Hybrid AI-Human Systems for Fair and Efficient Essay Grading: Research Topics Report

Jonas Hentschel

Introduction

Automated essay evaluation has emerged as a promising solution to address the scalability challenges of traditional manual grading, which suffers from grader fatigue and consistency issues Wilcox et al., 2016. While current AI-assisted systems demonstrate a good potential on objective evaluation criteria Yavuz et al., 2024, significant challenges remain in assessing subjective writing aspects and ensuring transparency Zhao et al., 2024. These limitations are surrounded by concerns about algorithmic bias and fairness in large language models, which are the primary AI system used by AI-assisted essay grading systems Navigli et al., 2023.

This research topics report examines hybrid AI-human systems as a potential framework for combining the efficiency of AI-assisted tools like LLMs and human insight for grading. The analysis focuses on three key aspects: (1) current grading methodologies and their limitations, (2) advancements in AI-assisted evaluation systems, and (3) approaches for maintaining fairness and transparency. The review will primarily focus how hybrid systems could reduce grader fatigue while maintaining evaluation quality Zupanc and Bosnic, 2017.

The findings will inform a master's thesis project addressing grading consistency in large-volume essay assessment. The following sections detail the methodology used for this review before presenting the thematic analysis of current research.

1 Research Preparation

1.1 Research Objective and Scope

The primary objective of this review is to investigate the potential of hybrid AI-human systems for improving and assisting the essay evaluation process. Building on previous work in automated essay scoring Lim et al., 2021 and recent advances in large language models Yavuz et al., 2024, this study focuses specifically on addressing two persistent challenges in educational assessment: grader fatigue and grading consistency.

The scope of this review includes three key dimensions of essay evaluation systems. Firstly, it aims identify specific pain points in manual assessment. Second, it analyzes current AI-assisted approaches, with the goal of assessing their reliability and potential application areas in the grading process. Third, it investigates hybrid systems that combine automated scoring with human oversight, which recent studies suggest may offer optimal balance between efficiency and quality.

1.2 Key Research Domains

The review organizes its investigation across five interconnected research domains that collectively address the essay evaluation pipeline. The foundation begins with current grading methodologies, while effective for evaluating nuanced writing qualities, suffers from scalability limitations.

Language model performance makes up the second domain, focusing on how systems like those described in Ye and Manoharan, 2021 achieve different levels of success across criteria. The third domain examines prompt engineering techniques that can enhance model reliability, while the fourth addresses critical bias mitigation strategies Navigli et al., 2023.

Finally, the hybrid systems domain combines these components, investigating frameworks where AI and human graders build a collaborative process to combat the negative effects of human based grading and fully automated processes. This structure ensures coverage of the topic while keeping the focus on the goal of developing a more effective hybrid intelligence framework.

1.3 Research Questions

The analysis is guided by a central research question examining how hybrid AI-human systems can improve grading consistency while reducing evaluator fatigue. This overarching question breaks down into specific sub-questions:

Research Domain	**Key Research Questions**
AI-driven essay grading methodologies	How have AI-based grading systems evolved, and what are their current capabilities and limitations?
Fairness and bias in automated grading	What mechanisms exist to detect and mitigate bias in AI-assisted grading?
Human-AI collaboration in grading	How can hybrid systems optimize the strengths of both AI and human evaluators?
Transparency and interpretability	What methods ensure the transparency of AI-generated grades for educators and students?
Scalability and real-world deployment	What challenges exist in deploying AI-assisted grading systems at scale in different educational contexts?

Table 1: Mapping of Research Domains to Research Questions

2 Search Methodology

2.1 Database Selection and Rationale

The literature search was conducted across three major academic databases selected for their comprehensive coverage of computer science and education research: Web of Science, IEEE Xplore, and ScienceDirect.

2.2 Search Queries

A structured search protocol was used, this included controlled vocabulary and Boolean operators to ensure reproducibility. The search strategy combined four conceptual clusters:

General Assessment Approaches:

```
("essay evaluation" OR "writing assessment" OR "essay scoring")
AND ("methods" OR "techniques" OR "approaches")
```

Automated Systems:

```
("automated grading" OR "AI evaluation" OR "computer scoring")
AND ("education" OR "writing" OR "essays")
```

Limitations and Challenges:

```
("grading limitations" OR "evaluation challenges" OR "assessment problems" OR "scoring
AND ("writing" OR "essay" OR "composition")
```

Comparative Studies:

```
("human grading" OR "teacher evaluation" OR "manual assessment")
AND ("computer grading" OR "AI scoring" OR "automated evaluation")
```

Quality and Reliability:

```
("grading consistency" OR "scoring reliability" OR "evaluation validity" OR "assessmen
```

These clusters were used in searches on the previously mentioned databases. The search was conducted iteratively, with initial results informing refinement of search terms to balance recall and precision. In more detail this means that after assessing initial results, in terms of the quantity results and the quality of the results by scanning through titles, the search terms were refined slightly to improve the quality of results.

