Athens Journal of Technology & Engineering 2026, 12: 1-21
https://doi.org/10.30958/ajte.X-Y-Z

AI vs. Human Evaluation of Student Research Papers in Dual University Education: A Comparative Analysis

By Franziska Schütz[*], *Anke Hutzschenreuter*[±] *& Till Hänisch*[○]

Dual university education combines theoretical study with practical industry experience. This structure creates challenges for traditional academic evaluation approaches, due to the strong focus on theory-practice integration. Despite advances in AI-based evaluation, no research has examined AI capabilities in this context requiring integration of theoretical depth alongside with practical application. This study provides the first systematic comparison of AI versus human evaluation in dual university contexts, analyzing 21 computer science student research papers from final-year students. Human evaluation employed three domain experts while AI evaluation utilized Claude Sonnet 3.7 with systematic multi-step prompting and role-based expertise simulation. Both evaluators used identical rubric-driven criteria. Statistical analyses included cosine similarity, Spearman correlation, and further criteria-specific analysis. Results revealed strong similarity agreement and moderate rank correlation between AI and human evaluation approaches. However, systematic differences emerged in three key areas: (1) AI demonstrated grading patterns within a smaller range, (2) AI showed limited contextual adaptation compared to extensive human weighting adjustments, and (3) project-type dependency, with theoretical work achieving higher agreement than practical implementations. Findings support complementary implementation rather than replacement. AI excels with explicit documentation but requires human oversight for practical projects. This research establishes empirical foundations for hybrid evaluation approaches in dual university contexts.

Keywords: *AI-based evaluation, human-AI comparison, dual university education, academic assessment, computer science, artificial intelligence, quality assurance in higher education*

Introduction

AI-based academic writing evaluation has demonstrated effectiveness in traditional university contexts where students focus primarily on theoretical analysis and abstract knowledge development. However, dual university programs operate under a different educational philosophy that creates evaluation challenges.

Dual university programs represent an educational paradigm designed to bridge the theory-practice divide in higher education. Students in these programs alternate between academic theory phases at university and practical industry phases with partner companies. This educational structure is designed to achieve knowledge

[*]Research Associate, DHBW Heidenheim, Germany.
[±]Professor, DHBW Heidenheim, Germany.
[○]Professor, DHBW Heidenheim, Germany.

transfer - the practical application of theoretical concepts in real-world professional contexts.

This theory-practice integration creates an academic paper profile that differs from traditional university work. Students produce research papers that emphasize practical application, real-world problem-solving, and implementation results alongside theoretical analysis. The evaluation of such work requires approaches capable of recognizing both academic rigor and practical implementation quality.

The practical orientation of dual university education introduces evaluation challenges. Traditional academic papers can be evaluated primarily through textual analysis of theoretical argumentation, literature synthesis, and methodological soundness. In contrast, dual university papers require evaluation of practical elements that may not be fully documented in text. These include implementation decisions based on real-world constraints, adaptation of theoretical frameworks to industry requirements, and practical problem-solving approaches that emerge from hands-on experience. These practical components often rely on implicit knowledge and contextual understanding that traditional text-based evaluation approaches may not capture effectively.

Given the proven capabilities of Large Language Models (LLMs) in text analysis and academic evaluation (Chiang et al. 2024, Usher 2025), AI-based evaluation presents a potential solution. However, while AI evaluation has shown effectiveness in traditional academic contexts focused on theoretical work, its capability to evaluate papers with practical implementation components remains unexplored. The question emerges whether AI can effectively assess the theory-practice integration that defines dual university educational outcomes.

This research examines whether AI-based evaluation can effectively evaluate student research papers in dual university contexts where practical application is a core educational objective. Using comparative methodology with identical evaluation criteria, we analyze AI and human evaluations of computer science papers that integrate theoretical knowledge with practical implementation work.

Literature Review

The development of automated essay scoring (AES) has progressed through several technological generations, each introducing new capabilities while revealing ongoing limitations. Traditional deep learning-based AES systems successfully capture local linguistic features and global document structure. However, these systems remain limited in their ability to assess semantic depth, domain-specific nuance and higher-order cognitive skills (Misgna et al. 2024).

The transition toward modern large language model (LLM) approaches has introduced new capabilities in contextual understanding and feedback generation. Recent comparative research provides insights into how these AI-based evaluation systems perform against human evaluators. These studies establish both methodological frameworks and empirical patterns that inform educational applications.

Multiple studies (Awidi 2024, Usher 2025, Wetzler et al. 2025) have established validated frameworks for comparing AI and human essay evaluations. They used identical rubrics for AI and human evaluators, conducted independent evaluations, and analyzed quantitative correlations and qualitative feedback patterns. This methodological convergence demonstrates that systematic comparison designs are both feasible and reliable for evaluation research.

These comparative studies reveal predictable differences between AI and human evaluation approaches. AI-based evaluation systems consistently assign higher average grades than human evaluators across multiple contexts (Usher 2025, Wetzler et al., 2025). This finding suggests systematic differences rather than random variation. (Wetzler et al. 2025) identified bias pattern where AI grading becomes more lenient at lower performance levels and stricter at higher levels. (Flodén 2025) confirmed that AI tends to give marginally higher scores and to avoid extreme grades. Despite grade differences, moderate positive correlations indicate agreement in relative ranking (Usher 2025).

Qualitative analysis reveals that AI and humans offer different but potentially complementary capabilities: AI-based evaluation excels at providing detailed, systematic feedback with consistent structure, while human feedback demonstrates contextual sensitivity, disciplinary expertise, and nuanced understanding of student intent (Awidi 2024, Flodén 2025, Usher 2025). Banihashem et al. (2024) found that peers provided higher quality feedback in problem identification, while AI excelled in descriptive feedback. These findings suggest hybrid rather than substitutive approaches.

Comparative studies confirm the generalizability of AI-human evaluation patterns across academic domains. The research spanned from social science domains such as psychology (Wetzler et al. 2025) up to technical context in engineering courses (Awidi 2024), instructional technologies (Usher 2025), and broader STEM disciplines (Flodén 2025). However, domain-specific research reveals limitations in AI's reasoning capabilities.

Studies in computer science and technical domains demonstrated that AI struggles with complex reasoning, contextual interpretation, and connecting steps to broader conceptual frameworks (Seßler et al. 2025, Tithi et al. 2025). This suggests limitations in higher-order analytical skills that are necessary for evaluating academic papers. (Smerdon 2024) confirms that implementation context significantly influences effectiveness.

Recent research provides insights into optimizing AI-based evaluation performance. Xie et al. (2024) found that systematic design and a multi-agent grading system enhances the AI grading process regarding consistency and accuracy. This aligns with the findings of (Wei et al. 2025) who demonstrated that concept-based rubrics significantly improve LLM assessment accuracy. Another factor for improving AI evaluation quality is sophisticated prompt engineering. Systematic prompt design significantly improves assessment quality (Tithi et al. 2025, Xie et al. 2024).

