Calibrating Algorithmic Judgment: A Hybrid Agentic Framework for Subjective Competency Assessment with Human-AI Collaboration

Daisy C. A. Silva * and Sérgio S. C. Silva

Military Institute of Engineering (IME), Rio de Janeiro, Brazil
Email: daisy.silva@ime.eb.br (D.C.A.S.); scardoso@ime.eb.br (S.S.C.S.)
*Corresponding author

***Abstract*—Automated Essay Scoring (AES) using Large Language Models (LLMs) presents a significant challenge due to its unstable behavior, which oscillates between excessive leniency on subjective criteria, such as argumentative depth, and undue severity on formal errors. This misalignment with nuanced human cognition undermines the pedagogical purpose of automated assessment. To address this issue, this paper proposes a hybrid agentic framework to calibrate algorithmic judgment, introducing the Hybrid Agentic Assessment System (SAHA) as its proof-of-concept implementation. The architecture of SAHA goes beyond the monolithic application of LLMs by employing multiple AI agents, each with a distinct evaluative persona. These agents engage in a deliberative debate to analyze the text from different perspectives, modeling the cognitive conflict inherent in human evaluation. This multi-agent deliberation aims to produce a more balanced and robust assessment. A key innovation of this framework is the targeted integration of human expertise, identifying specific points of disagreement among agents. This approach seeks to ensure that technology serves as an intelligent support in the educational process.**

***Keywords*—Automated Essay Scoring (AES), Large Language Models (LLMs), agentic Artificial Intelligence (AI), human-AI collaboration, algorithmic judgment**

I. Introduction

The growing integration of Artificial Intelligence (AI) in education promises to revolutionize processes previously restricted to human cognition, such as the evaluation of textual productions. Automatic Essay Scoring (AES), a research field established for decades, has found in Large Language Models (LLMs) a potential to analyze texts with a level of nuance previously unattainable [1, 2]. The promise is that of systems capable of providing feedback on a large scale, consistently and instantaneously, alleviating the workload of educators and enhancing the formative process for students.

However, the practical application of LLMs in AES reveals a fundamental tension. While on one hand the technology offers scalability, on the other it exposes the difficulty of machines in replicating the complex, subjective, and context-dependent discernment that characterizes human judgment. In previous research, the phenomenon of algorithmic leniency was identified and named, where LLMs, despite their technical proficiency, tend to overestimate the quality of subjective criteria such as argumentative depth [3–6].

The rationale for this research lies in going beyond the simple application of LLMs to propose AI architectures that actively manage their inherent biases. A misaligned evaluation, whether due to excessive leniency or rigor, fails in its pedagogical purpose. Inspired by this need, the framework for the calibration of algorithmic judgment is proposed. This framework is based on the innovative synergy of three technological pillars: Generative AI as the analysis engine, Agentic AI as a deliberative architecture that models different evaluative perspectives, and Conversational AI as an interface for effective collaboration between the machine and the inclusion of the human expert. The main innovation is to conceive of evaluation not as a monolithic task, but as a society of agents that deliberate, a paradigm aligned with the pursuit of a more explainable and reliable AI.

A. Problem and Research Questions

The central problem of this research is the fundamental instability and misalignment of LLM judgment in the assessment of subjective competencies. This problem manifests as a paradox documented in the literature: while studies point to algorithmic leniency [3–6], recent works such as Kundu and Barbosa [7] reveal the opposite, algorithmic severity, where LLMs strictly penalize formal errors that human evaluators tend to overlook. Therefore, the real challenge is not to correct an isolated bias, but rather to design a solution that can manage this instability and produce more robust, fair, and aligned evaluations.