2.3 Inclusion and Exclusion Criteria

The study employed the following selection criteria to ensure methodological quality while maintaining relevance to the research objectives:

- Peer-reviewed journal articles or conference proceedings
- Empirical studies with defined methodology

Exclusion criteria eliminated:

- Theoretical papers without applied results
- Studies limited to non-English language assessment
- Pre-print publications without peer review

These criteria were applied in two screening phases: first at the title/abstract level, then through full-text review.

3 Literature Screening Process

3.1 Screening Methodology

Figure 1, shows a flow chart of how many papers were selected where. The initial search step refers to the total number of papers that were identified as a result of the search queries in the selected databases. For better visual clarity the number of papers was split into the five search categories to better illustrate the amount of literature available for each domain. The papers that were found as a result of the search queries were then screened on a title basis to quickly assess papers that fit the research. This step was accompanied by occasionally reading the abstract of promising papers to further confirm their relevance to the goal of this report. This step yielded a total number of 34 promising papers. These papers were then read in their entirety as mentioned in step section 3.3.

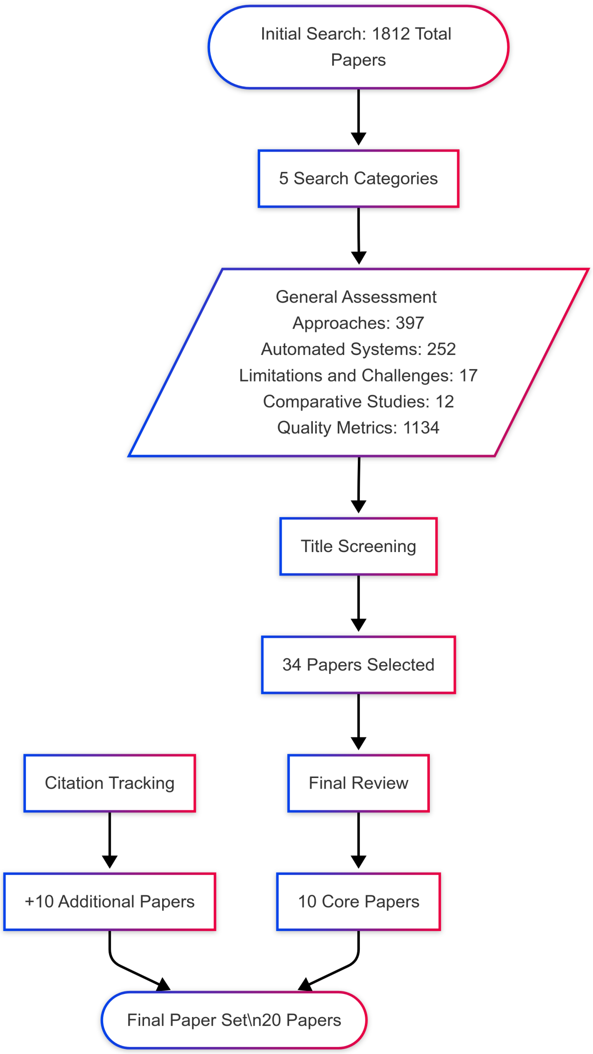

Figure 1: Screening flowchart documenting the selection process from initial search to final included studies. The diagram illustrates the application of inclusion/exclusion criteria at each stage.

3.2 Initial Screening Phase

The first screening phase applied the predefined inclusion criteria and scanning at the title and abstract level. This phase identified 34 publications, as potentially relevant.

3.3 Full-Text Review

The 34 selected publications underwent full-text analysis using refined criteria:

- Methodological rigor (study design, sample size)
- Technical validity (evaluation metrics, statistical analysis)
- Practical significance for this research (educational applicability)

This stage excluded 24 studies. Common reasons included insufficient reporting of evaluation metrics or studies not being fully relevant for the goal of this report.

3.4 Citation Tracking and Snowballing

To ensure comprehensive coverage, citation tracking identified 10 additional relevant studies. This snowballing technique helped mitigate potential database coverage limitations , particularly for recent LLM-based approaches not yet indexed in major databases.

