Наукові перспективи
Видавнича група

НАУКА і ТЕХНІКА СЬОГОДНІ

№ 1(55)
2026

З Україною в серці!

UDC 004.42/.8/.91

https://doi.org/10.52058/2786-6025-2026-1(55)-1929-1945

Parfonov Yurii Eduardovych Candidate of Technical Sciences, Senior Researcher, Associate Professor of the Information Systems Department, Simon Kuznets Kharkiv National University of Economics, Kharkiv, https://orcid.org/0000-0003-3943-3201

AUTOMATING THE PRE-SUBMISSION REVIEW OF ACADEMIC REPORTS VIA THE INTEGRATION OF FORMAL RULES AND LARGE LANGUAGE MODELS

Abstract. Providing high-quality feedback on structured academic reports remains a significant challenge in higher education. Manual pre-submission review of these documents is time-consuming, while existing automated tools are limited in their ability to perform detailed structural and semantic analysis. Although Large Language Models (LLMs) offer advanced capabilities, their unconstrained use in academic assessment is often compromised by systematic leniency bias and scoring inconsistency.

To overcome these limitations, the Human-Aligned Rubric and Review System (HARRS) was developed. It features a hybrid architecture that combines deterministic rule-based logic with a constrained generative feedback layer.

The study involved analyzing the current limitations of Automated Writing Evaluation (AWE), designing and implementing the architecture, and conducting empirical validation.

The HARRS performance was evaluated against expert human judgment and student utility metrics. The results demonstrated that the rule-based component accurately detects structural and formatting compliance. The hybrid pipeline achieved a high correlation ($r = 0.98$) with human expert ranking. Unlike unconstrained Baseline LLMs, which showed significant scoring inflation, HARRS effectively mitigates evaluative bias.

This study proves that anchoring generative AI within a rigid, human-defined rubric can transform the model into a reliable augmented feedback engine. HARRS ensures that disciplinary standards remain governed by faculty expertise, maintaining a human-in-the-loop philosophy. The findings suggest that this hybrid approach provides a trustworthy mechanism for enhancing the self-regulation of higher education students.

Keywords: academic report, pre-submission review, unified rubric, semantic feedback, hybrid architecture, rule-based layer, LLM semantic layer.

Парфьонов Юрій Едуардович кандидат технічних наук, старший науковий співробітник, доцент кафедри інформаційних систем, Харківський національний економічний університет ім. С. Кузнеця, Харків, https://orcid.org/0000-0003-3943-3201

АВТОМАТИЗАЦІЯ ПОПЕРЕДНЬОГО АНАЛІЗУ АКАДЕМІЧНИХ ЗВІТІВ ШЛЯХОМ ІНТЕГРАЦІЇ ФОРМАЛЬНИХ ПРАВИЛ ТА ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Анотація. Надання високоякісного зворотнього зв'язку щодо структурованих академічних звітів залишається значним викликом у сфері вищої освіти. Попередня перевірка таких документів вимагає великих затрат часу, а наявні автоматизовані інструменти мають обмежені можливості з виконання структурного та семантичного аналізу. Хоча великі мовні моделі (ВММ) і мають розширені можливості, їх неконтрольоване використання для академічного оцінювання часто ускладняється систематичною упередженістю та непослідовністю.

Щоб подолати ці обмеження, був розроблений застосунок Human-Aligned Rubric and Review System (HARRS). Він має гібридну архітектуру, яка поєднує детерміновану логіку на основі правил та контрольований генеративний шар зворотного зв'язку.

Дослідження включало аналіз поточних обмежень автоматизованого оцінювання текстів, проєктування і реалізацію гібридної архітектури, та проведення емпіричної валідації. Ефективність HARRS оцінювалася за результатами судження експертів та показників корисності для студентів. Отримані результати продемонстрували, що компонент на основі правил точно виявляє відповідність звітів вимогам до структури та форматування. Гібридний конвеєр досяг високої кореляції (r = 0,98) з рейтингом людських експертів. На відміну від неконтрольованих базових ВВМ, які значно завищували оцінки, HARRS ефективно зменшує їх упередженість.

Це дослідження доводить, що обмеження генеративного ШІ рамками визначеної людиною рубрики, може перетворити ВММ на надійний механізм розширеного зворотного зв'язку. Дотримуючись філософії «людина в циклі», HARRS забезпечує, що навчальні стандарти залишаються під контролем викладацького складу. Результати дослідження свідчать, що гібридний підхід забезпечує надійний механізм для підвищення саморегуляції студентів вищих навчальних закладів.

Ключові слова: академічний звіт, попередній аналіз, уніфікована рубрика, семантичний зворотний зв'язок, гібридна архітектура, шар на основі правил, семантичний шар ВММ.

Problem statement. Higher education programs often rely on structured academic reports (e.g., lab, capstone, and project reports). These reports mandate specific requirements for content relevance, sectional coherence, and proper formatting. Therefore, formal rules are essential for academic assessment because they establish a structured framework for disciplinary standards.

However, the manual pre-submission review* of these reports is time-consuming and inconsistent, especially in large courses. Current automated tools primarily focus on plagiarism detection and surface-level grammar/style correction. They lack robust support for two critical elements of structured reports - section-level structural compliance and deep semantic feedback during the pre-submission review.

The advent of LLMs offers a new opportunity to bridge this gap. They can generate sophisticated, semantic feedback on report drafts, helping students refine their work before submitting it for final review. However, relying solely on LLMs may lead to inconsistency and bias in grading formal criteria.

This work introduced a hybrid approach combining rule-based checks for document structure and formatting with section-specific recommendations generated by LLMs. It features a modular design that distinguishes auditable rules from probabilistic LLM feedback, allowing for transparent triage and instructor oversight.

