

# A Teacher-in-the-Loop Multi-Agent Framework for Transparent Essay Scoring and Feedback

Ali Keramati, Mark Warschauer,  
[a.kera@uci.edu](mailto:a.kera@uci.edu), [markw@uci.edu](mailto:markw@uci.edu),  
University of California, Irvine

**Abstract:** We present MADEST, a multi-agent debate essay scoring triangulation, as a teacher-centered AI tool for scoring and feedback. The system distributes evaluation across specialized agents and enables educators to interactively query and refine rubric-aligned assessments. Evaluated on the ASAP dataset, MADEST outperforms single-agent baselines and approaches human inter-rater reliability. By combining transparent scoring with teacher-in-the-loop interaction, MADEST reframes automated assessment as a collaborative process that augments teacher capacity and supports more consistent, timely feedback.

## Introduction

Essay writing is central to developing students' critical thinking and communication skills, yet providing high-quality, individualized feedback remains one of the most time-consuming challenges teachers face. Scoring essays requires interpreting complex, multidimensional rubrics, a task prone to inconsistency even among trained raters (Kayapinar, 2014; Huang & Whipple, 2023). In large classrooms, these demands make it difficult to deliver timely formative feedback that supports student revision and growth (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006). Recent advances in large language models (LLMs) have opened new possibilities for automating essay assessment (Li & Liu, 2024). Yet current single-agent approaches often evaluate writing superficially, lack transparency in their reasoning, and generate feedback with limited instructional value (Fallah et al., 2024). Critically, they offer teachers little control over the assessment process, functioning as black-box tools that sideline educator judgment rather than supporting it. We introduce MADEST, a multi-agent system that reimagines AES as a teacher-facing collaborative tool. Rather than replacing human judgment, MADEST is designed to augment teacher capacity by providing structured, rubric-aligned evaluations with transparent rationales and enabling educators to interact conversationally with the system to refine and validate outputs.

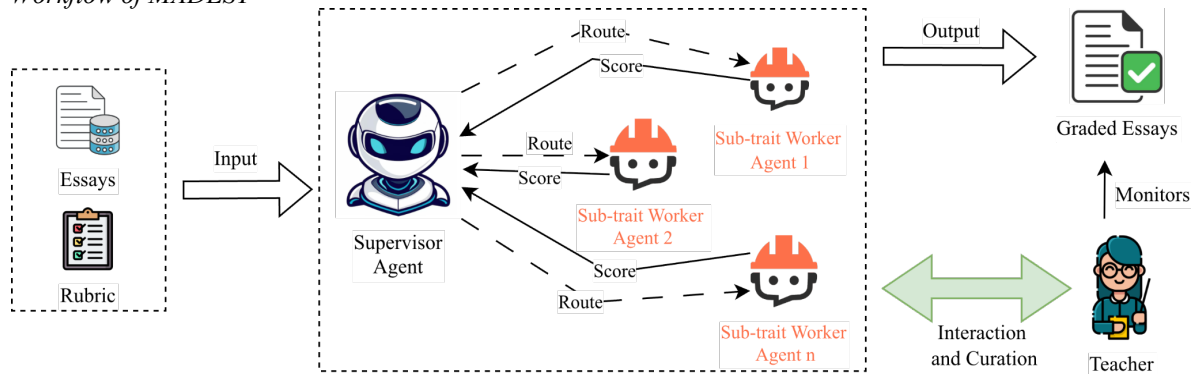
## Method and Data

The MADEST framework organizes essay evaluation as a multi-agent assessment pipeline that decomposes scoring into interpretable, rubric-aligned components. To evaluate MADEST, we used a portion of Essay Set 8 from the Automated Student Assessment Prize (ASAP) dataset (Crossley et al., 2025), originally consisting of 723 Grade 10 essays written in response to a narrative prompt. Each essay was independently scored by two human raters across six writing traits—Ideas and Content, Organization, Voice, Word Choice, Sentence Fluency, and Conventions—on a 1-6 scale. A holistic score is derived from these trait-level ratings. This multi-trait, multi-rater structure enables fine-grained evaluation of both analytic and holistic scoring performance, as well as model-human agreement.

