A Multi-Agent System for Feedback Quality Assurance: Evaluate, Revise, and Validate at Scale

Okan Bulut BULUT@UALBERTA.CA
University of Alberta, Edmonton, AB T6G 2G5, Canada

Cole Walsh CWALSH@ACUITYINSIGHTS.COM
Acuity Insights Inc., Toronto, ON M5V 2Y1, Canada

Abstract

Large language models (LLMs) have shown considerable promise for automatic feedback generation (AFG) in educational settings, yet their inaccuracies and unpredictable failures underscore the need for robust quality assurance (QA) before deployment at scale. Traditional QA approaches relying on human raters are inherently time-consuming, costly, and difficult to scale. This study extends the LLM-as-judge paradigm beyond evaluation to introduce a multi-agent QA system that not only assesses but autonomously revises LLM-generated feedback until predefined quality standards are met. The system comprises seven specialized agents operating in a closed evaluate–revise–re-evaluate loop. Skill-level feedback from a formative assessment measuring professional competencies ($n = 841$) is evaluated against six quality criteria using chain-of-thought prompting, and revised using role-based prompting where quality thresholds are not met. The system achieved a mean QA total of 5.99 out of 6 across all records, approved 99.0% of feedback within three iterations, and escalated only 1.0% of cases for human review. A single-agent baseline is compared against the proposed system to assess the contribution of the multi-agent architecture. The baseline performed comparably on five of the six criteria but failed markedly on second-person language (32.1% vs. 100%) and escalated 68.0% of records, compared to 1.0% in the proposed system, demonstrating that distributing QA responsibilities across specialized agents substantially improves both evaluation precision and revision effectiveness. A criterion-level stress test on 180 distorted records demonstrated a 100% correction rate for all detected violations and a 0% false positive rate, with detection rates ranging from 80.0% for actionability to 100% for second-person language. Human expert review of flagged and revised cases largely confirmed the system's judgments, though reviewers noted that certain linguistic corrections, particularly those involving punctuation, were unlikely to meaningfully affect student comprehension. These findings establish the viability of multi-agent LLM systems as scalable, automated QA and remediation pipelines for AFG.

Keywords: feedback, LLM, agentic AI, assessment, quality assurance, automatic feedback generation

1. Introduction

Effective feedback is typically characterized by being timely, personalized, and actionable – qualities that are notoriously difficult to deliver at scale, contributing to significant workload pressures for educators (Jomuad et al., 2021). To address this challenge, researchers have explored automatic feedback generation (AFG) methods for over a decade. Early AFG systems relied on expert knowledge encoded in predefined templates and error libraries, but rarely succeeded in delivering feedback tailored simultaneously to the task, the individual student, and the instructor's preferences (Cavalcanti et al., 2021).

The rapid advancement of LLMs has opened new possibilities for AFG. With their ability to understand context, generate fluent text, and follow instructions, LLMs offer a promising foundation for feedback systems that are highly adaptive to different tasks and individual responses (Mazzullo et al., 2025; Mazzullo and Bulut, 2025). Early studies applying LLMs to AFG have yielded encouraging results, with LLM-generated feedback approaching the quality of expert teachers in some contexts (Jacobsen and Weber, 2023; Steiss et al., 2024). However, these tools remain "useful but fallible" (Matelsky et al., 2023), with occasional inaccuracies, generic suggestions, and unpredictable failures that raise ethical concerns about their use in consequential educational settings (Mazzullo et al., 2025; Mazzullo and Bulut, 2025).

The fallibility of LLM-generated feedback underscores the critical need for robust QA strategies before such systems can be responsibly deployed at scale. Traditional approaches rely on human raters to assess outputs using rubrics (Mazzullo et al., 2025; van der Lee et al., 2019). While human evaluation remains the gold standard, it is inherently time-consuming, costly, and difficult to scale, creating a fundamental tension: the same conditions that make AFG appealing also make systematic human QA impractical (Chang et al., 2024).

A promising emerging approach is the use of LLMs themselves as evaluators – commonly referred to as "LLM-as-judge" (Zheng et al., 2023). Recent work has demonstrated that LLMs can evaluate text quality with reasonable alignment to human judgments across a variety of dimensions (Chiang and Lee, 2023; Zheng et al., 2023). However, existing LLM-as-judge applications in educational contexts are largely limited to evaluation; once a quality issue is identified, human intervention is still required to act on it. This limits the scalability gains that agentic approaches could otherwise provide.

