Evaluating an AI-Supported Revision Module for Project-Based Research in Teacher Education

Assylzhan Yessimbekova[1], Ainash Issabekova[2], Karlygash Almenbetova[1] and Araily Shakirova[3]

[1]Department of Pedagogy and Methodology of Primary Education, Pedagogy and Psychology Faculty, Zhetysu University named after I. Zhansugurov, Taldykorgan, Kazakhstan
[2]Department of Pedagogy and Psychology, Pedagogy and Psychology Faculty, Zhetysu University named after I. Zhansugurov, Taldykorgan, Kazakhstan
[3]Department of Pedagogy and Educational Management, Farabi University, Almaty, Kazakhstan

assylzhanyessimbekova@gmail.com
ainashissabekova38@gmail.com
almenbetovakarlygash@gmail.com
shakirovaaraily@gmail.com (corresponding author)

https://doi.org/10.34190/ejel.24.3.4432

Abstract: While Generative Artificial Intelligence (GenAI) allows new methods to support students in writing processes, its incorporation into university teaching needs well-designed pedagogy to ensure academic integrity and autonomy of learners, especially in Project-Based Research (PBR) courses in teacher education, where iterative feedback is important but restricted by the size of groups and limited resources. The current study aims to address the scalability problem in the feedback process by evaluating the AI-Supported Revision Module (AIRM). It is implemented as a scaffold designed according to a rubric and embedded into Moodle, using a local version of the LLaMA-2-7B model for generation of prompts based on instructor comments for revision in four modes: polishing, restructuring, justifying, and synthesizing. The purpose of the module is to provide targeted guidance to students rather than full-text generation and support revision while preserving authorship. A mixed-methods case study approach was used involving 158 primary education student teachers, out of which 132 submitted draft and revision versions, assessed using a six-criterion rubric. A mixed-design repeated measures ANOVA test revealed interaction effects for both Literature Review and Methodology at the level of $p < .01$, with effect sizes of $d = 0.58$ and $d = 0.52$. The results suggest more improvement in structure and synthesis for the AI group than the conventional group, while no significant differences were found for Data Justification and Interpretation, suggesting a boundary between procedural support and higher-order analytical reasoning. Process data analysis revealed active involvement of learners, evidenced by the fact that on average 14.2% of draft content was changed and 39.7% of AI suggestions were rejected. Qualitative data analysis revealed that learners utilized this AI-powered module mostly to increase text coherence and clarity while staying in control of making sense of the content, whereas teachers indicated a shift from superficial to more methodological feedback practices. Thus, the findings demonstrate that GenAI may be successfully implemented in the feedback process as an additional scaffold for revision while still respecting the agency of learners.

Keywords: Generative artificial intelligence, Project-based research, Teacher education, Primary teachers, Feedback

1. Introduction

The swift advancement of Artificial Intelligence (AI) and especially Generative AI (GenAI), powered by Large Language Models (LLMs), has fundamentally shifted academic writing at universities. More broadly, AI has been adopted in higher education to support learning, assessment, and feedback. GenAI analyzes context and provides structured feedback, extending support beyond error correction toward higher-level writing processes. Unlike earlier rule-based systems, GenAI generates context-sensitive suggestions based on probabilistic language modelling, which has expanded its applicability in academic support tasks. For teacher education, this change carries a particular relevance to Project-Based Research (PBR) – a method aimed at fostering skills necessary for future primary teachers (Kokotsaki, Menzies and Wiggins, 2016). PBR's success is frequently hampered by what is known as a "feedback paradox." Although iteration is an indispensable aspect of gaining research skills, the conventional approach of providing individual feedback poses several logistic barriers within larger classes. As noted by Dean et al. (2023), without multiple feedback cycles, revisions often remain surface-level, focusing on grammar rather than methodological alignment. With insufficient feedback, students will often find it difficult to advance past purely descriptive literature reviews or to explain their conclusions transparently.

Cite this article: Yessimbekova, A.,et al. 2026. "Evaluating an AI-Supported Revision Module for Project-Based Research in Teacher Education", *Electronic Journal of e-Learning*, 24(3), pp 33-48, https://doi.org/10.34190/ejel.24.3.4432

The above limitations call for the use of GenAI as a scaffolding mechanism. While commercial applications of GenAI (such as ChatGPT) function independently to generate generic responses, this study employs a targeted approach that is tailored to a specific subject. Unlike commercial "black-box" services, which raise concerns about academic integrity and authorial voice (Zawacki-Richter et al., 2019), the AI-Supported Revision Module (AIRM) was designed as a transparent pedagogical tool. Based on the local LLaMA-2-7B model, it protects user privacy while creating prompts based on the rubric. The module processes student text together with instructor feedback to generate targeted, rubric-aligned revision suggestions at the paragraph level. Instead of producing text, it focuses on detecting gaps where improvements can be made by correlating feedback from the instructor with methods and giving suggestions for revisions. Feedback is operationalised through four modes: polish, restructure, justify, and synthesize.

AIRM is conceptualised as a pedagogical modification of the feedback cycle. The focus of this evaluation is not to claim long-term cognitive transfer, but to determine how such an AI-driven scaffold influences the quality of the final research product and the depth of the conceptual revision process. The study was conducted with third-year pre-service primary teachers at a pedagogical university in Kazakhstan, enrolled in the "Project-Based Research Activity" course. By comparing draft–revision pairs across AI-assisted and traditional cohorts, the research addresses the following questions:

RQ1: How does the AI-Supported Revision Module function within the revision cycle of project-based research in teacher education?

RQ2: Which aspects of student research writing show the greatest improvement when revisions are AI-assisted?

RQ3: How do students and instructors perceive the benefits and risks of integrating AI into revision tasks?

2. Literature Review

2.1 Project-Based Learning in Teacher Education

Project-Based Learning (PBL) encourages active, inquiry-based learning rather than passive learning. The use of projects helps the pre-service teachers learn to conduct inquiries, and also to acquire the identity of research-oriented professionals (Tsybulsky and Muchnik-Rozanov, 2023). At the same time, PBL is demanding for novices. Students often struggle to define researchable questions, formulate clear aims and narrow broad topics (Costley et al., 2023). Literature reviews tend to remain descriptive rather than synthetic, limiting the identification of gaps (Gao and Chen, 2024), while methodological reasoning is frequently weakened by misalignment between aims and methods (Helle, Tynjälä and Olkinuora, 2006).

2.2 Feedback and Revision in Project-Based Research Writing

Revision is central to PBR writing, enabling learning through iterative drafting and feedback cycles (Lu, Yao and Zhu, 2023). Feedback improves coherence and reflection, while peer input adds diverse perspectives, provided that students possess sufficient feedback literacy (Lineback and Holbrook, 2023; Wei and Liu, 2024; Carless and Boud, 2018).

However, revision practices remain inconsistent: students often focus on surface-level corrections or misinterpret feedback due to limited metacognitive skills, and instructors in large courses are unable to provide multiple feedback cycles (Hanafi et al., 2025; Yu and Xie, 2025). This creates a “feedback paradox,” where revision is essential for learning but constrained by institutional conditions. Structured feedback loops, particularly draft–feedback–resubmission cycles, have been shown to improve outcomes by narrowing the gap between current and desired performance (Bjælde, Boud and Lindberg, 2025), especially when combined with multiple feedback sources and structured debriefing (Sahlan, 2025). These limitations highlight the need for scalable approaches that operationalise feedback processes within high-enrollment contexts.