MADEST implements a hierarchical multi-agent architecture composed of a Supervisor Agent and multiple specialized Worker Agents. Given a student essay and scoring rubric, the Supervisor Agent first parses the rubric and decomposes it into its constituent traits. Each trait is then assigned to a dedicated Worker Agent, which operates independently to evaluate a single dimension of writing. This decomposition reduces cross-dimensional interference and allows each agent to focus on specific rubric criteria. Each Worker Agent produces three outputs: (1) a numeric sub-trait score aligned with rubric definitions, (2) an interpretable rationale explaining the scoring decision, and (3) targeted, actionable feedback to support student revision. These outputs are generated in parallel across agents. The Supervisor Agent subsequently aggregates them into a structured evaluation report that includes analytic scores, dimension-level explanations, and a synthesized holistic assessment.

A central feature of MADEST is its teacher-in-the-loop interaction design. After the report is generated, teachers engage with the system through a conversational interface (Figure 1). Rather than passively accepting results, educators can request clarification of specific scores, examine underlying rationales, flag inconsistencies, or trigger re-evaluation of selected traits. Queries related to individual dimensions are routed to the corresponding Worker Agent, while the Supervisor Agent handles holistic or cross-trait inquiries. This interaction positions MADEST as a teacher-centered assessment tool, enabling educators to actively curate, validate, and refine AI-generated evaluations before feedback is shared with students.

**Figure 1**  
Workflow of MADEST



## Results

We evaluated MADEST against a conventional single-agent AES baseline using two backbone LLMs: GPT-4o-mini and GPT-5-mini. Model–human agreement on holistic scores was measured using quadratic weighted kappa (QWK),  $\pm 1$  agreement, and Spearman’s  $\rho$ , with the human–human agreement reported as a benchmark (Table 1). Across both backbones, MADEST consistently achieved stronger alignment with human raters than the baseline. With GPT-4o-mini, QWK improved from 0.367 to 0.521 against Rater 1 and from 0.312 to 0.449 against Rater 2.  $\pm 1$  agreement increased from 25.45% to 35.27% (Rater 1) and 26.42% to 32.37% (Rater 2), while Spearman’s  $\rho$  rose from 0.523 to 0.568 (Rater 1) and 0.471 to 0.505 (Rater 2). Even larger gains were observed with GPT-5-mini, where QWK climbed from 0.447 to 0.615 (Rater 1) and 0.378 to 0.548 (Rater 2);  $\pm 1$  agreement from 27.25% to 44.95% and 28.49% to 39.42%; and  $\rho$  from 0.569 to 0.618 and 0.522 to 0.560, respectively. The strongest configuration, MADEST with GPT-5-mini, approached human inter-rater reliability on all three metrics, with QWK = 0.615 versus the human benchmark of 0.624 ( $\Delta = 0.009$ ),  $\pm 1$  agreement of 44.95% versus 48.13% ( $\Delta = 3.18$  pp), and  $\rho = 0.618$  versus 0.626 ( $\Delta = 0.0077$ ).

**Table 1**  
Model vs. Human Agreement on Final (Holistic) Scores

Model	Method	QWK		$\pm 1$ Agreement		Spearman $\rho$	
		Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
GPT-4o-mini	SA	0.367	0.312	25.450	26.418	0.523	0.471
	MADEST	0.521	0.449	35.270	32.365	0.568	0.505
GPT-5-mini	SA	0.447	0.379	27.248	28.492	0.569	0.522
	MADEST	<b>0.615</b>	<b>0.548</b>	<b>44.952</b>	<b>39.419</b>	<b>0.618</b>	<b>0.560</b>
Human Agreement		<u>0.624</u>		<u>48.133</u>		<u>0.626</u>	

SA: Single Agent; IC: Ideas and Content; Org: Organization; Voc: Voice; WC: Word Choice; SF: Sentence Fluency; Conv: Conventions.