To guide the investigation, the following Research Questions (RQ) were formulated:

(1) RQ1: Is it possible to model the complexity of human evaluative judgment through an agentic framework

Manuscript received April 16, 2026; accepted June 22, 2026; published July 27, 2026.

that simulates a debate between different evaluative perspectives?
(2) RQ2: Does a system instantiated from this framework demonstrate superior alignment with human evaluators on subjective criteria, compared to monolithic LLM approaches?
(3) RQ3: How can the interaction and visualization of disagreements between agents optimize the collaboration and decision-making of the human expert in the human computation workflow?
(4) RQ4: What design principles for ethical and collaborative AI frameworks can be derived from the implementation and evaluation of the proposed system?

B. Research Objectives

To answer these questions, this research defines one general objective and four specific objectives. General objective: To design, develop, and validate the hybrid agentic framework that calibrates LLM judgment in text assessment, mitigating instability (leniency / severity) and increasing alignment with the assessment of human experts on subjective criteria. Specific objectives:
(1) To model and implement the Hybrid Agentic Assessment System (HAAS), a proof of concept of the proposed framework, equipped with multiple agents with distinct evaluative personas and a debate mechanism to generate evaluations with confidence scores.
(2) To quantitatively and qualitatively evaluate the effectiveness of HAAS in reducing misalignment with human evaluators, comparing its performance with that of monolithic LLM approaches.
(3) To analyze the dynamics of the human-AI collaboration provided by the framework, investigating the impact of the arbitration workflow on the cognitive load and trust of the human expert.
(4) To propose a set of guidelines for the design of ethical and collaborative AI frameworks for application in educational assessment contexts.

II. Literature Review

The existing literature positions the originality of the proposed framework.

A. Biases and Fairness in AES with LLM

Previous research has established algorithmic leniency [3–6], while other studies like Kundu and Barbosa's [7] have identified algorithmic severity. Together, these works form the basis of the problem that the framework aims to manage: the instability of AI judgment. This instability is, in itself, a facet of the broader problem of algorithmic fairness in education. Yang *et al.* [8] systematically investigate how AES models can exhibit predictive biases against students from different socioeconomic statuses, reinforcing the need to develop architectures that not only seek accuracy but also equity.

B. Architectures for AES

Current approaches, even the most advanced ones, are predominantly monolithic. They explore sophisticated techniques such as prompting strategies [9] or fine-tuning methods like p-tuning [10], which aim to optimize the performance of a single LLM. Although effective in their own right, these approaches do not propose a framework for internal deliberation between different algorithmic points of view, which is the core of this proposal.

C. Agentic AI

Applications of debate among LLM agents are emerging and have proven effective in enhancing performance on complex tasks [11]. This approach is enabled by increasingly sophisticated agent orchestration frameworks [12]. However, the current literature focuses on applying this paradigm to generic reasoning tasks, such as solving logical problems and answering questions. The application of this paradigm to model the cognitive conflict of subjective criteria in educational assessment, treating different biases as specialized agents in a debate, is a contribution that this research proposes to explore.

D. Human-AI Collaboration

Concepts like AIED Unplugged [13] highlight the importance of the human in the loop. The proposed framework advances by defining a more specific and optimized role for the human: that of an arbiter of disagreements, a more efficient collaboration mechanism.

The research gap is, therefore, the absence of a framework that combines agentic AI to simulate an evaluative debate with a human computation workflow to arbitrate the resulting disagreements, as a way to actively manage the biases and instability inherent in LLMs.

III. Materials and Methods

This research adopts Design Science Research (DSR) as its epistemological-methodological approach, which aims to generate knowledge through the construction and evaluation of innovative technological artifacts [14]. The main artifact to be developed is the framework for hybrid agentic assessment, whose proof of concept will be the SAHA system. DSR is the ideal methodology due to its intrinsic focus on solving real-world practical problems, seeking to align the rigor of academic research with relevance to practice. Specifically, the research will follow an iterative cycle of construction and evaluation, as recommended by DSR methodologies [15], to progressively refine the framework and its implementation.