This review of current literature reveals advances in AI-human comparative evaluation research with existing studies focusing on traditional educational settings. Domain-specific limitations highlight the importance of contextual

adaptation when implementing AI-based evaluation systems in specialized academic environments. However, no research has examined AI-based evaluation capabilities in dual educational contexts that emphasize practical application alongside theoretical knowledge. This research addresses this gap by conducting a systematic comparative analysis of AI and human evaluation in dual university contexts while providing domain-specific insights for computer science academic papers.

Methodology

This exploratory study employs a comparative experimental design to evaluate AI-based versus human evaluation of computer science student research papers. The comparison examines evaluation similarity and consistency using identical evaluation criteria to establish insights into AI evaluation capabilities in dual university contexts.

Data Collection and Sample

The data collection comprises a dataset of 21 student research papers from computer science students in their final year of Bachelor studies (5th and 6th semester). This sample size aligns with similar exploratory studies in AI evaluation research (Seßler et al. 2025) and enables an exploratory comparative study with detailed analysis and serves as a foundational investigation to establish methodology and preliminary findings for future larger-scale studies.

The student research papers integrate theoretical knowledge with practical application, reflecting typical dual university requirements. Papers range from 30-80 pages depending on practical implementation requirements, with an average length of 65 pages.

The papers cover diverse computer science topics, e.g., software development, AI, data science, IoT, software architecture, and ethics. The papers employ diverse methodological approaches that include theoretical elaboration. Most papers focus on practical application through concept design, implementation, and prototyping. The dataset also includes several purely theoretical papers to enable comparison with existing studies and validate the methodological approach.

AI Setup and Prompt Engineering Strategy

AI evaluation was conducted using Claude Sonnet 3.7, accessed via the web interface with project-based document management for consistent access to evaluation materials. This model was selected based on preliminary testing showing superior prompt adherence and consistency in complex instruction-following compared to alternatives (e.g., Gemini). According to Anthropic's privacy documentation, user inputs and outputs are not used to train their generative models (Anthropic n.d.), ensuring that student papers remain

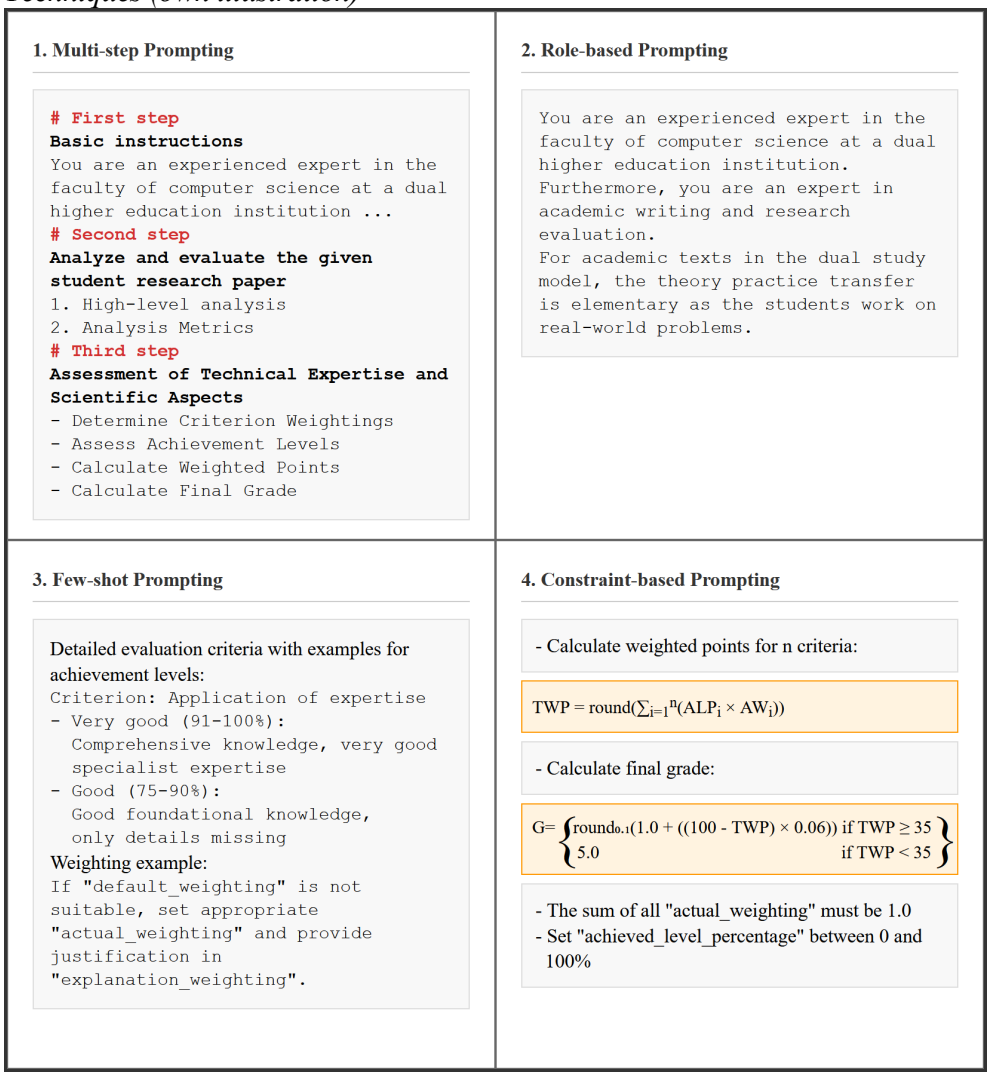
confidential and providing protection for students' intellectual property. Additionally, all papers were de-identified by removing sensitive information while preserving content quality and evaluation validity.

The prompting strategy follows established frameworks for educational AI applications (Tithi et al. 2025), employing a systematic prompting structure. Each of the four key prompt engineering techniques contributes to the structured and consistent evaluation process.

The four prompting techniques, as shown in Figure 1, are:

1. **Multi-step prompting**: The prompts were divided into three steps (see Panel 1 in Figure 1) that were executed sequentially to prevent premature evaluation. The three phases include: understand (comprehend instructions and paper context), analyze (conduct high-level assessment), and evaluate (apply detailed rubric scoring). Each step's output informs the next phase of the evaluation process, ensuring systematic and thorough analysis.
2. **Role-based prompting**: For consistent expert perspective, the AI was instructed to act as a computer science professor, senior researcher, and evaluation expert. The role assignment is illustrated in Panel 2 of Figure 1.
3. **Few-shot prompting**: To enhance understanding and consistency, we provided detailed evaluation criteria with explanations, examples of different achievement levels, specific instructions for weighting adaptations, and a grading report structure. Panel 3 of Figure 1 demonstrates this example-based guidance.
4. **Constraint-based prompting**: We implemented numerical restrictions for weightings and achievement levels and provided mathematical formulas for grade calculation to ensure consistent evaluation outcomes. These constraints are demonstrated in Panel 4 of Figure 1. Additionally, we used structured input/output formats including JSON schema for evaluation forms to ensure standardized output formatting.

