Agents in the Classroom: A Multi-Agent AI Environment for Developing Critical Thinking, Digital Competence, and Professional Readiness among Pre-Service Teachers

Nahed F. Abdel-Maksoud

Damietta University, Damietta, Egypt
nahed@du.edu.eg

Abstract

This study investigated the effect of a Multi-Agent AI-based learning environment on pre-service teachers' critical thinking, digital competence, and professional readiness for teaching. A randomized pretest–posttest controlled experimental design was employed. The participants were 100 fourth-year pre-service teachers enrolled in the General Education Program at the Faculty of Education, Damietta University, during the first semester of the 2022/2023 academic year. Following stratified random selection, participants were randomly allocated to an experimental group that used the Multi-Agent AI-based learning environment (n = 50) or a control group that studied the same Field Training course content through a conventional Moodle environment (n = 50). The 10-week intervention was delivered entirely online from October 1 to December 3, 2022, with one three-hour session each week. The experimental environment was implemented as a custom Moodle plugin connected to the OpenAI Completions API using the text-davinci-002 model. Its coordinated agents provided differentiated support for pedagogical planning, critical analysis, digital production, reflective practice, and professional decision-making. Data were collected using a critical-thinking situational test, a digital-competence performance test, a professional-readiness scale, a professional-readiness situational judgment test, a digital instructional product rubric, an agent-interaction log rubric, and an environment usability, trust, and satisfaction questionnaire. After controlling for pretest scores, the experimental group significantly outperformed the control group in critical thinking, $F(1, 97) = 99.22$, $p < .001$, $\eta^2_p = .506$; digital competence, $F(1, 97) = 71.99$, $p < .001$, $\eta^2_p = .426$; professional readiness measured by the scale, $F(1, 97) = 156.08$, $p < .001$, $\eta^2_p = .617$; and situationally assessed professional readiness, $F(1, 97) = 115.01$, $p < .001$, $\eta^2_p = .542$. The experimental group also achieved substantially higher digital product scores. These findings indicate that pedagogically differentiated and coordinated AI agents can strengthen the cognitive, digital, and professional dimensions of pre-service teacher preparation when embedded in authentic field-training activities.

Keywords: multi-agent artificial intelligence; pre-service teachers; critical thinking; digital competence; professional readiness; field training; Moodle.

1. Introduction

Teacher preparation is increasingly expected to equip future teachers for professional practice in learning environments shaped by digital transformation, rapidly expanding information resources, and increasingly intelligent educational technologies. This expectation extends beyond the operational use of digital tools. Pre-service teachers must be able to evaluate information critically, select technologies in accordance with pedagogical purposes, respond appropriately to complex classroom situations, and make professionally defensible decisions. These capabilities have become central to the quality of teacher preparation because contemporary teaching requires the integration of pedagogical judgment, disciplinary knowledge, digital competence, and reflective professional action rather than the isolated mastery of technological procedures (Darling-Hammond, 2006).

Digital transformation has also changed the nature of the learning opportunities available in teacher education. Online learning environments can provide flexible access to content, communication, assessment, and instructional resources; nevertheless, access alone does not guarantee intellectually demanding or professionally meaningful learning. Effective technology integration depends on how learning activities are designed, how learners are supported during complex tasks, and whether technology is connected to authentic pedagogical problems. Research on pre-service teacher education has consequently emphasized the need to integrate technology systematically across coursework, modeling, practice, reflection, and assessment instead of treating digital competence as a separate or predominantly technical component of teacher preparation (Tondeur et al., 2012).

Artificial intelligence offers a further level of educational support by enabling learning systems to interpret learner inputs, generate adaptive responses, automate selected forms of feedback, and provide differentiated assistance. In higher education, artificial intelligence has been applied to profiling and prediction, assessment and evaluation, adaptive systems, and intelligent tutoring. However, a systematic review by Zawacki-Richter et al. (2019) found that much of the higher-education literature was driven by technical perspectives and devoted comparatively limited attention to educators and pedagogical design. This imbalance is particularly important in teacher education, where the educational value of an AI system should be judged not only by the accuracy or speed of its responses but also by its capacity to support reasoning, professional reflection, and the responsible exercise of human judgment.

The educational use of AI therefore requires a pedagogically controlled design. AI-generated responses may offer immediate and context-sensitive assistance, but their instructional value depends on the nature of the prompts, the quality of the feedback, the learner's ability to evaluate the generated content, and the relationship between the technological support and the intended learning outcomes. UNESCO emphasized that AI should be used to strengthen human capabilities and advance equitable, inclusive, and human-centered education rather than displace teacher agency or reduce learning to automated decision-making (Miao et al., 2021). In teacher preparation, this principle means that AI should stimulate analysis and reflection while leaving responsibility for interpreting evidence and making professional decisions with the pre-service teacher.

A single intelligent tutor or conversational interface may provide broad assistance, yet a single undifferentiated source of support can obscure the distinct cognitive and professional demands embedded in complex learning tasks. Planning a lesson, evaluating the credibility of information, selecting a digital resource, reviewing an instructional product, and reflecting on a classroom problem require different forms of guidance. A multi-agent architecture provides a potentially stronger design by distributing these responsibilities across several specialized agents. In a multi-agent system, autonomous or semi-autonomous components operate within a shared environment and interact or coordinate their actions to address tasks that may be difficult to manage through one general-purpose component (Wooldridge, 2009).

In the present study, the multi-agent concept was operationalized pedagogically rather than as an unrestricted simulation of independent human-like entities. The environment used separate role-specific prompt templates within a custom Moodle plugin. Each agent was assigned a defined instructional function and a controlled feedback strategy. The agents supported pedagogical planning, critical questioning, digital production, professional reflection, and performance improvement, while a coordinating component directed requests toward the relevant type of assistance. This functional differentiation was intended to make AI support more transparent, task-specific, and aligned with the learning outcomes of the Field Training course.

Field training provides an especially appropriate context for examining such an environment because it represents the point at which pre-service teachers are expected to connect theoretical knowledge with the practical demands of teaching. Within field-training activities, students analyze learners and instructional contexts, formulate objectives, prepare lesson plans, select teaching strategies, develop resources, evaluate classroom situations, and reflect on their performance. These tasks cannot be performed adequately through procedural knowledge alone; they require critical interpretation, digital production, and context-sensitive professional judgment. Professional learning experiences are most valuable when they allow prospective teachers to engage with authentic problems of practice and reconsider their decisions in light of evidence and feedback (Korthagen, 2010).

Critical thinking constitutes one of the principal cognitive foundations of this process. It involves purposeful and self-regulatory judgment expressed through interpretation, analysis, evaluation, inference, explanation, and the monitoring of one's reasoning (Facione, 1990). In teacher preparation, critical thinking enables students to distinguish evidence from unsupported assertions, identify assumptions, evaluate alternative teaching responses, and justify decisions in complex situations. A Multi-Agent AI-based learning environment may contribute to this capability when agents pose competing interpretations, request justification, identify missing evidence, or challenge an initially accepted response. Its value would not arise from supplying a final answer, but from structuring a dialogue in which students must inspect, compare, revise, and defend their reasoning.

Digital competence represents a second essential outcome. It includes the confident, critical, responsible, and pedagogically purposeful use of digital technologies. For educators, digital competence encompasses professional engagement, the selection and creation of digital resources, technology-supported teaching and learning, digital assessment, learner empowerment, and the facilitation of learners' own digital competence (Redecker, 2017). Consequently, digital competence in this study was treated as observable performance rather than as perceived confidence alone. Students were required to perform relevant digital tasks and create an instructional product suitable for use in a teaching context.

Professional readiness for teaching constitutes the third outcome and integrates knowledge, skills, dispositions, judgment, and perceived capacity to assume authentic teaching responsibilities. It is reflected in a pre-service teacher's ability to plan instruction, respond to learner diversity, manage pedagogical challenges, employ assessment evidence, behave responsibly, and reflect on performance. Readiness is broader than academic achievement because a student may understand educational concepts without being able to apply them appropriately in a realistic professional situation. For this reason, the present study assessed professional readiness through both a multidimensional self-report scale and a situational judgment test, allowing perceived readiness to be considered alongside context-sensitive professional decision-making.

Although these three outcomes are conceptually distinct, they are functionally interdependent. A pre-service teacher cannot employ digital technologies responsibly without critically evaluating their relevance, accuracy, accessibility, and pedagogical value. Similarly, professional readiness increasingly involves the ability to select and use digital resources while retaining responsibility for instructional decisions. Falloon (2020) argued that teachers' digital competence should be understood comprehensively, incorporating technical proficiency, pedagogical application, ethical awareness, and professional judgment. Accordingly, an intervention that addresses these outcomes simultaneously may correspond more closely to the integrated character

of contemporary teaching than separate forms of training directed toward individual skills.

The existing literature before the end of 2022 provided substantial evidence of growing interest in AI-supported education, but several limitations remained. Research frequently examined prediction, automated assessment, personalization, or intelligent tutoring without adequately situating AI within authentic teacher-preparation practice. It also commonly focused on a single system function or a general learner outcome rather than examining whether differentiated AI support could simultaneously influence critical thinking, applied digital competence, and readiness for teaching. Moreover, the limited participation of educators in AI-in-higher-education research raised questions about the alignment of technological capabilities with pedagogical needs (Zawacki-Richter et al., 2019). The use of coordinated, pedagogically specialized agents in an Arabic-language field-training context therefore represented an insufficiently examined area.