2.3 Generative Systems in Higher Education Writing

The introduction of GenAI has reframed debates about student writing, shifting from initial concerns about plagiarism, loss of originality and fabricated references toward a more balanced evidence base. Systematic reviews and meta-analyses show that, when used as a supplement rather than a substitute, generative tools can improve fluency and organisation, although effects on higher-order reasoning remain mixed and context-dependent (Deng et al., 2025; Lo, 2023). Similar patterns are observed in higher education, where improvements in readability and efficiency are accompanied by variability in output quality and the need for critical verification (Bhullar, Joshi and Chugh, 2024; O’Dea, 2024). Experimental studies further indicate gains in clarity and structure

when AI is used for feedback (Mahapatra, 2024; Polakova and Ivenz, 2024) but also point to risks such as reduced creativity and shallow revisions (Niloy et al., 2024; Kim et al., 2025).

This limitation is relevant in education because research writing requires combining theoretical knowledge with practical applications; the methodology needs to be justified, and educational implications have to be discussed (Eysenbach, 2023). This indicates a limitation in human–AI collaboration between humans and artificial intelligence, since generative models are good at facilitating procedural activities in writing but cannot perform meaning-making tasks. Thus, the problem is not about how to employ the technology, but rather about how to create a proper scaffolding.

2.4 Integrity, Reliability and Reference Practices

A major limitation of generative systems is the unreliability of references, with studies showing frequent fabrication of citations in AI-generated texts (Rashidi et al., 2023), posing risks for academic writing where source credibility is essential (Chelli et al., 2024). Beyond this, overreliance on generative support may homogenise writing and weaken authorial voice, contributing to "cognitive offloading," where students uncritically adopt AI suggestions (Koo, 2023; Kooli, 2023; Nautiyal, Albrecht and Nautiyal, 2023). To mitigate these risks, students must develop evaluative judgement to critically assess automated guidance, particularly given the inherent variability of feedback processes (Ayçiçek, 2025). Recommended safeguards include DOI verification, disclosure of AI use and reflective tasks documenting how tools are applied (Walters and Wilder, 2023). However, these approaches remain largely theoretical, with limited empirical evidence on their effectiveness in authentic educational settings.

2.5 Positioning Within Technology-Integration Frameworks

Conceptual frameworks clarify how generative systems operate in education. The SAMR (Substitution–Augmentation–Modification–Redefinition) model defines technology adoption as substitution, augmentation, modification or redefinition (Hamilton, Rosenberg and Akcaoglu, 2016), with current evidence suggesting that generative tools primarily function at the levels of augmentation (improving grammar and style) and modification (supporting structural reorganisation and argument development), while redefinition remains rare in teacher education (Jiménez Sierra et al., 2023). The TPACK (Technological Pedagogical Content Knowledge) framework emphasises the alignment of technological, pedagogical and content knowledge (Mishra and Koehler, 2006), with research indicating that generative tools are most effective when embedded within pedagogical structures such as feedback cycles and aligned with disciplinary content (Lim et al., 2023). Across both frameworks, the role of agency is central: technology does not transform practice by itself, but depends on how it is integrated, perceived and critically engaged with. In this perspective, the value of AI lies in supporting a human-centred assessment transition, where technology facilitates student development and professional growth without displacing the learner's role as the primary agent in the writing process (Goode et al., 2026).

2.6 Summary of Gaps

However, PBL is crucial in teacher education and comes with constant challenges in problem identification, synthesis of literature and methodology justification. To address these challenges, feedback and revisions are needed but the process is hampered by organizational restrictions and lack of feedback literacy skills. Generative tools can enhance coherence and reflection within writing but there are concerns of invented citations, lack of originality and generalized writing.

The key limitation is the absence of scholarly studies that explain the functioning of generative tools during revision processes within teacher education. Studies have been more theoretical, focusing on either integrity and the technological potentials of the tool but not its real incorporation into feedback processes. Also, although SAMR and TPACK frameworks have been cited numerous times, very little attention has been paid to interpreting findings from the perspective of such frameworks.

3. Materials and Methods

3.1 Research Design

This study is a mixed-method case study that evaluated the effect of AIRM on a class of pre-service teachers. In a single course, students participated in two different conditions; namely, instructor-only feedback and instructor feedback with AIRM through Moodle. As randomization was not possible, equivalent baselines were set. Quantitative data included paired draft–revision submissions scored on six rubric criteria, along with process metrics (e.g., suggestions, acceptance rates, text modification, revision modes). Qualitative data included open-

ended student surveys and instructor reflections, analysed thematically. Triangulation of rubric gains, system logs and thematic findings provided convergent evidence on revision behaviour, writing improvement and practical feasibility. The primary outcome of the study is the improvement in writing quality measured through rubric-based draft–revision comparisons, while interaction metrics (e.g., suggestion uptake and ratings) are used as complementary process indicators to interpret student engagement.

3.2 Context and Participants

This research was conducted at Zhetysu University named after I. Zhansugurov in Kazakhstan as part of an obligatory course during the third year of studies on PBR activities among pre-service primary school teachers. Students are required to conduct small-scale research related to primary schools. There were 158 participants in total, of which 132 provided complete pairs of drafts and revisions (Table 1). Students were allocated at the seminar group level for timetable reasons, which created a quasi-experimental design with non-random assignment but controlled instructional conditions. Sections assigned to the AI-assisted condition accessed AIRM in Moodle after receiving instructor feedback, while all other instructional conditions remained identical across groups. To minimise contamination, sections operated in partitioned Moodle spaces without cross-access to materials, while baseline comparability was established based on draft rubric scores and demographic variables (age, gender). No systematic differences between seminar groups or instructors were observed. All instructors followed a shared rubric and feedback protocol to ensure consistency.

Table 1: Participant Profile

	Participation			**Demographics**		
Delivery type	Enrolled students	Complete draft–revision pairs	Excluded submissions	Average age	Female (%)	Male (%)
Traditional (instructor comments only)	76	62	14	20.3	85.5	14.5
AI-assisted (with revision module)	82	70	12	20.5	83	17
Total	158	132	26	20.4	84	16

3.3 AI-Supported Revision Module

The AIRM was developed as an integrated extension of Moodle to support structured revision of PBR reports. It followed an iterative design process aligning LLM constraints with institutional rubric requirements, extending instructor feedback with targeted and transparent recommendations. Technically, AIRM used a locally hosted LLaMA-2-7B (temperature = 0.2, max tokens = 512) fine-tuned on institutional academic writing guides and program materials. The model was deployed locally without external API calls to ensure data privacy and full control over prompts and outputs. Predefined prompts generated concise, actionable guidance across four modes: polish (clarity and style), restructure (organisation), justify (strengthening arguments) and synthesise (connecting literature and concepts). Each prompt incorporated three elements: (1) the relevant segment of student text, (2) instructor feedback, and (3) task instructions aligned with a revision mode. Example prompt templates, including the orchestration step and mode-specific guidance, are provided in Appendix 1. The internal architecture and processing pipeline of the module are presented in Figure 1. The architecture restricts outputs to feedback-linked suggestions rather than full-text rewrites, ensuring student authorship.