Figure 1. *Systematic Prompt Engineering Framework demonstrating Four Key Techniques (own illustration)*



Additionally, prompt effectiveness was validated by having the AI reflect on instructions and confirm understanding of evaluation criteria and rubric interpretations before beginning each evaluation. This validation process ensured consistent interpretation across all evaluations.

Human Evaluation Process

Human evaluation was conducted by three domain experts, with each paper evaluated by one expert. The experts are professors and academic staff with expertise in computer science and academic assessment, each with at least 6 years of evaluation experience. Experts were assigned to papers based on their subject matter knowledge. All experts used identical rubrics to ensure comparability of evaluations.

Evaluation Framework

For the evaluation, we used the structured DHBW evaluation form (DHBW Heidenheim, 2025), which provides measurable components through explicitly defined categorized criteria. The criteria are grouped into two sections:

1. **Technical and conceptual execution**: Subject-related work, application of expertise, application of methods and tools, feasibility and practical relevance of results, creativity, and economic assessment.
2. **Academic work and methodology**: Independence and initiative, systematic approach, documentation, literature review, and use and integration of sources.

Each criterion provides specific descriptions for the five achievement levels to guide consistent evaluation (see Figure 2). While suggested weightings are provided, evaluators may adjust them if justified. Any changes must be documented to ensure transparency and traceability.

Figure 2. *Example Criteria with Descriptions for Achievement Levels (own illustration based on (DHBW Heidenheim, 2025))*

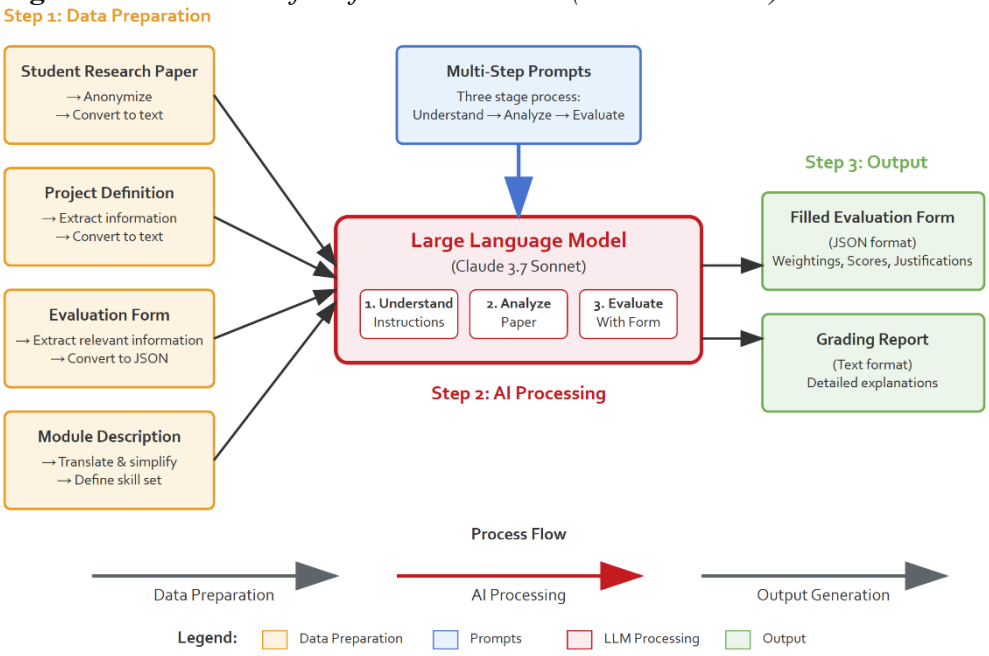
Achievement level	Application of expertise	Systematic approach
Fail 0 - 49 %	Limited knowledge of the state of the art	No recognizable systematic approach in procedure and results
Sufficient 50 - 57 %	Basic knowledge but with significant gaps in specialist expertise	Considerable deficiencies in systematic approach regarding methods, priority setting, content structure
Satisfactory 58 - 74 %	Basic knowledge but with moderate gaps in specialist expertise	Methods comprehensibly applied and pursued, priorities partially sensibly set, content structure comprehensible
Good 75 - 90 %	Good foundational knowledge, only details missing	Methods consistently applied, priorities sensibly set, concentration on essentials, clear content structure
Very good 91 - 100 %	Comprehensive knowledge, very good specialist expertise	Methods systematically applied, priorities sensibly set, clear focus on essentials, clear content structure

This study recognizes that academic evaluation by human experts is inherently subjective. The structured evaluation criteria aim to maximize consistency and transparency while acknowledging the inherent limitations of any single evaluator approach.

Process Workflow

As shown in Figure 3, the study follows a systematic three-step process that directly implements the multi-step prompting framework described in *AI setup and Prompt Engineering Strategy*.

Figure 3. *Process Workflow for AI Evaluation (own illustration)*



1. **Data preparation**: This step processes all input materials for structured AI evaluation:
   - Student research papers (from the sample described in Section *Data Collection and Sample*) are anonymized and converted to text format
   - Project definitions and module descriptions are extracted and translated for contextual understanding
   - DHBW evaluation form (detailed in Section *Evaluation Framework*) is converted to JSON format for structured AI input
   - All materials are prepared according to the constraint-based prompting requirements (see Section: *AI setup and Prompt Engineering Strategy*)
2. **AI processing**: Sequential execution of the three-phase prompting strategy. Each phase incorporates the prompt engineering strategy outlined in Section *AI setup and Prompt Engineering Strategy*:
   - Phase 1 (Understand): AI comprehends evaluation instructions and paper context
   - Phase 2 (Analyze): High-level assessment using role-based prompting techniques
   - Phase 3 (Evaluate): Detailed rubric scoring with few-shot prompting guidance
3. **Output generation**: Production of structured evaluation results (matching the format described in Section *Evaluation Framework*):

- Completed JSON evaluation forms with weightings, achievement levels, and justifications
- Structured grading reports in German and English with detailed criterion-specific explanations
- All outputs follow the mathematical constraints and formatting requirements established in Section *AI setup and Prompt Engineering Strategy*

Statistical Analysis

To compare AI and human evaluations, we conduct statistical analysis and visualize the results. We analyze similarity, grade distributions, and ranking patterns to assess agreement between evaluators. Additionally, we conduct criteria-specific analysis to examine evaluation patterns for individual criteria.

For similarity comparison, we treat the evaluations as vectors where each criterion score represents a dimension. We calculate the cosine similarity to measure the angle between evaluation vectors (ranging from 0 to 1). Additionally, we calculate the normalized Euclidean distance to measure the magnitude of differences between evaluation vectors (ranging from 0 to 1). Cosine values close to 1 (>0.9) indicate high agreement, while smaller Euclidean distances (<0.3) indicate similar evaluation patterns.