4 Detailed Literature Analysis

4.1 Current Essay Grading Methods and Their Limitations

4.1.1 Evolution of Traditional Grading Methodologies

Traditional essay grading has evolved through several methodological approaches over the past five decades. Early standardized rubric-based assessment emerged in the 1970s, with holistic scoring predominating until the 1990s, when analytic rubrics gained popularity for their ability to provide specific feedback across multiple dimensions Jonsson and Svingby, 2007. Contemporary human grading typically employs multi-trait analytic rubrics that evaluate essays across 4-6 distinct dimensions including content, organization, language use, and mechanics (Baker, 2014; Weigle, 2002).

Human evaluation remains the standard for assessing complex writing dimensions. Experienced graders excel at evaluating nuanced subjective qualities such as argumentation depth, creative expression, and logical structure. Research by Eckes, 2008 demonstrates that expert human raters achieve strong reliability in evaluating argument quality and rhetorical effectiveness.

4.1.2 Documented Limitations of Manual Assessment

Despite its strengths, manual grading faces several well-documented challenges:

Cognitive Load and Grader Fatigue: Studies have shown significant declines in grading consistency when evaluators assess a large number of essays in a short period Klein and El, 2003. This fatigue effect manifests as both decreased attention to scoring criteria and increased variation in standards application Wilcox et al., 2016.

Consistency Challenges: Even with standardized rubrics, inter-rater reliability in essay assessment rarely exceeds 0.7-0.8 Jonsson and Svingby, 2007. Research by Tsai, 2012 documents how this reliability further decreases when evaluating subjective dimensions like creativity or argument quality. These subjective assessment criteria have been shown to pose a significant challenge when trying to maintain inter-rater consistency.

Time and Resource Intensity: The grading process of college level essays is a time consuming process as it needs to be thorough and detailed. This time commitment creates significant scalability challenges Zupanc and Bosnic, 2017 in large educational settings, where courses may generate hundreds of submissions requiring evaluation within constrained time frames. Leading to potential drop-offs in grading consistency and feedback quality.

4.2 Advancements in AI-Driven Essay Evaluation Systems

4.2.1 Historical Development of Automated Essay Scoring (AES)

The evolution of AI-driven essay evaluation spans nearly six decades, progressing through distinct technological paradigms. Early systems from the 1960s and 1970s, like Project Essay Grade (PEG), relied primarily on surface-level proxy measures such as essay length, word length, and punctuation Page, 1966. These systems demonstrated moderate correlations with human scores but failed to assess meaning or content quality.

The 1990s saw the emergence of the second generation of AES systems, exemplified by programs like E-rater and Intelligent Essay Assessor (IEA), which incorporated more sophisticated natural language processing techniques. E-rater employed syntactic variety analysis and discourse structure identification Burstein et al., 1998, while IEA introduced latent semantic analysis to evaluate content and vocabulary usage.

The third generation of AES incorporated machine learning approaches that could identify patterns from human-scored essays. Systems like Bayesian Essay Test Scoring System (BETSY) and LightSIDE Hussein et al., 2019 used supervised learning algorithms trained on large corpora of pre-scored essays to improve accuracy across diverse writing tasks . These systems achieved correlations exceeding 0.8 with human raters on standardized assessments but continued to struggle with evaluating creativity, argument quality, and factual accuracy Lim et al., 2021

The current generation of AES (2016-present) employs deep learning approaches and transformer-based language models. These systems leverage transfer learning from large datasets to develop more nuanced understanding of text. Neural network approaches like the two-stage deep neural network (TDNN) proposed by Jin et al., 2018 can model complex relationships between essay features and quality indicators. Most recently, large language models (LLMs) like GPT-4 and Claude have demonstrated promising capabilities in essay assessment even without task-specific training Yavuz et al., 2024.

4.2.2 Key Technological Approaches

Modern AI-driven essay evaluation has evolved through several technological approaches. Feature engineering models extract linguistic elements like syntactic complexity and vocabulary diversity, using statistical algorithms to correlate these features with quality scores. Systems like SAGE (Semantic Automated Grader for Essays) Lim et al., 2021 integrate numerous features across linguistic dimensions, providing consistent evaluation of mechanical aspects, though they require substantial expertise to develop and often struggle to adapt across different writing contexts. End-to-end neural networks have moved beyond manual feature engineering by learning directly from text. The TDNN architecture processes essays through hierarchical convolutional and recurrent neural networks to capture both local and global text characteristics Jin et al., 2018. The emergence of Large Language Models (LLMs) represents the latest advancement in automated essay scoring. Models like GPT-4, Claude, and LLaMA can evaluate writing without task-specific training by leveraging their general language understanding capabilities combined with effective prompting techniques Yavuz et al., 2024. Hybrid systems combine multiple approaches or integrate AI with human judgment to create more comprehensive evaluation tools. SAGE employs both feature engineering and semantic analysis to evaluate both mechanical and content-related aspects of writing Lim et al., 2021.

