TacpAgent: Enhancing Student Engagement in Classroom Exercises Through LLM-Generated Feedback

Wanlu Zhang[1], Jia Zhu[1*], Xi Yang[2], Weijie Shi[3], Yue Cui[3], Xinle Dai[1], Jiewen Sun[1]

[1]Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University
[2]Lviv Polytechnic National University
[3]Hong Kong University of Science and Technology Hong Kong SAR
No.688 Yingbin Avenue, Jinhua,Zhejiang Province,321004,China
zhangwanlu@zjnu.edu.cn, jiazhu@zjnu.edu.cn

Abstract

Classroom exercises are imperative for reinforcing learning; however, in conventional instruction, students frequently lack timely and personalized feedback. To address this, we present TacpAgent(Teaching Agent for Classroom Practice),a generative LLM-based agent that delivers detailed, individualized guided feedback to promote self-reflection through a prompt-template framework.In the context of this study, Classroom Practice is defined as the set of structured classroom exercises used for formative assessment.TacpAgent takes as input the teacher-prepared exercises and reference answers, together with students' submitted responses, confidence levels, and brief explanations, then leverages the DeepSeek LLM with designed prompt templates to generate guided feedback and recommend relevant textbook sections for targeted review. We conducted a three-month quasi-experimental study with two high school classes (N=87). The study compared TacpAgent-supported exercises with traditional paper-based exercises. The experimental group showed significantly higher quiz scores ($F=18.516$, $p<0.001$) and improved emotional ($p<0.001$) and behavioral engagement ($p=0.003$). In contrast, the control group demonstrated no significant changes. The results suggest that TacpAgent may enable scalable, personalized formative assessment in classroom settings and provide practical guidance for integrating generative AI into everyday teaching.

Introduction

Classroom exercises are a critical component of instruction. They help students consolidate knowledge, identify misconceptions, and provide teachers with a basis for adjusting pedagogy (Yang et al. 2021). However, in conventional exercise-based teaching, time and resource constraints often prevent educators from delivering timely, individualized, and comprehensive feedback(Wulansari, Kyaw et al. 2022), a limitation particularly evident in large-class settings(Schaffer et al. 2017). As a result, many students fail to promptly correct errors, which undermines their learning outcomes.

Studies indicate that students rarely engage in proactive textbook reading but show higher motivation when preparing for practice tests or exams(Doorn, Janssen, and O'Brien 2010).This suggests that classroom exercises could serve as an effective entry point for deeper learning if feedback is timely and promotes active reflection.Unfortunately, feedback is often static and delayed(Bangert-Drowns et al. 1991).Students tend to focus narrowly on scores rather than conceptual understanding, further limiting effectiveness(Scott and Husain 2021).

Recent advances in artificial intelligence have opened up new frontiers for educational technology. Large language models (LLMs) can generate highly tailored explanatory feedback, offering a promising pathway for connecting practice exercises directly to the relevant textbook knowledge base(Tan et al. 2023; Zhu et al. 2025). However, the efficacy and practical value of these LLM-based solutions in operational, real-world classrooms, especially under the typical constraints of daily instruction, have not yet been sufficiently explored by empirical work(Dai et al. 2025).

To enhance feedback immediacy and personalization, we developed TacpAgent, an LLM-based teaching assistant that provides guided feedback instead of simple correctness indicators. TacpAgent supports learning in two primary ways: (i) guiding students to relevant textbook sections for problem-solving knowledge; and (ii) generating targeted textual guidance based on detected student errors. To ensure precise, reflective feedback aligned with student understanding, the system requires users to submit a brief rationale and report their confidence levels.

This study compares the effectiveness of Tac-pAgent versus traditional paper-based exercises in problem-solving classes, focusing on their impact on academic performance and classroom participation. Classroom participation is the second core indicator, chosen because it reflects students' attention/engagement and is a key behavioral variable for predicting learning outcomes, especially in formative assessment scenarios where the teaching tools' actual impact is more evident(Fredricks, Blumenfeld, and Paris 2004; Bond et al. 2020).

The study addresses the following research questions:

- **RQ1:** Does TacpAgent improve students' learning outcomes compared to traditional paper-based exercises?
- **RQ2:** Does TacpAgent enhance students' engagement in classroom exercises compared to traditional approaches?

*Corresponding Author

Background

Feedback Research in Educational Contexts

Research in educational and cognitive psychology has shown that feedback type and timing significantly affect learning outcomes(Hattie and Timperley 2007).Empirical studies suggest that correct/incorrect prompts alone are insufficient for conceptual understanding. Guided feedback promotes integration, transfer, and retention of knowledge(Gawronski et al. 2008; Zhu et al. 2022).

The "three questions" model of effective feedback emphasizes goal alignment, current status, and next actions, underscoring the need for feedback to support cognitive processing and self-regulation beyond binary correctness judgments(Hattie and Timperley 2007). Longitudinal studies further show that cognitive diagnostic feedback tailored to students' knowledge attributes outperforms conventional feedback, especially for challenging tasks (Tang and Zhan 2021).