In this study, we extend the LLM-as-judge paradigm by introducing a multi-agent QA system that not only evaluates LLM-generated feedback but also autonomously revises it until quality standards are met. Specifically, we deploy a seven-agent pipeline, comprising agents for context enrichment, evaluation, consistency checking, revision routing, revision, orchestration, and escalation, to assess and improve skill-level feedback produced by a formative assessment measuring professional competencies. In doing so, this study contributes a scalable, criterion-based QA and remediation methodology for AFG systems.

2. Methods

2.1. Instrument and Dataset

This study used the Professional Skills Development assessment developed by Acuity Insights[1]. This formative assessment presents learners with a series of hypothetical social or workplace dilemmas and asks how they would respond, drawing on their lived experiences. Upon completion, learners receive a report with LLM-generated personalized feedback to strengthen their professional skills.

The dataset included responses from 841 students in the United States and Canada who responded to 30 items related to one of the following skills: critical thinking, social and ethical responsibility, intrapersonal skills, or interpersonal skills. Students consent to the use of their data for research purposes before starting the assessment. Students' written

1. See https://acuityinsights.com/products/professional-skills-development/

responses were used to generate item-level feedback, which was then aggregated at the skill level, yielding four blocks of feedback per student. This aggregated feedback included direct quotes from the student's answers and summarized strengths and potential gaps.

2.2. Agentic QA System

The multi-agent QA system was implemented using Qwen3-235B via the Vulcan API (Alliance Canada). The pipeline comprises seven specialized agents operating in a closed evaluate–revise–re-evaluate loop, with a maximum of three iterations per record (see Figure 1). The **Context Enrichment Agent** first assembles the full student context for each skill-level feedback record, aggregating item-level responses and parsing item-level feedback into readable text. The **Consistency Agent** runs in parallel with the **Evaluator Agent** and serves a distinct diagnostic function. The Evaluator Agent assesses the skill-level feedback against six quality criteria using chain-of-thought (CoT) prompting (Wei et al., 2022). The Consistency Agent evaluates whether the skill-level summary is coherent with the item-level feedback that collectively informed it. This cross-level check is necessary because a skill-level summary can be linguistically correct, personalized, and actionable, passing all six QA criteria, while still misrepresenting the balance of item-level assessments, contradicting specific item feedback, or drawing disproportionately on a single item at the expense of others. The **Revision Routing Agent** inspects the evaluation results and builds a targeted revision plan identifying which criteria failed and providing specific fix guidance for each. The **Reviser Agent** rewrites the feedback to address only the failing criteria, using role-based prompting in which the model is assigned the persona of a senior assessor writing feedback for a professional skills assessment. The **Orchestrator Agent**, operating under a QA manager persona, coordinates the loop and determines whether to approve, continue revising, or escalate a record. Finally, the **Escalation Agent** flags records that fail to converge within the iteration limit, routing them to a human review queue with a structured report of persistent failures.

2.3. Single-Agent Baseline

To contextualize the contribution of the multi-agent architecture, a single-agent baseline was developed that replicates the same QA pipeline within one unified LLM call per iteration. Rather than distributing responsibilities across seven specialized agents, the baseline system consolidates all tasks into a single structured prompt. The model receives the full student context, the current skill-level feedback, and all evaluation criteria simultaneously, and is asked to produce a single JSON response containing QA scores, consistency scores, a revision decision, a list of failed criteria, and a revised version of the feedback where needed. The evaluate–revise loop operates identically to the multi-agent system, with the same maximum iteration limit, approval logic, and escalation conditions. The single-agent baseline differs from the proposed system in three key respects: it uses a single generic structured prompt rather than criterion-specific prompting strategies; it processes evaluation and consistency checking sequentially within one call rather than in parallel; and it does not benefit from the targeted revision guidance produced by the dedicated Revision Routing Agent.