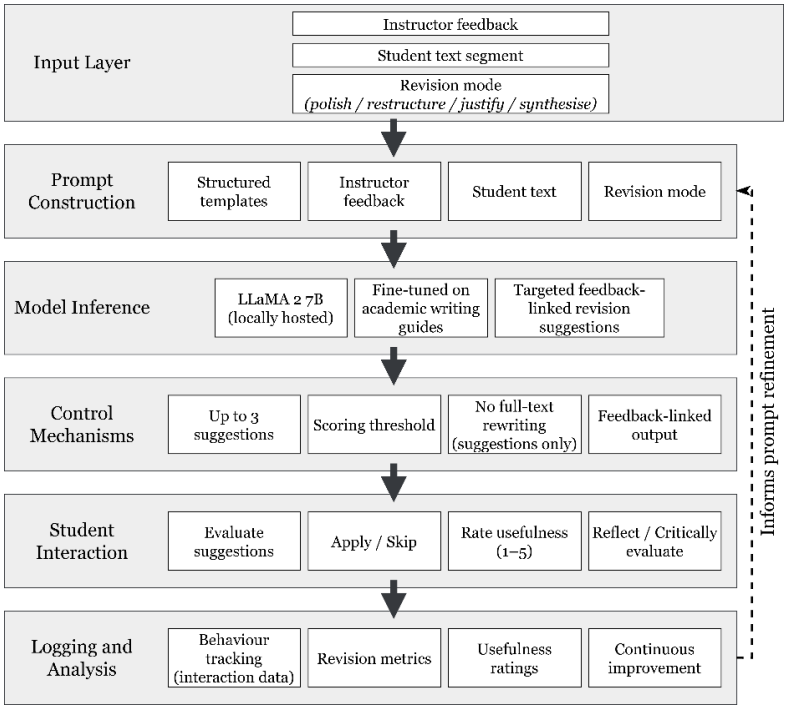

Figure 1: System Architecture of the AI-Supported Revision Module

When students accessed AIRM after receiving instructor feedback, relevant sections of the draft were identified based on instructor feedback and highlighted within the interface. Segmentation was performed at paragraph level and anchored to instructor comments to ensure consistent alignment. In cases where feedback referred to sub-paragraph elements, the nearest paragraph was used as the minimal segmentation unit to preserve contextual coherence. For each highlighted section, the system generated short, targeted suggestions indicating required improvements, without rewriting the text. Suggestions were based on the selected text, instructor feedback and instructional materials (e.g., academic writing guides) uploaded by the instructor. The overall workflow of the module within the revision cycle is illustrated in Figure 2.

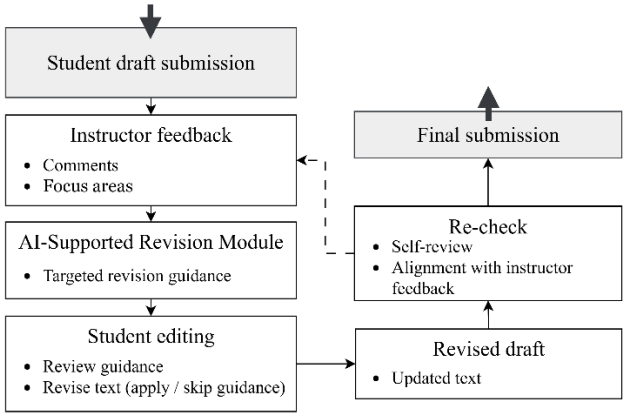

Figure 2: Workflow of the AI-Supported Revision Module Within the LMS Revision Cycle

The interface, shown in Figure 3, provided up to three suggestions for each section, allowing the user to either accept, skip, or evaluate its usefulness on a scale from one to five. Applying a suggestion indicated independent revision by the student; no automatic text replacement occurred. This interaction design was intended to support selective uptake of suggestions, ensuring active evaluation rather than passive acceptance.

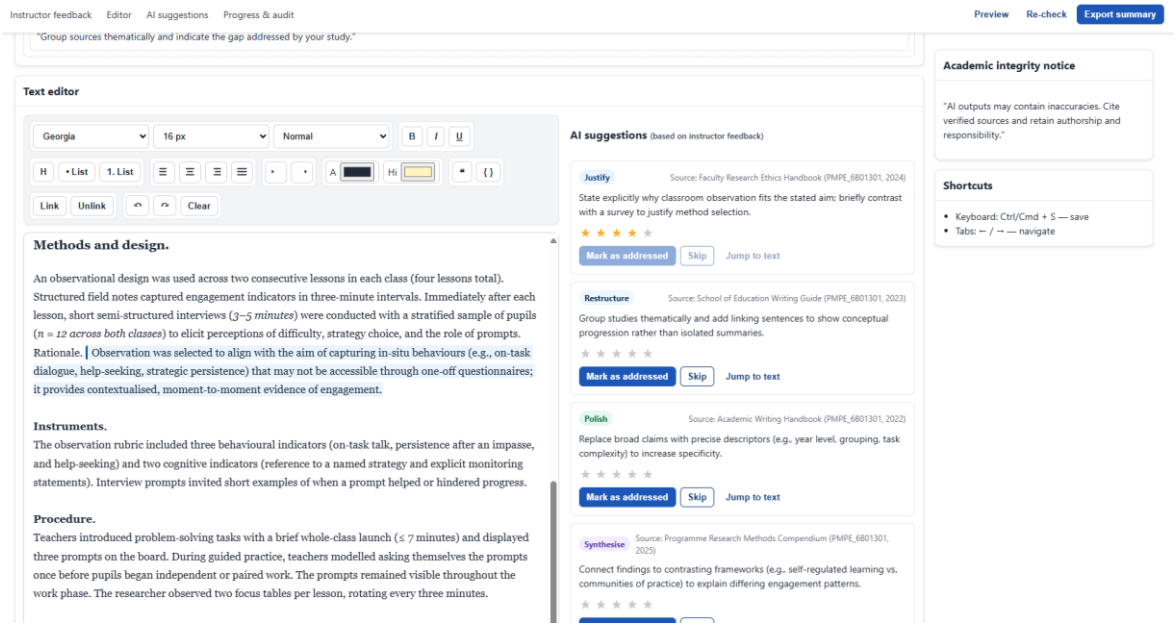

Figure 3: Student Interaction Interface and Suggestion Management in AIRM

In order to avoid cognitive overload, the number of suggestions that could be generated was restricted to three per section and determined by a threshold explicitly drawn from criteria specified within the rubric (Appendix 2). This threshold was operationalised via semantic similarity scoring between student text and rubric descriptors (embedding-based comparison), combined with rule-based heuristics (e.g., missing rationale, weak citation patterns, aim–method misalignment) to determine whether additional suggestions were required. For weaker segments, rule-based heuristics identified gaps (e.g., missing rationale, weak citation patterns, aim–method misalignment) to trigger mode-specific guidance. For example, the absence of an explicit methodological rationale triggered a "justify" suggestion, while disconnected paragraph transitions prompted a "restructure" recommendation. This ensured alignment between automated feedback and assessment standards.

Students retained full control over the revision process. Suggestions could be applied or skipped and optionally rated on a 1–5 usefulness scale. After revision, a report summarised text modifications and mode distribution. The four modes were aligned with key aspects of research writing: polish captured clarity and style, restructure targeted organisation in the literature review and methodological reasoning, justify strengthened methodological argumentation and data justification, while synthesise supported integration of sources in the literature review and interpretation. To safeguard integrity, each session began with a disclaimer reminding students to critically evaluate AI outputs and maintain authorship.

3.4 Data Collection

Multiple data sources were collected to evaluate the effects of the AI-Supported Revision Module across performance, process and perception. The core dataset consisted of 132 paired student submissions (draft and revised versions) from both cohorts.

All drafts and revisions were evaluated by two independent raters using a six-criterion rubric covering core dimensions of PBR. The raters were blinded to group allocation. Each of the 132 projects was assessed twice (draft and final), yielding 264 ratings. Inter-rater agreement was assessed with a two-way random-effects ICC (ICC(2,1), absolute agreement, single measures) with 95% confidence intervals. Overall reliability was good (ICC = .82, 95% CI [.76, .87]). Final scores were calculated as the mean of the two raters' evaluations.