4.3 Challenges and Limitations of AI Models

4.3.1 Explainability and Transparency

The lack of transparency in AI evaluation systems represents a fundamental challenge for educational applications. As models have grown more complex, their decision-making processes have become increasingly opaque, creating what researchers term the black box problem in educational assessment Zhao et al., 2024.

Early rule-based systems offered straightforward explanations for their evaluations, but modern deep learning approaches sacrifice transparency for performance. Research by Kabra et al., 2022 demonstrates that educators and students express significant distrust toward assessment systems that cannot articulate clear reasoning for assigned scores. This transparency gap undermines both the pedagogical value of feedback and stakeholder confidence in AI-driven evaluation.

Recent research has explored several approaches to enhance explainability:

Attention Visualization techniques highlight text regions that significantly influence the model's scoring decisions. The neural network framework developed by Alikaniotis et al., 2016 visualizes these discriminative regions, providing insights into how models prioritize different textual elements. While helpful for identifying what text the model focuses on, these approaches often fail to explain why certain parts of the text receive attention.

Natural Language Explanations represent the most promising direction for enhancing transparency. Recent work with LLMs demonstrates their ability to generate human-readable explanations for scoring decisions. However, these explanations sometimes rationalize decisions post-hoc rather than reflecting the actual reasoning process, creating the illusion of transparency without genuine explainability Yavuz et al., 2024.

The transparency challenge extends beyond technical considerations to ethical concerns about algorithmic accountability in high-stakes assessment contexts. Meaningful transparency requires not only explanations of how scores are determined but also disclosure of system limitations and potential biases.

4.3.2 Bias and Fairness

Algorithmic bias in AI evaluation systems presents a significant ethical challenge. Multiple studies have documented systematic disparities in how these systems assess writing from different demographic groups:

Linguistic Bias affects non-native English speakers and writers using non-standard dialects. Automated scoring systems consistently assigned lower scores to grammatically correct essays written in non-mainstream English varieties. Similarly, studies show that essays employing culturally distinct rhetorical structures receive lower evaluations despite communicating effectively Braun and Clarke, 2021.

Content Biases appear when systems evaluate topics differently based on their representation in training data. Essays addressing minority experiences or perspectives from non-dominant cultural traditions receive lower scores due to their statistical dissimilarity from training examples. These biases raise significant concerns about reinforcing existing educational inequities through automated assessment Navigli et al., 2023.

Mitigation Strategies take several forms. Adversarial training approaches aim to create models that cannot distinguish between demographic groups, preventing them from incorporating demographic signals into scoring decisions. Data augmentation techniques expand training sets with diverse writing samples to create more inclusive evaluation standards. Most promisingly, human-in-the-loop approaches incorporate expert oversight to identify and correct systematic biases in AI evaluations.

4.4 Hybrid Evaluation Approaches

The integration of AI into traditional grading methods seems like a promising solution to address the limitations of both methods. Hybrid models can combine the efficiency of AI-driven scoring with the subjective judgment of human graders, particularly for creative aspects of writing Kabra et al., 2022. Additionally, hybrid systems can leverage human feedback to recursively improve AI models, enhancing both their accuracy and fairness. Recent studies, such as Escalante et al., 2023, have explored the efficacy of AI-generated feedback in educational contexts, finding that a blended approach—combining AI and human feedback—can maintain learning outcomes while improving efficiency and scalability. This approach is particularly effective in English as a New Language (ENL) contexts, where AI tools like ChatGPT have shown promise in providing consistent and timely feedback without compromising educational quality.

4.5 Note on Literature Limitations

It is important to acknowledge that the research literature on AI-assisted essay evaluation, particularly for hybrid systems, remains limited in several respects. The most sophisticated AI approaches using large language models have emerged only in the past 2-3 years, resulting in a relatively small body of peer-reviewed research examining their applications in educational assessment. Many promising approaches are still in early stages of development and testing. Existing studies often rely on small sample sizes or limited implementation contexts, constraining the generalizability of findings. Additionally, many studies evaluate systems using pre-existing datasets rather than in authentic classroom environments, potentially overlooking important implementation challenges.Commercial automated essay scoring systems often operate as proprietary "black boxes," with limited peer-reviewed publication of their internal methodologies or performance characteristics. This creates challenges for comprehensive evaluation of state-of-the-art approaches.