The distribution for AI and human grades is shown in a boxplot to identify the spectrum of both evaluators and to discover patterns. We convert the raw grades into ordinal ranks for a rank-based comparison that is less sensitive to scale differences or outliers. The correspondence between AI and human ranks is visualized in a rank correlation plot, where each research paper is represented as a point. To quantify the relationship between the two rankings, we compute the Spearman rank-order correlation coefficient (Spearman's ρ). This non-parametric measure is well-suited for ordinal data and reflects how well the relationship between the two variables can be described by a monotonic function. We use thresholds and interpretation established in research (≥ 0.7 strong agreement, 0.4-0.7 moderate, <0.4 weak agreement) (Dancey & Reidy 2017).

For detailed comparison, we conduct criteria-specific analysis examining deviation per criterion for detailed comparison. The individual paper evaluations are visualized in interactive charts using line, bar, and radar charts to compare AI and human rating for the individual criteria.

These statistical approaches were selected to provide both overall agreement measures and detailed dimensional analysis. Cosine similarity and Spearman correlation assess overall agreement, while criteria-specific comparisons enable detailed examination of evaluation patterns.

Results

The results of our comparative study for AI-based and human evaluations are organized around three key analytical dimensions: descriptive statistics (grade distribution, similarity and correlation analysis), criteria-specific analysis, and detailed case study analysis examining specific evaluation scenarios.

Overview of Key Findings

The comparative analysis between AI and human evaluation revealed both substantial agreement and systematic differences in evaluation approaches. AI and human evaluators demonstrated strong overall similarity with 81% of evaluations achieving cosine similarity >0.9 and moderate rank correlation (Spearman's ρ=0.69). This indicates consistent relative performance ranking across evaluation approaches. However, three systematic differences emerged that distinguish AI and human evaluation patterns:

1. Constrained grading behavior[1]: AI demonstrated a smaller grading range (1.8-2.6) compared to human evaluators who utilized the full grading range for passing (1.0-4.0).
2. Limited contextual adaptation: Human evaluators adjusted criterion weightings in 90.5% of evaluations based on project-specific characteristics. Whereas AI applied non-standard weightings in only 19.0% of cases, indicating restricted ability to adapt evaluation emphasis to individual work contexts.
3. Project type dependencies: Evaluation agreement varied by project type: While theoretical research papers achieved high AI-human alignment (cosine similarity >0.9), the largest disagreements were shown by practical implementation projects.

The following sections provide detailed statistical analysis of these patterns and examine specific evaluation scenarios that illustrate the differences.

Descriptive Statistics

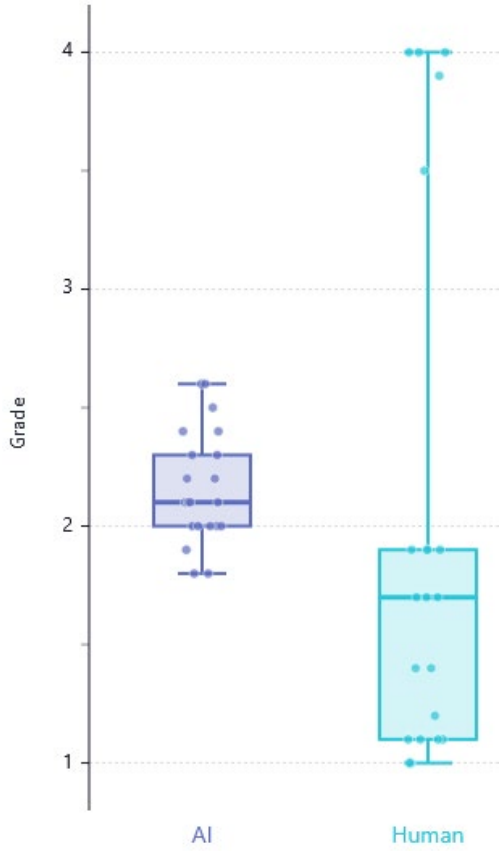
The descriptive statistical analysis examines the patterns in AI and human evaluation approaches across three key dimensions: grade distribution characteristics, evaluation similarity measures, and rank correlation patterns. This analysis provides understanding of evaluator agreement and differences before examining detailed criteria-specific patterns.

The statistical analyses were conducted using a custom web-based analysis tool developed specifically for this study (available at https://github.com/schuetz-dhbw/eval_vis). The tool implements standard statistical formulas. All visualization charts were generated using this tool to ensure consistency in analysis and presentation.

The boxplot shown in Figure 4 illustrates the grade distribution for AI and human evaluations. AI demonstrated grading patterns within a smaller range compared to human evaluators. Human evaluations utilized the full grading range (1.0-4.0) with a median of 1.7, while AI evaluations clustered in a smaller range (1.8-2.6) with a median of 2.1. This pattern indicates that AI tends to avoid both extreme excellence (grades below 1.5) and borderline performance ratings (grades above 3.0).

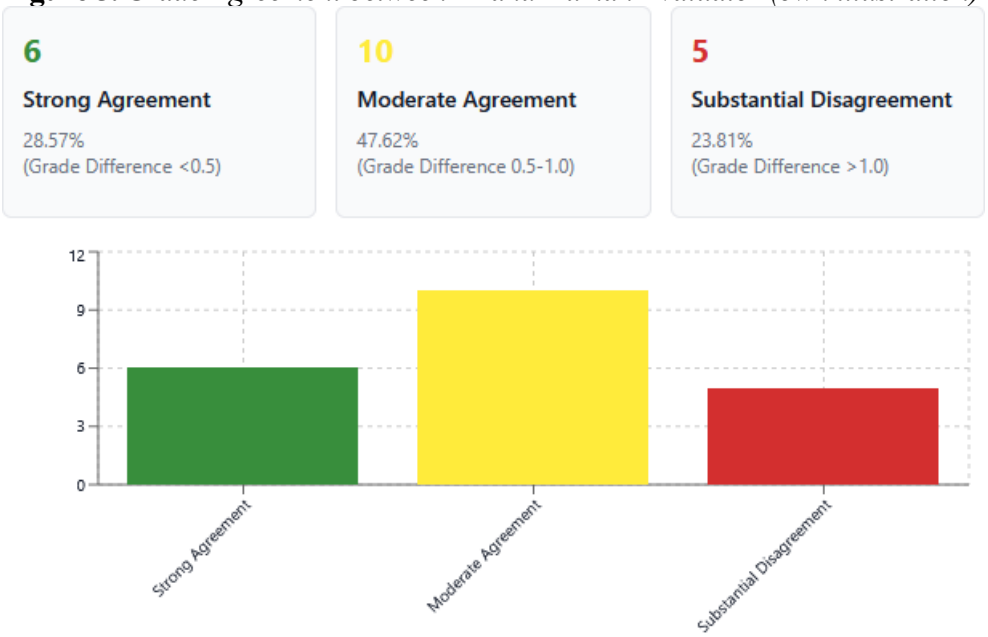
[1]This study employed the German grading system. The grades range from 1.0 (excellent) to 4.0 (sufficient/pass), with 5.0 representing failure. This system allows for precise differentiation in academic performance assessment.