A further methodological need concerned the measurement of outcomes. Reliance exclusively on self-report measures may produce an incomplete account of an AI-supported intervention because favorable perceptions do not necessarily demonstrate improved performance. The present study addressed this issue through a multidimensional assessment strategy that combined situational tests, a performance test, a professional-readiness scale, independent evaluation of digital products, interaction-log analysis, and an evaluation of usability, trust, and satisfaction. This approach made it possible to examine not only what students believed about their readiness but also how they performed and responded to professionally relevant tasks.

Against this background, the study investigated whether embedding specialized and coordinated AI agents within an online Field Training course could produce measurable development in pre-service teachers' critical thinking, digital competence, and professional readiness for teaching. The study also examined the quality of the digital instructional products produced by students and used interaction-log and user-evaluation data to provide a fuller account of how the experimental environment functioned.

1.1 Problem Statement

Fourth-year pre-service teachers are approaching entry into professional teaching and are therefore expected to demonstrate the capacity to analyze teaching situations critically, employ digital technologies competently, and make informed professional decisions. Conventional online learning environments can provide content, assignments, communication, and assessment, but they may not consistently offer the differentiated, immediate, and role-specific support required while students complete complex field-training tasks. General feedback may be insufficient when one student requires assistance with lesson planning, another needs critical questioning, and another needs technical or evaluative support for a digital product.

The problem addressed by this study was therefore the need for an empirically tested learning environment capable of coordinating multiple forms of AI support around the integrated demands of teacher preparation. More specifically, it remained unclear whether a Multi-Agent AI-based learning environment embedded in an online Field Training course would produce greater development than a conventional Moodle environment in pre-service teachers' critical thinking, digital competence, and professional readiness for teaching.

The study consequently sought to answer the following principal question:

What is the effect of a Multi-Agent AI-based learning environment on the development of critical thinking, digital competence, and professional readiness for teaching among fourth-year pre-service teachers at the Faculty of Education, Damietta University?

1.2 Purpose of the Study

The primary purpose of the study was to determine the effect of a Multi-Agent AI-based learning

environment, compared with a conventional Moodle learning environment, on:

1. critical thinking among pre-service teachers;
2. performance-based digital competence;
3. self-reported professional readiness for teaching;
4. situationally assessed professional readiness for teaching; and
5. the quality of students' digital instructional products.

The study additionally sought to describe students' interactions with the specialized AI agents and evaluate the perceived usability, trustworthiness, and satisfaction associated with the experimental environment.

1.3 Research Questions

1. Does participation in the Multi-Agent AI-based learning environment significantly improve pre-service teachers' critical thinking after controlling for their pretest scores?
2. Does participation in the Multi-Agent AI-based learning environment significantly improve pre-service teachers' digital competence after controlling for their pretest scores?
3. Does participation in the Multi-Agent AI-based learning environment significantly improve pre-service teachers' professional readiness for teaching after controlling for their pretest scores?
4. Does participation in the Multi-Agent AI-based learning environment significantly improve pre-service teachers' performance in professional teaching situations after controlling for their pretest scores?
5. Do students in the experimental group produce higher-quality digital instructional products than students in the control group?
6. What patterns characterize experimental-group students' interactions with the Multi-Agent AI system?
7. How do experimental-group students evaluate the usability, trustworthiness, and overall satisfaction of the environment?

1.4 Research Hypotheses

H1. After controlling for pretest scores, the experimental group will achieve significantly higher posttest critical-thinking scores than the control group.

H2. After controlling for pretest scores, the experimental group will achieve significantly higher posttest digital-competence scores than the control group.

H3. After controlling for pretest scores, the experimental group will achieve significantly higher posttest professional-readiness scale scores than the control group.

H4. After controlling for pretest scores, the experimental group will achieve significantly higher posttest professional-readiness situational judgment scores than the control group.

H5. The experimental group will achieve significantly higher digital instructional product scores than the control group.

1.5 Significance of the Study

The study contributes theoretically by connecting multi-agent instructional support with three interdependent dimensions of teacher preparation: critical thinking, digital competence, and professional readiness. Rather than conceptualizing AI as a single tutoring or content-generation tool, it examines a differentiated architecture in which multiple agents perform pedagogically bounded roles within a shared learning environment.

Methodologically, the study employs complementary measurement approaches, including situational judgment, performance assessment, self-report, product evaluation, and digital interaction logs. This design reduces dependence on a single source of evidence and allows the intervention's effects to be evaluated across perceived capability, demonstrated performance, and recorded engagement.

Practically, the study presents an implementable model for integrating Multi-Agent AI into Moodle without replacing the instructor or transferring professional judgment to the technology. The environment may inform the design of field-training courses and other practice-oriented teacher-education experiences that require timely, differentiated, and scalable support.

2. Theoretical Background and Related Literature

2.1 Multi-Agent Artificial Intelligence in Education

Artificial intelligence in education encompasses systems that perform functions commonly associated with human reasoning, such as interpreting learner responses, identifying patterns, generating feedback, recommending learning resources, and adapting instructional support. Its educational significance does not derive from automation alone, but from the possibility of providing timely assistance that responds to the learner's task, progress, and demonstrated needs. Intelligent tutoring research has shown that technology-supported guidance can improve learning when it includes appropriate task selection, targeted feedback, and opportunities for learners to revise their understanding (Ma et al., 2014).

A Multi-Agent AI-based learning environment extends this principle by distributing instructional responsibilities among several specialized agents. Each agent operates according to a defined role, while coordination mechanisms maintain coherence among their actions. This differs from a single general-purpose assistant that responds to all requests through the same broad interaction pattern. Role differentiation enables one agent to focus on pedagogical planning, another on critical questioning, another on digital production, and another on reflective professional evaluation. The architecture is therefore suited to learning tasks that involve several interconnected but cognitively distinct forms of support.

From a pedagogical perspective, the importance of specialization lies in the relationship between the learner's immediate need and the form of assistance provided. Feedback on a lesson plan requires attention to the alignment of objectives, activities, and assessment. Evaluating an argument requires questions about evidence, assumptions, and alternative interpretations. Reviewing a digital instructional product requires criteria concerning usability, accessibility, technical quality, and pedagogical value. Combining all these functions in an undifferentiated response may increase cognitive complexity and reduce the clarity of the guidance. Specialized agents can instead provide bounded and task-relevant feedback while the coordinating layer preserves continuity across the learning process.

The environment developed for this study did not treat the agents as independent authorities capable of making final professional decisions. They operated as instructional supports whose responses were subject to evaluation by the student. This distinction is important because educational AI can produce incomplete, inaccurate, or contextually inappropriate outputs. Human oversight must therefore remain central, particularly when the learner is preparing for a profession involving responsibility for other individuals. Baker and Hawn (2022) emphasized that the consequences of educational algorithms depend on design choices, data, context, and patterns of use. Accordingly, the agents in this study were designed to promote analysis and revision rather than encourage passive acceptance of generated responses.

2.2 Pedagogical Foundations of the Environment

2.2.1 Social Constructivist Learning

The environment was informed by the view that knowledge develops through active engagement, dialogue, interpretation, and the progressive reconstruction of understanding. Learning support is most useful when it assists learners with tasks that they cannot yet complete independently but can accomplish through appropriately structured guidance. The concept of the zone of proximal

development provides a relevant explanation for adaptive assistance: support should respond to the learner's current capability and be directed toward increasingly independent performance (Vygotsky, 1978).

Within the experimental environment, AI-agent interaction functioned as a form of digitally mediated scaffolding. Students remained responsible for producing lesson plans, analyzing situations, designing resources, and revising their work. The agents provided questions, prompts, explanations, alternatives, and criterion-referenced observations. Consequently, the learning process was not reduced to receiving generated products; it required students to interpret the assistance and apply it to their own professional tasks.

2.2.2 Cognitive Apprenticeship

Cognitive apprenticeship explains professional learning as a process through which normally hidden forms of expert thinking are made visible to the learner. Modeling, coaching, scaffolding, articulation, reflection, and exploration enable novices to understand not only what an expert does but also the reasoning underlying professional action (Collins et al., 1989). This principle is particularly relevant to field training because many teaching decisions depend on tacit judgments that may not be apparent in formal descriptions of teaching methods.

The agents were therefore configured to externalize aspects of pedagogical reasoning. The planning agent demonstrated how to examine alignment among objectives, activities, resources, and assessment. The critical agent drew attention to assumptions, evidence, and possible alternatives. The professional-readiness agent encouraged students to anticipate classroom consequences and justify their decisions. The reflection agent prompted comparison between intended and actual performance. These functions approximated selected cognitive-apprenticeship processes while preserving the instructor's responsibility for academic oversight.

2.2.3 Formative Feedback

Feedback contributes to learning when it clarifies the intended performance, identifies the current state of the learner's work, and indicates productive next steps. Its effectiveness depends on specificity, timing, comprehensibility, and the learner's opportunity to act on it. Feedback that merely identifies an error may have limited value, whereas information that directs attention to the task, the process used, and strategies for improvement can support deeper learning (Hattie & Timperley, 2007).

Immediate feedback was a central affordance of the experimental environment. Students could request assistance while developing their products rather than waiting until a task had been completed and formally graded. The agents' responses were connected to the current task and were intended to stimulate revision. This feedback cycle—production, review, reflection, and improvement—was expected to strengthen the quality of students' decisions and products over the 10-week intervention.

Feedback was nevertheless constrained to avoid replacing the learner's reasoning. Shute (2008) cautioned that formative feedback must be appropriately timed, focused, and manageable. Excessive explanation may overload learners, whereas overly directive feedback may reduce constructive mental activity. The agent prompts therefore emphasized concise, criterion-based guidance and questions that encouraged students to reconsider their work before receiving more explicit recommendations.