Despite these limitations, the available literature provides sufficient foundation for identifying key challenges and promising directions for hybrid AI-human assessment systems, while highlighting areas requiring further research.

5 Thematic Analysis

5.1 Recurring Research Themes

Three dominant themes emerge from the literature on automated essay evaluation. First, the *transparency gap* in AI systems persists as a major adoption barrier. While automated graders achieve high scoring accuracy, their decision-making processes remain opaque to educators, particularly for subjective scoring criteria Zhao et al., 2024.

Second, the *subjectivity challenge* reveals a fundamental limitation of current systems. Automated evaluators match human performance on grammar and structure (ICC ¿ 0.80) but struggle with assessing creativity and argument quality (ICC ¡ 0.50) Jin et al., 2018. This performance gap suggests hybrid systems must strategically allocate evaluation tasks.

Third, studies consistently report a *scaling paradox* where systems optimized for specific domains perform poorly when generalized. The TDNN architecture Jin et al., 2018 demonstrates this challenge, achieving strong results on trained prompts but requiring significant adaptation for new evaluation contexts.

5.2 Methodological Trends

The reviewed studies employ a range of methodological approaches to develop and evaluate AI-driven essay evaluation systems:

- **Quantitative Analysis:** Many studies, such as Fazal et al., 2013 and Jin et al., 2018, use quantitative methods to measure the performance of AI models, including metrics like intraclass correlation (ICC) and accuracy.
- **Qualitative Analysis:** Studies like Braun and Clarke, 2021 employ qualitative methods to explore the perceptions and preferences of users, such as students and teachers, regarding AI-generated feedback.
- **Experimental Designs:** Several papers, including Escalante et al., 2023, use experimental designs to compare the efficacy of AI-generated feedback with human feedback.
- **Case Studies:** Case studies, as seen in Lim et al., 2021, evaluate the practical application of AI-driven systems in real-world educational settings.

5.3 Identified Research Gaps

Four key areas require further investigation:

- Actual time savings in hybrid implementations
- Educator training requirements for effective system use
- Methods for maintaining consistency in continuous learning systems

These gaps collectively suggest that while hybrid systems show theoretical promise, their practical implementation requires more rigorous study.

6 Proposed Research Topic

6.1 Research Objectives and Questions

The final master project will focus specifically on addressing two critical challenges in essay evaluation: grader fatigue and inconsistency in assessment. The primary research question guiding this work is: **How can advances in hybrid AI-human systems and LLMs improve grading consistency and reduce grading fatigue of essay evaluation systems?**

Traditional manual grading methods, while effective for evaluating nuanced aspects of writing, often lead to fatigue when educators assess large volumes of essays. This fatigue directly contributes to inconsistency in grading, as evaluators' judgment can vary significantly throughout the grading process. The project aims to develop and test a hybrid AI-human framework that specifically targets these twin challenges, providing empirical evidence on how such systems can maintain consistent evaluation standards while alleviating the cognitive burden on human graders.

6.1.1 Addressing Key Research Questions

The literature review provides substantial insights that directly address each of the key research questions identified in Table 1. These findings inform both the conceptual framework and methodological approach of the proposed research.

AI-driven essay grading methodologies: The evolution of automated grading systems has progressed through distinct technological generations, from early rule-based approaches to the latest transformer-based language models. Contemporary systems like GPT-4 and Claude demonstrate promising capabilities without task-specific training Yavuz et al., 2024. However, a consistent performance gap persists between objective criteria (grammar, mechanics) and subjective dimensions (creativity, argumentation). This differential performance suggests that hybrid systems should leverage AI for objective dimensions while preserving human judgment for subjective assessment.

Fairness and bias in automated grading: Research has documented systematic biases affecting non-native English speakers and writers using non-standard dialects. Essays employing culturally distinct rhetorical structures often receive lower evaluations despite communicating effectively. Promising mitigation strategies include adversarial training approaches, data augmentation with diverse writing samples, and human-in-the-loop oversight to identify and correct systematic biases Navigli et al., 2023.