Figure 4. *Boxplot Analysis for Grade Distribution (own illustration)*



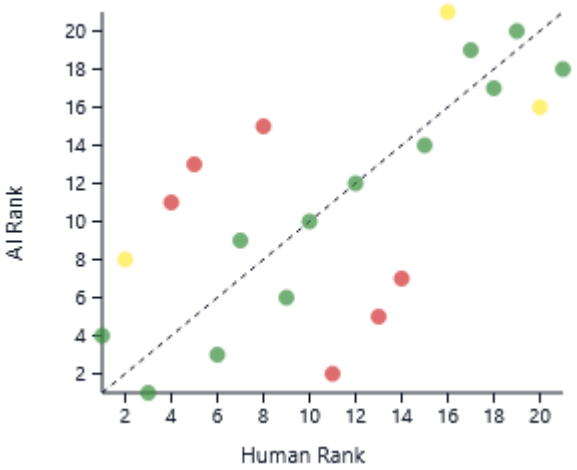
Despite these distribution differences, the grade comparisons of the individual evaluations revealed a more nuanced pattern of evaluator agreement. The grade differences between AI and human evaluations ranged from 0.1 to 1.7 points. Figure 5 shows the grade agreement percentage, with the colors indicating the agreement level (green: high agreement, yellow: moderate agreement, red: low agreement). Strong agreement (grade difference <0.5) was observed in 28.6% of cases. Most evaluations (47.6%) show a moderate agreement with grade differences between 0.5 and 1.0, while substantial disagreement is displayed for 23.8%. This distribution indicates that large discrepancies occur in specific cases rather than systematically across all evaluations.

Figure 5. *Grade Agreement between AI and Human Evaluator (own illustration)*



To examine the differences independently of the exact grades, we converted the grades into ordinal ranks for a rank-based comparison. The rank correlation analysis shown in the scatter plot in Figure 6 illustrates the correspondence between AI and human ranks. Each point shows the AI and human rank and is colored to indicate the agreement level (green: high agreement, yellow: moderate agreement, red: low agreement). Several papers maintained identical or very similar rankings between evaluators, while others showed more substantial rank differences.

Figure 6. *Rank Correlation Analysis comparing AI and Human Evaluator Rankings (own illustration)*

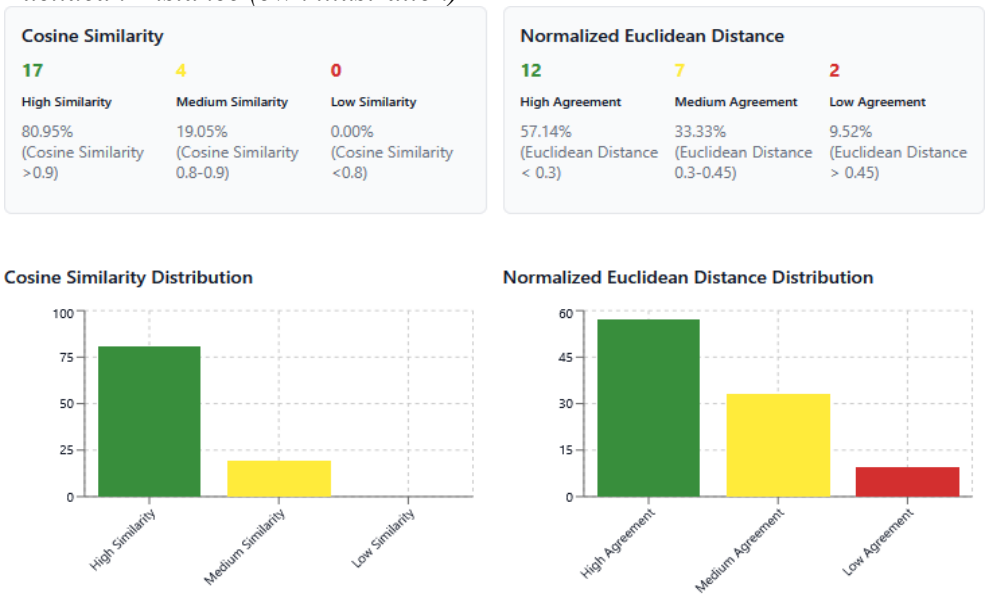


The Spearman rank order correlation coefficient is calculated by comparing the rank positions between AI and human evaluators. This quantifies the relationship between the two ranking systems with values ranging from -1 to +1. The Spearman

correlation coefficient ρ = 0.69 determines the correlation strength. This indicates moderate positive correlation between AI and human rankings (Dancey & Reidy 2017). While evaluators generally agree on relative performance rankings, individual papers showed varying degrees of rank agreement, providing insights into the types of work where evaluation approaches diverge.

To achieve a more comprehensive understanding of the evaluation differences, we did similarity analysis over the evaluations to compare AI and human evaluation pattern. Similarity analysis between AI and human evaluation vectors revealed strong overall agreement across the majority of papers, as shown in Figure 7. The colors indicate the agreement level (green: high agreement, yellow: moderate agreement, red: low agreement).

Figure 7. *Similarity and Agreement Levels based on Cosine Similarity and normalized Euclidean Distance (own illustration)*



Using cosine similarity measurements, 81% of evaluations achieved high similarity (>0.9), while 19% demonstrated medium similarity (0.8-0.9) with the lowest similarity occurred for practical implementation projects requiring implicit knowledge assessment. The analysis of the normalized Euclidean distances revealed a similar agreement level: This analysis showed that 57.1% of evaluations had a high agreement level with normalized distances <0.3. One third of the papers had a normalized Euclidean distance between 0.3 and 0.45, which corresponds to a moderate agreement level. With 9.5% of evaluations having a low agreement level (>0.45), this analysis suggests that evaluators captured similar evaluation patterns while differing primarily in magnitude rather than direction of judgment.

Papers with cosine similarities above 0.95 consistently showed normalized distances below 0.28, while the four papers with medium similarity classifications (cosine similarity < 0.9) corresponded to normalized distances above 0.38. This consistent inverse relationship between cosine similarity and normalized Euclidean distance validates both measures. Furthermore, it reveals that evaluation disagreements

are limited to specific project types, particularly those involving hands-on development work where contextual understanding beyond written documentation becomes relevant for accurate evaluation.

Detailed similarity analysis of the individual criteria reveals criterion-specific agreement patterns. To identify which evaluation criteria contributes to the observed similarities and differences, we calculated cosine similarity separately for each criterion across all 21 papers. For each criterion, the 21 AI scores were treated as one vector and the 21 human scores as another vector. Then we calculated the cosine of the angle between these vectors. This approach shows the agreement strength between AI and human for each individual dimension.