Feedback richness also matters: high-information feedback, which combines task-level accuracy with self-monitoring of attention, emotion, or motivation, is more effective than simple rewards/punishments(Narciss and Huth 2004; Rittle-Johnson, Fyfe, and Loehr 2016). Timing interacts with learning goals: immediate feedback aids skill practice and error correction, while delayed feedback benefits long-term retention and conceptual transfer(Kulhavy and Anderson 1972; Van der Kleij, Feskens, and Eggen 2015). Additionally, students' confidence and self-regulation moderate feedback effectiveness: confident learners benefit from simple cues, whereas less confident learners rely on detailed explanations linked to foundational knowledge (Butler, Karpicke, and Roediger III 2008; Bjork, Dunlosky, and Kornell 2013).

In light of the aforementioned research context, this study integrates guided feedback derived from a large language model (LLM) with confidence scores and answer summaries. It then employs this integration to student exercises through TacpAgent. The objective of the design is to provide scalable, personalized, and highly informative feedback, while promoting students' active monitoring and adjustment of learning strategies in the classroom. This addresses the shortcomings of traditional classroom feedback in terms of personalization and timeliness.

The Application of AI in Educational Feedback

The emergence of artificial intelligence (AI), particularly large language models (LLMs), has transformed the generation and delivery of educational feedback. Compared with traditional human-based feedback, AI offers immediacy, personalization, and scalability: learners can quickly correct errors, receive guidance tailored to their knowledge and cognitive needs, and benefit from coverage suitable for large classes(Luckin and Holmes 2016; Holmes, Bialik, and Fadel 2019). LLMs have demonstrated the ability to produce guided feedback across domains, including reading comprehension(Xiao et al. 2023), programming(Finnie-Ansley et al. 2022), and science education (Graesser et al. 2018), supporting self-regulated learning through natural language interaction(Panadero, Andrade, and Brookhart 2018).

Empirical studies indicate that AI feedback can enhance performance and metacognitive strategy use. For example, AI-based formative feedback improved task performance and strategy utilization(Winkler and Söllner 2018), and AI-supported feedback in argumentative writing enhanced writing quality and critical thinking(Banihashem et al. 2024). Notably, LLMs can identify students' comprehension gaps from their responses and generate context-relevant guidance (Zawacki-Richter et al. 2019; Guo et al. 2025b), addressing the persistent challenge of insufficient personalized feedback in large-class settings.

Nevertheless, AI feedback faces challenges. Alignment with curriculum or textbook systems is often incomplete, risking cognitive confusion(Holmes and Tuomi 2022). Students may over-rely on AI guidance, potentially limiting independent inquiry and critical thinking (Luckin 2018). Existing systems also often lack self-reflection prompts, reducing students' ability to translate feedback into learning actions(Ochoa, Huang, and Charlton 2024). Moreover, stability, interpretability, and teacher intervention strategies remain areas for improvement in authentic classrooms(Kendapadi et al. 2024).

To address these limitations, TacpAgent integrates students' answers, their confidence in their responses, and their reasoning processes. Confidence scores facilitate the system's capacity to assess students' self-assessed mastery of knowledge, thereby enabling the provision of detailed, tailored feedback and guidance. The reasoning of students serves as a basis for analyzing their cognitive processes, helping the system accurately identify misconceptions. The LLM system integrates these inputs with textbook content tags in the question bank, providing students with targeted guidance toward relevant learning materials. The objective of this design is to deliver precise, immediate, and targeted feedback, thereby enhancing classroom participation, proactivity, and effective interaction.

TacpAgent Design and Implementation

The intelligent agent TacpAgent, developed in this study, aims to provide a more interactive and guided learning experience than traditional paper-based exercises. This improves student engagement and learning effectiveness. Its design objectives are as follows:

- Incorporating self-monitoring elements into the answer process to guide students to reflect on the basis for their answers and their level of confidence.
- Generate personalized guided feedback based on student responses to promote active correction and continuous exploration.
- Offer multiple pathways for subsequent learning to meet the immediate needs of different learners.

The functions and interaction process of TacpAgent are as follows (Figure 1):

- Teachers configure a question bank for specific course knowledge points. Each question is labeled with knowledge point tags, textbook chapter tags, and reference answers.