Additional data were collected for the AI-assisted cohort. Module logs (N = 70 sessions) recorded suggestions generated, uptake (applied vs. skipped), percentage of text modified and changes across four modes (polish, restructure, justify, synthesise). In addition, 64 students completed voluntary surveys on benefits and limitations, and four instructors provided perspectives on feasibility and teaching implications. Survey participation was voluntary, which may introduce self-selection bias. Table 2 summarises all data sources.

Table 2: Overview of Data Sources

Data Source	Records / N	Purpose of Use
Draft–Revision Pairs	132 pairs	To measure improvement across six rubric criteria
Independent Rubric Scores	264 ratings	To ensure objective evaluation of drafts/revisions
Module Logs	70 sessions	To analyse interaction patterns (modes, % edits)
Student Surveys	64 forms	To capture perceptions of benefits and challenges
Instructor Surveys	4 forms	To assess feasibility and teaching implications

3.5 Data Analysis

Within each cohort, draft and revision scores were compared using paired t-tests across six rubric criteria. Between cohorts, differences in gain scores were analysed using independent t-tests and repeated-measures ANOVA. ANOVA tested the interaction between stage (draft vs. revision) and group (AI-assisted vs. traditional). All tests were two-tailed with the significance threshold set at $\alpha = .05$. Effect sizes (Cohen's d) were calculated using pooled Standard Deviations (SD) of draft and revision scores. For ANOVA, effect sizes are reported as partial eta squared (η^2_p), providing an indication of both statistical significance and practical importance. Assumptions of normality and homogeneity of variance were checked and met prior to analysis. All statistical analyses were conducted using JASP (version 0.95.4.0).

Qualitative data were analysed using thematic analysis to capture latent patterns and subjective experiences in student–AI interaction. Open-ended responses (n = 64 students; n = 4 instructors) were independently coded by two researchers using a hybrid framework combining deductive categories (polish, restructure, justify, synthesise) with inductive themes (e.g., cognitive load, feedback redundancy). Coding reliability was ensured through initial calibration and iterative consensus. Responses in Russian and Kazakh were translated and verified by bilingual researchers. Code frequencies and illustrative quotations are reported to support interpretation and triangulate quantitative findings.

The triangulation among rubric score improvements, log data, and qualitative themes yielded convergent and supplementary insights into how the module affected revision processes and results. The summary of the data collection process is presented in Figure 4.

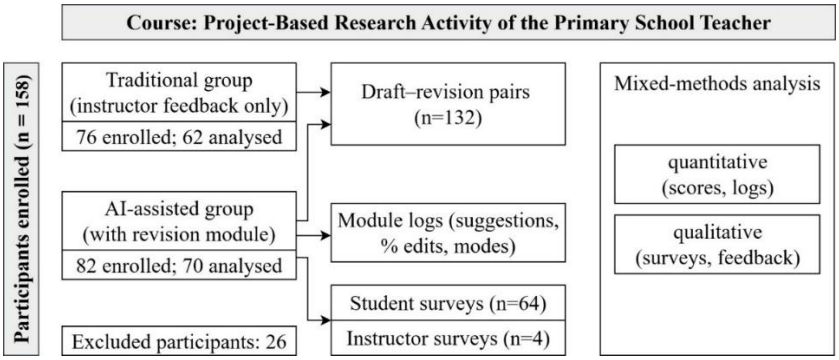

Figure 4: Study Design and Data Sources of the Research

3.6 Ethical Considerations

The study was conducted in a natural classroom environment without experimental manipulation beyond access to the AI-Supported Revision Module. All participation was voluntary and informed consent was obtained from every participant. Students could withdraw at any time without academic consequences, and participation did not affect course grades or formal assessment outcomes. No grades were assigned for the use of the module or participation in the study. All drafts and revisions were anonymised, and survey data were reported in aggregate form. The module itself was explicitly framed as a supportive tool rather than a replacement for student work. Students retained full control over revisions, including accepting or skipping suggestions, ensuring full authorship of the final text. Each session began with a disclaimer that AI outputs may be inaccurate and require critical evaluation. Students were instructed not to treat system outputs as authoritative sources. These measures ensured transparency, academic integrity and alignment with recognised ethical standards, including informed consent, data protection and voluntary participation.

4. Results and Findings

4.1 Quantitative Outcomes: Reliability and Overall Performance

The technical validity of the assessment was determined using the inter-rater reliability for 264 independent evaluations. A two-way random-effects intraclass correlation coefficient (ICC) for absolute agreement (single measures) demonstrated good reliability across the six criteria, with an overall ICC of 0.821 (95% CI [0.755, 0.869]).

Regarding overall student progress across the entire sample (N = 132), a paired samples t-test confirmed statistically significant improvements from draft to revision for all rubric criteria (Table 3). The largest effect sizes were observed in Literature Review (d = 0.58) and Methodology (d = 0.52), both indicating a medium-to-large impact. Improvements in analytical dimensions, such as Data Justification (d = 0.21) and Interpretation (d = 0.19), yielded smaller effect sizes, representing the baseline growth in higher-order research competencies (Mekheimer, 2025).

Table 3: Descriptive Statistics and Overall Growth from Draft to Revision (N = 132 Pairs)

Criterion	Draft Mean (SD)	Revision Mean (SD)	t (131)	Cohen's d	p
Problem Formulation	3.00 (1.35)	3.50 (1.64)	12.42	0.33	< .001
Literature Review	2.90 (1.12)	3.71 (1.61)	10.65	0.58	< .001
Methodology	2.80 (1.27)	3.61 (1.78)	10.44	0.52	< .001
Data Justification	3.10 (1.26)	3.40 (1.58)	2.20	0.21	.029
Interpretation	3.20 (1.29)	3.50 (1.85)	2.21	0.19	.029
Reflection	3.30 (1.39)	3.71 (1.48)	16.94	0.28	< .001

4.2 Comparative Analysis: AI-Assisted vs. Traditional Cohorts

To assess the added value of the module, a mixed-design repeated measures ANOVA was conducted to compare the performance of the AI-assisted group (n = 70) and the traditional group (n = 62). Descriptive statistics for both cohorts are presented in Table 4, including means and SD for each stage.

Table 4: Descriptive Statistics by Group and Stage (n = 132)

Group	Criterion	Draft Mean (SD)	Revision Mean (SD)	Gain (Δ)
Traditional	Problem Formulation	3.17 (1.32)	3.56 (1.64)	+0.39
	Literature Review	2.86 (1.19)	3.45 (1.73)	+0.59
	Methodology	3.03 (1.29)	3.62 (1.74)	+0.59
	Data Justification	3.13 (1.20)	3.43 (1.59)	+0.30
	Interpretation	3.36 (1.28)	3.61 (1.83)	+0.25
	Reflection	3.38 (1.36)	3.73 (1.45)	+0.35
AI-assisted	Problem Formulation	2.85 (1.37)	3.44 (1.65)	+0.59
	Literature Review	2.94 (1.05)	3.93 (1.47)	+0.99
	Methodology	2.60 (1.22)	3.59 (1.83)	+0.99
	Data Justification	3.07 (1.31)	3.37 (1.58)	+0.30
	Interpretation	3.06 (1.28)	3.41 (1.87)	+0.35
	Reflection	3.24 (1.42)	3.69 (1.52)	+0.45

The ANOVA results (Table 5) revealed a highly significant interaction effect between Stage and Group for Literature Review ($F(1, 130) = 7.506$, $p = .007$, with a partial eta squared of 0.055) and Methodology ($F(1, 130) = 7.197$, $p = .008$, with a partial eta squared of 0.052).