Figure 8. *Radar Chart showing the Cosine Similarities per Criterion*

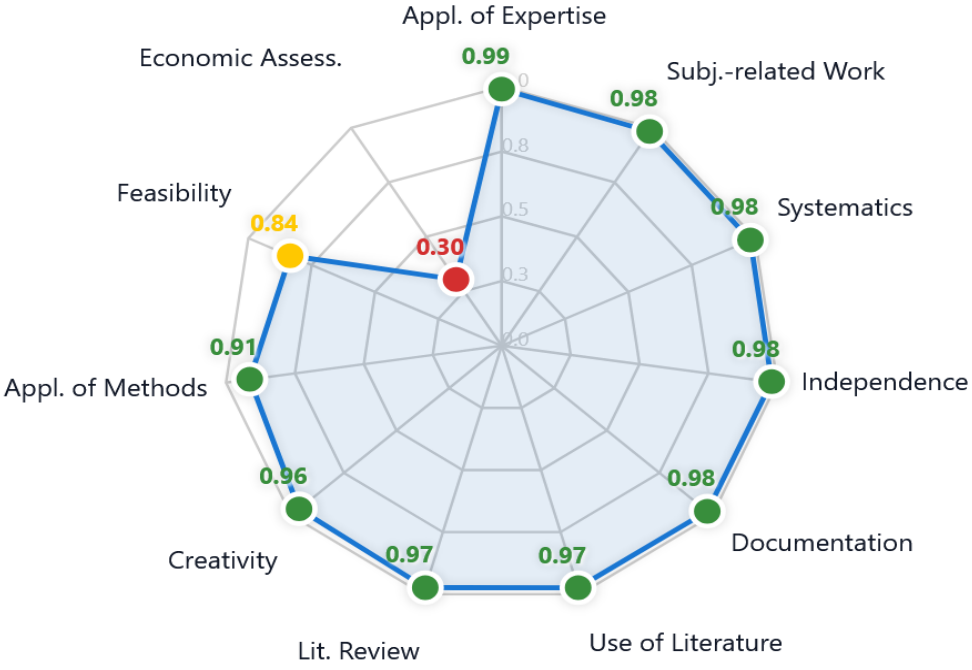


Figure 8 presents the criterion-specific cosine similarity scores, revealing variation in AI-human agreement across evaluation dimensions. AI demonstrates capability in evaluating traditional academic criteria such as documentation, literature use, and systematic approaches (cosine similarity > 0.97). Feasibility showed moderate similarity (0.84), suggesting some evaluator differences in evaluating practical implementation. However, AI struggles with contextual evaluation like Economic Assessment (0.30). This indicates differences in how AI and human evaluators approach the relevance of Economic Assessment. This criterion-specific pattern explains the overall moderate correlation ($\rho = 0.69$), as agreement in most criteria is offset by poor performance in context-dependent assessments.

Criteria-Specific Analysis

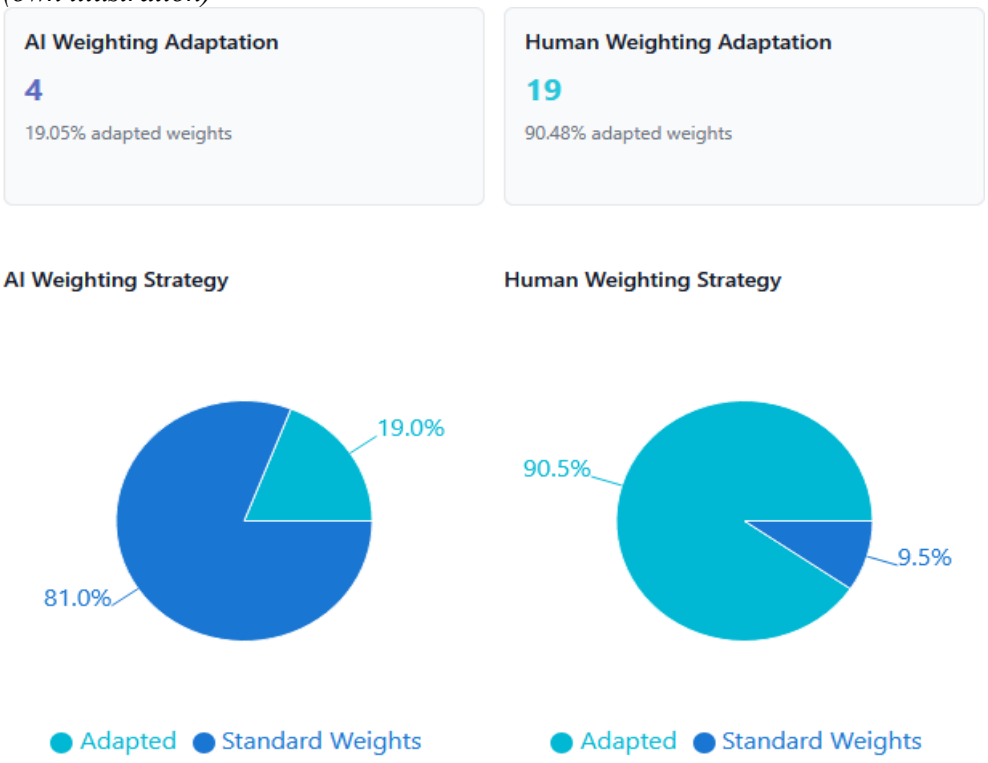
To gain a deeper understanding of how AI and human evaluators differ in their evaluation behavior, we analyzed the results at the criterion level. This includes

differences in performance scoring, criterion weighting, and the resulting weighted contributions to the final grade.

We examined the scoring differences across criteria to identify where evaluators disagreed most on performance levels. Only those criteria that had assigned a weighting >0 by both evaluators were included in this comparison. The largest discrepancy was observed for the criterion Creativity, with an average scoring difference of 19.4%. This suggests that AI and human evaluators apply different standards when assessing innovative aspects of student work. Similarly, the Literature Review criterion showed a scoring difference of 18.8%, indicating divergent interpretations regarding the depth and quality of scholarly engagement.

Beyond performance scoring, a major difference emerged in how evaluators prioritized criteria. As shown in Figure 9, human evaluators adapted the weighting of evaluation criteria in 90.5% of cases, frequently tailoring the emphasis based on project-specific characteristics. In contrast, AI adjusted its weightings in only 19% of evaluations. The highest disagreement was found in the weighting of the criterion Economic Assessment. While AI provides consistent criterion application, human evaluators often considered it irrelevant for the given project context and therefore excluded it. This reflects a key difference in evaluation strategy: AI provides consistent criterion application across projects, whereas human evaluators demonstrate flexibility and context-awareness.