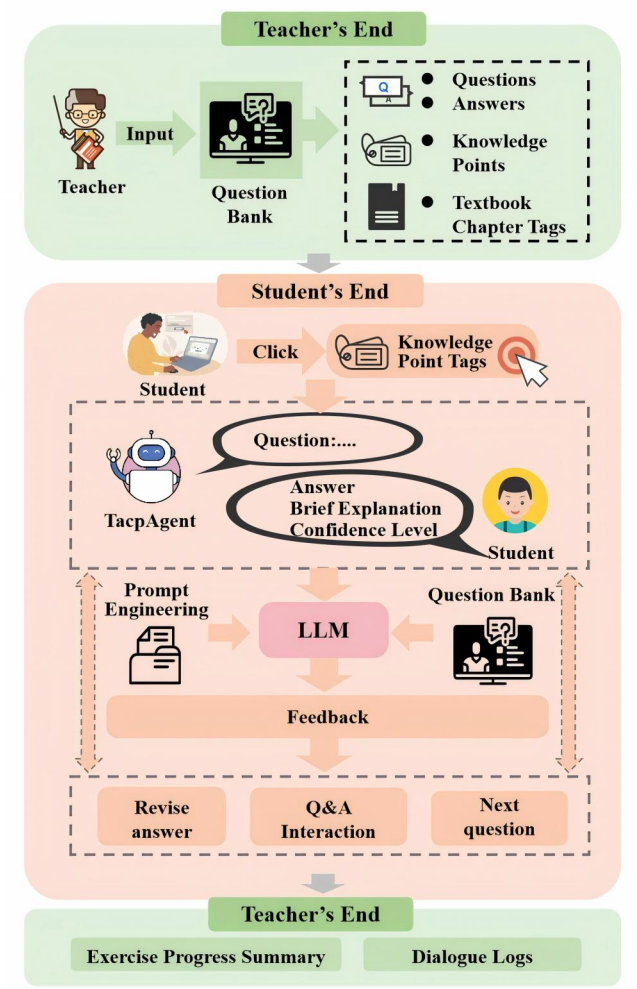

Figure 1: Functional Architecture and Interaction Flowchart

- In practice mode, students answer questions one by one. They must also self-assess their confidence level and provide a brief written explanation of their reasoning.
- In feedback mode, the AI generates customized feedback based on the student's answer, confidence level, and rationale. This feedback includes correctness verification, guiding questions, and textbook-based learning recommendations.
- Students can choose to proceed to the next question, modify the current answer, or initiate further interactive questioning in the feedback interface to obtain more in-depth support.
- Teachers can receive real-time records of student answers and interaction logs for subsequent analysis and targeted interventions in teaching.

Practice Task

The practice task module is designed to closely resemble real classroom teaching scenarios. Before class, teachers select an appropriate number of practice questions from textbooks and exercise books, based on the requirements of the course content, and enter them into the tool. Teachers determine the number of questions flexibly according to the actual progress of the teaching. Each question is accompanied by a standard answer, a corresponding textbook chapter index, and knowledge point tags. During the practice process, students click on the knowledge point tags in the interface (Figure 2-a) to access related questions. After independently thinking, the students enter their answers in the dialog box. Before submission, the tool requires students to complete two key metacognitive behaviors: first, they must briefly explain their problem solving approach in text (Figure 2-b); second, they must select their confidence level in the current answer from three options: "guessing," "certain," or "somewhat confident" (Figure 2-c). This design deeply embeds metacognitive reflection into the routine practice process by requiring immediate explanations and a confidence self-assessment, while ensuring that all practice content aligns strictly with the teacher's instructional progress and the textbook curriculum. The entire interaction process is linear: after confirming their answers and explanations, the students submit their responses, and the tool generates targeted feedback based on their answers and confidence ratings.

Feedback Generation Module

The intelligent feedback module of this tool is based on the DeepSeek-V3 API, as it has been demonstrated to excel in complex reasoning and contextualized educational content generation(Liu et al. 2024). The feedback generation process first constructs a structured contextual information framework. This includes the question content, correct answers, textbook chapter tags, students' submitted answers and reasoning, as well as their self-assessed confidence levels. This information is subsequently entered into a meticulously designed prompt framework, wherein predefined feedback prompts guide the model to generate feedback that aligns with the educational tone and instructional objectives. In instances where incorrect responses are provided, the tool does not directly provide the correct answer. Instead, it utilizes guiding questions and textbook references to encourage students to reflect and revise independently (Figure 3). If students input content that is not related to the question, the tool provides attitude-based reminders to guide them back to the task.

The prompt word templates were optimized through multiple rounds of collaboration and iteration with subject teachers (Table 1) to ensure alignment with course objectives while meeting students' personalized feedback needs. The model utilizes input information to evaluate the accuracy of responses and dynamically integrates this evaluation with confidence levels, thereby selecting the most appropriate prompt template. In conclusion, the model produces feedback text in accordance with the specified prompt requirements. Each feedback message is constrained to a maximum length of 400 characters, while maintaining logical rigor, clear structure, and pedagogical depth.