The research followed a cycle that consisted of analyzing existing AWE limitations, designing the hybrid architecture, implementing it through prompt engineering, and conducting empirical validation. This was justified by the need to engineer a practical solution to the problems of inconsistency and resource intensity inherent in manual pre-submission review.

The study presented a novel approach based on the following research questions:

RQ1: How accurately can rule-based checks detect structural and formatting compliance compared to expert assessment?

RQ2: To what extent does the hybrid architecture align with expert judgments in discriminative ability and overall scoring levels?

RQ3: Does the HARRS provide pedagogical utility while successfully mitigating systematic evaluative bias?

Recent studies analysis. Historically, the automation of writing assessment has been mainly influenced by two key areas: AWE and Automated Assignment Grading (AAG). AWE systems utilize Natural Language Processing (NLP) techniques to provide prompt, objective feedback on surface linguistic features, including grammar, syntax, and vocabulary. This encourages students to revise and improve their writing accuracy. [1, 2].

However, traditional AWE and AAG models face significant limitations when applied to the formative review of structured academic reports. These systems often face criticism for promoting a narrow view of writing by emphasizing only

"correctness". They struggle to evaluate higher-level thinking, content depth, and compliance with specific organizational elements [1-5]. Additionally, their assessment primarily relies on statistical or predefined linguistic features [4], but they do not effectively measure the content mastery and structural communication essential for technical reports [6].

Furthermore, existing research mainly focuses on general essays, leaving a clear gap in robust support for section-level compliance and the semantic requirements of structured report formats.

* - The paper uses the term 'Pre-Submission Review' to denote the formative quality check applied to assignment drafts before official grading, distinguishing it from the summative focus of Automated Essay Grading and the prior-knowledge focus of educational pre-assessment.

The recent integration of LLMs into automated grading systems has addressed some of AWE's semantic limitations and demonstrated the ability to provide richer, more contextual, and human-like feedback on content and coherence [1, 2]. Yet, that introduced a new set of risks. Research indicates that LLM-based evaluators often lack the consistency required for objective assessment. They tend to assign higher median scores than human evaluators and demonstrate low intra-rater reliability when evaluating the same content at different runs [7 - 9]. Such issues, worsened by how small prompt changes can significantly affect model outputs, undermine transparency and reliability needed for enforcing strict educational criteria [10, 11].

While traditional AWE and pure LLM approaches have limitations [4, 8], the field is moving toward hybrid models that combine automated scoring with human-defined criteria.

Several researchers have developed methods to utilize formal rules to govern or constrain automated evaluation systems [4, 12]. This typically involves using a rule-based engine to verify essential requirements, serving as a gatekeeper before conducting more intricate analysis. It guarantees that objective, non-negotiable criteria (e.g., file type, minimum length, presence of key sections) are assessed accurately, minimizing the unreliability of LLMs [4].

Others have explored using rubrics [13] and checklists [14] not only for final scoring, but also as input prompts to guide the feedback generation process of sophisticated models. By defining precise traits and criteria for each section of a report, these systems aim to align the computational model's evaluation (whether statistical or LLM-based) with the institutional grading scheme [13, 15].

However, many existing hybrid systems, including those that use rubrics, remain focused on summative scoring rather than providing targeted, formative feedback during the pre-submission revision phase [13, 16, 17]. Current research extensively describes the use of LLMs to generate feedback, but it is not necessarily focused on the pre-submission, compliance-oriented stage [16 - 18].

There is a lack of literature detailing modular hybrid systems for structured academic reports. Existing studies fall short in addressing the dual compliance challenge: verifying adherence to structural rules, such as figure captions, referencing styles, and section placement [14], and ensuring that section content aligns with its purpose.

In synthesis, the current landscape provides robust tools for grammar correction and sophisticated semantic dialogue. However, no single validated system effectively integrates a rule-based compliance engine, a reliable LLM-based analyzer for semantic feedback, and a focus on the Pre-Submission Review phase for structured Academic Reports.

This paper addresses this critical gap by presenting an integrated hybrid approach that automates the pre-submission review of academic reports, thereby enhancing consistency and capacity in higher education assessment.

Research objective. The study aims to show that anchoring AI within a human-defined rubric yields more accurate and consistent reviews than standard models. To achieve this, it develops a hybrid system that integrates automatic rule-checking with AI-driven feedback for academic reports. Ultimately, the goal is to provide a reliable tool that helps students improve their work while maintaining high academic standards.

Methodology. This section details the systematic approach used to develop and validate HARRS. This automated review tool integrates a rule-based layer (RBL) for formal scoring and an LLM semantic layer (LLM-SL) for formative feedback.

The study employed a mixed-methods research design within the framework of the Design Science Research paradigm. Statistical agreement between automated and human scoring was established via the quantitative RBL, while the practical value of the LLM-SL's qualitative critiques was confirmed through student survey data. This design is justified because the system serves a dual purpose: ensuring formal compliance and providing actionable pedagogical support.

The primary data source consisted of a collection of N = 500 university-level Computer Science lab reports, written in accordance with a standardized style guide. The whole corpus was used for initial RBL tuning and iterative prompt engineering for the LLM-SL.

A subset of n=50 reports (a Reference Review Set) was randomly selected from the corpus for expert annotation. Human experts reviewed these reports using a unified rubric to establish "the Ground Truth" for quantitative validation.

To select reports that conformed to a specific institutional framework and rubric, the work used purposive sampling. This approach ensured the system was tested against a standardized set of academic criteria, enabling the RBL to identify formal violations and the LLM-SL to assess semantic levels (0–4) as specified by the faculty.