2.2.4 Active and Constructive Learning

The environment also reflected the distinction among passive, active, constructive, and interactive engagement. Learning becomes progressively deeper when students move from receiving information to manipulating it, generating new ideas, explaining their reasoning, and engaging in substantive interaction (Chi,

2009). In the present study, students were required to produce lesson plans, evaluate teaching situations, design instructional resources, explain choices, respond to challenges, and revise products. AI support was thus embedded in productive activity rather than presented as a separate demonstration of technological capability.

2.2.5 Cognitive Load Management

Complex teaching tasks place simultaneous demands on content knowledge, pedagogical reasoning, digital production, and professional judgment. If assistance is poorly structured, it can add unnecessary processing demands and distract from the central task. Cognitive load theory distinguishes the intrinsic complexity of the learning material from avoidable demands created by ineffective presentation or task design (Sweller et al., 2011).

The multi-agent design sought to manage these demands through functional segmentation. Students accessed the agent relevant to the current need rather than receiving all forms of assistance at once. The coordinating layer organized access to the agents, while Moodle presented the instructional sequence, resources, assignments, and assessment criteria. This design was expected to reduce extraneous processing and allow students to direct their cognitive resources toward the professional problem being addressed.

2.3 Structure of the Multi-Agent AI-Based Learning Environment

The experimental environment incorporated six coordinated agent functions.

2.3.1 Pedagogical Planning Agent

The pedagogical planning agent supported the construction and review of lesson plans. Its prompts focused on learner characteristics, learning outcomes, content organization, instructional strategies, learning activities, resources, differentiation, and assessment. Rather than generating a complete lesson plan without learner involvement, the agent was configured to review submitted elements, identify gaps, and request justification for major pedagogical decisions.

2.3.2 Critical Inquiry Agent

The critical inquiry agent was responsible for promoting analysis and evaluation. It asked students to distinguish claims from evidence, identify assumptions, examine the credibility and relevance of information, consider alternative interpretations, and anticipate the consequences of instructional choices. This role was designed to prevent the environment from functioning exclusively as an answer-generation tool.

2.3.3 Digital Competence Agent

The digital competence agent provided assistance with selecting, producing, evaluating, and improving digital instructional resources. Its feedback addressed technical functionality, instructional alignment, accessibility, copyright awareness, information credibility, digital safety, and usability. Students were expected to apply the guidance directly while developing their digital instructional products.

2.3.4 Professional Readiness Agent

The professional readiness agent presented or analyzed teaching situations requiring professionally defensible responses. Its prompts encouraged students to consider learner needs, classroom management, assessment evidence, professional responsibility, communication, and ethical consequences. The agent was intended to connect theoretical principles with decisions resembling those encountered in teaching practice.

2.3.5 Product Evaluation Agent

The product evaluation agent used criteria aligned with the digital product rubric to provide formative observations before final submission. It did not assign the official research score. The final products were evaluated independently by trained human raters. This separation was necessary to prevent the intervention system from serving simultaneously as the support mechanism and the definitive evaluator of its own outcomes.

2.3.6 Coordination Agent

The coordination agent routed requests to the relevant support function and maintained continuity among task stages. It helped students identify whether a problem was primarily pedagogical, critical, technical, evaluative, or professionally situated. The coordination function reduced overlap among agents and supported a coherent progression from planning to production, evaluation, reflection, and revision.

2.4 Multi-Agent AI and Critical Thinking

Critical thinking is not developed simply by exposing students to large quantities of information. It requires tasks that demand interpretation, comparison, evaluation, justification, and reflection. Technology can support critical thinking when it creates opportunities for inquiry, structured dialogue, problem solving, and the examination of competing explanations. Abrami et al. (2015) found that critical-thinking instruction is more effective when critical thinking is taught explicitly and learners engage in dialogue, authentic problems, or mentoring.

The critical inquiry agent incorporated these principles by requiring students to justify their choices and consider alternatives. For example, when a student selected a teaching strategy, the agent could ask what evidence supported the choice, whether the strategy was appropriate for learner characteristics, what limitations it might present, and what alternative could be used. This interaction was intended to shift attention from producing a formally acceptable answer to examining the quality of the reasoning supporting it.

The pedagogical potential of such interaction rests on the design of the prompts. An agent that immediately provides a complete response may reduce the need for independent reasoning. By contrast, an agent that poses sequenced questions can make the learner's assumptions visible and create opportunities for self-correction. The environment therefore used questioning, requests for evidence, counterexamples, and comparison among alternatives before providing directive recommendations.

Collaborative and dialogic learning arrangements have generally shown positive associations with critical thinking because they expose learners to alternative positions and require the articulation of reasoning. A 2022 meta-analysis reported that collaborative learning produced positive effects on critical thinking, creative thinking, and metacognitive skills, although effects varied across instructional designs and learning contexts (Ramdani et al., 2022). In the present environment, the agent dialogue did not replace peer or instructor interaction; it provided an additional structured opportunity for students to externalize and reconsider their reasoning.

2.5 Multi-Agent AI and Digital Competence

Digital competence in teacher education is frequently weakened when training focuses on isolated software procedures without connecting technology to instructional purposes. Knowing how to operate a digital tool does not necessarily mean that a pre-service teacher can determine when it should be used, evaluate its quality, adapt it to learners, or address ethical and accessibility considerations. Instefjord and Munthe (2017) found that professional digital competence must be integrated coherently into teacher education rather than treated as peripheral technical knowledge.

The experimental environment addressed digital competence through authentic production tasks. Students selected tools, developed instructional resources, organized content, applied design principles, and revised products in response to formative guidance. The digital competence agent provided technical assistance, but its recommendations were connected to pedagogical criteria. This integration was consistent with the technological pedagogical content knowledge perspective, which views meaningful technology

integration as the interaction of technological, pedagogical, and content knowledge rather than the independent accumulation of separate competencies (Mishra & Koehler, 2006).

Performance assessment was particularly important because self-reported digital confidence may not correspond closely to demonstrated competence. A student may feel confident using common applications but experience difficulty when required to produce an accessible resource, resolve a technical problem, or align a digital activity with learning outcomes. The study therefore used a digital-competence performance test and an independently evaluated digital product to capture applied competence.

The environment was also designed to promote responsible use. Students were expected to evaluate AI-generated suggestions, respect intellectual property, protect personal information, verify sources, and avoid submitting unexamined generated content. Digital competence was consequently treated as a combination of performance, critical awareness, responsibility, and pedagogical purpose.

2.6 Multi-Agent AI and Professional Readiness for Teaching

Professional readiness is multidimensional. It involves more than confidence or willingness to enter the profession; it includes the ability to perform expected responsibilities and respond appropriately to situations involving learners, curriculum, assessment, classroom interaction, and professional ethics. Readiness develops through opportunities to connect theoretical learning with practice, receive feedback, reflect on performance, and revise professional decisions.

Field training is central to this development because it places students in situations requiring the integration of knowledge and action. However, the quality of field-training experiences can vary depending on access to supervisors, timing of feedback, and the range of situations encountered. A digital environment cannot reproduce all dimensions of school-based practice, but it can supplement field training through structured cases, simulated decisions, planning tasks, product development, and guided reflection.

Research on pre-service teachers' perceptions of preparedness has indicated that perceived readiness may vary across professional domains and may reveal areas in which prospective teachers remain uncertain about their capacity to meet professional standards. Willis et al. (2022) therefore emphasized the value of examining preparedness in relation to professional expectations, teacher agency, and the practical demands of teaching.

The professional-readiness and reflection agents were expected to influence this outcome through repeated engagement with authentic teaching decisions. Students examined classroom situations, anticipated consequences, evaluated possible responses, and connected their choices to pedagogical and professional principles. Over time, such guided practice was expected to strengthen both perceived capacity and performance in context-sensitive professional situations.

2.7 Relationship Among the Study Variables

The three dependent variables were expected to develop through related mechanisms. Critical thinking supported students' evaluation of pedagogical alternatives and AI-generated recommendations. Digital competence enabled them to transform pedagogical decisions into functional instructional resources. Professional readiness required them to combine critical judgment and digital capability with an understanding of learners, teaching responsibilities, and ethical practice.

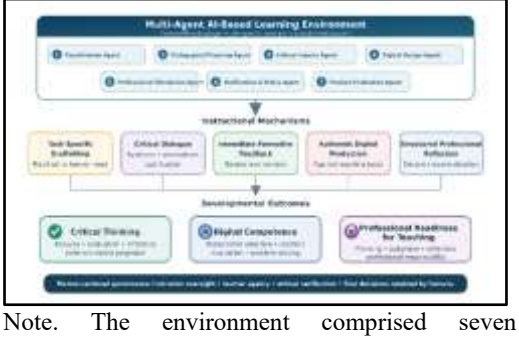
The independent variable—the Multi-Agent AI-based learning environment—was therefore conceptualized as a coordinated system of instructional affordances rather than a single technological tool. Role-specific support, immediate formative feedback, guided questioning, authentic tasks, iterative production, and reflective prompts were expected to operate together. Critical thinking, digital competence, and professional readiness were treated as separate outcomes in the statistical analysis, but the

conceptual model recognized their functional interdependence.