Category	Description
Right+Certain	Affirm the correctness and confidence, highlighting deep understanding of{{knowledge_point}}. Provide a step-by-step explanation: {{steps}}. pointing out the{{key_point}}. Offer motivational and enthusiastic encouragement. Suggest an advanced exploration related to{{related_knowledge}}, such as{{extension_example}}.
Right+somewhat confident	Acknowledge correctness but point out possible weak spots{{weak_point}}. Encourage use of a "Three-step Self-check": 1.Review common mistakes like{{common_errors}} 2.Double-check your reasoning 3.Attempt a similar variant problem:{{variant_example}}. Provide constructive guidance to boost confidence.
Right+guessing	Caution that correctness seems guessed, possibly hiding gaps in{{weak_point}}. Break down the problem requirements with a guided outline (without looking at the answer):{{steps}} Encourage a serious learning attitude rather than guesswork. Review{{textbook}}, p.{{page}}, section "{{chapter}}" for key points.
Wrong+Certain	Warn that high confidence with a wrong answer indicates misunderstanding. Identify the main error {{error_point}}. Suggest reflection with guiding questions. Review {{textbook}}, p.{{page}}, section "{{chapter}}" to clarify your thinking.
Wrong+somewhat confident	Explain the partial misunderstanding and locate the {{error_point}}. Provide guiding questions to encourage reflection and correction. Review {{textbook}}, p.{{page}}, section "{{chapter}}" for more detail.
Wrong+guessing	Indicate that the answer shows random guessing and unstable understanding. Provide a patient step-by-step breakdown with illustrative examples (without giving the final answer). Include guiding questions for self-reflection. Review {{textbook}}, p.{{page}}, section "{{chapter}}".
Irrelevant answer	First, determine whether the irrelevant response {{answer}} relates to {{question}}. If yes, respond; if no, classify as "off-task." If off-task, emphasize the need to focus on the learning {{topic}} While pointing out irrelevance, encourage the student to maintain curiosity and positive thinking.

Table 1: Feedback prompt framework designed by answer accuracy and confidence level, with templates refined through multiple teacher iterations to support personalized and well-structured instructional feedback.

Methods

Study Design

The study incorporates two **independent variables**:

- **Control Group (Traditional)**: The pedagogical approach incorporates a combination of paper-based exercises, complemented by teacher and peer discussions, with supplementary support from textbook materials.
- **Experimental Group (TacpAgent)**: TacpAgent offers exercises that provide immediate feedback, in conjunction with teacher (human and agent) and peer discussions, and are supported by textbook materials.

The **dependent variables** encompass two dimensions:

- **Classroom engagement**: The scale was developed by Wang(Wang, Bergin, and Bergin 2014) and encompasses four dimensions: emotional engagement, cognitive engagement, behavioral engagement, and disengagement.
- **Learning outcomes**: The assessment was conducted by measuring the pre-test and post-test scores related to the course content.

The three-month study consisted of sequential phases: (1) A preparation phase (Month 1) involving TacpAgent training, system testing, and baseline data collection; (2) An implementation phase (Months 2-3) where experimental and control groups completed bi-weekly classroom exercises using their respective methods (TacpAgent vs. traditional paper-based) while maintaining implementation fidelity; and (3) A final evaluation phase (Month 3) comprising post-tests, engagement surveys, and system data collection to assess outcomes.

Participants

The study participants were two parallel classes of first-year high school students from an eastern Asian high school, totaling 87 students. There were 49 students in the experimental class and 38 students in the control class. The two classes were similar in terms of course progress and academic performance; both offered the required Information Technology 1 course. The students were between 15 and 16 years old, with 47 males (54%) and 40 females (46%). All experimental courses were conducted in a computer lab, and each student was provided with a computer.

Prior to the study, informed consent was obtained from the students and their guardians. An independent samples t-test of the pre-test scores revealed no significant difference in

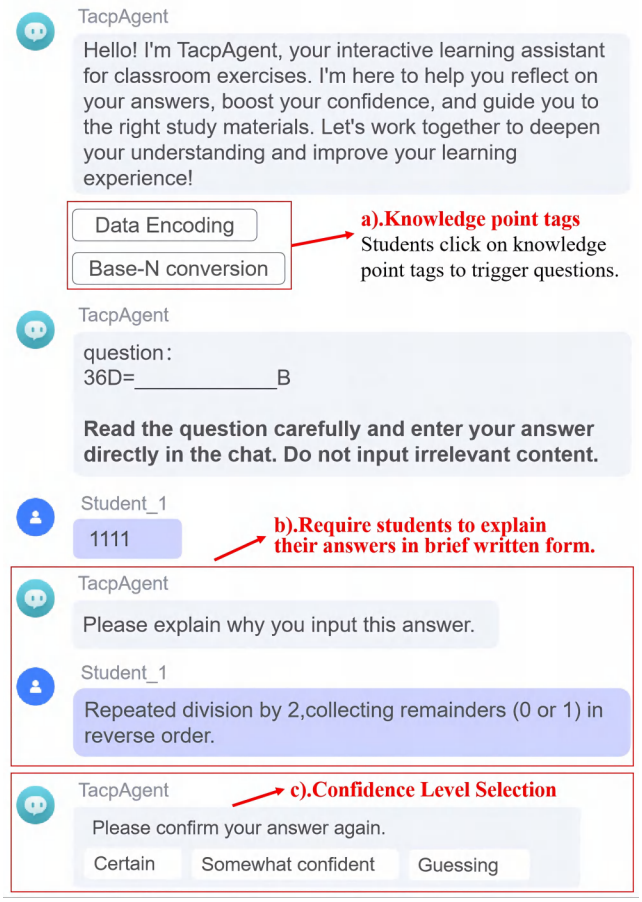

Figure 2: TacpAgent Student Practice Interface.