Data collection and system processing followed a multi-stage pipeline:

1. Pre-processing.

A custom parsing pipeline extracted text from DOCX files and segmented reports into structural blocks based on rubric-defined section titles.

2. Hybrid Processing.

RBL identified formal violations using deterministic logic and produced a JSON payload that included statuses and a rule-based score. This payload was passed to LLM-SL, powered by a cost-effective API such as Groq, to generate formative feedback. HARRS then combined the RBL's structured output with the LLM-SL's natural-language critique into a unified Comprehensive Feedback Report. In this report, the formal percentage score was definitive, while the semantic feedback provided actionable guidance.

3. Utility Assessment.

Following the system's review, a post-task survey was administered to students to collect qualitative data on the helpfulness and actionability of the LLM-SL's recommendations.

The system's performance was analyzed through Quantitative Scoring Validation (Agreement with Experts) and Qualitative Feedback Analysis (Utility-Driven). The RBL's scoring accuracy on the Reference Review Set reports was assessed using Mean Absolute Error (MAE), Pearson's r, and Analysis of Variance (ANOVA). MAE measured how closely the system's percentage scores matched human "Ground Truth". Pearson's r was evaluated to determine whether the scores consistently reflected expert judgment. To confirm the RBL's reliability, an ANOVA was conducted to assess the extent to which the system's average scores differed from those of human experts.

The outcome of the LLM-SL was evaluated based on its primary objective of providing constructive, formative feedback. Post-task survey results were analyzed to measure student perceptions of the actionable recommendations. This analysis focused on the "helpfulness for revision" metric, validating the system's effectiveness as a pre-submission review tool rather than a mere grading automaton.

To protect student privacy, all reports were anonymized to remove Personally Identifiable Information before being ingested into the system. The study also included a bias audit protocol during the development phase. The system was tested on reports with varying non-academic characteristics (e.g., irregular length or formatting) to confirm the LLM-SL's feedback remained fair and strictly focused on rubric criteria.

System Architecture. HARRS is founded upon a Unified Rubric that dictates the requirements for all academic reports. The system's core innovation lies in the architectural division of labor, where the rubric is systematically partitioned to guide the operation of two distinct layers: the deterministic RBL and the semantic LLM-SL.

HARRS operates as a two-stage sequential pipeline where the output of the RBL serves as contextual input for the LLM-SL, grounding the qualitative review in verified compliance data.

The UML Component Diagram (Figure 1) shows the system's structural elements and their contractual interactions.

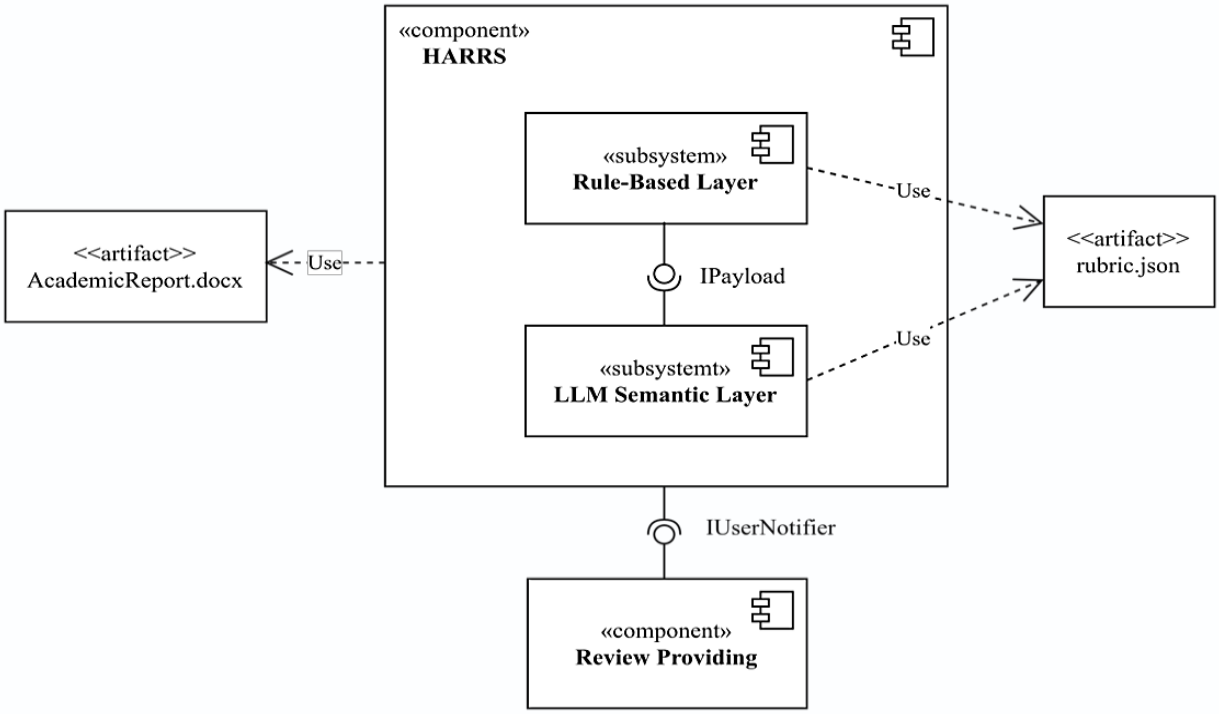

Fig. 1. HARRS architecture

The artifact AcademicReport.docx represents the input academic report and is connected via a Use dependency to the central HARRS component.

The main component, HARRS, comprises two key subsystems that represent the core processing pipeline: RBL and LLM-SL. The RBL subsystem performs deterministic checks and generates the structured context. The LLM-SL subsystem performs qualitative analysis using the context. RBL provides an interface, labeled IPayload. It is consumed by the LLM Semantic Layer, signifying that the structured JSON payload from the RBL is the critical contextual input for the LLM-SL.