Human-AI collaboration in grading: Studies indicate that optimal collaboration leverages the complementary strengths of AI and human evaluators rather than treating AI as a replacement for human judgment. Effective collaboration requires clear task allocation based on performance characteristics and transparent communication of AI decisions to human graders. Additionally, research by Escalante et al., 2023 demonstrates that blending AI and human feedback can maintain learning outcomes while improving efficiency. These findings suggest that the hybrid system should establish a clear division of responsibilities and provide mechanisms for human graders to understand and override automated assessments when necessary.

Transparency and interpretability: The "black box" problem in AI evaluation presents a significant challenge Zhao et al., 2024. Educators and students express distrust toward systems that cannot articulate clear reasoning for scores. Research into attention visualization techniques and natural language explanations offers promising approaches for enhancing transparency. The hybrid system must incorporate these techniques to generate explanations that are both technically accurate and educationally valuable, ensuring that automated assessments contribute to the learning process rather than merely accelerating grading.

Scalability and real-world deployment: Implementing hybrid systems in educational contexts presents several practical challenges, including integration with existing workflows, educator training requirements, and maintaining consistency during continuous learning Lim et al., 2021. The literature reveals a "scaling paradox" where systems optimized for specific domains perform poorly when generalized, suggesting the need for adaptable architectures. To address these challenges, the hybrid system must balance innovation with practical constraints, leveraging existing language model capabilities through strategic integration rather than developing entirely new AI technology.

These findings collectively inform a framework that addresses the central research question by suggesting specific ways in which hybrid AI-human systems can improve grading consistency while reducing evaluator fatigue. The proposed research will build upon these insights to develop and empirically evaluate a practical implementation within the scope of a master's thesis project.

6.2 Conceptual Framework for the Hybrid AI-Human System

Based on the literature review findings, the proposed hybrid system will incorporate several key elements designed to address the identified limitations of current grading approaches. The research findings provide a foundation for developing specific system components that address the challenges identified throughout the literature review.

6.2.1 Filtering Criteria from Research Findings

The literature analysis revealed several crucial considerations that must guide the development of an effective hybrid system. Research by Zhao et al., 2024 and Kabra et al., 2022 emphasizes that transparency in AI scoring decisions is essential for user trust and pedagogical value. Without clear explanations of how assessments are generated, educators and students alike demonstrate significant skepticism toward automated evaluation.

The consistent performance gap between objective and subjective assessment criteria poses another significant challenge. While AI systems demonstrate strong performance on objective criteria like grammar and structure, they struggle considerably with subjective elements such as creativity and argumentation quality Jin et al., 2018. This discrepancy suggests that any effective hybrid system must strategically allocate evaluation tasks according to the relative strengths of human and automated assessment.

Studies focusing on cognitive load, particularly those by Klein and El, 2003 and Wilcox et al., 2016, document how grader fatigue impacts evaluation quality over extended grading sessions. These findings indicate the need for intelligent distribution of grading tasks and cognitive load management within the system design. Additionally, the research on algorithmic bias by Navigli et al., 2023 and Braun and Clarke, 2021 highlights the importance of addressing potential discrimination, particularly for non-native English speakers and students employing diverse rhetorical structures.

6.2.2 Proposed System Components

Filtering the research insights into design priorities, the proposed hybrid system will consist of several core components that address the identified challenges.

The system architecture will be organized around three fundamental principles derived from the literature. First, the framework will implement a strategic division of labor between AI and human graders based on evaluation dimensions. This approach acknowledges the different performance of automated systems across assessment criteria, as documented by Jin et al., 2018. The system will assign initial evaluation of more objective elements (such as grammar, structure, and mechanics) to the AI component while preserving human judgment for subjective dimensions that require nuanced understanding of creativity and argumentation quality. This division aims to maximize efficiency while maintaining assessment integrity.

Second, the framework will incorporate transparency mechanisms to address the black box problem identified in Zhao et al., 2024 and Kabra et al., 2022. The system will provide clear explanations for AI-generated assessments, helping educators understand the basis for automated scoring and allowing them to make informed decisions about when to accept or override these evaluations. This transparency layer serves not only to build user trust but also to enhance the educational value of the assessment process by generating meaningful feedback that students can use to improve their writing. This may be further enhanced by comparing the assessment of an automated evaluation by the AI with the human grader evaluation. The user may then engage in a conversation about potential differences in scoring and prompt the AI to reason about decisions.

Third, the system will include fatigue management features that directly address the consistency challenges documented by Klein and El, 2003 and Wilcox et al., 2016. By intelligently structuring the assessment workflow and providing consistency checks during extended grading sessions, the framework aims to mitigate the cognitive load that leads to evaluation drift.