The theoretical foundation that informs the design of this artifact is structured on three pillars:

Automated Essay Scoring (AES): The literature on AES [2, 8] highlights a frontier: the difficulty of LLM models in simultaneously considering different dimensions of textual quality. This limitation results in unstable judgments, which oscillate between algorithmic leniency [3–6] and algorithmic severity [7]. The proposed framework addresses this gap not by trying to eliminate the bias, but by modeling it as part of a deliberative process,

an approach that seeks a more faithful representation of the complexity of judgment.

Multi-agent systems and Agentic AI: The theory of multi-agent systems, where autonomous agents interact to solve problems, provides the architectural basis for this framework [16]. The rise of LLMs has revitalized this field, enabling the creation of generative agents capable of complex reasoning, planning, and interaction, simulating human behaviors in digital environments [17]. The central innovation of this research is the application of the emerging technique of debate between agents [11], which has proven effective in refining reasoning and reducing biases in logic and planning tasks. The proposal is to adapt this paradigm to model the cognitive conflict inherent in subjective assessment, treating different evaluative perspectives as specialized agents. The implementation will rely on modern frameworks like AutoGen [12], which are specifically designed to orchestrate complex and collaborative conversations among multiple LLM agents.

Human Computation and Hybrid Intelligence: The concept that human-machine collaboration is superior to full automation for complex and nuanced tasks [18] guides the design of this framework. It will function as an augmented intelligence system, using AI to identify points of greatest complexity and direct human cognition to where it is most valuable: the arbitration of subjectivity. This model of targeted human computation not only optimizes the expert's time but also positions the human as an essential and irreplaceable component of the information system, ensuring a final layer of validation and common sense.

The epistemological-methodological approach of this project is DSR, which aims to generate knowledge through the construction and evaluation of innovative technological artifacts [14]. The main artifact to be developed is a framework for hybrid agentic assessment, whose proof of concept will be the SAHA system. The DSR methodology is ideal for its focus on solving practical problems, and an iterative cycle of construction and evaluation will be followed to progressively refine the framework and its implementation [15].

A. Phase 1: Framework Conception and SAHA Development (v1)

In this phase, the core of the system will be built: the multiagent architecture.

1) Definition of the framework components

- Agent Component: Definition of the agents' personas, directly inspired by the conflicting findings on algorithmic biases [3–7].
- Initial proposal: Holistic Agent (focus on coherence and flow of ideas, tending to simulate human leniency) and Normative Agent (focus on grammatical rules and structure, tending to simulate the severity of a machine, machinic severity).
- Deliberation component: Definition of a debate protocol where agents interact to refine the judgment, a technique that has proven effective in enhancing reasoning in LLMs [11]. The debate will be managed by a Moderator Agent, responsible for synthesizing the arguments.
- Human arbitration component: Definition of the interface and the interaction flow with the expert, following the principles of augmented intelligence.

2) SAHA implementation

Utilization of Python and agent orchestration frameworks, such as AutoGen [12] or LangChain [19]. Specific prompts for each persona and the debate mechanism will be implemented, resulting in a provisional evaluation and a disagreement index, calculated based on the variance of the scores and the semantic divergence of the justifications.

B. Phase 2: Integration of Human Computation into SAHA (v2)

In this phase, the human component is integrated into the pipeline, transforming the system into a hybrid intelligence solution.

- Arbitration Interface Design: Development of a web interface that presents the points of conflict and the arguments of each agent to the expert in a clear and visual way, optimizing their cognitive load.
- Arbitration and Calibration Workflow: Implementation of the pipeline where the human resolves disagreements. Their feedback (the final decision and, optionally, a brief justification) will be recorded in a structured manner, creating a dataset for future analyses and for refining the agents' prompts.

C. Phase 3: Framework Evaluation through SAHA

This phase is dedicated to the rigorous evaluation of the artifact, as advocated by DSR.

1) Corpus and baseline

Utilization of essay corpora with ground truth (National High School Examination or National Secondary Education Examination (ENEM)), ensuring the validity of the reference data. The baseline for comparison will be a state-of-the-art LLM operating in a monolithic mode, replicating the approach of previous research.