Figure 1 presents the conceptual model that guided the design and implementation of the study. The model proposes that the Multi-Agent AI-based learning environment contributes to the development of the targeted outcomes through five interrelated instructional mechanisms: task-specific scaffolding, critical dialogue, immediate formative feedback, authentic digital production, and structured professional reflection. Seven specialized AI agents operate within a coordinated Moodle-based architecture to provide differentiated support during field-training activities. The model also incorporates a human-centered governance layer in which instructor oversight, learner agency, ethical verification, and human responsibility for final decisions remain integral to the learning process.

Figure 1

Conceptual Model of the Multi-Agent AI-Based Learning Environment and the Study Outcomes



Note. The environment comprised seven coordinated agents embedded in Moodle: a coordination agent, pedagogical planning agent, critical inquiry agent, digital design agent, professional simulation agent, verification and ethics agent, and product evaluation agent. AI-generated support was formative and advisory; students retained responsibility for evaluating recommendations, while the instructor retained responsibility for academic supervision and consequential assessment.

As illustrated in Figure 1, the anticipated effects of the learning environment are not attributed to exposure to artificial intelligence alone. Rather, the model assumes that development occurs through students' active engagement with task-specific support, evidence-based dialogue, iterative feedback, authentic digital production, and structured reflection on professional decisions. These mechanisms are expected to operate jointly in developing critical thinking, digital competence, and professional readiness for teaching. The model therefore positions the AI agents as coordinated sources of formative assistance rather than autonomous substitutes for students' judgment or the instructor's professional authority.

2.8 Research Gap

The literature available by December 31, 2022 established the potential of intelligent tutoring, adaptive technology, formative feedback, authentic learning, and digital competence frameworks. Nevertheless, four gaps remained relevant.

First, much AI-in-education research examined a single tutoring or analytical function. Less attention was directed toward coordinated environments in which specialized agents supported different dimensions of a complex professional task.

Second, studies of educational AI commonly concentrated on achievement, engagement, prediction, or system acceptance. The simultaneous development of critical thinking, demonstrated digital competence, and professional readiness among pre-service teachers remained insufficiently examined.

Third, teacher-education research frequently assessed digital competence or readiness through self-report instruments. Fewer studies combined situational tests, performance tasks, independently evaluated products, interaction logs, and user-experience measures within the same controlled intervention.

Fourth, limited empirical evidence was available from Arabic-language teacher-education contexts in which AI support was embedded in an authentic Field Training course. The present study addressed these gaps through a randomized controlled educational experiment involving a custom Moodle-integrated Multi-Agent AI environment.

2.9 Operational Definitions

Multi-Agent AI-based learning environment. An online learning environment implemented through a custom Moodle plugin and connected to the OpenAI Completions API using text-davinci-002. The environment employed role-specific prompt templates representing coordinated pedagogical agents that supported planning, critical inquiry, digital production, professional readiness, product improvement, and interaction routing.

Critical thinking. The pre-service teacher's ability to interpret information, analyze evidence and assumptions, evaluate claims and alternatives, draw reasoned conclusions, explain decisions, and reconsider judgments in teaching-related situations. Operationally, it was represented by the score obtained on the Critical-Thinking Situational Test.

Digital competence. The ability to select, use, create, evaluate, and improve digital resources in a pedagogically appropriate, responsible, and technically effective manner. Operationally, it was represented by performance on the Digital Competence Test and supported by the score assigned to the final digital instructional product.

Professional readiness for teaching. The integrated capacity and perceived preparedness to undertake teaching responsibilities involving planning, instruction, assessment, classroom decision-making, professional communication, ethical conduct, and reflective practice. Operationally, it was represented by scores on the Professional Readiness for Teaching Scale and the Professional Readiness Situational Judgment Test.

Pre-service teachers. Fourth-year students enrolled in the General Education Program at the Faculty of Education, Damietta University, who were preparing to enter the teaching profession and were registered in the Field Training course during the 2022/2023 academic year.

2.10 Delimitations of the Study

The study was delimited to:

- 100 fourth-year pre-service teachers at the Faculty of Education, Damietta University;
- the General Education Program specializations of sciences, mathematics, languages, and social studies;
- the Field Training course;
- a 10-week, fully online intervention conducted from October 1 to December 3, 2022;
- a custom Moodle 4.0 environment connected to text-davinci-002;
- critical thinking, digital competence, professional readiness, digital product quality, AI-agent interaction, and environment evaluation as the measured outcomes; and
- the technological and pedagogical conditions available during the first semester of 2022/2023.

3. Method

3.1 Research Design

The study employed a randomized pretest–posttest controlled experimental design with two parallel groups. Participants were assigned in a 1:1 ratio to either an experimental group that used the Multi-Agent AI-based learning environment or a control group that accessed the same Field Training course through a conventional Moodle environment without AI-agent support.

Pretest measures of critical thinking, digital competence, and professional readiness for teaching were administered before the intervention. The corresponding posttest measures were administered after completion of the 10-week intervention. Digital instructional products were evaluated at the end of the intervention. The interaction-log rubric and the usability, trust, and

satisfaction questionnaire were applied only to the experimental group because these instruments concerned features that were not available in the control condition.

3.2 Participants and Sampling

The target population comprised 1,253 fourth-year pre-service teachers enrolled in the General Education Program at the Faculty of Education, Damietta University, during the 2022/2023 academic year. According to the enrollment records available at the beginning of the first semester, three students were excluded from the accessible population because of enrollment deferral, formal withdrawal, or discontinuation. The accessible population therefore consisted of 1,250 students.

A proportionate stratified random sampling procedure was used to represent the principal academic specializations within the General Education Program. The participating specializations were grouped into four strata: sciences, mathematics, languages, and social studies. Following selection, eligible participants were allocated in a 1:1 ratio to the experimental and control groups. Random allocation was stratified by academic specialization and gender to reduce the likelihood of systematic baseline imbalance.

The final sample included 100 participants, with 50 assigned to each group. All participants completed the study, and no withdrawal occurred after allocation. Participants were 21–22 years old. The experimental group included 12 men and 38 women, with a mean age of 21.40 years (SD = 0.50). The control group included 14 men and 36 women, with a mean age of 21.50 years (SD = 0.51). The overall sample comprised 26 men and 74 women and had a mean age of 21.45 years (SD = 0.50).

Table 1

Demographic Characteristics of the Participants

Characteristic	Experimental group (n = 50)	Control group (n = 50)	Total (N = 100)
Men, n	12	14	26
Women, n	38	36	74
Age, M	21.40	21.50	21.45
Age, SD	0.50	0.51	0.50
Age range	21–22	21–22	21–22

As shown in Table 1, the two groups were demographically comparable. The small difference in the proportions of men and women reflected the gender distribution of students enrolled in the participating teacher-education program rather than differential allocation to the study conditions.

Table 2 Distribution of Participants by Academic Specialization

Academic specialization	Experimental group (n = 50)	Control group (n = 50)	Total	Percentage
Sciences: physics, chemistry, and biology	14	15	29	29%
Mathematics	12	11	23	23%
Languages: Arabic and English	14	14	28	28%
Social studies: history and geography	10	10	20	20%
Total	**50**	**50**	**100**	**100%**

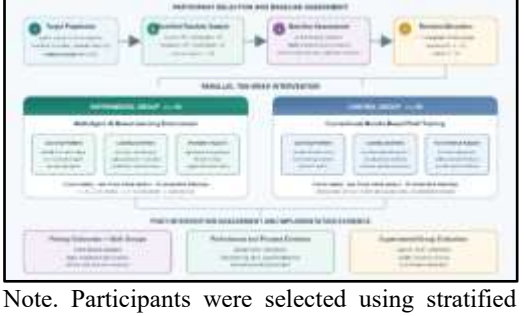
Table 2 demonstrates comparable representation of the four specialization strata. Languages and social studies were distributed equally across the conditions, while the science and mathematics strata differed by only one participant. This distribution reduced the likelihood that academic specialization systematically confounded the estimated intervention effects.

A randomized two-group pretest–posttest design was employed to examine the effect of the multi-Agent AI-based learning environment. The accessible population comprised 1,250 fourth-year pre-service teachers enrolled at the Faculty of Education, Damietta University, during the 2022/2023 academic year. A stratified random sample of 100 students was selected to represent science, mathematics, languages, and social studies specializations. Following baseline assessment, the participants were randomly assigned in a 1:1 ratio to an experimental group and a control group, with 50 students in each group.

Both groups studied the same Field Training course content during a ten-week period extending from 1 October to 3 December 2022. The intervention consisted of one three-hour online session each week. The same instructor taught both groups, and the course content, learning activities, instructional time, assessment schedule, and intended outcomes were held constant. The principal experimental difference was the availability of coordinated multi-Agent AI support to the experimental group. Figure 2 summarizes the experimental design and implementation sequence.

Figure 2

Experimental Design and Ten-Week Implementation Sequence



Note. Participants were selected using stratified random sampling and subsequently assigned randomly to the experimental and control groups. The experimental group used the custom Multi-Agent AI plugin embedded within Moodle 4.0, whereas the control group used the standard Moodle environment without AI-agent support. Both groups received the same Field Training course content from the same instructor for three hours weekly over ten consecutive Saturdays. Product performance was evaluated in both groups, whereas the AI interaction-log rubric and the usability, trust, and satisfaction questionnaire were applied to the experimental group. No participant withdrew during the study.

As illustrated in Figure 2, the experimental procedures progressed through four connected stages: participant selection, baseline assessment, random allocation, and parallel implementation followed by post-intervention assessment. Stratification ensured representation of the four principal disciplinary clusters within the sample: science, mathematics, languages, and social studies. Random allocation was subsequently used to distribute the 100 participants equally between the experimental and control conditions.