Upon review completion, the HARRS component requires the IUserNotifier interface to publish the final integrated review content. This interface is supplied by the Review Providing component, which handles the delivery of the final review to the user interface or system of record.

An artifact rubric.json connects externally to the main component via a Use dependency to both RBL and LLM-SL subsystems, indicating they rely on this single configuration file for their operational criteria.

The rubric includes two complementary elements: Quantitative Scoring and Qualitative Feedback. The first one defines objective criteria for formal compliance (section headers, sentence count, etc.). It directly informs the development and scoring logic of the RBL. The second element defines semantic and discursive quality levels. It is transformed into explicit instructions and exemplary feedback styles for the LLM-SL's prompt engineering.

The RBL is the initial processing stage responsible for rapid, deterministic checks against the Quantitative Scoring element of the Unified Rubric. It performs structural analysis, validates mandatory elements (such as section presence and figure count), and calculates objective metrics.

Upon completion, the RBL serializes the results into a JSON payload with three fields: rule_based_score (a numerical compliance score), status (such as 'excellent' or 'insufficient'), and rule_based_comment (feedback on violations).

LLM-SL is responsible for conducting a nuanced qualitative review of the report. The layer prioritizes cost-effective, high-speed LLM API endpoints for development and iteration. It performs semantic analysis against the Qualitative Feedback element of the Unified Rubric, focusing on argumentation, coherence, subject matter accuracy, and the use of sophisticated language.

The core mechanism is Contextual Prompt Injection. The JSON payload generated by the RBL is embedded directly into the LLM-SL's system prompt. This ensures that the LLM's future semantic analysis and qualitative feedback align with the formal compliance status already established by the RBL. For example, if the RBL detects a missing section, the LLM-SL is explicitly instructed to incorporate this fact into its final qualitative summary.

The UML Activity Diagram (Figure 2) illustrates the strict, sequential workflow of the HARRS system, partitioned across the User System and HARRS.

The processing of an academic report using HARRS involves the following steps:

1. The process begins with the User System executing the Submit Academic Report action.

2. Control flows to HARRS, beginning the processing with Receive & Parse Report.

3. The Execute RBL activity is performed. This step utilizes the <<dataStore>> rubric.json to access the Quantitative Scoring criteria. The execution output is the JSON Payload artifact.

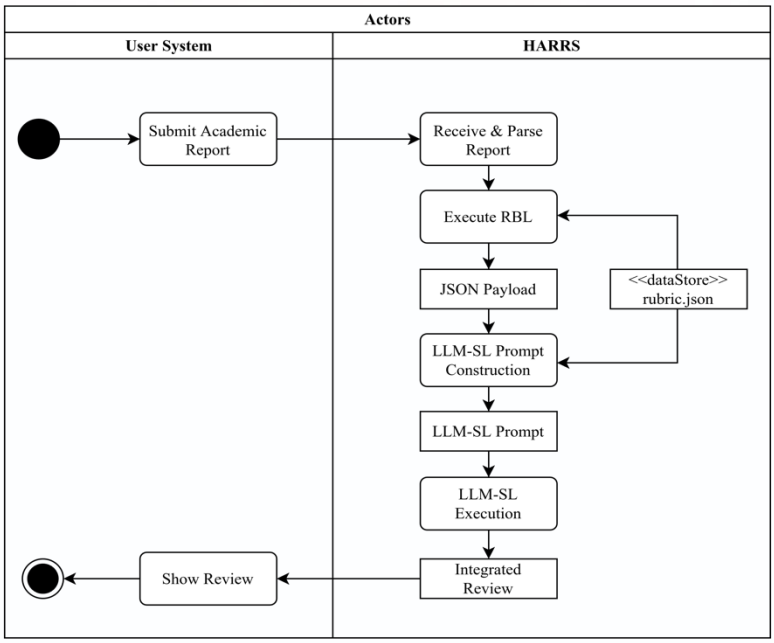

Fig. 2. UML Activity Diagram

4. The LLM-SL Prompt Construction activity is executed. This activity consumes the JSON Payload artifact and concurrently accesses the <<dataStore>> rubric.json to retrieve the Qualitative Feedback criteria, embedding both as context into the final LLM-SL Prompt artifact.

5. The LLM-SL Prompt is consumed by the LLM-SL Execution activity, which generates the Integrated Review artifact.

6. The final Review artifact flows back across the partition boundary to the User System, which performs the Show Review activity to complete the process.

Experimental Evaluation. This section details the design, core quantitative metrics, and the study's results, which validated the performance of HARRS against human expert reviewers and a conventional LLM baseline.

The study was structured as a statistical comparison of scoring consistency across three primary rating groups: HARRS, a general-purpose Baseline LLM with minimal instruction, and Human Reviewers. This design was crucial for quantifying HARRS alignment, specifically the unified RBL/LLM-SL architecture based on the Unified Rubric, relative to an unguided automated baseline and expert standards.

"The Ground Truth" was substantiated by Human Reviewers, consisting of two experienced instructors per sampled course. To secure reliable scoring, a two-step process was followed. Firstly, Rater Training outlined how to apply the Unified Rubric across all report dimensions (Formal, Structural, Semantic). During Calibration, reviewers independently scored five anchor reports, with their consensus score (mean or median) set as the standard.

The experimental protocol subjected all reports to a few distinct review processes. First, HARRS executed a Pre-Submission Review, producing a formal quantitative score and detailed semantic feedback via its RBL and LLM-SL layers. Simultaneously, Baseline LLM was evaluated with a minimal instruction prompt (e.g., "Score this report based on general academic standards") to determine an automated performance floor. Finally, Human Reviewers independently conducted a Pre-Submission Review of the entire dataset using the same Unified Rubric to establish «the Ground Truth» scores.