2) Comparative study

The corpus will be evaluated by three methods: (a) SAHA with human calibration (the complete proposal), (b) SAHA without human intervention (to isolate the effect of agentic AI), and (c) the baseline (for comparison with the state of the art).

3) Data analysis

Quantitative: Analysis of agreement with the ground truth using standard metrics for AES: Spearman's Correlation (r), to assess consistency in the ordering of scores, and the Intraclass Correlation Coefficient (ICC), to measure reliability and absolute agreement between evaluators.

Qualitative: Application of Thematic Content Analysis [20] on the justifications generated by SAHA. The objective will be to identify emerging codes related to the richness, balance, and transparency of the judgment, verifying if the debate produced a more balanced and analytically deeper evaluation.

D. Phase 4: Generalization and Dissemination

The final phase focuses on user validation and the generalization of findings:

User Study: A usability study will be conducted with evaluators (teachers). Quantitative instruments, such as the System Usability Scale (SUS) [21], and qualitative methods, through semi-structured interviews, will be applied to collect data on usability, perceived trust in the framework, and cognitive load.

Dissemination: Publication of the results of the conceptual framework and its empirical evaluation in high-quality journals and conferences in the areas of information systems, artificial intelligence, and education.

IV. Result and Discussion

The execution of this research aims to generate tangible results and multifaceted contributions to the areas of Artificial Intelligence, Education, and Information Systems.

A. Results

A Conceptual and Methodological Framework: The primary result will be the formalization of the framework for the design of hybrid and agentic assessment systems. This framework will include the deliberation architecture among agents, the workflow for human arbitration, and the metrics to identify algorithmic disagreement.

A Functional Proof of Concept (SAHA): The implementation of the SAHA system as a functional software artifact, empirically validating the technical feasibility and effectiveness of the proposed framework.

Empirical Evidence of Calibration: Quantitative and qualitative data demonstrating that the framework, when instantiated, produces evaluations that are statistically more aligned and qualitatively more balanced with the judgment of human experts, compared to monolithic LLM approaches.

A Set of Design Principles for Collaborative AI: A set of design guidelines (design principles) for the development of ethical and collaborative AI systems in assessment contexts, focused on transparency, explainability, and the optimization of human-computer interaction.

B. Contributions

Theoretical Contribution: The proposition of a conceptual model for the field of AES that addresses the instability of AI judgment not as an error to be eliminated, but as a characteristic to be modeled and managed through a deliberative process. This represents a paradigm shift from the search for the perfect AI judge to AI as a committee of experts.

Contribution to Information Systems (IS): The project contributes an architectural pattern for Decision Support Systems (DSS) in semi-structured domains. The SAHA framework serves as a model for the design of augmented intelligence systems, where information generated by AI is qualified by confidence indices (disagreement), optimizing human decision-making in areas such as content moderation, sentiment analysis, and document screening.

Methodological contribution: The validation of a workflow for hybrid intelligence in educational assessment, which can be generalized to other domains that require the arbitration of subjective judgments at scale.

Technological contribution: The proposed framework itself and its reference implementation, SAHA, serving as a basis for future research and development of new-generation assessment tools.

Social and ethical contribution: By making the AI evaluation process more transparent and keeping the human as the final arbiter, the project offers a path for the development of a more responsible, fair, and trustworthy AI for the field of education, increasing the confidence of educators and students in the technology.

V. Conclusion

This research is positioned at the forefront of knowledge about AI in education, addressing one of its most complex challenges: the alignment of machine judgment with human discernment. Instead of pursuing a monolithic AI model that emulates human cognition, this research proposes a pragmatically more robust and ethically more defensible path: a framework that replaces the paradigm of the single evaluator with a collective intelligence of agents in collaboration with experts.