The scope of the project necessitates certain practical limitations. Rather than attempting to develop novel AI models, the framework will leverage existing language model capabilities through strategic prompting and integration techniques. The focus will remain on creating an effective interface between automated and human assessment rather than advancing the underlying AI technology itself. Similarly, while bias mitigation represents an important concern, the project will address this primarily through thoughtful system design and awareness rather than attempting comprehensive algorithmic solutions that would exceed the project's scope.

This more focused approach aligns with the primary research objective of addressing grader fatigue and inconsistency while creating a framework that could be expanded in future research. The resulting system will serve as a testbed for evaluating how hybrid assessment approaches affect these specific challenges in educational contexts. And the potential advances that are necessary to build trust in automated systems.

6.3 Methodology

The research methodology centers on a comparative experimental design that directly measures consistency and fatigue across different grading approaches. First, a set of student essays will be collected, representing topics familiar to professors/teachers participating in the study. These essays will then be evaluated under three distinct conditions:

traditional manual grading, fully automated AI grading, and the proposed hybrid AI-human system.

To measure consistency, multiple graders will assess each essay, and inter-rater reliability will be quantified using Cohen's Kappa or Intraclass Correlation Coefficient (ICC). These statistical measures will provide concrete evidence of how AI assistance affects grading consistency compared to purely manual approaches. The hypothesis is that AI-assisted grading will demonstrate higher inter-rater reliability, indicating more consistent evaluation across different graders.

Fatigue assessment will form the second critical component of the methodology. Graders will complete standardized workload and fatigue assessments before and after grading sessions. In addition, a time series analysis of the scoring patterns will be performed to identify fatigue-related changes in the evaluation standards that typically occur during extended grading sessions. The study will measure concrete metrics such as time to grade, grader satisfaction, and self-reported cognitive load to quantify fatigue reduction.

The methodology will include specific evaluation of each system component. Task allocation effectiveness will be measured through comparative analysis of time spent on different grading dimensions and educator satisfaction ratings. Interface usability will be assessed through standardized usability metrics and qualitative feedback on explanation quality and accessibility. Fatigue reduction will be evaluated using NASA Task Load Index measurements and statistical analysis of grading consistency over time.

Based on these measurements, the hybrid AI-human framework may be refined to maximize both consistency and fatigue reduction.

6.4 Expected Contributions

This research will make focused contributions to educational assessment by providing empirical evidence on how hybrid AI-human systems specifically impact grading consistency and fatigue. Unlike broader studies on automated essay scoring, this project will deliver several concrete advances in the field of educational assessment technology.

First, the project will develop a validated task allocation framework determining optimal division of responsibilities between AI and human graders based on empirical performance data. This framework will identify specific essay dimensions where automated evaluation performs reliably and those where human judgment remains essential, creating an efficient division of labor that maximizes both time savings and assessment quality.

Second, the research will establish quantifiable metrics on cognitive load reduction and their correlation with grading consistency, providing evidence-based guidelines for managing grader fatigue. By documenting how different system configurations affect educator cognitive load and evaluation consistency, the study will generate practical recommendations for educational institutions implementing similar systems.

Third, the project will produce a transparent explanation model that demonstrably improves educator trust and student understanding of assessment criteria. This model will translate complex AI decision processes into clear, actionable feedback that supports the pedagogical goals of writing assessment.

Finally, the project will create a generalizable architecture for hybrid grading systems that can be adapted to different educational contexts and assessment needs. This architecture will provide a foundation for future developments in AI-assisted educational assessment, allowing for customization to specific institutional requirements and grading practices.

The most significant expected contribution is the development of a validated framework that demonstrably reduces grader fatigue while maintaining or improving evaluation consistency. This framework will include specific implementations detailing how to effectively balance AI and human input during different stages of the essay evaluation process. By focusing exclusively on these two critical challenges, the research aims to provide a practical solution that addresses real pain points faced by educators when evaluating large volumes of written work.

References

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks [54th Annual Meeting of the Association-for-Computational-Linguistics (ACL), Berlin, GERMANY, AUG 07-12, 2016]. In K. Erk & N. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics, vol 1* (pp. 715–725).