To contextualize these results, we compared MADEST with prior AES approaches using average QWK across raters. As shown, MADEST achieves an average QWK of 0.582, substantially outperforming prior single-agent methods such as Mansour et al., 2024 (QWK = ChatGPT: 0.313, Llama: 0.297). While supervised models such as Jiang et al., 2022 (QWK = 0.677) and Xie et al., 2022 (QWK = 0.779) report higher performance, these approaches rely on task-specific training. In contrast, MADEST operates in a zero-shot setting without fine-tuning, showing its advantage as a flexible and practical tool for real-world classroom use.

## Discussions and Conclusions

The results show that MADEST substantially improves alignment with human raters compared to single-agent approaches, approaching human-level agreement without task-specific training. Consistent gains across models and raters indicate that the multi-agent design provides a robust and generalizable approach to holistic essay scoring. Beyond performance, MADEST functions as a teacher-centered assessment tool. By generating rubric-aligned scores with interpretable rationales, it makes evaluation more transparent and supports teacher judgment. The conversational interface further enables educators to query and refine outputs, reframing automated scoring as a collaborative human–AI process rather than a black-box prediction. While MADEST demonstrates strong performance, automated scoring remains imperfect, and teacher oversight is essential.



## References

- Crossley, S. A., Baffour, P., Burleigh, L., & King, J. (2025). A large-scale corpus for assessing source-based writing quality: ASAP 2.0. *Assessing Writing*, 65, 100954. doi:10.1016/j.asw.2025.100954
- Fallah, A., Keramati, A., Nazari, M. A., & Fatemeh Sadat Mirfazeli. (2024). Automating Theory of Mind Assessment with a LLaMA-3-Powered Chatbot: Enhancing Faux Pas Detection in Autism. 365–372. <https://doi.org/10.1109/iccke65377.2024.10874775>
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Huang, J., & Whipple, P. B. (2023). Rater variability and reliability of constructed response questions in New York state high-stakes tests of English language arts and mathematics: implications for educational assessment policy. *Humanities and Social Sciences Communications*, 10(1). <https://doi.org/10.1057/s41599-023-02385-4>
- Jiang, Z., Gao, T., Yin, Y., Liu, M., Yu, H., Cheng, Z., & Gu, Q. (2023, July). Improving Domain Generalization for Prompt-Aware Essay Scoring via Disentangled Representation Learning. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 12456–12470). doi:10.18653/v1/2023.acl-long.696
- Kayapinar, U. (2014). Measuring Essay Assessment: Intra-rater and Inter-rater Reliability. *Eurasian Journal of Educational Research*, 14(57). <https://doi.org/10.14689/ejer.2014.57.2>
- Li, W., & Liu, H. (2024). Applying large language models for automated essay scoring for non-native Japanese. *Humanities & Social Sciences Communications*, 11(1). <https://doi.org/10.1057/s41599-024-03209-9>
- Mansour, W. A., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024, May). Can Large Language Models Automatically Score Proficiency of Written Essays? In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 2777–2786). Retrieved from <https://aclanthology.org/2024.lrec-main.247/>
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative Assessment and Self-regulated learning: a Model and Seven Principles of Good Feedback Practice. *Studies in Higher Education*, 31(2), 199–218. <https://doi.org/10.1080/03075070600572090>
- Xie, J., Cai, K., Kong, L., Zhou, J., & Qu, W. (2022, October). Automated Essay Scoring via Pairwise Contrastive Regression. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, ... S.-H. Na (Eds), *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 2724–2733). Retrieved from <https://aclanthology.org/2022.coling-1.240/>

## Acknowledgments

This work is based upon work supported by the National Science Foundation under Grant No. 2315294.