(a) Knowledge tags (click to load questions); (b) Open-ended explanation field (mandatory self-explanation);(c) Confidence selection (3 levels: Guessing/Somewhat confident/Certain)

baseline performance between the experimental and control groups (t=-0.593, p>0.05), thus meeting the comparability criteria for the experiment.

Intervention implementation

The teacher who implemented the intervention in both groups was not a member of the research or author team, thereby minimizing the potential for unconscious bias toward either experimental condition. Furthermore, to ensure impartiality and control for the 'teacher effect,' the instructional activities for both groups were rigidly standardized.

Both classes were taught by the same teacher, with identical course content, learning objectives, homework, and assessment methods. The only difference was the classroom exercise tool.

Crucially, the set of exercises for each bi-weekly session was drawn from the same pool of teacher-prepared questions, ensuring that both the TacpAgent-supported group and the traditional paper-based group completed an identical number of the same practice problems, maintaining content

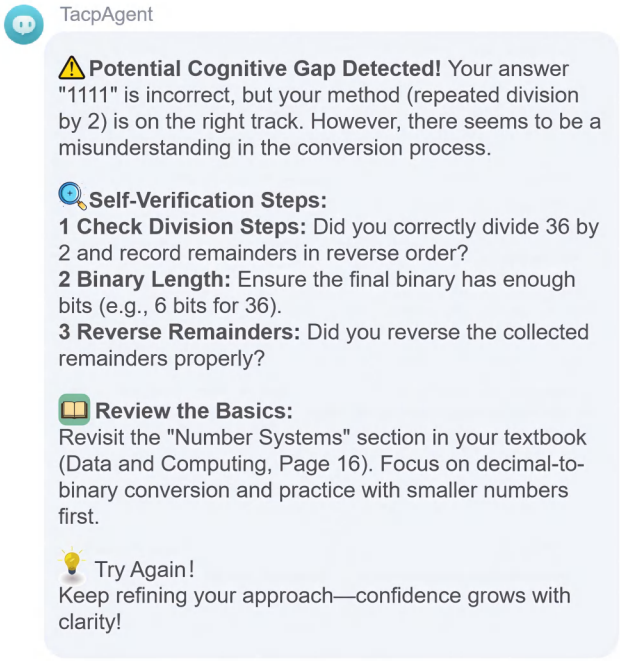

Figure 3: TacpAgent Intelligent Feedback Interface (for Incorrect Answer with Somewhat Confidence)

validity.

This design ensured variable isolation, with two functional differences serving as the independent variable. Furthermore, time allocated for the core practice task was strictly controlled and identical for both groups (Table 2), thus maintaining temporal fidelity.

Measurement instruments

Classroom Participation. Classroom participation was measured using a 24-item scale adapted from Wang(Wang, Bergin, and Bergin 2014), originally derived from the Classroom Engagement Scale. To align with the experimental context, references to "classroom activities" in the item descriptions were replaced with "TacpAgent-based classroom exercises" or "traditional paper-based exercises." The scale uses a 5-point Likert format (1=never, 5=always) and assesses four dimensions.

Internal consistency coefficients for this study were α=0.91 (emotional), α=0.89 (behavioral), α=0.87 (cognitive), and α=0.82 (disengagement). The questionnaire was administered once before the intervention (end of preparation phase) and once after (end of implementation phase) via the Questionnaire Star platform during class (10–15minutes). Items were presented in random order to minimize order effects.

Learning Outcomes. Learning outcomes were assessed using course exams (one pretest and one posttest) designed by three senior teachers, covering the core knowledge and skills of the Information Technology 1 course. Exams were reviewed by experts and achieved a KR-20 coefficient of

Phase	Group	Teacher Activities	Student Activities
Task Introduction (3 min)	Experimental/Control	Outline objectives, distribute tasks, explain rules.	Note objectives, listen to instructions.
Practice (25 min)	Experimental	Monitor without direct tutoring; observe and record performance; handle issues.	Self-practice via TacpAgent.
	Control	Monitor without direct tutoring; respond to raised-hand queries with guiding prompts only (e.g., "Which step is unclear?", "Review last lesson's method", "See p.X"); no direct answers; record requests.	Self-practice with worksheets.
Summary&Q&A (10 min)	Experimental/Control	Present answers; students self-check/correct; address common difficulties.	Submit most challenging problem.

Table 2: Comparison of Teacher and Student Activities

0.87, indicating high internal consistency and content validity.

Data Analysis. All statistical analyses were performed using SPSS 27.0. Selection of statistical tests was guided by dependent variable characteristics and initial normality tests. Pre-test scores were compared using independent samples t-tests. As learning outcome data met parametric assumptions, differences were analyzed using ANCOVA for post-test comparisons, including pretest scores as covariates. Baseline differences in participation were examined using the Mann-Whitney U test. Within-group pre-post differences in participation were assessed using the Wilcoxon signed-rank test. Non-parametric tests were used because classroom participation data did not meet normality assumptions.