The core metric for measuring the accuracy of the automated scores was MAE. A lower MAE indicated greater proximity to expert judgment.

The main findings were visually summarized using a comparative bar chart that displays MAE for HARRS versus the Baseline LLM relative to the Human Consensus (Figure 3).

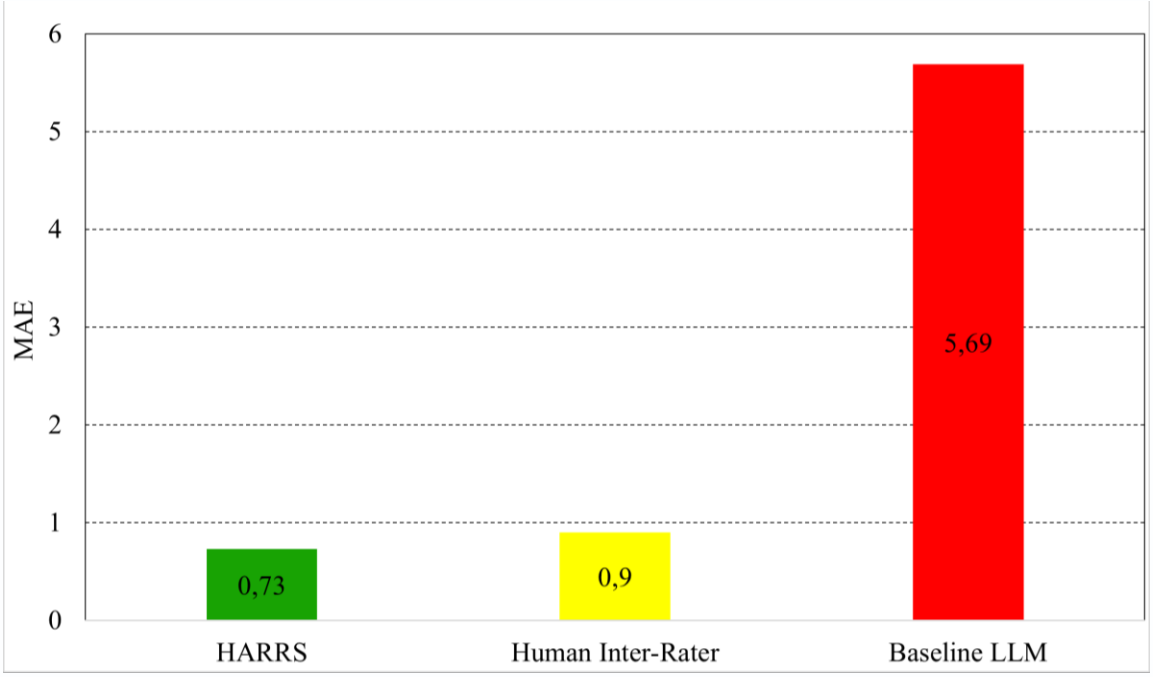

Fig. 3. MAE Comparison to Human Consensus

As shown in Figure 3, HARRS demonstrated a significantly lower error rate compared to Baseline LLM. It achieved an MAE of 0.73, representing more than an 87% reduction in error compared to the Baseline LLM's MAE of 5.69. Furthermore, the HARRS MAE (0.73) was found to be lower than the Human Inter-Rater MAE (0.9).

To assess the linear relationship between the HARRS scores and the Human Consensus scores, Pearson's Correlation Coefficient (***r***) was used. A corresponding scatter plot is presented in Figure 4.

The data points are distributed in close proximity to the line of ideal agreement (Y=X) throughout the 0–24 scoring range. The distance between the individual points and the diagonal remains consistently small, with no visible patterns of systematic deviation at specific score intervals.

These visual observations were consistent with the calculated Pearson's r coefficient of 0.98 for the HARRS system. For comparison, Pearson's r coefficient for the Baseline LLM was calculated at 0.46.