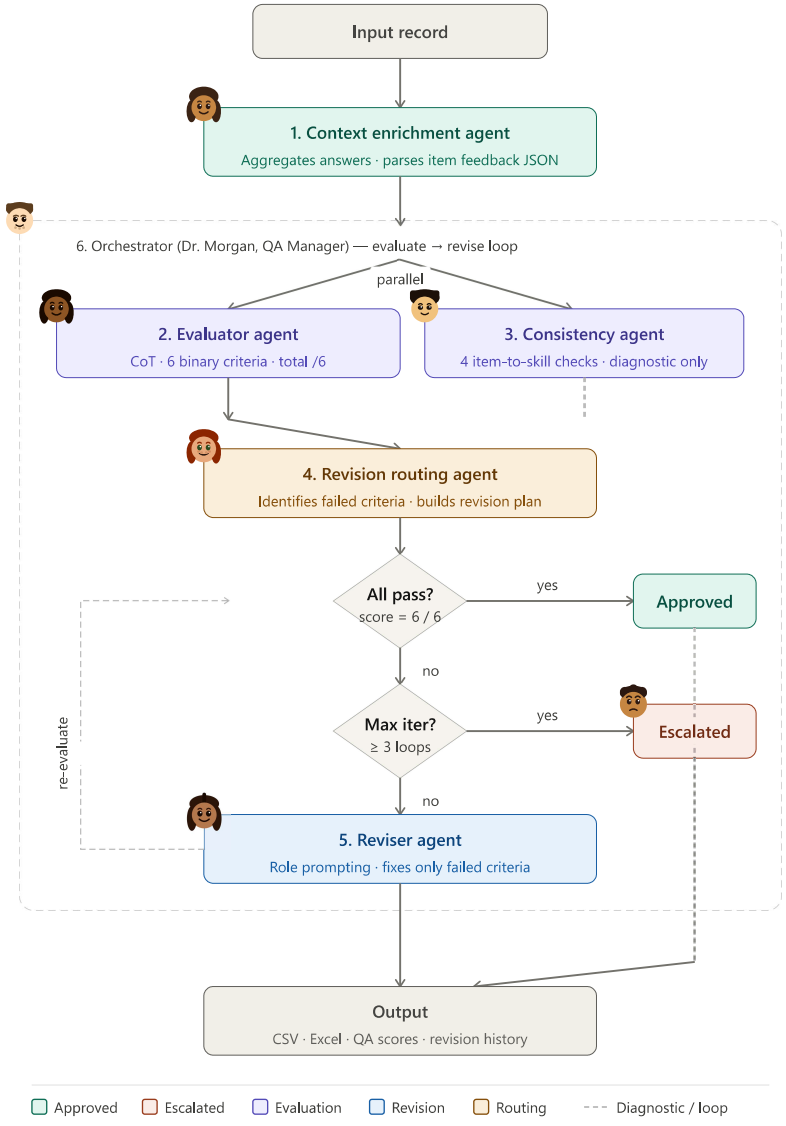

Figure 1: Architecture of the Multi-Agent Quality Assurance System.

2.4. Evaluation Criteria and Validation

Each skill-level feedback record was evaluated against six binary quality criteria drawn from prior research (Mazzullo et al., 2025; Mazzullo and Bulut, 2025): (1) **linguistic quality**: grammar and punctuation correctness, excluding direct quotes from students; (2) **factual accuracy**: whether feedback accurately reflects the student's answer and includes direct quotes; (3) **personalization**: whether feedback is tailored to the student rather

than generic; (4) **actionability**: whether feedback includes concrete, implementable improvement suggestions; (5) **affective tone**: whether tone is balanced, neither excessively praising nor critical; and (6) **second-person language**: whether the student is addressed using "you/your." To validate the automated evaluation, human experts independently reviewed all records where the revision loop was triggered or where the case was escalated for human review, using the same six quality criteria.

To assess detection and correction ability, a stress test was conducted on 180 clean records, each distorted so that exactly two of the six quality criteria were violated, with each criterion violated in 60 records across all 15 criterion pairs (12 repetitions each) for a fully balanced design. Distortions were criterion-specific: linguistic quality violations introduced grammar and punctuation errors; factual accuracy violations removed direct quotes and student-specific references; personalization violations replaced second-person address with generic third-person language; actionability violations replaced concrete suggestions with vague platitudes; affective tone violations injected harsh or flattering language; and second-person language violations converted "you/your" to "the student/their". The full agentic pipeline was run on each distorted record, and correction rates (the proportion of violated records passing after the revision loop) and false-positive rates were computed per criterion.