Table 5: Summary of Repeated Measures ANOVA (Interaction Effects: Stage x Group)

Criterion	SS	df	MS	F	p	η^2_p
Problem Formulation	0.696	1	0.696	4.718	.032	0.035
Literature Review	2.685	1	2.685	7.506	.007	0.055
Methodology	2.685	1	2.685	7.197	.008	0.052
Data Justification	0.000	1	0.000	0.000	.995	<.001
Interpretation	0.164	1	0.164	0.131	.718	0.001
Reflection	0.175	1	0.175	4.861	.029	0.036

Significant interaction effects were also found for Problem Formulation ($p = .032$) and Reflection ($p = .029$). However, no significant interaction was observed for Data Justification ($F(1, 130) < 0.001$) or Interpretation ($F(1, 130) = 0.131$, $p = .718$), where both cohorts demonstrated statistically equivalent progress. Figure 5 illustrates these interaction patterns, highlighting the steeper slope of improvement for the AI-assisted group in Literature Review and Methodology, contrasting with the parallel trajectories observed in analytical criteria.

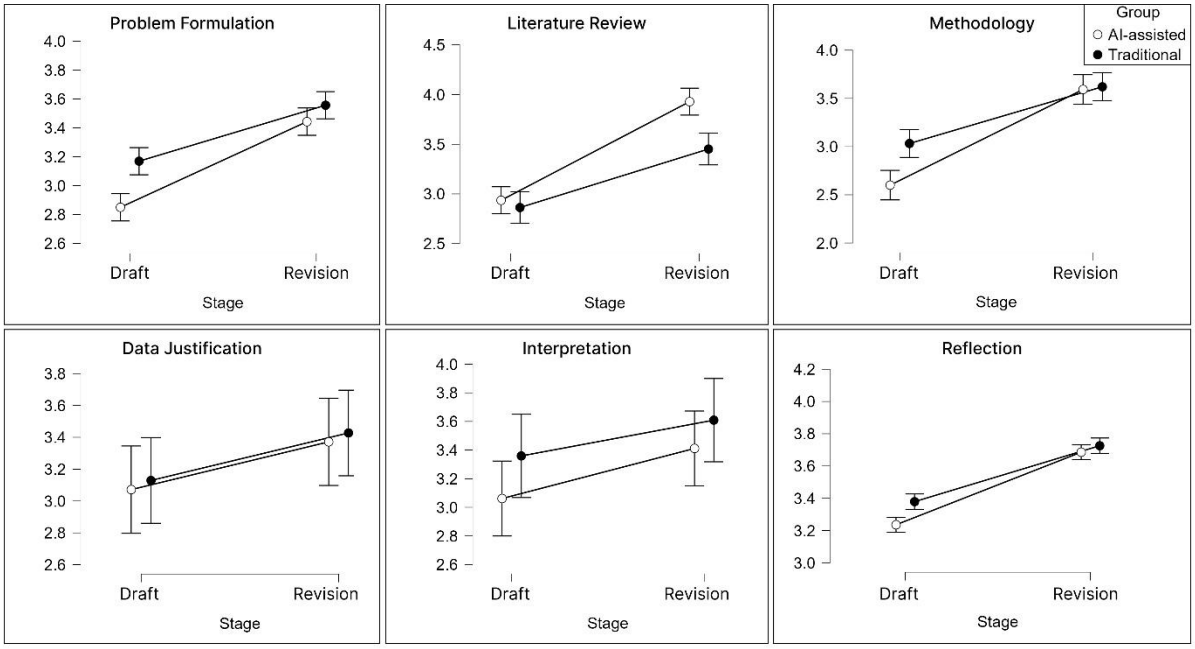

Figure 5: Interaction Effect Between Stage (Draft vs. Revision) and Condition (AI-Assisted vs. Traditional) on Mean Rubric Scores

4.3 Process Metrics: Student Interaction with AIRM

To understand the mechanisms behind the score improvements, interaction logs from the AI-assisted cohort ($n = 70$) were analysed. On average, the system generated 6.8 suggestions per report ($SD = 2.24$). A key finding was student agency: students demonstrated a selective approach, accepting 60.3% of recommendations ($M = 4.1$, $SD = 1.45$) and rejecting 39.7% ($M = 2.7$, $SD = 0.91$).

The system also captured utility ratings for each suggestion. Of the 476 generated suggestions, 412 (86.5%) received an optional rating on a 1–5 scale, yielding an overall mean of 3.9 ($SD = 0.8$). Perceived usefulness was highest for the Polish mode ($M = 4.2$, $SD = 0.6$), followed by Restructure ($M = 3.9$, $SD = 0.7$), Justify ($M = 3.6$, $SD = 0.9$), and Synthesise ($M = 3.4$, $SD = 1.1$). The descriptive statistics for these interaction metrics are summarised in Table 6.

Table 6: Interaction Metrics from AIRM System Logs (n = 70)

Metric	Mean (SD)	Range	Total
Suggestions Generated	6.8 (2.24)	3–12	476
Suggestions Accepted	4.1 (1.45)	2–8	287
Suggestions Rejected	2.7 (0.91)	1–5	189
Text Modified (Word Count)	281.4 (109.7)	125–519	19,700

The volume of revision was significantly higher in the AI-assisted group compared to the traditional cohort. Students using the module modified an average of 14.2% (SD = 5.02) of their draft text, whereas the traditional group, relying on peer and self-correction, modified only 8.5% (SD = 3.2) of their content. This difference suggests that the AI-assisted feedback acted as a more effective "nudge," prompting students to engage in more extensive rewriting, particularly in sections where the system identified structural or logical gaps. As shown in Figure 6, the modification rate was not uniform across the report sections. The highest percentages of modified text were observed in the Literature Review and Methodology sections, which directly correlates with the highest score gains reported in Section 4.2.

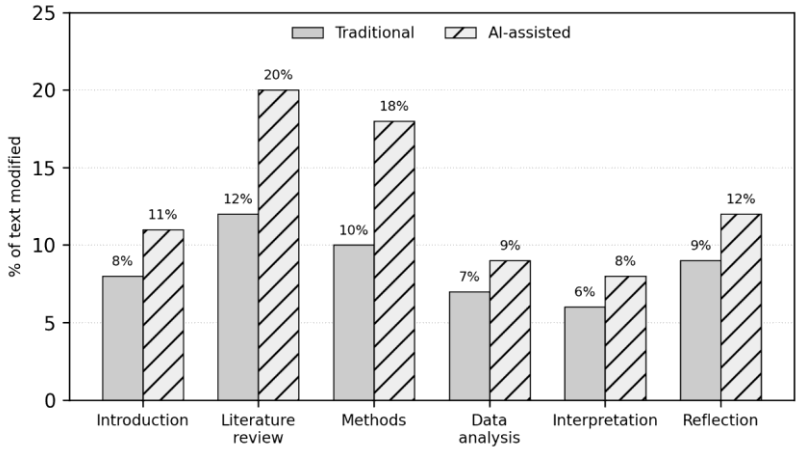

Figure 6: Percentage of Text Modified by Section and Condition

4.4 Qualitative Insights: Thematic Analysis of Feedback

To capture the subjective experiences of the participants, qualitative data from open-ended survey responses (N = 64 students) and instructor reflections (n = 4) were analysed. Thematic coding was performed by two independent researchers on a corpus of 118 individual comments. An initial inter-coder agreement of 88% was achieved, with discrepancies resolved through consensus. The analysis identified distinct patterns of engagement, with frequencies indicating the relative prominence of themes, categorised into Perceived Benefits and Reported Challenges (Table 7).