Result

Learning Outcomes

To address **RQ1**, we first confirmed baseline comparability. We conducted an independent samples t-test to compare the pre-test scores of the experimental group (M=25.94,SD=6.72) and the control group (M=25.08,SD=6.70). The results indicated no significant difference between the two groups (t=-0.593,p>0.05), confirming baseline equivalence.

Based on the test for between-subjects effects, we found that the slopes of the regression lines for all groups were equal (F(1, 83)=0.207, p>0.05). The homogeneity of variance test also met the requirements (F(1, 83)=0.693, p>0.05), allowing us to conduct an ANCOVA analysis.

As shown in Table 3, there was a significant difference in post-test performance on learning outcomes (F(1, 84)=18.516, p<0.001, η^2=.181).Specifically, the experimental group achieved a higher post-test mean score compared to the control group.This indicates that students who used TacpAgent-supported exercises achieved significantly higher learning outcomes than students in the control group.

Group	N	$M(SD)$	$M(SE)$	F	η^2
Experimental	49	31.39 (6.36)	31.07 (0.53)	**18.516****	0.181
Control	38	27.18 (7.23)	27.59 (0.61)		

Table 3: ANCOVA results for post-test learning outcomes

These gains can be attributed to TacpAgent's structured metacognitive workflow, which required students to rate their confidence and explain their reasoning before receiving feedback. This process likely enhanced cognitive processing by promoting explicit reflection on problem-solving strategies, thereby strengthening conceptual understanding and error awareness. In contrast, the control group's paper-based exercises provided delayed, generalized feedback, lacking the immediacy and personalization of AI-generated responses.

Classroom Engagement

To address **RQ2**, baseline differences in classroom engagement between the experimental group (EG) and control group (CG) were examined using the Mann–Whitney U test. No significant differences were found in total engagement, emotional engagement, behavioral engagement, cognitive engagement,or disengagement, confirming group equivalence prior to the intervention (Table 4).

Dimension	Group Median		Mann–Whitney U	z	p
	EG (n=49)	CG (n=38)			
Total score	87	86	916	-0.12	0.89
Emotional engagement	20	20	900	-0.26	0.78
Behavioral engagement	29	30.5	793.5	-1.18	0.23
Cognitive engagement	26	29.5	889	-0.36	0.71
Disengagement	10	9.5	862	-0.59	0.55

Table 4: Results of Mann–Whitney U Test for Pre-test Classroom Engagement

Within-group pre–post comparisons were conducted using the Wilcoxon signed-rank test. In the control group, only cognitive engagement increased significantly, with no significant changes in emotional engagement, behavioral engagement, or disengagement (Table 5). In contrast, the experimental group exhibited significant improvements in emotional engagement, behavioral engagement, and cognitive engagement, while disengagement remained unchanged(Table 6).

These results suggest that TacpAgent-supported exercises enhanced students' engagement across emotional, behavioral, and cognitive dimensions, likely driven by interactive

Dimension	Median Pre-test	Median Post-test	Median Difference	z	p
Emotional engagement	20.0	19.0	-1	0.537	0.591
Behavioral engagement	30.5	32.0	1.5	1.566	0.117
Cognitive engagement	29.5	32.0	2.5	2.015	**0.044***
Disengagement	9.5	11.0	1.5	1.324	0.186

Table 5: Results of Wilcoxon Signed-Rank Test for Pre–Post Classroom Engagement (Control Group)

Dimension	Median Pre-test	Median Post-test	Median Difference	z	p
Emotional engagement	20.0	23.0	3	4.666	**0.000****
Behavioral engagement	29.0	33.0	4	4.333	**0.000****
Cognitive engagement	26.0	32.0	6	2.778	**0.005****
Disengagement	10.0	11.0	1	0.68	0.494

Table 6: Results of Wilcoxon Signed-Rank Test for Pre–Post Classroom Engagement (Experimental Group)

feedback and confidence-based prompts that encouraged active participation and deeper cognitive processing. The control group's improvement was limited to cognitive engagement, reflecting teacher guidance but lacking the sustained motivational and behavioral reinforcement provided by the AI-supported workflow.

Discussion

RQ1: Impact on Learning Outcomes

Experimental results show students using TacpAgent achieved significantly higher post-test scores compared to the traditional paper-based group. This aligns with prior research emphasizing the advantages of immediate, personalized feedback over delayed or generic responses(Van der Kleij, Feskens, and Eggen 2015). TacpAgent's effectiveness can be attributed to three key mechanisms: reducing the temporal gap between error and correction, enabling timely conceptual adjustment ; mandating metacognitive reflection through self-explanation and confidence ratings, which strengthens error awareness and strategy monitoring(Bjork, Dunlosky, and Kornell 2013); and employing interactive, guided feedback that guides deeper cognitive processing rather than passive score reception. These findings corroborate existing evidence on the benefits of AI-generated feedback(Banihashem et al. 2024), but TacpAgent's design introduces two novel contributions. First, unlike open-ended AI tools(Guo et al. 2025a), TacpAgent integrates tightly with teacher-curated question banks and textbook chapter tags, ensuring alignment with curriculum objectives and reducing cognitive dissonance from unvetted LLM outputs. This addresses a critical gap in AI education tools, where misalignment with instructional content often undermines utility(Holmes and Tuomi 2022). Second, the combination of confidence-based prompts and textbook linking fosters a dual-loop learning process: students not only correct errors but also contextualize them within structured knowledge frameworks, a feature underexplored in prior AI feedback systems.