3. Results

3.1. Feedback Quality Evaluation

Across all 841 evaluated skill-level feedback records, the agentic QA system assigned consistently high-quality scores. The mean total score out of 6 was 5.99 ($SD = 0.08$), with scores ranging from 5 to 6. These findings suggest that the AI-generated feedback was of near-uniformly high quality, with only a small fraction of records failing on any criterion before revision. The left part of Figure 2 presents pass rates for each of the six quality criteria. Results were at or near ceiling across the board. Factual accuracy, personalization, actionability, affective tone, and second-person language all achieved a pass rate of 100%, indicating that feedback consistently reflected the content of students' responses, addressed them directly, and provided balanced, actionable guidance. Linguistic quality was the only criterion falling below 100%. The records that failed on linguistic quality were subsequently flagged for revision by the routing agent. Following revision, the overall approval rate reached 99.0%, with only 8 records (1.0%) escalated for expert review after three attempts for revisions. Quality was consistently high across all four skill dimensions.

3.2. Revision Loop Performance

The right part of Figure 2 shows that the majority of records (777 out of 841, 92.3%) were approved in a single iteration, requiring no revision. A further 43 records (5.1%) required one revision cycle and were approved in two iterations, and 21 records (2.5%) required two revision cycles before reaching a decision. Of the 64 records that entered the revision loop (7.5% of all records), 56 were successfully revised and approved, yielding an overall approval rate of 99.0%. The remaining 8 records (1.0%) were escalated to a human review queue after failing to converge within three iterations, all due to persistent failures on linguistic quality, which was the most sensitive to direct student quotes embedded in the feedback.

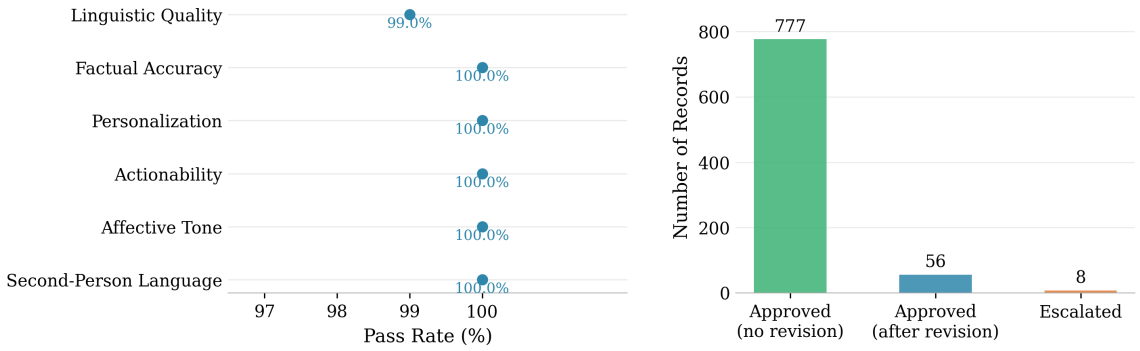

Figure 2: Multi-Agent System: Pass Rates for Quality Criteria (left) and Number of Iterations Required to Reach Approval or Escalation (right).

Human experts who reviewed the flagged and revised cases broadly agreed with the quality issues identified by the agentic system and the revisions it produced. However, they noted that a subset of the linguistic corrections (particularly those involving punctuation) would have a negligible practical impact on how clearly the feedback communicates to students. While technically improving surface-level correctness, these revisions were judged unlikely to meaningfully alter student comprehension or the perceived quality of the feedback, suggesting that the system's linguistic quality threshold may be set at a higher level of strictness than is necessary for the practical purposes of formative assessment delivery.

3.3. Comparison with Single-Agent Baseline

To assess the contribution of the multi-agent architecture, results were compared against the single-agent baseline, which consolidates all QA tasks into one LLM call per iteration (see Figure 3). The baseline achieved high pass rates on five of the six quality criteria: factual accuracy (97.7%), actionability (98.8%), personalization (97.5%), affective tone (96.4%), and linguistic quality (92.3%), which is broadly comparable to the multi-agent system. However, the baseline failed markedly on the second-person language criterion, achieving only a 32.1% pass rate, compared to 100% in the multi-agent system. This represents the most substantial performance gap between the two architectures and suggests that, when all evaluation criteria are handled within a single prompt, the model de-prioritizes certain criteria, particularly surface-level linguistic conventions, in favor of higher-order dimensions such as factual accuracy and actionability. The pipeline outcome breakdown further illustrates the limitations of the single-agent approach. Of 841 records, only 50 (5.9%) were approved without revision, 188 (22.4%) were approved after revision, and 572 (68.0%) were escalated after failing to converge within three iterations. This rate is markedly higher than that of the multi-agent system, which escalated only 8 records (1.0%).