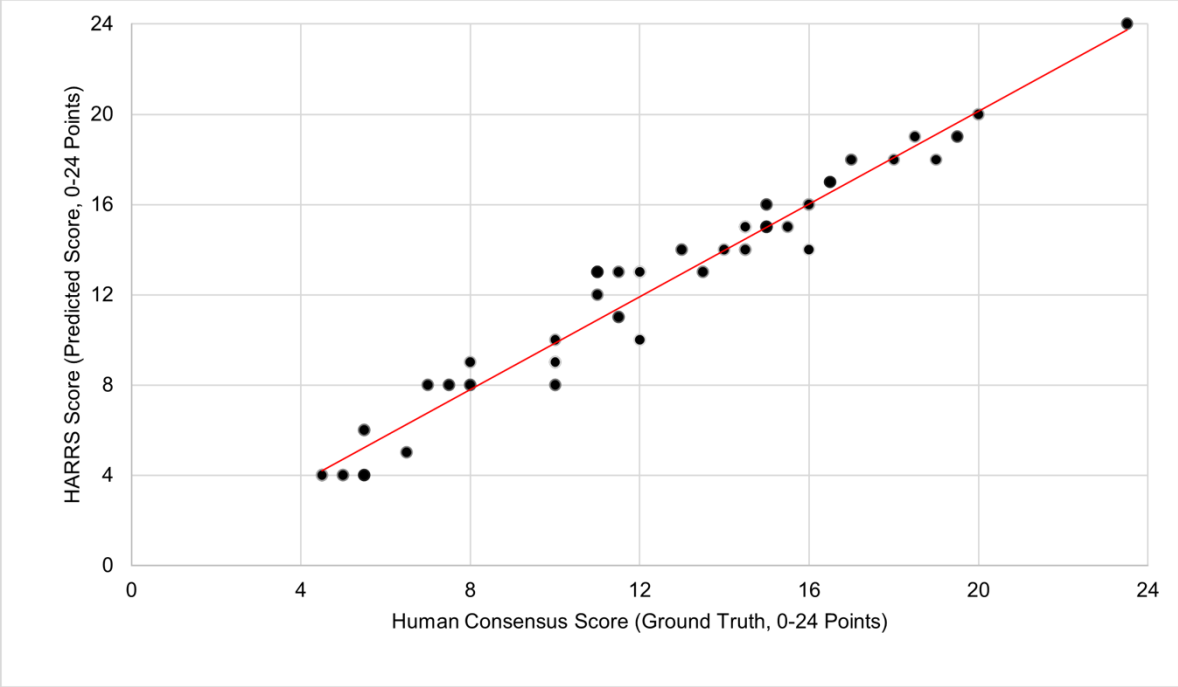

Fig. 4. Alignment HARRS Scores with Human Consensus

To evaluate the differences in mean scores across HARRS, Baseline LLM, and Human Reviewers, a one-way ANOVA was performed at a significance level of $\alpha = 0.05$. Subsequently, a Tukey's Honestly Significant Difference (HSD) post hoc test was conducted to identify specific pairwise differences while controlling Type I error. These statistical outputs are summarized in Table 1.

Table 1

Statistical Comparison of Score Means

Metric	HARRS	Baseline LLM	Human Consensus
Mean Score	12.28	17.62	12.35
Standard Deviation	4.95	4.37	4.73
Pearson's r	0.98	0.46	N/A

Metric	HARRS	Baseline LLM	Human Consensus
One-Way ANOVA	$F(2, 147) = 21.35, p < 0.001$		
Post-Hoc Comparison (Tukey HSD)	HARRS ≈ Human (NS) Baseline ≠ Human ($p < 0.001$) HARRS ≠ Baseline ($p < 0.001$)		

The ANOVA yielded a highly significant result ($F(2, 147) = 21.35, p < 0.001$), indicating that the mean scores are not equal across the three groups.

The Tukey's HSD test used the pooled standard error to compare mean differences between each pair of groups.

Discussion. The experimental results show that the HARRS architecture effectively mitigates systematic leniency bias and low discriminative precision, the primary limitations of unconstrained LLMs in academic assessment. By anchoring generative logic to a rule-based Rubric Engine, the system achieves statistical parity with expert human judgment. In addressing RQ1, the findings demonstrate that rule-based checks are highly effective at replicating expert assessments of structural and formatting compliance. The 87% reduction in MAE compared to Baseline LLM resulted in a final HARRS MAE of 0.73. This suggests that objective criteria are best handled by deterministic logic rather than probabilistic generation. Such high accuracy shows the system provides a reliable "first-pass" review, catching formal errors that often consume significant time during manual reviews.

Regarding RQ2, the stark contrast in Pearson's r between HARRS (0.98) and Baseline LLM (0.46) provides strong evidence of "evaluative grounding." While unconstrained models struggle to distinguish among varying levels of report quality, HARRS's near-perfect correlation suggests it correctly replicates the qualitative hierarchy established by experts. Furthermore, the ANOVA and Tukey's HSD results indicate that the difference between HARRS and Human Consensus is not statistically significant. This demonstrates that the system functions as a "synthetic expert," effectively mimicking the average judgment of human experts. This finding shows that statistical parity is attainable when models are constrained by structured, external rubrics rather than relying solely on internal training data.

In response to RQ3, the data confirm that the hybrid approach successfully mitigates the "leniency bias" observed in the Baseline LLM (mean score of 17.62 versus 12.35 for Human Consensus). The highly significant difference between the Baseline and Human Consensus ($p < 0.001$) indicates that unconstrained models are prone to systematic scoring inflation. HARRS addresses this by transforming the LLM from an evaluator into an augmented feedback engine.

High student utility ratings indicate that the system bridges the gap between quantitative rigor and qualitative depth. Even when the semantic phrasing deviated slightly from human nuance, students found the actionable feedback helpful for revision. This demonstrates that HARRS is a powerful pre-submission aid that enhances student self-regulation without compromising academic standards.

The effective alignment of the HARRS with expert human judgment suggests a fundamental shift in how such systems can be integrated into higher education.

The most significant theoretical implication of the results is that "generative accuracy" in AI is a product of architectural constraints rather than model scale. While previous research has cautioned against the "Rating Roulette" [7] and the inherent inconsistency of LLMs in evaluative tasks, this study demonstrates that a hybrid, rule-based approach can stabilize AI judgment to the point of "synthetic expertise." This offers a new perspective on the relationship between human-defined rules and machine learning evaluation. It suggests that for AI to be trustworthy in high-stakes environments, the evaluative authority must remain with the human-defined rubric rather than the model's internal weights [12].

From a practical standpoint, the high student utility scores and low error rates indicate that HARRS can serve as a reliable "pre-evaluator."

For students, this means a heightened capacity for self-regulation; they no longer need to wait for instructor feedback to identify structural or formatting gaps.

For teachers, deploying such a system shifts the workload from "compliance checking" to "mentorship." Because the system handles the objective, rule-based aspects of the review with high accuracy (RQ1), they can focus on the more abstract, rhetorical dimensions of student work that AI still struggles with.

The findings also have profound implications for institutional policies on "algorithmic fairness" [9]. By achieving statistical parity with human experts, HARRS proves that automated systems can reduce the "leniency bias" that often leads to inflated scores. However, the use of such a system must be managed by a "human-in-the-loop" philosophy, where faculty remain the governing body of the Rubric Engine.