The experimental group accessed the Field Training course through a custom Moodle 4.0 plugin incorporating one coordination agent and six specialist agents. The environment supported pedagogical planning, critical inquiry, digital production, professional simulation, ethical verification, and formative product evaluation. Students received agent-generated guidance, revised their work iteratively, and generated interaction records documenting requests, responses, and revision events.

The control group accessed the same course content and completed equivalent field-training tasks through the standard Moodle environment without multi-Agent AI assistance. Conventional support consisted of instructor explanations and ordinary Moodle-based feedback. The duration, dates, online delivery mode, instructor, instructional content, task requirements, and assessment schedule were standardized across the two conditions. This procedural equivalence was intended to isolate the availability of coordinated multi-Agent AI support as the principal difference between the groups.

At the end of the intervention, the critical thinking, digital competence, and professional readiness instruments were administered to both groups. The

educational digital-product rubric was also applied to the products produced by students in both conditions. In the experimental group, additional implementation evidence was obtained from the interaction-log analysis rubric and the usability, trust, and satisfaction questionnaire. Product and interaction-record evaluations were performed by two independent raters without access to students' identities or group-related personal information.

3.3 Educational Context

The intervention was embedded in the Field Training course completed by fourth-year pre-service teachers. The course was selected because it required students to integrate pedagogical knowledge, digital production, critical judgment, and professional decision-making in authentic or simulated teaching tasks.

The course content addressed the following areas:

1. analysis of teaching contexts and learner characteristics;
2. formulation of measurable learning outcomes;
3. lesson planning and instructional alignment;
4. selection of teaching strategies and learning activities;
5. design and production of digital instructional resources;
6. digital assessment and feedback;
7. analysis of classroom and professional teaching situations;
8. responsible and critical use of artificial intelligence;
9. reflective evaluation of teaching performance; and
10. development and revision of a final digital instructional product.

The same instructor taught both groups throughout the intervention. The learning outcomes, core content, weekly schedule, task requirements, instructional time, and assessment expectations were held constant. The principal difference was the availability of coordinated multi-Agent AI support in the experimental condition.

3.4 The Multi-Agent AI-Based Learning Environment

3.4.1 Technical Architecture

The experimental environment was implemented as a custom plugin integrated into Moodle 4.0. Moodle managed user authentication, course organization, content delivery, assignments, discussion, assessment access, and activity records. The custom plugin provided agent selection, prompt-template management, interaction routing, response presentation, and the recording of AI-agent interactions.

The plugin communicated with the OpenAI Completions API through the /v1/completions endpoint and used the text-davinci-002 model. Because the legacy Completions architecture did not use the role structure subsequently associated with chat interfaces, the agents were operationalized through separate role-specific instruction templates. Each template specified the agent's instructional role, response boundaries, feedback approach, and required output structure.

The standard generation parameters were configured as follows:

- model: text-davinci-002;
- temperature: 0.7;
- top_p: 1.0;
- maximum response length: 500 tokens;
- frequency penalty: 0;
- presence penalty: 0.

These settings were standardized to maintain a reasonable balance between response stability and the generation of alternative pedagogical suggestions.

The environment was hosted on a virtual private server running Ubuntu Linux. The server included four virtual processing cores, 8 GB of RAM, and a MySQL database. Communication with the API was conducted through encrypted HTTPS requests. API credentials were stored on the server and were not transmitted to students' browsers.

3.4.2 Agent Structure

The environment included the following coordinated agent roles:

Table 3

Coordinated agent roles

Agent	Principal function	Typical support
Coordination agent	Identifying the nature of the request and routing it to the relevant agent	Task classification, agent recommendation, continuity across stages
Pedagogical planning agent	Supporting lesson planning and instructional alignment	Objectives, activities, teaching strategies, differentiation, and assessment
Critical inquiry agent	Promoting analysis and evaluation	Evidence checking, assumption identification, comparison, and justification
Digital design and competence agent	Supporting digital production and technology selection	Technical guidance, usability, accessibility, and pedagogical suitability
Professional simulation agent	Supporting responses to teaching situations	Classroom decisions, professional conduct, learner needs, and consequences
Verification and ethics agent	Encouraging responsible AI and information use	Source verification, privacy, intellectual property, bias, and human oversight
Product evaluation agent	Providing formative rubric-aligned feedback	Identification of strengths, omissions, and possible revisions

The agents did not assign official research scores. The final products and interaction logs were evaluated independently by human raters. This separation prevented the intervention system from serving as the sole evaluator of the outcomes it was designed to influence.

3.4.3 Learning Analytics and Interaction Records

The plugin recorded a pseudonymous student identifier, timestamp, agent selected, task context, response status, and relevant interaction indicators. Records included the sequence of agent use, requests for evidence, comparisons among responses, acceptance or modification of recommendations, revisions to products, and reflective comments.

Automated platform events and experimental logins preceding the intervention were excluded. The raw logs were preserved separately, and an anonymized copy was prepared for analysis. The number of messages or duration of platform access was not treated independently as evidence of learning quality; interpretation was based on patterns of purposeful interaction and documented task performance.

3.4.4 Access Control and Treatment Separation

Each participant accessed the environment through an individually registered Moodle account. The plugin verified the authenticated Moodle identity before processing a request and linked the interaction to a pseudonymous study identifier. API credentials remained inaccessible to students.

The two groups were enrolled in separate Moodle course spaces. The agent plugin was enabled only for the experimental course. The control course contained the same core content and equivalent activities but did not display agent interfaces or provide AI-generated feedback. Moodle access logs were used to monitor unusual concurrent access patterns and maintain separation between the treatment conditions.

The experimental environment was implemented as a custom multi-agent plugin embedded within Moodle 4.0. Its technical architecture integrated the learner and instructor interfaces, a multi-agent orchestration layer, an external AI-processing service, a structured interaction-record system, and a human-supervision mechanism. The seven-agent system comprised one coordination agent and six specialist agents, each governed by a role-specific prompt template aligned with a defined educational

function. Figure 3 illustrates the principal architectural layers and the flow of interaction among the student, Moodle, the coordinated agents, the AI-processing service, the interaction records, and instructor oversight.

Figure 3

Technical Architecture of the Multi-Agent AI-Based Learning Environment

Note. The seven-agent system comprised one coordination agent and six specialist agents: pedagogical planning, critical inquiry, digital design, professional simulation, verification and ethics, and product evaluation. The agents operated through role-specific prompt templates within a custom Moodle 4.0 plugin connected to the OpenAI Completions API using text-davinci-002. AI-generated responses were formative and advisory; students retained responsibility for their submitted work, and the instructor retained authority over supervision, follow-up, professional judgment, and formal assessment.

As shown in Figure 3, students accessed the environment through an authenticated Moodle workspace using coded identifiers. They submitted field-training tasks, questions, drafts, digital products, and reflective responses through the agent-interaction panel. The coordination agent interpreted each request, identified the type of assistance required, routed it to the appropriate specialist agent, and integrated the resulting guidance before returning it to the student.

The six specialist agents provided differentiated forms of support. The pedagogical planning agent assisted with lesson and activity design; the critical inquiry agent prompted students to examine evidence and alternative interpretations; the digital design agent supported tool selection, digital production, and accessibility; the professional simulation agent presented classroom scenarios requiring professional judgment; the verification and ethics agent prompted the examination of claims, potential bias, and responsible use; and the product evaluation agent generated criteria-based formative recommendations.

The generated guidance entered an iterative learner-controlled feedback cycle. Students reviewed each recommendation and could accept, modify, or reject it before revising and resubmitting their work. The system recorded the coded identifier, timestamp, selected agent, request, generated response, and revision event within the Moodle/MySQL data layer. These records supported analysis of engagement and revision patterns without transferring academic decision-making authority to the system.

Instructor oversight remained active throughout the intervention. The dashboard enabled the instructor to review interaction patterns, identify students requiring additional assistance, provide targeted follow-up, and exercise professional judgment. Consequently, the environment functioned as a coordinated formative-support system rather than an autonomous instructional or assessment system.

3.5 Control Condition

The control group studied the same Field Training content using conventional Moodle functions. These included digital content pages, uploaded learning materials, assignments, discussion forums, announcements, and instructor feedback. Control-group participants completed equivalent planning, analysis, assessment, and digital-production tasks within the same weekly schedule. The control group did not have access to the specialized AI agents, automated AI-generated

formative feedback, agent coordination, or AI-interaction analytics. Instructor contact and scheduled learning time were equivalent across the groups.

3.6 Implementation Procedure

The intervention was delivered entirely online over 10 consecutive weeks during the first semester of 2022/2023, from October 1 to December 3, 2022. Students attended one three-hour online session each week, producing a total intervention duration of 30 hours.

The intervention content was organized into eight instructional units delivered over ten consecutive weeks. The first week was devoted to orientation, baseline assessment, and training in the use of the learning environment. Weeks 2–9 addressed the instructional units through authentic field-training situations and progressively more complex applied tasks. The final week was allocated to the presentation of the capstone project, professional reflection, and post-intervention assessment. Each instructional week produced a tangible performance artifact that subsequently contributed to the students' digital professional teaching portfolios. Figure 4 presents the temporal progression of the intervention and the principal applied output associated with each week.