RQ2: Impact on Classroom Engagement

The results indicated that TacpAgent significantly enhanced students' engagement in emotional, behavioral, and cognitive aspects, while the control group only saw a modest increase in cognitive engagement. Neither group showed a significant change in disengagement.This suggests TacpAgent's feedback mechanism effectively sustains attention and motivation, unlike traditional paper-based practice.

These findings align with prior research. TacpAgent's self-explanation step strengthens metacognitive processing(Hausmann and VanLehn 2010),and its adaptive feedback sustains attention(Panadero, Andrade, and Brookhart 2018). However,unlike some studies reporting that AI reduces disengagement(Holmes and Tuomi 2022), this study found no significant reduction.

Several factors may explain these findings. Positive and negative engagement are not symmetrical. Positive engagement is more readily enhanced through short-term interventions, while reducing negative behaviors requires deeper motivational changes(Fredricks, Blumenfeld, and Paris 2004). TacpAgent's self-explanation and confidence steps may have imposed additional cognitive demands. Reflective prompts were sometimes redundant for those with very high or very low confidence levels. Moreover, lengthy or complex textbook references could lead to cognitive overload and superficial engagement. These factors explain why disengagement did not decrease.The modest increase in cognitive engagement in the control group is understandable (due to teacher guidance and exam pressure).However, the lack of immediate and personalized feedback likely limited improvements in emotional and behavioral engagement.Overall, results support that engagement dimensions evolve at different rates: positive engagement is enhanced short-term through adaptive feedback, while reducing disengagement requires more sustained and multi-faceted interventions.

Conclusion and Future Work

This study demonstrates that integrating TacpAgent into classroom exercises significantly enhances students' learning performance and engagement.By delivering personalized, metacognitive feedback, the agent effectively addresses the limitations of traditional paper-based practice and validates the substantial promise of AI-supported pedagogy. However, the results are constrained by variable isolation.Since the experimental benefits stem concurrently from feedback immediacy, personalized quality, and interactive Q&A, attributing gains to a single factor is challenging. Future work must employ a dismantling study design to rigorously deconstruct these coupled variables. Separating the independent impact of the agent's core components via ablation or mediation analysis will greatly strengthen the persuasiveness of the research conclusions.

Acknowledgments

We acknowledge the support of the National Natural Science Foundation of China under Grant (No. 62577050).

References

Bangert-Drowns, R. L.; Kulik, C.-L. C.; Kulik, J. A.; and Morgan, M. 1991. The instructional effect of feedback in test-like events. *Review of educational research*, 61(2): 213–238.

Banihashem, S. K.; Kerman, N. T.; Noroozi, O.; Moon, J.; and Drachsler, H. 2024. Feedback sources in essay writing: peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21(1): 23.

Bjork, R. A.; Dunlosky, J.; and Kornell, N. 2013. Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, 64(1): 417–444.

Bond, M.; Buntins, K.; Bedenlier, S.; Zawacki-Richter, O.; and Kerres, M. 2020. Mapping research in student engagement and educational technology in higher education: A systematic evidence map. *International journal of educational technology in higher education*, 17(1): 2.

Butler, A. C.; Karpicke, J. D.; and Roediger III, H. L. 2008. Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(4): 918.

Dai, X.; Wen, Z.; Jiang, J.; Liu, H.; and Zhang, Y. 2025. How Students Use AI Feedback Matters: Experimental Evidence on Physics Achievement and Autonomy. *arXiv preprint arXiv:2505.08672*.

Doorn, D. J.; Janssen, S.; and O'Brien, M. 2010. Student Attitudes and Approaches to Online Homework. *International Journal for the scholarship of teaching and learning*, 4(1): n1.

Finnie-Ansley, J.; Denny, P.; Becker, B. A.; Luxton-Reilly, A.; and Prather, J. 2022. The robots are coming: Exploring the implications of openai codex on introductory programming. In *Proceedings of the 24th Australasian computing education conference*, 10–19.

Fredricks, J. A.; Blumenfeld, P. C.; and Paris, A. H. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research*, 74(1): 59–109.

Gawronski, B.; Deutsch, R.; Mbirkou, S.; Seibt, B.; and Strack, F. 2008. When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of experimental social psychology*, 44(2): 370–377.