Table 7: Thematic Coding of Student and Instructor Feedback

Category	Theme	Frequency	Illustrative Quote	Source
Perceived benefits	Clarity and organisation	40	"The module helped me make my text clearer and more structured, especially in the introduction."	Student
	Metacognitive reflection	24	"I liked the final report because it showed me what I had actually changed, not just what I thought I changed."	Student
	Value of new perspectives	8	"The suggestion gave me a new idea for connecting literature."	Student
	Shift in instructor role	4	"The module freed me from correcting grammar, so I could focus on methodology with students."	Instructor
	Selective adoption of suggestions	12	"I skipped some advice because it did not fit my argument, but others were very useful."	Student

Category	Theme	Frequency	Illustrative Quote	Source
Reported Challenges	Cognitive demand	16	"Sometimes there were too many points at once and I did not know where to start."	Student
	Feedback redundancy	10	"This was similar to what the teacher already wrote, so I did not use it."	Student
	Critical evaluation requirement	3	"Students must learn not to trust the AI blindly, but to use it as support."	Instructor

Note: Frequency = number of mentions

Students consistently described the module as a tool for improving clarity and organisation, suggesting that it functioned primarily as a structural scaffold. As one student noted: "The AI didn't just fix my grammar; it highlighted where my methodology was fragmented, which made it easier to align my steps with the research aims." Many students emphasised the role of the revision report in supporting metacognitive awareness, particularly in recognising the extent of their own revisions. Another participant reflected: "Seeing the summary of my changes at the end was eye-opening; it made me realize how much I had actually improved the text compared to my first draft."

However, students also reported cognitive challenges, indicating that the volume of automated suggestions could at times hinder the revision process. One student remarked: "At times, receiving three different suggestions for one paragraph was overwhelming, and I struggled to decide which one to prioritize." Finally, while Critical Evaluation (n = 3) was less frequently mentioned in surveys, the qualitative evidence confirms that student agency was active and deliberate. This was best captured by a student who explained their selective adoption: "I had to be careful; I skipped suggestions that tried to change the meaning of my results, as the AI didn't know the specific classroom context of my project." These insights, aligned with the 40% rejection rate in Section 4.3, demonstrate that the module fostered a critical dialogue rather than a passive acceptance of technology.

5. Discussion

5.1 Summary of Key Findings

The results show that the AIRM greatly improved students' research writing through a more stringent revision strategy. According to the mixed-design ANOVA outcomes (Section 4.2), the treatment was found to be effective in achieving significant improvement in two important dimensions – Literature Review ($\eta^2_p = 0.055$, d = 0.58) and Methodology ($\eta^2_p = 0.052$, d = 0.52). These effects suggest that the module supported idea development within a rubric framework. As argued by Bjælde, Boud and Lindberg (2025), the most effective feedback activities are those that move beyond mere information delivery and instead facilitate a continuous "draft–feedback–resubmission" cycle. In our study, the module transformed instructor comments from static advice into actionable, iterative prompts. This is evidenced by the process metrics: the AI-assisted group modified 14.2% of their draft text, which is a nearly 70% increase compared to the 8.5% modification rate observed in the traditional cohort. This discrepancy confirms that the module acted as a "catalyst," preventing the "minimalist revision trap" where students change only minor surface elements while leaving the underlying conceptual flaws untouched.

Furthermore, the significant improvements in methodological reasoning support the findings of Sahlan (2025), who emphasizes that integrating tutor feedback with structured reflection enhances the pedagogical competence of pre-service teachers. By utilizing the Restructure and Synthesise modes, students were nudged to align their research aims with their methodological choices more precisely. Operating at the SAMR Modification level, the system redesigned revision tasks to foster deeper inquiry rather than merely substituting feedback.

The concentration of gains in structural and logical dimensions (Literature Review and Methodology) highlights a strategic alignment between the inherent strengths of LLMs and the procedural needs of novice researchers. These results also suggest that the AIRM effectively managed the "procedural load" of research writing. By providing a logical framework for structural synthesis, the module enabled students to overcome the cognitive barriers of organizing complex information, allowing them to produce more coherent narratives. This alignment between pedagogy and technology (TPACK) shows that the module's effectiveness comes from its integration into instructor-led feedback cycles.

5.2 Conceptual Boundaries and the Analytic Gap in AI Scaffolding

Results reveal an “analytic gap,” evidenced by non-significant performance gains in Data Justification and Interpretation across both cohorts. Unlike the structural improvements observed in other criteria, these analytical dimensions yielded no significant interaction effect ($p > .05$). This gap highlights a boundary between procedural support and higher-order synthesis, distinguishing AIRM’s scaffolded approach from authorship-blurring open tools.

The absence of growth in interpretation serves as empirical evidence against concerns regarding unreflective AI adoption or automated content generation. As highlighted by Tlili et al. (2023), the integration of generative tools often carries the risk of "hallucinations" that can undermine a student’s critical engagement. However, the plateau in these specific results demonstrates that the AIRM functioned strictly as a cognitive scaffold rather than a generative substitute. By restricting the model’s access to raw primary datasets, the system design ensured that students remained the sole intellectual owners of their findings. PBR requires situated knowledge in order to achieve interpretative results since there is no way for any generalized model to match the knowledge of a specific pedagogical situation existing in Kazakhstan. In this case, the "analytic gap" justifies the use of the module as an ethical tool because it successfully handles the "procedural load" required by the structure, and at the same time retains the "epistemic load" in meaning creation for the human researcher.

5.3 Selective Adoption and the Manifestation of Student Agency

The interaction logs reveal a complex pattern of engagement that contradicts the premise of unreflective AI adoption. The 39.7% rejection rate suggests that students acted as critical evaluators rather than passive recipients. This selective adoption aligns with the "active role" of students in the feedback process advocated by Bjælde, Boud and Lindberg (2025), where learning occurs through the deliberate evaluation and application of suggestions. Rejecting AI suggestions manifests evaluative judgment as a core research competence rather than scaffold failure.

The qualitative data presented in Table 7 further elucidates these thematic priorities. The high frequency of mentions regarding Clarity and Organisation (40 instances) compared to Justification and Argumentation (12 instances) reflects a hierarchical approach to revision. Students utilized the AIRM primarily to resolve "first-order" structural and linguistic barriers, thereby creating the necessary cognitive space to refine their internal arguments. Furthermore, the integration of AI guidance provides a distinct contrast to traditional peer assessment mechanisms. While Ayçiçek (2025) highlights the inherent variability and subjectivity often associated with peer-driven feedback, the AIRM offered a consistent methodological "grid." However, the 14.2% text modification rate indicates that this stability did not lead to standardisation; instead, it served as a catalyst for individualised rewriting. By positioning the student as the final authority over each suggested revision, the module ensured that the authorial voice remained intact. Consequently, the revision process was transformed into a critical dialogue, where the ability to discern and reject irrelevant advice became as pedagogically valuable as the adoption of helpful prompts.

5.4 Pedagogical Transitions and the Reconfiguration of Instructional Roles

The implementation of the AIRM facilitated a shift in instructional dynamics of the research-based course. Qualitative insights from the participating instructors indicated a transition from serving as primary providers of surface-level corrections to assuming roles as high-level methodological mentors. This redistribution of instructional effort was characterized by the delegation of repetitive structural and grammatical feedback to the AI module, which consequently permitted instructors to dedicate more time to complex, inquiry-driven dialogue.