Figure 4

Ten-Week Learning Journey and Applied Outputs

Note. The intervention was delivered through one three-hour online session each Saturday from 1 October to 3 December 2022. Week 1 included study orientation, informed consent, baseline assessment, and training in the use of the environment. Weeks 2–9 covered the eight instructional units, while Week 10 included the capstone-project presentation, final professional reflection, and post-intervention assessment. The weekly artifacts were progressively incorporated into the students' digital professional teaching portfolios.

As Figure 4 illustrates, the intervention followed a progressive sequence rather than presenting the instructional units as isolated topics. The orientation week established the procedural and technical conditions required for participation. The next three weeks developed the conceptual foundations necessary for responsible engagement with AI-supported teaching, including understanding the changing role of the teacher, evaluating educational evidence, and locating and assessing digital information and data.

Weeks 5–7 moved from foundational knowledge to design and AI-supported practice. Students designed digital instructional content, examined the structure and operation of Multi-Agent AI systems, and developed digital assessment and feedback materials. These units required students to translate conceptual understanding into usable teaching products while considering instructional alignment, accessibility, verification, fairness, and responsible use.

The final instructional phase emphasized professional application. During Week 8, students analyzed classroom problems, generated alternatives, evaluated expected consequences, and documented justified professional decisions. Week 9 focused on professional identity, reflective practice, portfolio construction, career preparation, and continuing professional growth. The intervention culminated in Week 10 with the presentation of a digital professional teaching portfolio and the administration of the post-intervention instruments.

Across Weeks 2–9, students followed a recurring learning cycle beginning with an authentic professional situation. They initially analyzed the situation independently, consulted the relevant AI agents, compared the evidence and

recommendations presented, and then made a justified decision. Each cycle concluded with the production or revision of a digital artifact, criteria-based feedback, and a concise professional reflection. This recurring structure provided continuity across the instructional units while progressively integrating critical thinking, digital competence, and professional readiness for teaching.

Table 4

Weekly Content, Applied Activities, and Performance Outputs of the Intervention

Week	Main focus	Experimental-group activity	Control-group activity
1	Orientation and baseline assessment	Environment orientation, pretests, and agent-use demonstration	Moodle orientation and pretests
2	Teaching context analysis	Agent-supported analysis of learners and teaching contexts	Instructor-guided analysis through Moodle
3	Learning outcomes and planning	Planning-agent support and iterative lesson-plan review	Conventional planning materials and instructor feedback
4	Teaching strategies	Comparison and justification of strategies using planning and critical agents	Strategy-selection task through Moodle
5	Digital instructional resources	Digital-design agent support for resource development	Resource-development task without agents
6	Assessment and feedback	Agent-supported design of digital assessment and feedback	Equivalent assessment-design task
7	Critical teaching situations	Critical inquiry and professional simulation	Moodle-based teaching cases
8	Teaching performance reflection	Simulated decisions and structured agent-supported reflection	Conventional written reflection and instructor feedback
9	Final digital product	Iterative development using design, verification, and evaluation agents	Product development using conventional Moodle resources
10	Final assessment	Product submission, posttests, log extraction, and environment evaluation	Product submission and posttests

Table 4 shows that both groups received the same amount of instructional exposure and addressed equivalent Field Training outcomes. Treatment differentiation concerned the source and structure of support rather than content coverage or instructional time.

3.7 Research Instruments

Seven instruments were used.

3.7.1 Critical-Thinking Situational Test

The Critical-Thinking Situational Test consisted of 30 teaching-related situations. Each situation was followed by four response alternatives, one of which represented the most defensible critical response. Correct responses received one point and other responses received zero, producing scores from 0 to 30. The allotted time was 45 minutes.

The situations addressed interpretation, analysis, evaluation of evidence, recognition of assumptions, inference, comparison of alternatives, explanation, and reflective reconsideration. The instrument was administered before and after the intervention.

3.7.2 Digital Competence Performance Test

The Digital Competence Performance Test included 10 practical tasks. Each task was scored from 0 to 4 using task-specific performance criteria, resulting in a possible total of 0–40. Students were allowed 120 minutes.

The tasks assessed:

- digital information search and source evaluation;
- organization of digital information;
- production of an accessible digital resource;
- design of an interactive activity;
- construction of digital assessment and feedback;
- interpretation of learning data;
- digital professional communication and collaboration;
- privacy, safety, and intellectual property;
- critical evaluation of AI-generated content; and
- digital problem solving and tool selection.

3.7.3 Professional Readiness for Teaching Scale

The Professional Readiness for Teaching Scale contained 48 statements distributed across 12 dimensions, with four statements representing each dimension. Responses were recorded on a five-point Likert scale ranging from 1 to 5. Eight negatively worded statements were reverse-coded before calculating the total score. Possible scores ranged from 48 to 240, with higher scores indicating greater perceived readiness. Completion required approximately 15–20 minutes.

The dimensions addressed instructional planning, implementation, classroom management, assessment, decision-making, communication, collaboration, digital practice, ethical responsibility, reflective practice, professional resilience, and continuing professional development.

3.7.4 Professional Readiness Situational Judgment Test

The Professional Readiness Situational Judgment Test consisted of 24 situations representing 12 professional skills, with two situations for each skill. Each situation presented four possible responses that were scored from 0 to 3 according to their professional appropriateness. Total scores ranged from 0 to 72. The recommended completion time was 45–50 minutes.

This instrument complemented the self-report scale by assessing students' judgments in realistic teaching situations rather than relying exclusively on perceived readiness.

3.7.5 Digital Instructional Product Rubric

The rubric evaluated the final digital instructional product using 12 analytic criteria. Each criterion was scored from 0 to 4, producing a total score from 0 to 48. The criteria addressed instructional alignment, content quality, organization, digital design, interactivity, assessment, technical functionality, accessibility, usability, source documentation, originality and revision, and responsible use of AI.

Products were anonymized and scored independently. The mean of the two raters' scores was used in the main analysis.

3.7.6 Multi-Agent AI Interaction Log Analysis Rubric

The log rubric contained 24 indicators distributed across eight dimensions:

1. organized participation;
2. functional use of agents;
3. evidence seeking and source verification;
4. critical comparison of agent outputs;
5. retention of human control over decisions;
6. use of feedback for performance improvement;
7. ethical and secure use; and
8. transfer to professional practice and reflection.

Each indicator was scored from 0 to 3, producing a total score from 0 to 72. The unit of initial analysis was the student's interaction within a learning task, while the final rating represented the overall pattern of performance across the intervention.

3.7.7 Usability, Trust, and Satisfaction Questionnaire

The questionnaire consisted of 30 statements distributed across 10 dimensions, with three statements per dimension. It assessed learnability, navigation, agent-role clarity, agent coordination, perceived usefulness, recommendation interpretability, appropriate trust, human control, overall satisfaction, and intention to use the environment in the future.

Responses were provided on a five-point scale ranging from 1 to 5. Five negatively worded statements were reverse-coded. Total scores ranged from 30 to 150, with 90 representing the theoretical neutral point. The questionnaire required approximately 10–15 minutes and was administered only to the experimental group after the intervention.

3.8 Content Validity and Pilot Study

The initial versions of the content, learning environment, and research instruments were reviewed by 11 specialists in educational technology, computer science, and curriculum and instruction. Reviewers examined the relevance of the items, alignment with the operational definitions, clarity of language, appropriateness for fourth-year pre-service teachers, adequacy of scoring criteria, and correspondence between the instruments and the intervention outcomes. Items and instructions were revised in response to their observations before pilot administration.

A pilot study was conducted with 40 students who belonged to the target population but were not included in the main sample. The pilot study examined clarity, completion time, scoring procedures, score variability, internal consistency, and the functionality of digital administration.

Table 5

Internal-Consistency Estimates Obtained in the Pilot Study

Instrument	Items/tasks	Score range	Reliability coefficient
Critical-Thinking Situational Test	30	0–30	KR-20 = .899
Digital Competence Performance Test	10	0–40	α = .912
Professional Readiness for Teaching Scale	48	48–240	α = .948
Professional Readiness Situational Judgment Test	24	0–72	α = .966
Digital Instructional Product Rubric	12	0–48	α = .949
Multi-Agent AI Interaction Log Rubric	24	0–72	α = .965
Usability, Trust, and Satisfaction Questionnaire	30	30–150	α = .853

The coefficients ranged from .853 to .966, indicating acceptable to excellent internal consistency for research purposes. Reverse-worded items in the Professional Readiness Scale

and the environment-evaluation questionnaire were recoded before calculating their total reliability coefficients.

3.9 Inter-Rater Reliability

Two independent raters scored the digital products and interaction logs. Agreement was examined on a subsample of 20 cases using a two-way random-effects model, absolute agreement, and 95% confidence intervals. Single-measure coefficients represented the reliability of one rating, whereas average-measure coefficients represented the reliability of the mean of two ratings.

Table 6

Inter-Rater Reliability of the Performance-Based Instruments

Instrument	Cases	ICC model	ICC (2,1)	95% CI	F(19, 19)	ICC(2,2)	95% CI
Digital Instructional Product Rubric	20	Two-way random, absolute agreement	.982	[.956, .993]	106.432***	.991	[.977, .996]
Multi-Agent AI Interaction Log Rubric	20	Two-way random, absolute agreement	.989	[.972, .996]	169.378***	.994	[.986, .998]

Note. ICC = intraclass correlation coefficient; CI = confidence interval.

**p < .001.

Both instruments demonstrated excellent absolute agreement. The average measure coefficients exceeded .99, supporting the use of the mean ratings in the main analyses.