This guarantees that pedagogical standards are an extension of the institution's existing curriculum rather than a "black-box" alternative [15]. Ultimately, these results provide a roadmap for scaling personalized feedback in higher education without compromising the academic integrity of the assessment process [18].

Despite the strong performance metrics, this study has certain constraints. First, the current validation was performed on a specific set of academic reports; the generalizability of the Unified Rubric across highly diverse or multi-disciplinary domains remains to be fully explored. Second, although the system reached statistical parity with human experts, relying on only one LLM backend suggests HARRS's "model-agnostic' abilities should be tested across different models.

Future work will focus on further improving feedback quality by iteratively adjusting the system's prompts and evaluative logic to better capture nuanced academic conventions. Additionally, the rule-based component will be refined to detect more sophisticated formatting and stylistic deviations. Future research should also test the system across various LLM backends. This will confirm that the Rubric Engine remains the primary factor influencing quality, regardless of the underlying generative model.

Conclusions. This research addressed the key challenge of providing scalable, consistent pre-submission feedback on structured academic reports, a process often limited by high instructor workloads and the variability of manual reviews. While LLMs offer promising semantic capabilities, their unconstrained application in academic contexts is frequently plagued by leniency bias and a lack of structural rigor. The study successfully demonstrated that a hybrid architecture, combining deterministic rule-based logic with generative feedback, effectively bridges this gap.

The empirical validation confirmed that HARRS provides a robust solution to these systemic issues. By separating auditable structural compliance from probabilistic semantic recommendations, the system achieved high accuracy in detecting formal errors while maintaining near-perfect alignment with human expert judgment. The results indicated that the hybrid approach effectively transformed LLM into a reliable augmented feedback engine that mirrors the standards of human consensus.

This work demonstrated that automated reviews can extend beyond surface-level grammar checks to provide in-depth, section-level guidance without compromising academic integrity. By maintaining a "human-in-the-loop" philosophy, HARRS ensures that disciplinary standards remain guided by faculty expertise while the system scales the labor of formative assessment. Future efforts will focus on extending both the rule-based validation logic and the automated feedback generation to accommodate a broader range of disciplinary conventions, ensuring the system remains a transparent and trustworthy aid for student self-regulation.

References:

1. Shi, H. & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, *36*(2), 187-209. https://doi.org/10.1017/S0958344023000265.

2. Xu, J., & Zhang, S. (2022). Understanding AWE Feedback and English Writing of Learners with Different Proficiency Levels in an EFL Classroom: A Sociocultural Perspective. The Asia-Pacific Education Researcher, 31(4), 357–367. https://doi.org/10.1007/s40299-021-00577-7.

3. Chen, M., & Cui, Y. (2022). The effects of AWE and peer feedback on cohesion and coherence in continuation writing. Journal of Second Language Writing, 57, 100915. https://doi.org/10.1016/j.jslw.2022.100915.

4. Yeung, S. (2025). A comparative study of rule-based, machine learning, and large language model approaches in automated writing evaluation (AWE). Proceedings of the 15th International Learning Analytics and Knowledge Conference, 984–991. https://doi.org/10.1145/3706468.3706566.

5. Wilson, J., & Roscoe, R. D. (2020). Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy. Journal of Educational Computing Research, 58(1), 87–125. https://doi.org/10.1177/0735633119830764.

6. Bylander, J., & Gustafsson, M. (2021). Improved content mastery and written communication through a lab-report assignment with peer review: An example from a quantum engineering course. European Journal of Physics, 42(2), 025701. https://doi.org/10.1088/1361-6404/abcb57.

7. Haldar, R., & Hockenmaier, J. (2025). Rating Roulette: Self-Inconsistency in LLM-As-A-Judge Frameworks. Findings of the Association for Computational Linguistics: EMNLP 2025, 24986–25004. https://doi.org/10.18653/v1/2025.findings-emnlp.1361.

8. Albuquerque Da Silva, D., Eduardo De Mello, C., & Garcia, A. (2024). Analysis of the Effectiveness of Large Language Models in Assessing Argumentative Writing and Generating Feedback: Proceedings of the 16th International Conference on Agents and Artificial Intelligence, 573–582. https://doi.org/10.5220/0012466600003636.

9. Emirtekin, E. (2025). Large Language Model-Powered Automated Assessment: A Systematic Review. Applied Sciences, 15(10), 5683. https://doi.org/10.3390/app15105683.

10. Huang, Y., & Wilson, J. (2025). Evaluating LLM-Based Automated Essay Scoring: Accuracy, Fairness, and Validity. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress,* 71-83. https://aclanthology.org/2025.aimecon-wip.9.pdf

11. Homiar, A. et al. (2025). Development and evaluation of prompts for a large language model to screen titles and abstracts in a living systematic review. BMJ Mental Health, 28(1), e301762. https://doi.org/10.1136/bmjment-2025-301762.

12. Zhang, Y. et al. (2024). The Fusion of Large Language Models and Formal Methods for Trustworthy AI Agents: A Roadmap (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2412.06512.

13. Jonsson, A., Panadero, E., Pinedo, L., & Fernández-Castilla, B. (2025). Using rubrics for formative purposes: Identifying factors that may affect the success of rubric implementations. Assessment in Education: Principles, Policy & Practice, 32(2), 192–211. https://doi.org/10.1080/0969594X.2025.2486947.

14. Chelvan, I. T., Smith, M. A., & Ghosh, S. (2021). Reflect, Review, and Revise: Using Checklists to Improve Students' Lab Reports. 2021 IEEE International Professional Communication Conference (ProComm), 85–90. https://doi.org/10.1109/ProComm52174.2021.00022.