This transition aligns with the principles of human-centered assessment articulated by Goode et al. (2026), where the focus of evaluation shifts from mere measurement toward the holistic support of student transitions and professional development. In this framework, the AI scaffold functioned as a preliminary layer of support, ensuring that drafts reached a certain level of structural maturity before human intervention. Consequently, the interaction between mentor and student was elevated, focusing on methodological nuances and pedagogical implications rather than basic clarity.

Furthermore, the integration of the module encouraged a more rigorous form of professional reflection. As suggested by Sahlan (2025), combining tutor feedback with structured revision tools is essential for enhancing the competence of pre-service teachers, as it necessitates a deliberate "debriefing" of one's own writing. The instructors observed that this model effectively addressed the logistical challenges of high-enrollment programs,

allowing for a scalable feedback mechanism that does not compromise the depth of individualised guidance. This pedagogical modification suggests that the value of AI in teacher education lies in its ability to optimize human expertise, ensuring that mentors intervene at the most critical stages of the inquiry process.

5.5 Research Limitations and Future Directions

This study has several limitations. First, it evaluates scaffolded performance rather than independent skill transfer, as no post-test was conducted after removing AIRM support. Second, the quasi-experimental design without random assignment introduces potential selection bias despite baseline equivalence. Third, the findings are context-specific, as the study was conducted within a single pedagogical university in Kazakhstan with pre-service primary teachers, in a multilingual setting where academic writing is often performed in English as a Foreign Language (EFL). These contextual, cultural, and linguistic factors may have influenced revision patterns and the effectiveness of AI-supported feedback, limiting generalisability. Fourth, a possible novelty effect may have increased student engagement. Fifth, the results depend on the specific configuration of the LLaMA-2-7B model and prompt design, which may not transfer to other systems. Finally, despite implemented safeguards, the risk of cognitive offloading and over-reliance on AI remains (Tlili et al., 2023).

Future research should examine long-term skill transfer through longitudinal designs and delayed post-tests without AI support. Replication across diverse cultural, linguistic and disciplinary contexts is needed to test generalisability. Further work should also explore adaptive prompting strategies to better support higher-order analytical reasoning and address the identified "analytic gap," while maintaining student agency and authorship.

6. Conclusion

The current research shows that it is possible to deploy the AIRM framework in a research course, thus providing an efficient example of feedback provision in the context of teacher education. By connecting a locally deployed language model with institutional rubrics and instructor intentions, this research goes further than the chatbot paradigm by creating a feedback-oriented scaffold that ensures effective revision without any creation of output text. This module has positively impacted the structure and methodology of students' essays, leaving the limits of human authorship untouched.

In relation to the research questions, the findings indicate that AIRM functioned as a structured scaffold within the revision cycle of project-based research (RQ1), leading to the most substantial improvements in literature review and methodological alignment (RQ2). At the same time, both students and instructors perceived the module as a supportive tool that enhances clarity and feedback processes while preserving learner agency and critical judgement (RQ3).

Evaluation suggests that although AIRM acts as a potent instigator of meaningful change, the extent of influence is dictated by the dynamics of a purposeful synergy and not automation per se. With the development of the "analytical gap" and the high rejection rate of AI recommendations, it is evident that students acted as the primary agents of change in the process, performing an evaluative function of judgment and "feedback literacy." Evaluation lends credence to the human-in-the-loop model, where instructional load is partially transferred to AI, enabling educators to focus on higher-level teaching processes while informing the design of scalable, human-centred feedback systems in teacher education and beyond.

AI Statement: During the preparation of this manuscript, the authors used Grammarly for language editing and proofreading purposes only. No generative AI tool was used for content generation, data analysis, interpretation, or reference generation.

Ethics Statement: The study was conducted in accordance with institutional ethical guidelines. Formal ethical approval was not required as the study involved normal educational practice, anonymised data, and no impact on grading or student progression.

Conflict of Interest: The authors declare no conflicts of interest.

References

Ayçiçek, B. (2025). Peer assessment analysis via the many-facet Rasch model. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 21*(2), pp.631–646. https://doi.org/10.17860/mersinefd.1642676

Bhullar, P.S., Joshi, M. and Chugh, R. (2024). ChatGPT in higher education: A synthesis of the literature and a future research agenda. *Education and Information Technologies*, *29*, pp.21501–21522. https://doi.org/10.1007/s10639-024-12723-x

Bjælde, O.E., Boud, D. and Lindberg, A.B. (2025). Designing feedback activities to help low-performing students. *Active Learning in Higher Education, 26*(1), pp.123–137. https://doi.org/10.1177/14697874231212820

Carless, D. and Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), pp.1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Chelli, M., Descamps, J., Lavoué, V., Trojani, C., Azar, M., Deckert, M., Raynier, J., Clowez, G., Boileau, P. and Ruetsch-Chelli, C. (2024). Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, *26*, e53164. https://doi.org/10.2196/53164

Costley, J., Zhang, H., Courtney, M., Shulgina, G., Baldwin, M. and Fanguy, M. (2023). Peer editing using shared online documents: The effects of comments and track changes on student L2 academic writing quality. *Computer Assisted Language Learning, 38*(4), pp.865–891. https://doi.org/10.1080/09588221.2023.2233573

Dean, C.G.P., Grossman, P., Enumah, L., Herrmann, Z. and Schneider Kavanagh, S. (2023). Core practices for project-based learning: Learning from experienced practitioners in the United States. *Teaching and Teacher Education, 133*, 104275. https://doi.org/10.1016/j.tate.2023.104275

Deng, R., Jiang, M., Yu, X., Lu, Y. and Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Eysenbach, G. (2023). The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Medical Education, 9*, e46885. https://doi.org/10.2196/46885

Gao, J. and Chen, W. (2024). Developing culturally-situated student feedback literacy through multi-peer feedback giving: An online community-based approach. *Language Awareness, 34*(1), pp.19–42. https://doi.org/10.1080/09658416.2024.2321894

Goode, E., Valentine, J., Krautloher, A. and Anand, P. (2026). Assessment for student transitions and success: A scoping review of assessment principles in Australian enabling education. *Higher Education Research & Development, 45*(3), pp.554-571. https://doi.org/10.1080/07294360.2025.2543409

Hamilton, E.R., Rosenberg, J.M. and Akcaoglu, M. (2016). The substitution augmentation modification redefinition (SAMR) model: a critical review and suggestions for its use. *TechTrends, 60*(5), pp.433–441. https://doi.org/10.1007/s11528-016-0091-y

Hanafi, I., Kheder, K., Sabouni, R., Gorra Al Nafouri, M., Hanafi, B., Alsalkini, M., Kenjrawi, Y., Albkhetan, H. and Alhalabi, M. (2025). Improving academic writing in a low-resource country: A systematic examination of online peer-run training. *Teaching and Learning in Medicine, 37*(3), pp.388–402. https://doi.org/10.1080/10401334.2024.2332890

Helle, L., Tynjälä, P. and Olkinuora, E. (2006). Project-based learning in post-secondary education – Theory, practice and rubber sling shots. *Higher Education, 51*(2), pp.287–314. https://doi.org/10.1007/s10734-004-6386-5

Jiménez Sierra, Á.A., Ortega Iglesias, J.M., Cabero-Almenara, J. and Palacios-Rodríguez, A. (2023). Development of the teacher's technological pedagogical content knowledge (TPACK) from the Lesson Study: A systematic review. *Frontiers in Education*, *8*, 1078913. https://doi.org/10.3389/feduc.2023.1078913