3.10 Ethical Considerations

No formal institutional ethics committee approval was obtained for this classroom-based educational study. Nevertheless, established ethical principles for research involving human participants were followed. Before enrollment, students received a clear explanation of the study's purpose, procedures, expected duration, voluntary nature, and data-handling arrangements. Written or electronic informed consent was obtained from all participants.

Students were informed that they could decline participation or withdraw without academic penalty and without any effect on their course grades. Research-instrument scores were maintained separately from formal course assessment. Participation decisions and responses were not used in determining academic grades.

Personal identifiers were replaced with study codes before analysis. Reports contained aggregated results, and access to raw interaction records was restricted to the research process. Students were also informed that their interactions with the AI agents could be logged and analyzed for research purposes. No participant withdrew after random allocation.

3.11 Data Analysis

Data were screened for missing values, impossible scores, duplicate identifiers, and influential outliers. Descriptive statistics were calculated for all outcomes. Internal consistency was examined using Cronbach's alpha or KR-20 as appropriate, while inter-rater reliability was evaluated using intraclass correlation coefficients.

Baseline equivalence was examined using independent-samples t tests for continuous variables and chi-square tests for categorical variables. The principal post-intervention outcomes were analyzed using analysis of covariance, with group as the fixed factor, the corresponding posttest score as the dependent variable, and the pretest score as the covariate.

The assumptions of linearity, homogeneity of regression slopes, normality of residuals, independence, and homogeneity of variance were examined. Because evidence of residual heteroscedasticity was identified for some outcomes, heteroscedasticity-consistent HC3

standard errors were used as a robustness procedure. The Holm method was applied to control familywise error across the principal outcome tests.

For each ANCOVA, adjusted means, adjusted mean differences, 95% confidence intervals, F statistics, p values, and partial eta squared were reported. The final digital product scores were compared using Welch's independent-samples t test because the product had no pretest equivalent and equality of variance could not be assumed. Hedges' g was reported as the standardized effect size.

Agent-interaction and environment-evaluation scores were analyzed descriptively within the experimental group. Correlations were treated as exploratory associations and were not interpreted as evidence of causation. Statistical significance was evaluated using a two-sided α level of .05.

4. Results

4.1 Data Screening and Preliminary Analysis

The dataset contained 100 complete cases, with 50 participants in each condition. No duplicate participant identifiers, duplicated response profiles, impossible scores, or missing values were identified for the primary pretest and posttest outcomes or the digital product score. The 50 missing entries for the interaction-log and satisfaction variables corresponded to the control group and were structurally expected because only experimental-group participants used the multi-Agent AI environment.

No multivariate outliers were identified at the conservative $p < .001$ criterion. Inspection of skewness and kurtosis indicated that the principal pretest and posttest distributions were sufficiently regular for the planned analyses. Ceiling effects were limited: 4% of the experimental group achieved the maximum critical-thinking posttest score, 2% achieved the maximum digital-competence score, and 6% achieved the maximum digital product score. No substantial floor effects were observed.

Tests of the group-by-pretest interaction were nonsignificant for critical thinking, $p = .896$; digital competence, $p = .474$; the Professional Readiness Scale, $p = .881$; and the Professional Readiness Situational Judgment Test, $p = .716$. The homogeneity-of-regression-slopes assumption was therefore supported by all four ANCOVA models. Evidence of unequal residual variance was observed for some outcomes. Consequently, conventional ANCOVA results were supplemented with HC3 heteroscedasticity-consistent standard errors and confidence intervals. The robust analyses produced the same substantive conclusions as the conventional models.

4.2 Baseline Equivalence

Pretest means were compared before examining intervention effects.

Table 7

Baseline Comparisons between the Experimental and Control Groups

Outcome	Experimental group M (SD)	Control group M (SD)	Welch's t	df	p	Hedges' g
Critical thinking	13.74 (3.24)	14.08 (3.36)	−0.516	97.88	.607	−0.102
Digital competence	17.22 (4.22)	17.58 (3.95)	−0.440	97.59	.661	−0.087
Professional readiness scale	130.40 (15.62)	129.66 (16.13)	0.233	97.90	.816	0.046
Professional readiness SJT	35.10 (5.78)	35.24 (7.15)	−0.108	93.86	.915	−0.021

Note. SJT = situational judgment test.
As shown in Table 7, no statistically significant baseline differences were identified. The absolute standardized differences were smaller than 0.11 for all measures, indicating negligible pre-intervention differences between the conditions.

4.3 Descriptive Statistics for the Primary Outcomes

Table 8
Pretest and Posttest Descriptive Statistics by Study Group

Outcome	Group	Pretest M (SD)	Posttest M (SD)	Mean gain
Critical thinking	Experimental	13.74 (3.24)	20.86 (4.63)	7.12
	Control	14.08 (3.36)	16.28 (4.01)	2.20
Digital competence	Experimental	17.22 (4.22)	28.58 (6.17)	11.36
	Control	17.58 (3.95)	21.80 (4.91)	4.22
Professional readiness scale	Experimental	130.40 (15.62)	175.10 (23.62)	44.70
	Control	129.66 (16.13)	140.72 (17.42)	11.06
Professional readiness SJT	Experimental	35.10 (5.78)	51.98 (8.75)	16.88
	Control	35.24 (7.15)	40.70 (8.02)	5.46

The descriptive results in Table 8 show improvement in both groups. However, the experimental group demonstrated consistently larger gains across the four outcomes. Because direct comparisons of raw gain scores do not fully account for baseline-outcome relationships, the hypotheses were tested using ANCOVA with the corresponding pretest score as the covariate.

4.4 Overall Effects of the Multi-Agent AI-Based Learning Environment

Table 9
ANCOVA Results for the Primary Posttest Outcomes

Outcome	Adjusted experimental M	Adjusted control M	Adjusted difference	HC3 95% CI	F(1, 97)	p	Partial η^2
Critical thinking	21.04	16.10	4.95	[3.95, 5.95]	99.22	< .001	.506
Digital competence	28.74	21.64	7.11	[5.42, 8.80]	71.99	< .001	.426
Professional readiness scale	174.73	141.09	33.64	[28.24, 39.04]	156.08	< .001	.617
Professional readiness SJT	52.05	40.63	11.42	[9.28, 13.56]	115.01	< .001	.542

Note. Adjusted means were estimated at the overall mean of the corresponding pretest. HC3 confidence intervals use heteroscedasticity-consistent standard errors. SJT = situational judgment test. Holm-adjusted conclusions remained significant for all four outcomes.
Table 9 demonstrates statistically significant adjusted differences favoring the experimental group on every primary outcome. The partial eta-squared values ranged from .426 to .617, indicating substantial differences under conventional effect-size guidelines.

4.5 Effect on Critical Thinking

After controlling for critical-thinking pretest scores, the experimental group obtained a significantly higher adjusted posttest mean than the control group, $F(1, 97) = 99.22$, $p < .001$, partial $\eta^2 = .506$. The adjusted difference was 4.95 points, HC3 95% CI [3.95, 5.95].

The HC3 robust test confirmed the result, $t(97) = 9.84$, $p < .001$. The model explained approximately 74.7% of the variance in critical-thinking posttest scores, including variance associated with the pretest covariate and group membership. Hypothesis 1 was therefore supported.

4.6 Effect on Digital Competence

After controlling for digital-competence pretest scores, the adjusted posttest mean was significantly higher for the experimental group, $F(1, 97) = 71.99$, $p < .001$, partial $\eta^2 = .426$. The adjusted difference between the conditions was 7.11 points, HC3 95% CI [5.42, 8.80].

The robust test yielded the same conclusion, $t(97) = 8.35$, $p < .001$. The complete model explained approximately 59.5% of the variance in digital-competence posttest performance. Hypothesis 2 was supported.

4.7 Effect on Professional Readiness Measured by the Scale

The experimental group obtained a significantly higher adjusted Professional Readiness Scale score after controlling for baseline readiness, $F(1, 97) = 156.08$, $p < .001$, partial $\eta^2 = .617$. The adjusted group difference was 33.64 points, HC3 95% CI [28.24, 39.04].

The HC3 robust estimate remained statistically significant, $t(97) = 12.37$, $p < .001$. The overall model explained approximately 75.5% of the posttest variance. Hypothesis 3 was therefore supported.

4.8 Effect on Situationally Assessed Professional Readiness

After controlling for pretest performance, the experimental group achieved significantly higher scores on the Professional Readiness Situational Judgment Test, $F(1, 97) = 115.01$, $p < .001$, partial $\eta^2 = .542$. The adjusted difference was 11.42 points, HC3 95% CI [9.28, 13.56].

The robust analysis confirmed this difference, $t(97) = 10.59$, $p < .001$. The model explained approximately 72.7% of the variance in posttest scores. Hypothesis 4 was supported.

The convergence between the scale and situational judgment results indicates that the observed difference was not limited to participants' perceptions of readiness; it was also evident in their responses to professionally situated teaching problems.

4.9 Digital Instructional Product Quality

Levene's test indicated unequal variances for the final digital product scores, $p = .023$. Accordingly, Welch's independent-samples t test was used.

Table 10

Comparison of Final Digital Instructional Product Scores

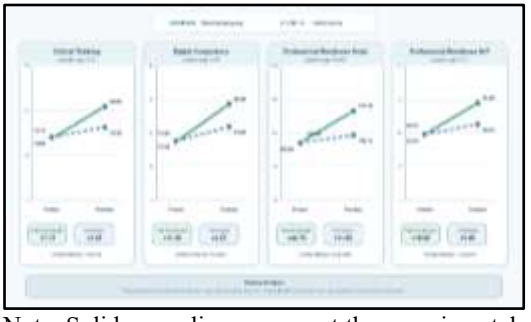
Group	n	M	SD	Minimum	Maximum
Experimental	50	39.18	5.07	27	48
Control	50	27.76	7.50	15	42

The experimental group produced significantly higher-quality digital instructional products than the control group, Welch's $t(86.00) = 8.92$, $p < .001$. The unstandardized difference was 11.42 points. The standardized difference was large, Hedges' $g = 1.77$. Hypothesis 5 was therefore supported.