15. Akiba, D., & Garte, R. (2024). Leveraging AI Tools in University Writing Instruction: Enhancing Student Success While Upholding Academic Integrity. Journal of Interactive Learning Research, 35(4), 467–480. https://doi.org/10.70725/355152wkijve.

16. Lee, S. S., & Moore, R. L. (2024). Harnessing Generative AI (GenAI) for Automated Feedback in Higher Education: A Systematic Review. Online Learning, 28(3). https://doi.org/10.24059/olj.v28i3.4593.

17. Steinert, S., Avila, K. E., Ruzika, S., Kuhn, J., & Küchemann, S. (2024). Harnessing large language models to develop research-based learning assistants for formative feedback. Smart Learning Environments, 11(1), 62. https://doi.org/10.1186/s40561-024-00354-1.

18. Sölch, M., Dietrich, F. T. J., & Krusche, S. (2025). Direct Automated Feedback Delivery for Student Submissions based on LLMs. Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering, 901–911. https://doi.org/10.1145/3696630.3727247.

Література:

1. Shi, H. & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, *36*(2), 187-209. https://doi.org/10.1017/S0958344023000265.

2. Xu, J., & Zhang, S. (2022). Understanding AWE Feedback and English Writing of Learners with Different Proficiency Levels in an EFL Classroom: A Sociocultural Perspective. The Asia-Pacific Education Researcher, 31(4), 357–367. https://doi.org/10.1007/s40299-021-00577-7.

3. Chen, M., & Cui, Y. (2022). The effects of AWE and peer feedback on cohesion and coherence in continuation writing. Journal of Second Language Writing, 57, 100915. https://doi.org/10.1016/j.jslw.2022.100915.

4. Yeung, S. (2025). A comparative study of rule-based, machine learning, and large language model approaches in automated writing evaluation (AWE). Proceedings of the 15th International Learning Analytics and Knowledge Conference, 984–991. https://doi.org/10.1145/3706468.3706566.

5. Wilson, J., & Roscoe, R. D. (2020). Automated Writing Evaluation and Feedback: Multiple Metrics of Efficacy. Journal of Educational Computing Research, 58(1), 87–125. https://doi.org/10.1177/0735633119830764.

6. Bylander, J., & Gustafsson, M. (2021). Improved content mastery and written communication through a lab-report assignment with peer review: An example from a quantum engineering course. European Journal of Physics, 42(2), 025701. https://doi.org/10.1088/1361-6404/abcb57.

7. Haldar, R., & Hockenmaier, J. (2025). Rating Roulette: Self-Inconsistency in LLM-As-A-Judge Frameworks. Findings of the Association for Computational Linguistics: EMNLP 2025, 24986–25004. https://doi.org/10.18653/v1/2025.findings-emnlp.1361.

8. Albuquerque Da Silva, D., Eduardo De Mello, C., & Garcia, A. (2024). Analysis of the Effectiveness of Large Language Models in Assessing Argumentative Writing and Generating Feedback: Proceedings of the 16th International Conference on Agents and Artificial Intelligence, 573–582. https://doi.org/10.5220/0012466600003636.

9. Emirtekin, E. (2025). Large Language Model-Powered Automated Assessment: A Systematic Review. Applied Sciences, 15(10), 5683. https://doi.org/10.3390/app15105683.

10. Huang, Y., & Wilson, J. (2025). Evaluating LLM-Based Automated Essay Scoring: Accuracy, Fairness, and Validity. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress,* 71-83. https://aclanthology.org/2025.aimecon-wip.9.pdf

11. Homiar, A. et al. (2025). Development and evaluation of prompts for a large language model to screen titles and abstracts in a living systematic review. BMJ Mental Health, 28(1), e301762. https://doi.org/10.1136/bmjment-2025-301762.

12. Zhang, Y. et al. (2024). The Fusion of Large Language Models and Formal Methods for Trustworthy AI Agents: A Roadmap (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2412.06512.

13. Jonsson, A., Panadero, E., Pinedo, L., & Fernández-Castilla, B. (2025). Using rubrics for formative purposes: Identifying factors that may affect the success of rubric implementations. Assessment in Education: Principles, Policy & Practice, 32(2), 192–211. https://doi.org/10.1080/0969594X.2025.2486947.

14. Chelvan, I. T., Smith, M. A., & Ghosh, S. (2021). Reflect, Review, and Revise: Using Checklists to Improve Students' Lab Reports. 2021 IEEE International Professional Communication Conference (ProComm), 85–90. https://doi.org/10.1109/ProComm52174.2021.00022.

15. Akiba, D., & Garte, R. (2024). Leveraging AI Tools in University Writing Instruction: Enhancing Student Success While Upholding Academic Integrity. Journal of Interactive Learning Research, 35(4), 467–480. https://doi.org/10.70725/355152wkijve.

16. Lee, S. S., & Moore, R. L. (2024). Harnessing Generative AI (GenAI) for Automated Feedback in Higher Education: A Systematic Review. Online Learning, 28(3). https://doi.org/10.24059/olj.v28i3.4593.

17. Steinert, S., Avila, K. E., Ruzika, S., Kuhn, J., & Küchemann, S. (2024). Harnessing large language models to develop research-based learning assistants for formative feedback. Smart Learning Environments, 11(1), 62. https://doi.org/10.1186/s40561-024-00354-1.

18. Sölch, M., Dietrich, F. T. J., & Krusche, S. (2025). Direct Automated Feedback Delivery for Student Submissions based on LLMs. Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering, 901–911. https://doi.org/10.1145/3696630.3727247.