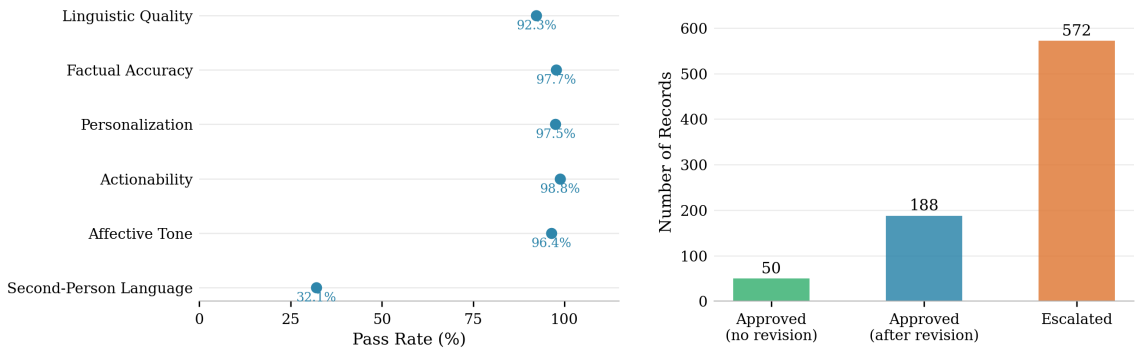

Figure 3: Single-Agent Baseline: Pass Rates for Quality Criteria (left) and Number of Iterations Required to Reach Approval or Escalation (right)

3.4. Stress Test Results

The agentic system corrected all violations detected across the six criteria. However, not all injected violations triggered a revision. Of the 180 distorted records, 18 (10.0%) were approved without revision, indicating that the distortion went undetected. Of the remaining 162 records that entered the revision loop, 160 (88.9%) required one revision cycle and were approved in two iterations, while 2 (1.1%) required two revision cycles before approval. Since each unrevised record contained exactly two violated criteria, these 18 records correspond to 36 undetected criterion-level violations. Figure 4 presents the distribution of these missed violations across criteria. Detection rates varied across criteria: second-person language achieved a perfect detection rate (100%), followed by personalization (98.3%) and affective tone (90.0%). Linguistic quality (88.3%), factual accuracy (83.3%), and actionability (80.0%) showed the highest miss rates, suggesting that the distortion functions for these criteria produced subtler violations that the evaluator agent did not consistently flag as failing. No record was escalated, and the overall approval rate was 100%.

Figure 5 presents the outcome breakdown by criterion pair. Revision was required for the vast majority of pairs, with the personalization, second-person language, factual accuracy, and affective tone combinations consistently requiring revision across all 12 records. The pairs least likely to trigger revision were those involving factual accuracy and actionability together, where half of the records passed without revision, suggesting that co-occurring failures on these two criteria were occasionally harder for the distortion functions to render unambiguously detectable.

4. Discussion and Conclusion

This study introduced and evaluated a multi-agent QA system that extends the LLM-as-judge paradigm from passive evaluation to autonomous remediation of LLM-generated educational feedback. Four findings stand out.

The multi-agent pipeline produced near-ceiling quality scores across all six criteria before any revision, consistent with evidence that well-prompted LLMs can serve as reliable

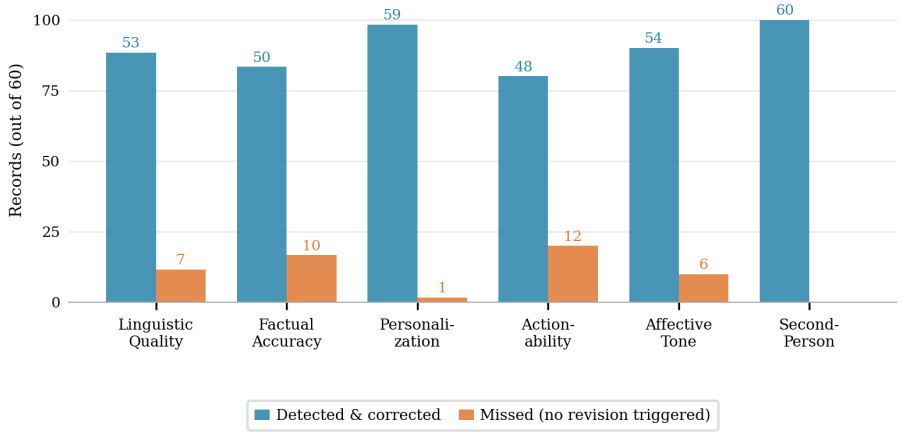

Figure 4: Detected-Corrected Violations versus Missed Violations per Quality Criterion.