Baker, N. L. (2014). "get it off my stack": Teachers' tools for grading papers [Feedback in Writing: Issues and Challenges]. *Assessing Writing*, *19*, 36–50. https://doi.org/https://doi.org/10.1016/j.asw.2013.11.005

Braun, V., & Clarke, V. (2021). One size fits all? what counts as quality practice in (reflexive) thematic analysis? *QUALITATIVE RESEARCH IN PSYCHOLOGY*, *18*(3, SI), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). Enriching automated essay scoring using discourse marking. *Discourse Relations and Discourse Markers.* https://aclanthology.org/W98-0303/

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Escalante, J., Pack, A., & Barrett, A. (2023). Ai-generated feedback on writing: Insights into efficacy and enl student preference. *INTERNATIONAL JOURNAL OF EDUCATIONAL TECHNOLOGY IN HIGHER EDUCATION*, *20*(1). https://doi.org/10.1186/s41239-023-00425-2

Fazal, A., Hussain, F. K., & Dillon, T. S. (2013). An innovative approach for automatically grading spelling in essays using rubric-based scoring. *JOURNAL OF COMPUTER AND SYSTEM SCIENCES*, *79*(7, SI), 1040–1056. https://doi.org/10.1016/j.jcss.2013.01.021

Hussein, M. A., Hassan, H. A., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*. https://doi.org/10.7717/peerj-cs.208

Jin, C., He, B., Hui, K., & Sun, L. (2018). Tdnn: A two-stage deep neural network for prompt-independent automated essay scoring [56th Annual Meeting of the Association-for-Computational-Linguistics (ACL), Melbourne, AUSTRALIA, JUL 15-20, 2018]. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th annual meeting of the association for computational linguistics (acl), vol 1* (pp. 1088–1097).

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/https://doi.org/10.1016/j.edurev.2007.05.002

Kabra, A., Bhatia, M., Singla, Y. K., Li, J. J., & Shah, R. R. (2022). Evaluation toolkit for robustness testing of automatic essay scoring systems [5th ACM India Joint International Conference on Data Science and Management of Data (CoDS-COMAD) / 9th ACM IKDD CODS Conference / 26th COMAD Conference, Bangalore, INDIA, JAN 08-10, 2022]. *PROCEEDINGS OF THE 5TH JOINT INTERNATIONAL CONFERENCE ON DATA SCIENCE & MANAGEMENT OF DATA, CODS COMAD 2022*, 90–99. https://doi.org/10.1145/3493700.3493765

Klein, J., & El, L. (2003). Impairment of teacher efficiency during extended sessions of test correction. *European Journal of Teacher Education*, *26*, 379–392. https://api.semanticscholar.org/CorpusID:145559783

Lim, C. T., Bong, C. H., Wong, W. S., & Lee, N. K. (2021). A comprehensive review of automated essay scoring (aes) research and development. *PERTANIKA JOURNAL OF SCIENCE AND TECHNOLOGY*, *29*(3), 1875–1900. https://doi.org/10.47836/pjst.29.3.27

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *ACM JOURNAL OF DATA AND INFORMATION QUALITY*, *15*(2). https://doi.org/10.1145/3597307

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238–243. Retrieved April 1, 2025, from http://www.jstor.org/stable/20371545

Tsai, M.-h. (2012). The consistency between human raters and an automated essay scoring system in grading high school students' english writing. *Action in Teacher Education*, *34*, 328–335. https://doi.org/10.1080/01626620.2012.717033

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Wilcox, K. C., Jeffery, J. V., & Gardner-Bixler, A. (2016). Writing to the common core: Teachers' responses to changes in standards and assessments for writing in elementary schools. *READING AND WRITING*, *29*(5, SI), 903–928. https://doi.org/10.1007/s11145-015-9588-6

Yavuz, F., Celik, O., & Celik, G. Y. (2024). Utilizing large language models for efl essay grading: An examination of reliability and validity in rubric-based assessments. *BRITISH JOURNAL OF EDUCATIONAL TECHNOLOGY*. https://doi.org/10.1111/bjet.13494

Ye, X., & Manoharan, S. (2021). Performance comparison of automated essay graders based on various language models [IEEE International Conference on Computing (ICOCO), ELECTR NETWORK, NOV 17-19, 2021]. *2021 IEEE INTERNATIONAL CONFERENCE ON COMPUTING (ICOCO)*, 152–157. https://doi.org/10.1109/ICOCO53166.2021.9673585

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM TRANSACTIONS ON INTELLIGENT SYSTEMS AND TECHNOLOGY*, *15*(2). https://doi.org/10.1145/3639372

Zupanc, K., & Bosnic, Z. (2017). Automated essay evaluation with semantic analysis. *KNOWLEDGE-BASED SYSTEMS*, *120*, 118–132. https://doi.org/10.1016/j.knosys.2017.01.006