By modeling the instability of algorithmic judgment (leniency vs. severity) as a debate between different perspectives, and by positioning the human expert as the final arbiter in moments of greatest ambiguity, this research not only seeks a technically more accurate solution but also explores a more synergistic and transparent model of human-AI interaction.

It is expected that the proposed framework and the results obtained with the SAHA system will contribute to the advancement of the theory and practice of automated assessment. More than that, the aim is to offer a path for the development of tools that enhance the human capacity to judge, ensuring that technology serves as an intelligent support, and not as an opaque substitute, in the educational process.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Daisy C. A. Silva conceived the research and the SAHA system; Sérgio S. C. Silva supervised the study and analyzed the data; all authors contributed to the writing and final approval of the manuscript.

Funding

This research is supported by the Military Institute of Engineering (IME).

Acknowledgment

The authors would like to thank the Graduate Program for the support.

REFERENCES

[1] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: A systematic literature review," *Artificial Intelligence Review*, vol. 55, no. 3, pp. 2495–2527, 2022.

[2] M. Uto, "A review of deep-neural automated essay scoring models," *Behaviormetrika*, vol. 48, no. 2, pp. 459–484, 2021.

[3] D. Silva, C. Mello, and A. C. Garcia, "Analysis of the effectiveness of large language models in assessing argumentative writing and generating feedback," in *Proc. ICAART*, 2024, pp. 573–582.

[4] D. C. A. da Silva, C. E. de Mello, and A. C. B. Garcia, "Exploring the role of large language models in evaluating argumentative writing in military school education," in *Proc. International Conference on Agents and Artificial Intelligence*, 2024, pp. 287–301.

[5] D. Silva, C. Ferreira, S. Silva, and J. Cardoso, "Analysis of the effectiveness of LLMs in handwritten essay recognition and assessment," in *Proc. ICAART*, 2025, pp. 776–785.

[6] D. C. Albuquerque da Silva, C. L. Ferreira, and S. d. S. Cardoso Silva, "Cross-domain analysis of the effectiveness of LLMs in essay assessment: From the technical rigor of IME to the large-scale context of ENEM," in *Proc. International Conference on Agents and Artificial Intelligence*, 2025, pp. 447–464.

[7] A. Kundu and D. Barbosa, "Are large language models good essay graders?" arXiv Preprint, arXiv:2409.13120, 2024.

[8] K. Yang *et al*., "Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability," in *Proc. AAAI Conference on Artificial Intelligence*, 2024.

[9] M. Stahl *et al*., "Exploring LLM prompting strategies for joint essay scoring and feedback generation," in *Proc. 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, 2024, pp. 283–298.

[10] Y. Song *et al.*, "Automated essay scoring and revising based on open-source large language models," *IEEE Transactions on Learning Technologies*, vol. 17, 2024.

[11] T. Liang *et al*., "Encouraging divergent thinking in large language models through multi-agent debate," arXiv Preprint, arXiv:2305.19118, 2023.

[12] Q. Wu *et al*., "AutoGen: Enabling next-gen LLM applications via multi-agent conversations," in *Proc. First conference on language modeling*, 2024.

[13] S. Isotani *et al*., "AIED unplugged: Leapfrogging the digital divide to reach the underserved," in *Proc. Artificial Intelligence in Education Conference*, 2023.

[14] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, pp. 75–105, 2004.

[15] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007.

[16] M. Wooldridge and N. R. Jennings, "Intelligent agents: Theory and practice," *The Knowledge Engineering Review*, vol. 10, no. 2, pp. 115–152, 1995.

[17] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proc. 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–22.

[18] L. Von Ahn, "Human computation," Ph.D. dissertation, Carnegie Mellon University, 2005.

[19] H. Chase. (2022). LangChain. [Online]. Available: https://github.com/langchain-ai/langchain

[20] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[21] J. Brooke, "SUS: A quick and dirty usability scale," *Usability Evaluation in Industry*, vol. 189, no. 194, pp. 4–7, 1996.