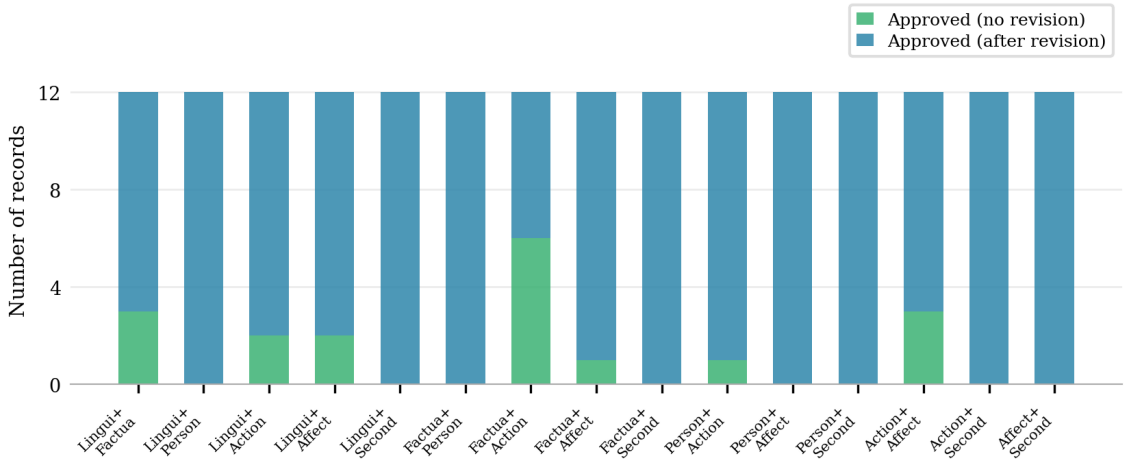

Figure 5: Outcome Breakdown by Violated Criterion Pair.

evaluators of text quality (Zheng et al., 2023). This suggests the upstream AFG system already produces high-quality feedback in most cases, with the QA pipeline serving as a safety net rather than a major corrective mechanism. Linguistic quality was the sole criterion with failures before revision and the sole driver of escalation, underscoring the challenge of evaluating surface-form correctness in feedback that intentionally embeds informal student quotes; human experts confirmed that many flagged cases involved punctuation issues with negligible communicative impact.

Second, the revision loop resolved most identified quality failures efficiently: of the 63 records requiring revision, 55 were approved within two iterations, with no additional failures introduced into previously passing criteria. This targeted behavior reduces the risk of revision degrading already-satisfactory feedback, a common failure mode when LLMs are

asked to freely rewrite text (Celikyilmaz et al., 2021). The escalated records, all involving persistent linguistic quality failures tied to quoted student text, represent a well-defined failure mode that a post-hoc rule excluding quoted passages from linguistic scoring could eliminate without further model calls.

Third, the comparison with the single-agent baseline indicates the value of the multi-agent architecture. While the baseline performed comparably on five of the six criteria, it failed markedly on second-person language (32.1% vs.100%) – notable given that this criterion requires only consistent use of "you/your" and is arguably the most straightforward to satisfy. This suggests that consolidating all responsibilities into a single prompt causes the model to de-prioritize surface-level properties in favor of higher-order dimensions such as factual accuracy and actionability, a problem the multi-agent architecture avoids through dedicated prompting strategies per stage. This difference is most visible in escalation rates: 572 records (68.0%) for the single-agent baseline versus 8 (1.0%) for the multi-agent system, consistent with evidence that orchestrated multi-agent systems sustain accuracy under scale while single-agent accuracy falls sharply as task complexity increases (Klang et al., 2026).

These findings have practical implications for the deployment of LLM-based QA systems in educational settings: a single-agent approach may appear adequate when evaluated on aggregate pass rates, masking critical failures on specific criteria that have a direct impact on how feedback is experienced by students. Architectural choices that distribute responsibility, enforce targeted evaluation, and guide revision at the criterion level seem to be a substantive determinant of system reliability. At the same time, the benefits of multi-agent systems over single-agent systems diminish as LLM capabilities improve (Gao et al., 2025), and it is possible that more capable frontier models would narrow the gap observed here.