Figure 9. *Pie Charts showing Weighting Adaptation for AI and Human Evaluators (own illustration)*

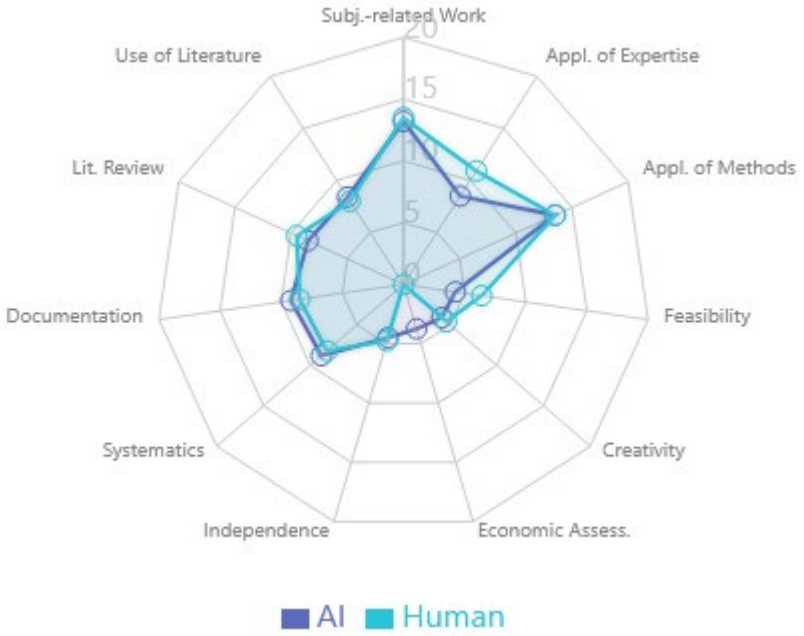


Case Study Analysis

Two case studies highlight the conditions under which AI and human evaluation approaches converge or diverge most, revealing key factors that influence evaluation accuracy.

High Agreement Case

Figure 10. *Radar Chart showing the distribution of weighted Points across all Criteria for AI (violet) and Human (turquoise) with nearly Identical Radars indicating the High Agreement in the Evaluation*

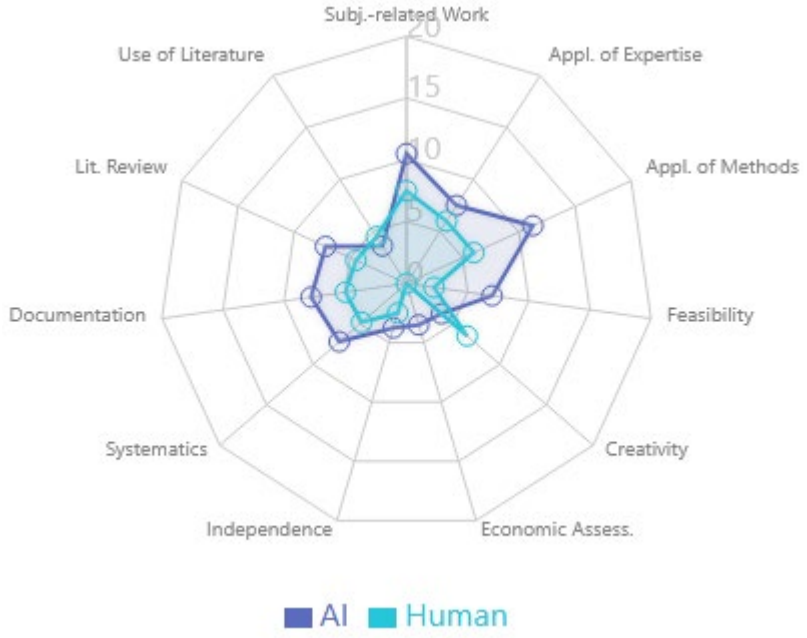


As shown in Figure 10, this research project achieved exceptional alignment (grade difference: 0.1, cosine similarity: 0.96) across all evaluation dimensions. This is an example of a typical project in the dual education context as it combines theoretical elaboration with practical implementation. The practical part was documented and explained in detail in the paper, which made a comprehensive evaluation based on the textual work possible.

Both AI and human evaluator recognized comparable strengths in methodology, documentation, and systematic approach, with near-perfect agreement on criterion relevance and weighting. The strong consensus demonstrates AI's capability to accurately assess even practical projects when evaluation can be based on explicit textual evidence and clearly documented methodologies.

Low Agreement Case

Figure 11. *Radar Chart showing the Distribution of weighted Points across all Criteria for AI (violet) and Human (turquoise) with very different Radars Indicating the Low Agreement in the Evaluation*



This practical implementation project showed the largest evaluation discrepancy (grade difference: 1.4, cosine similarity: 0.84) in the dataset (see Figure 11). This example was a group project with a very large practical component consisting of concept development, implementation, documentation, and user guide. AI evaluation was limited by insufficient documentation, as the paper failed to clearly describe the implementation conditions, process, and outcomes.

While AI assigned consistently high scores to technical criteria, the human evaluator appeared to account for undocumented implementation challenges. This led to lower ratings in areas such as feasibility and methodological rigor, which suggests concerns that AI could not assess from documentation alone.

These cases reveal that AI evaluation not only varies by project type but also by paper quality. AI evaluation strengths include application of stated rubric criteria, detailed justifications, and systematic feedback structure. However, AI evaluation is limited by the difficulty of evaluating undocumented practical achievements, understanding real-world applications, and the tendency to apply all criteria regardless of project-specific relevance.

Discussion

This comparative study of AI versus human evaluation in dual university contexts reveals both convergent patterns with existing literature and novel insights.

Our findings contribute empirical evidence about AI-based evaluation capabilities while informing implementation strategies for the dual university setting.

Confirmation of Established Patterns and Novel Insights

Our moderate positive correlation ($\rho = 0.69$) between AI and human rankings aligns with findings from traditional educational contexts (Awidi 2024, Usher, 2025). This confirms that AI-based evaluation maintains reasonable agreement in relative performance ranking across educational contexts. The observed grading distribution differences confirm previous findings (Flodén 2025, Wetzler et al. 2025). AI demonstrated grading patterns within a smaller range (1.8-2.6) compared to human evaluators (1.0-4.0). These convergent findings establish that AI-human evaluation differences are predictable and systematic.

Beyond these confirmatory findings, our research contributes three novel insights specific to dual university contexts that have not been systematically examined in prior literature. Our study extends these findings by quantifying contextual adaptation differences and identifying project type dependencies. Human evaluators demonstrated weighting adjustments in 90.5% of evaluations, while AI applied non-standard weightings in only 19.0% of cases. This gap represents a limitation in AI's ability to adapt evaluation approaches based on project context. This capability matters in dual education contexts where projects vary in their theory-practice integration requirements.

Our case study analysis revealed patterns where AI evaluation accuracy depends on project type. Theoretical research projects achieved high AI-human alignment (grade difference: 0.1, cosine similarity: 0.96). Practical implementation projects showed larger disagreements (grade difference: 1.4, cosine similarity: 0.84). This pattern indicates AI performs well when evaluation can rely on explicit textual evidence but struggles with implicit knowledge assessment required for practical projects.

The criteria-specific analysis identified areas of difference, while providing quantification of these limitations in dual university contexts. The largest scoring differences were found for creativity assessment (19.4%) and literature review evaluation (18.8%). These findings align with domain-specific research (Seßler et al., 2025; Tithi et al., 2025), demonstrating AI limitations in complex reasoning and contextual interpretation within technical domains.

