Askell Tom Henighan Dawn Drain Ethan Perez Nicholas Schiefer Zac Hatfield-Dodds Nova DasSarma Eli Tran-Johnson Scott Johnston Sheer El-Showk Andy Jones Nelson Elhage Tristan Hume Anna Chen Yuntao Bai Sam Bowman Stanislav Fort Deep Ganguli Danny Hernandez Josh Jacobson Jackson Kernion Shauna Kravec Liane Lovitt Kamal Ndousse Catherine Olsson Sam Ringer Dario Amodei Tom Brown Jack Clark Nicholas Joseph Ben Mann Sam McCandlish Chris Olah Jared Kaplan Saurav Kadavath, Tom Conerly. Language models (mostly) know what they know. *https://arxiv.org/abs/2207.05221*, 2022.

[11] Avinash Madasu Vasudev Lal Phillip Howard Sungduk Yu, Man Luo. Is your paper being reviewed by an llm? benchmarking ai text detection in peer review. *https://arxiv.org/abs/2502.19614*, 2026.

[12] Nick Ryder Melanie Subbiah Jared Kaplan Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry Amanda Askell-Sandhini Agarwal Ariel Herbert-Voss-Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray Benjamin Chess Jack Clark Christopher Berner Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei Tom B. Brown, Benjamin Mann. Language models are few-shot learners. *https://arxiv.org/abs/2005.14165*, 2020.

[13] Hancheng Cao Binglu Wang Daisy Ding Xinyu Yang Kailas Vodrahalli Siyu He Daniel Smith Yian Yin Daniel McFarland James Zou Weixin Liang, Yuhui Zhang. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *https://doi.org/10.48550/arXiv.2310.01783*, 2023.