Graesser, A. C.; Fiore, S. M.; Greiff, S.; Andrews-Todd, J.; Foltz, P. W.; and Hesse, F. W. 2018. Advancing the science of collaborative problem solving. *Psychological science in the public interest*, 19(2): 59–92.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, H.; Zhu, J.; Di, S.; Shi, W.; Chen, Z.; and Xu, J. 2025b. DioR: Adaptive Cognitive Detection and Contextual Retrieval Optimization for Dynamic Retrieval-Augmented Generation. *arXiv preprint arXiv:2504.10198*.

Hattie, J.; and Timperley, H. 2007. The power of feedback. *Review of educational research*, 77(1): 81–112.

Hausmann, R. G.; and VanLehn, K. 2010. The effect of self-explaining on robust learning. *International Journal of Artificial Intelligence in Education*, 20(4): 303–332.

Holmes, W.; Bialik, M.; and Fadel, C. 2019. *Artificial intelligence in education promises and implications for teaching and learning*. Center for Curriculum Redesign.

Holmes, W.; and Tuomi, I. 2022. State of the art and practice in AI in education. *European journal of education*, 57(4): 542–570.

Kendapadi, A.; Zaman, K.; Menon, R. R.; and Srivastava, S. 2024. Interact: Enabling interactive, question-driven learning in large language models. *arXiv preprint arXiv:2412.11388*.

Kulhavy, R. W.; and Anderson, R. C. 1972. Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology*, 63(5): 505.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Luckin, R. 2018. *Machine Learning and Human Intelligence. The future of education for the 21st century*. UCL institute of education press.

Luckin, R.; and Holmes, W. 2016. Intelligence unleashed: An argument for AI in education.

Narciss, S.; and Huth, K. 2004. How to design informative tutoring feedback for multimedia learning. *Instructional design for multimedia learning*, 181195.

Ochoa, X.; Huang, X.; and Charlton, A. 2024. Unpacking the Complexity: Why Current Feedback Systems Fail to Improve Learner Self-Regulation of Participation in Collaborative Activities. *Journal of Learning Analytics*, 11(2): 246–267.

Panadero, E.; Andrade, H.; and Brookhart, S. 2018. Fusing self-regulated learning and formative assessment: A roadmap of where we are, how we got here, and where we are going. *The Australian Educational Researcher*, 45(1): 13–31.

Rittle-Johnson, B.; Fyfe, E. R.; and Loehr, A. M. 2016. Improving conceptual and procedural knowledge: The impact of instructional content within a mathematics lesson. *British Journal of Educational Psychology*, 86(4): 576–591.

Schaffer, H. E.; Young, K. R.; Ligon, E. W.; and Chapman, D. D. 2017. Automating individualized formative feedback in large classes based on a directed concept graph. *Frontiers in psychology*, 8: 260.

Scott, T.; and Husain, F. N. 2021. Textbook Reliance: Traditional Curriculum Dependence Is Symptomatic of a Larger Educational Problem. *Journal of Educational Issues*, 7(1): 233–248.

Tan, K.; Pang, T.; Fan, C.; and Yu, S. 2023. Towards applying powerful large ai models in classroom teaching: Opportunities, challenges and prospects. *arXiv preprint arXiv:2305.03433*.

Tang, F.; and Zhan, P. 2021. Does diagnostic feedback promote learning? Evidence from a longitudinal cognitive diagnostic assessment. *AERA Open*, 7: 23328584211060804.

Van der Kleij, F. M.; Feskens, R. C.; and Eggen, T. J. 2015. Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research*, 85(4): 475–511.

Wang, Z.; Bergin, C.; and Bergin, D. A. 2014. Measuring engagement in fourth to twelfth grade classrooms: the Classroom Engagement Inventory. *School Psychology Quarterly*, 29(4): 517.

Winkler, R.; and Söllner, M. 2018. Unleashing the potential of chatbots in education: A state-of-the-art analysis. In *Academy of management proceedings*, volume 2018, 15903. Academy of Management Briarcliff Manor, NY 10510.

Wulansari, R. E.; Kyaw, Z. Y.; et al. 2022. The Adventure of Formative Assessment with Active Feedback in The Vocational Learning: The Empirical Effect for Increasing Students' Achievement. *Journal of Technical Education and Training*, 14(1): 54–62.

Xiao, C.; Xu, S. X.; Zhang, K.; Wang, Y.; and Xia, L. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, 610–625.

Yang, C.; Luo, L.; Vadillo, M. A.; Yu, R.; and Shanks, D. R. 2021. Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological bulletin*, 147(4): 399.

Zawacki-Richter, O.; Marín, V. I.; Bond, M.; and Gouverneur, F. 2019. Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International journal of educational technology in higher education*, 16(1): 1–27.

Zhu, J.; Guo, H.; Shi, W.; Chen, Z.; and De Meo, P. 2025. Radio: Real-time hallucination detection with contextual index optimized query formulation for dynamic retrieval augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 26129–26137.

Zhu, Y.; Leong, V.; Hou, Y.; Zhang, D.; Pan, Y.; and Hu, Y. 2022. Instructor–learner neural synchronization during elaborated feedback predicts learning transfer. *Journal of Educational Psychology*, 114(6): 1427.