Implications for Dual University Implementation

The moderate correlation between AI-based and human evaluations ($\rho = 0.69$), combined with project type dependencies, suggests strategic implementation opportunities. AI-based systems could provide evaluation support for theoretical research projects and structured assignments with explicit documentation. However, contextual adaptation limitations indicate AI evaluation requires human oversight, particularly for practical implementation projects.

The grading patterns differences and limited weighting adaptation raise considerations for maintaining academic standards. The 90.5% disagreement rate in Economic Assessment relevance demonstrates potential risks of inappropriate

criterion application. This suggests that AI-based systems require domain-specific validation before deployment.

These findings support hybrid implementation approaches that leverage AI evaluation capabilities while addressing limitations through human oversight. AI-based student self-evaluation tools could provide detailed feedback before submission, potentially improving academic paper quality. Secondary opinion systems for human evaluators could enhance consistency and fairness while maintaining human authority for contextual adaptation and final decisions.

Limitations and Future Research

This exploratory study is limited by the sample size of 21 papers and focus on computer science. AI evaluation consistency across multiple runs was not systematically tested. Furthermore, the use of a single LLM prevents assessment of consistency across different AI models. The single-expert-per-paper design prevents direct assessment of inter-rater reliability among human evaluators, while reflecting real-world practice. As the study focuses on comparing evaluation quality between AI and human raters for future development of an AI-based student self-evaluation tool, it does not consider the students' perspectives on this approach. These limitations provide direction for future investigations.

Future research should focus on expanded validation studies with larger sample sizes and multiple AI models to confirm the identified patterns. A larger sample size allows for more precise insights into the impact of different project types within the computer science domain. Furthermore, cross-domain validation beyond computer science would enhance understanding of AI evaluation impact. Longitudinal studies examining student outcomes could assess educational effectiveness. Enhanced prompting strategies, particularly chain-of-thought prompting, could improve AI evaluation transparency and potentially address contextual adaptation limitations (Tithi et al. 2025). Automated grading report analysis would enable systematic extraction of qualitative insights across larger datasets. The development of an AI-based self-evaluation tool could be used to examine the students' perspectives on AI-based evaluation and consider their acceptance towards this approach.

The findings establish a foundation for developing AI-human collaborative assessment approaches that optimize both evaluation methods' strengths while mitigating their limitations. This supports more effective and efficient evaluation processes in dual university education.

Conclusion

This study presents the first systematic investigation of AI evaluation capabilities in dual university contexts. Our analysis addresses the central research objective to examine whether AI can effectively evaluate student research papers in dual university contexts. The results demonstrate that AI achieves strong overall similarity and moderate rank correlation with human evaluators. This indicates that AI can maintain consistency with expert evaluation in dual educational settings. However, our

investigation also revealed systematic limitations: While the results demonstrate the potential of AI-based systems to support academic evaluation, their effective use depends on strategic integration as complementary tools rather than replacements for human evaluation.

The analysis shows that AI evaluation often aligns with human evaluation, particularly in cases where the evaluation criteria can be addressed through explicit textual documentation. In contrast, practical projects involve feasibility considerations or technical decisions that may be only partially documented or implicit. These limitations highlight the need for human oversight for evaluating projects with context-dependent or practice-oriented components. A possible approach is to combine AI with human evaluation in a hybrid model. Such a hybrid approach enables expanded feedback mechanisms and cross-validation while preserving human responsibility for contextual adaptation and final grading decisions.

This research fills a gap by providing the first empirical evidence of AI evaluation capabilities in dual university contexts. It quantifies contextual adaptation limitations and identifies project type dependencies. The findings provide a foundation for AI-based evaluation research in a practical context beyond traditional university settings.

References

Anthropic (n.d.) *How do you use personal data in model training?* Anthropic Privacy Center. https://privacy.anthropic.com/en/articles/10023555-how-do-you-use-personal-data-in-model-training

Awidi IT (2024) Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence*, *6*, 100226. https://doi.org/10.1016/j.caeai.2024.100226

Banihashem SK, Kerman NT, Noroozi O, Moon J, Drachsler H (2024) Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, *21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

Chiang C-H, Chen W-C, Kuan C-Y, Yang C, Lee H (2024) *Large Language Model as an Assignment Evaluator: Insights, Feedback, and Challenges in a 1000+ Student Course*. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2489–2513. https://doi.org/10.18653/v1/2024.emnlp-main.146

Dancey CP, Reidy J (2017) *Statistics without maths for psychology* (Seventh Edition). Pearson.

DHBW Heidenheim (2025) *Bewertungsformular-PA-BA-SA-Technik*. https://www.heidenheim.dhbw.de/service-einrichtungen/dokumente-downloads/informatik

Flodén J (2025) Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, *51*(1), 201–224. https://doi.org/10.1002/berj.4069

Misgna H, On B-W, Lee I, Choi GS (2024) A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, *58*(2), 36. https://doi.org/10.1007/s10462-024-11017-5

Seßler K, Bewersdorff A, Nerdel C, Kasneci E (2025) *Towards Adaptive Feedback with AI: Comparing the Feedback Quality of LLMs and Teachers on Experimentation Protocols* (No. arXiv:2502.12842). arXiv. https://doi.org/10.48550/arXiv.2502.12842

Smerdon D (2024) AI in essay-based assessment: Student adoption, usage, and performance. *Computers and Education: Artificial Intelligence*, *7*, 100288. https://doi.org/10.1016/j.caeai.2024.100288

Tithi SD, Ramesh AK, DiMarco C, Tian X, Alam N, Fazeli K, Barnes T (2025) *The Promise and Limits of LLMs in Constructing Proofs and Hints for Logic Problems in Intelligent Tutoring Systems* (No. arXiv:2505.04736). arXiv. https://doi.org/10.48550/arXiv.2505.04736

Usher M (2025) Generative AI vs. instructor vs. peer assessments: A comparison of grading and feedback in higher education. *Assessment & Evaluation in Higher Education*, 1–16. https://doi.org/10.1080/02602938.2025.2487495

Wei Y, Pearl D, Beckman M, Passonneau RJ (2025) *Concept-based Rubrics Improve LLM Formative Assessment and Data Synthesis* (No. arXiv:2504.03877). arXiv. https://doi.org/10.48550/arXiv.2504.03877

Wetzler EL, Cassidy KS, Jones MJ, Frazier CR, Korbut NA, Sims CM, Bowen SS, Wood M. (2025) Grading the Graders: Comparing Generative AI and Human Assessment in Essay Evaluation. *Teaching of Psychology*, *52*(3), 298–304. https://doi.org/10.1177/00986283241282696

Xie W, Niu J, Xue CJ, Guan N (2024) *Grade Like a Human: Rethinking Automated Assessment with Large Language Models* (No. arXiv:2405.19694). arXiv. https://doi.org/10.48550/arXiv.2405.19694