Kim, J., Yu, S., Detrick, R. and Li, N. (2025). Exploring students' perspectives on generative AI-assisted academic writing. *Education and Information Technologies, 30*(1), pp.1265–1300. https://doi.org/10.1007/s10639-024-12878-7

Kokotsaki, D., Menzies, V. and Wiggins, A. (2016). Project-based learning: A review of the literature. *Improving Schools, 19*(3), pp.267–277. https://doi.org/10.1177/1365480216659733

Koo, M. (2023). The importance of proper use of ChatGPT in medical writing. *Radiology, 307*(3), e230312. https://doi.org/10.1148/radiol.230312

Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability, 15*, 5614. https://doi.org/10.3390/su15075614

Lim, W.M., Gunasekara, A., Pallant, J.L., Pallant, J.I. and Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education, 21*(2), 100790. https://doi.org/10.1016/j.ijme.2023.100790

Lineback, J.E. and Holbrook, E. (2023). Engaging in a collaborative space: Exploring the substance and impact of peer review conversations. *Assessment & Evaluation in Higher Education, 49*(4), pp.556–571. https://doi.org/10.1080/02602938.2023.2290978

Lo, C.K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*, 410. https://doi.org/10.3390/educsci13040410

Lu, Q., Yao, Y. and Zhu, X. (2023). The relationship between peer feedback features and revision sources mediated by feedback acceptance: The effect on undergraduate students' writing performance. *Assessing Writing, 56*, 100725. https://doi.org/10.1016/j.asw.2023.100725

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*, 9. https://doi.org/10.1186/s40561-024-00295-9

Mishra, P. and Koehler, M.J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), pp.1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Mekheimer, M. (2025). Generative AI-assisted feedback and EFL writing: a study on proficiency, revision frequency and writing quality. *Discover Education, 4*, 170. https://doi.org/10.1007/s44217-025-00602-7

Nautiyal, R., Albrecht, J.N. and Nautiyal, A. (2023). ChatGPT and tourism academia. *Annals of Tourism Research, 99*, 103544. https://doi.org/10.1016/j.annals.2023.103544

Niloy, A.C., Akter, S., Sultana, N., Sultana, J. and Rahman, S.I.U. (2024). Is ChatGPT a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning, 40*(2), pp.919–930. https://doi.org/10.1111/jcal.12929

O'Dea, X. (2024). Generative AI: Is it a paradigm shift for higher education? *Studies in Higher Education, 49*(5), pp.811–816. https://doi.org/10.1080/03075079.2024.2332944

Polakova, P. and Ivenz, P. (2024). The impact of ChatGPT feedback on the development of EFL students' writing skills. *Cogent Education, 11*(1), 2410101. https://doi.org/10.1080/2331186X.2024.2410101

Rashidi, H.H., Fennell, B.D., Albahra, S., Hu, B. and Gorbett, T. (2023). The ChatGPT conundrum: Human-generated scientific manuscripts misidentified as AI creations by AI text detection tool. *Journal of Pathology Informatics, 14*, 100342. https://doi.org/10.1016/j.jpi.2023.100342

Sahlan, M. (2025). Integrating self-assessment, peer assessment, and tutor feedback with structured debriefing to enhance pre-service teachers' teaching competence: Empirical evidence from Indonesia. *Qualitative Research Journal*, pp.1–17. https://doi.org/10.1108/QRJ-05-2025-0184

Tlili, A., Shehata, B., Adarkwah, M.A., Bozkurt, A., Hickey, D.T., Huang, R. and Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*(1), 15. https://doi.org/10.1186/s40561-023-00237-x

Tsybulsky, D. and Muchnik-Rozanov, Y. (2023). The contribution of a project-based learning course, designed as a pedagogy of practice, to the development of preservice teachers' professional identity. *Teaching and Teacher Education, 124*, 104020. https://doi.org/10.1016/j.tate.2023.104020

Walters, W.H. and Wilder, E.I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports, 13*, 14045. https://doi.org/10.1038/s41598-023-41032-5

Wei, Y. and Liu, D. (2024). Incorporating peer feedback in academic writing: A systematic review of benefits and challenges. *Frontiers in Psychology, 15*, 1506725. https://doi.org/10.3389/fpsyg.2024.1506725

Yu, H. and Xie, Q. (2025). Generative AI vs. teachers: Feedback quality, feedback uptake, and revision. *Language Teaching Research Quarterly, 47*, pp.113–137. https://doi.org/10.32038/ltrq.2025.47.07

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), 39. https://doi.org/10.1186/s41239-019-0171-0

Appendix 1

All prompts were executed deterministically within the same template structure to ensure consistency of outputs across users and sessions.

Table 8: Prompt Templates for AI-Supported Revision Guidance

Prompt Type	Prompt Template
Orchestrator	Task: Analyse student text in relation to instructor feedback and supporting materials, and generate structured revision guidance. Inputs: text = [TEXT]; feedback = [TEXT]; materials = [INSTITUTIONAL MATERIALS] Instructions: identify relevant segments; classify issues based on rubric-aligned criteria (clarity, coherence, argumentation, synthesis); map each issue to a revision mode (polish / restructure / justify / synthesise); generate targeted guidance per segment. Output (JSON): { "segments": [{ "text_span": "...", "mode": "...", "issue": "...", "suggestion": "..."}] } Constraints: no full-text rewriting; guidance only; strongly align with feedback and materials. Example output: { "segments": [{ "text_span": "…", "mode": "justify", "issue": "claim not supported", "suggestion": "add empirical reference linking to aim" }] }
Restructure	Task: Analyse text structure using feedback. Identify issues in flow and organisation. Suggest how to reorder or connect ideas. Constraints: no rewriting; concise guidance only.
Justify	Evaluate strength of argumentation. Identify unsupported claims or weak reasoning. Suggest how to strengthen justification and link to aims. Constraints: no rewriting; no full sentences for insertion.

Appendix 2

Table 9: Rubric for Evaluating Project-Based Research Reports

Criterion	1 – Minimal	2 – Emerging	3 – Developing	4 – Proficient	5 – Advanced
Problem Formulation	Problem absent or incomprehensible	Problem vague, aim unclear	Problem identifiable but	Problem clear and mostly	Problem precise, fully aligned with methodology,

Criterion	1 – Minimal	2 – Emerging	3 – Developing	4 – Proficient	5 – Advanced
			partly aligned with methods	aligned with aims	demonstrates originality
Literature Review	Fewer than 3 references; no academic sources	3–5 sources; descriptive, weak relevance	6–8 sources; some organisation, limited synthesis	9–12 sources; partial synthesis, mostly relevant and recent	≥ 13 sources; strong synthesis (>60% thematic discussion), clear gap identified
Methodology	Method missing or irrelevant	Method present but described superficially	Method described with some rationale, partial alignment	Method explained clearly, rationale present, aligned with aims	Method rigorously explained, rationale well supported, fully aligned with aims
Data Justification	Data absent or irrelevant	Data mentioned without justification	Data described but weakly linked to aims	Data appropriate and linked to aims with adequate justification	Data fully justified, strongly linked to aims, supported by credible references
Interpretation	No interpretation beyond restating results	Minimal interpretation, weak link to aims	Interpretation present but descriptive	Interpretation clear, with critical linkage to aims	Interpretation comprehensive, critical and connected to broader literature
Reflection	No reflection on process or outcomes	Minimal reflection, superficial	Reflection present but limited in depth	Reflection balanced, covers strengths and weaknesses	Reflection extensive, critical and linked to professional growth