Fourth, the stress test results reveal important nuances in the multi-agent system's detection sensitivity. While the overall correction rate was 100% for all detected violations, and false positive rates were 0% across all criteria, the 18 records (10%) that passed without revision despite containing injected violations highlight the limits of the evaluator's sensitivity. Actionability, factual accuracy, and linguistic quality showed the highest miss rates, whereas second-person language achieved perfect detection. These findings suggest that detection performance is not solely a function of model capability but also of how clearly a violation manifests in the surface form of the text.

Several limitations warrant acknowledgment. First, the study uses a single assessment instrument and student population, so generalizability to other AFG systems, subject domains, age groups, or educational levels remains to be established, and whether system-approved feedback is pedagogically effective in terms of student learning gains, reflection, or skill development remains an open question; classroom-based pilot studies with outcome-oriented evaluation would help address both issues. Second, the Consistency Agent currently serves a diagnostic function only and does not drive revisions; extending the revision loop to address item-to-skill coherence failures represents a natural next step that would further differentiate the architecture from single-agent alternatives. Lastly, the stress test relied on programmatic distortion functions of inherently bounded realism, as natural quality failures may be more varied and contextually entangled than the injected distortions.

References

Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027, 2021. doi: 10.1016/j.caeai.2021.100027.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey, 2021. URL https://doi.org/10.48550/arXiv.2006.14799.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024. doi: 10.1145/3641289.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations?, 2023. URL https://arxiv.org/abs/2305.01937.

Mingyan Gao, Yanzi Li, Banruo Liu, Yifan Yu, Phillip Wang, Ching-Yu Lin, and Fan Lai. Single-agent or multi-agent systems? why not both?, 2025. URL https://arxiv.org/abs/2505.18286.

L. J. Jacobsen and K. E. Weber. The promises and pitfalls of ChatGPT as a feedback provider in higher education: An exploratory study of prompt engineering and the quality of AI-driven feedback. *Preprint*, 2023. doi: 10.31219/osf.io/cr257.

Perlito D. Jomuad, Leah Mabelle M. Antiquina, Eusmel U. Cericos, Joicelyn A. Bacus, Juby H. Vallejo, Beverly B. Dionio, Jame S. Bazar, Joel V. Cocolan, and Analyn S. Clarin. Teachers' workload in relation to burnout and work performance. *International Journal of Educational Policy Research and Review*, 8(2):48–53, 2021. doi: 10.15739/IJEPRR.21.007.

Eyal Klang, Mahmud Omar, Ganesh Raut, Reem Agbareia, Prem Timsina, Robert Freeman, Nicholas Gavin, Lisa Stump, Alexander W Charney, Benjamin S Glicksberg, et al. Orchestrated multi agents sustain accuracy under clinical-scale workloads compared to a single agent. *npj Health Systems*, 3(1):23, 2026. doi: 10.1038/s44401-026-00077-0.

Jordan K. Matelsky, Felipe Parodi, Tony Liu, Richard D. Lange, and Konrad P. Kording. A large language model-assisted education tool to provide feedback on open-ended responses, 2023.

Elisabetta Mazzullo and Okan Bulut. Automated feedback generation for open-ended questions: Insights from fine-tuned llms. In Sheng Li, Zhongmin Cui, Jiasen Lu, Deborah Harris, and Shumin Jing, editors, *Proceedings of Machine Learning Research*, volume 264, pages 103–120. PMLR, 15–16 Dec 2025. URL https://proceedings.mlr.press/v264/mazzullo25a.html.

Elisabetta Mazzullo, Okan Bulut, Cole Walsh, Gill Sitarenios, and Alexander MacIntosh. *Fine-Tuning GPT-3.5-Turbo for Automatic Feedback Generation*, page 40–47. Association for Computing Machinery, New York, NY, USA, 2025. ISBN 9798400706295. URL https://doi.org/10.1145/3672608.3707735.

Jacob Steiss, Tamara Tate, Steve Graham, Jazmin Cruz, Michael Hebert, Jiali Wang, Youngsun Moon, Waverly Tseng, Mark Warschauer, and Carol Booth Olson. Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, 91:101894, 2024. doi: 10.1016/j.learninstruc.2024.101894.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-8643.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. doi: 10.48550/arXiv.2201.11903.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena, 2023. URL https://doi.org/10.48550/arXiv.2306.05685.