Table 10 presents the descriptive statistics for the experimental and control groups across the pretest and posttest measurements. To facilitate visual interpretation of the direction and magnitude of change within each group, the means were plotted separately for critical thinking, digital competence, the professional readiness scale, and the professional readiness situational judgment test. Because the four instruments had different theoretical score ranges, each outcome was displayed in an independent panel using its original scoring metric. Figure 5 presents the pretest-to-posttest trajectories for the two groups across the primary outcomes.

Figure 5

Pretest and Posttest Means for the Experimental and Control Groups across the Primary Outcomes



Note. Solid green lines represent the experimental group, and dashed blue lines represent the control group. Each panel uses the original theoretical score range of its corresponding instrument: 0–30 for critical thinking, 0–40 for digital competence, 48–240 for the professional readiness scale, and 0–72 for the professional readiness situational judgment test (SJT). Values shown beside the data points are unadjusted arithmetic means. Gain values represent the difference between the posttest and pretest means within each group. Each group comprised 50 participants.

As shown in Figure 5, the experimental and control groups began the intervention with closely comparable mean scores across all four primary outcomes. The pretest differences were small: 0.34 points for critical thinking, 0.36 points for digital competence, 0.74 points for the professional readiness scale, and 0.14 points for the professional readiness situational judgment test. This descriptive pattern is consistent with the baseline analyses reported in Table 6, which should remain the formal basis for judging pre-intervention equivalence.

Both groups demonstrated higher means at posttest; however, the trajectories were consistently steeper for the experimental group. The experimental group's mean critical-thinking score increased from 13.74 to 20.86, representing a gain of 7.12 points, compared with a gain of 2.20 points in the control group. Consequently, the posttest difference between the groups reached 4.58 points.

A comparable pattern was observed for digital competence. The experimental group improved by 11.36 points, from 17.22 at pretest to 28.58 at posttest, whereas the control group improved by 4.22 points, from 17.58 to 21.80. The resulting unadjusted posttest difference was 6.78 points in favor of the experimental group.

The largest raw-score change occurred on the professional readiness scale, whose wider theoretical range should be considered when interpreting its numerical magnitude. The experimental group increased from 130.40 to 175.10, producing a mean gain of 44.70 points. In comparison, the control group increased from 129.66 to 140.72, producing a gain of 11.06 points. The unadjusted posttest difference was therefore 34.38 points.

The situational judgment measure provided further performance-based evidence of professional readiness. The experimental group's mean increased from 35.10 to 51.98, corresponding to a gain of 16.88 points, whereas the control group's mean increased from 35.24 to 40.70, corresponding to a gain of 5.46 points. The posttest difference between the groups was 11.28 points.

Collectively, the descriptive trajectories indicate greater pretest-to-posttest improvement in the experimental group across all primary outcomes. Nevertheless, these visual and descriptive differences should not be interpreted independently as evidence of statistical significance. The formal conclusions regarding the effect of the intervention must be based on the inferential analyses, adjusted group comparisons, confidence intervals, and effect sizes reported in the

4.10 Experimental-Group Interaction and Environment Evaluation

4.10.1 Multi-Agent AI Interaction Logs

The experimental group obtained a mean interaction-log score of 56.76 (SD = 6.63) out of 72, equivalent to 78.83% of the maximum score. Scores ranged from 43 to 72. Only 4% achieved the maximum, indicating no substantial ceiling effect.

This result suggests that experimental-group participants generally demonstrated organized and purposeful interaction with the agents, including agent selection, source verification, comparison of outputs, retention of human control, use of feedback, responsible interaction, and professional reflection. Because the control group did not use the agents, this outcome was not subjected to a between-group comparison.

4.10.2 Usability, Trust, and Satisfaction

The mean total questionnaire score was 123.58 (SD = 13.88) out of 150, corresponding to 82.39% of the maximum possible score. Scores ranged from 98 to 150, and 2% of participants achieved the maximum. The mean item response was 4.12 out of 5, exceeding the theoretical neutral value of 3.

The result indicates a generally favorable evaluation of the environment's usability, role clarity, agent coordination, usefulness, interpretability, appropriate trust, human control, satisfaction, and potential future use. These data describe participants' perceptions and should not be interpreted independently as evidence of intervention effectiveness.

Table 11

Experimental-Group Scores for Agent Interaction and Environment Evaluation

Measure	n	Possible range	M	SD	Observed range	Percentage of maximum
Multi-Agent AI Interaction Log Rubric	50	0–72	56.76	6.63	43–72	78.83 %
Usability, Trust, and Satisfaction Questionnaire	50	30–150	123.58	13.88	98–150	82.39 %

Table 11 shows that participants demonstrated relatively high-quality use of the agent system and evaluated the environment favorably. The higher relative questionnaire score should not be interpreted as superior to the interaction score because the two instruments measure different constructs and use different scoring systems.

4.11 Exploratory Associations within the Experimental Group

Exploratory Pearson correlations were calculated within the experimental group to examine associations between interaction quality and post-intervention outcomes.

Table 12

Associations of Agent-Interaction Quality with Experimental-Group Outcomes

Outcome	Correlation with interaction-log score, r	p
Critical-thinking posttest	.550	< .001
Digital-competence posttest	.412	.003
Professional-readiness scale posttest	.511	< .001
Professional-readiness SJT posttest	.365	.009
Digital instructional product	.476	< .001
Usability, trust, and satisfaction	.492	< .001

As shown in Table 12, higher-quality interaction with the Multi-Agent AI system was positively associated with each post-intervention outcome and with students' evaluation of the environment. The associations ranged from moderate to moderately strong. The largest association was observed with critical thinking, followed by professional readiness measured through self-report.

These correlations are exploratory and do not establish that higher interaction scores caused higher outcome scores. Students with stronger

initial capability or greater engagement may also have used the agents more effectively.

4.12 Summary of Hypothesis Testing

Table 13

Summary of Hypothesis Decisions

Hypothesis	Outcome	Statistical conclusion	Decision
H1	Critical thinking	Significant adjusted difference favoring the experimental group	Supported
H2	Digital competence	Significant adjusted difference favoring the experimental group	Supported
H3	Professional readiness scale	Significant adjusted difference favoring the experimental group	Supported
H4	Professional readiness SJT	Significant adjusted difference favoring the experimental group	Supported
H5	Digital instructional product quality	Significant difference favoring the experimental group	Supported

Taken together, the findings indicate that the experimental condition was associated with significantly stronger outcomes across critical thinking, demonstrated digital competence, perceived professional readiness, situational professional judgment, and digital instructional product quality.

5. Discussion

This study examined whether a Multi-Agent AI-based learning environment embedded in an online Field Training course could improve pre-service teachers' critical thinking, digital competence, and professional readiness for teaching. The results consistently favored the experimental condition across the four primary outcomes and the independently evaluated digital instructional product. These effects remained statistically significant after controlling for baseline scores and using heteroscedasticity-consistent standard errors. The convergence of the findings across self-report, situational judgment, practical performance, product evaluation, and recorded interaction provides a stronger basis for interpretation than reliance on a single measurement method.

The findings should not be interpreted as demonstrating that the language model alone produced the observed development. The intervention combined several elements: specialized agent roles, immediate formative feedback, critical questioning, authentic Field Training tasks, iterative digital production, structured reflection, and instructor supervision. The estimated effects therefore apply to the complete instructional environment rather than to text-davinci-002 as an isolated technological component.

5.1 Development of Critical Thinking

The adjusted critical-thinking difference supported the first hypothesis and represented a substantial intervention effect. This result may be explained by the role assigned to the critical inquiry agent. Students were not consistently given final solutions; instead, they were asked to identify evidence, expose assumptions, compare alternatives, examine the relevance of information, and justify pedagogical choices. These processes correspond closely to the reasoning operations assessed by the situational test.

The result also reflects the instructional design of the weekly tasks. Participants applied critical thinking to lesson planning, technology selection, AI-generated information, assessment evidence, and professional teaching situations. Critical thinking was therefore practiced within meaningful disciplinary and professional activity rather than presented as a decontextualized set of rules. Lai

(2011) argued that critical thinking develops more effectively when learners receive explicit instruction, encounter opportunities for inquiry and dialogue, and apply reasoning within a relevant knowledge domain.

The multi-agent structure may have strengthened this process by exposing students to functionally different perspectives. A planning recommendation could be examined by the critical agent, reviewed for digital feasibility by the design agent, and reconsidered in relation to professional consequences by the simulation agent. This arrangement created opportunities to compare and reconcile responses rather than accept the first available recommendation.

Nevertheless, the result does not show that unrestricted AI use necessarily enhances critical thinking. A system that supplies polished answers without requiring evaluation could have a different effect. The present finding is more specifically attributable to an environment in which AI assistance was structured around questioning, evidence, comparison, and learner-controlled decision-making.

5.2 Development of Digital Competence

The second hypothesis was supported by the adjusted digital-competence results. Experimental-group students demonstrated stronger performance in searching for and evaluating information, organizing digital resources, designing activities and assessment, interpreting learning data, addressing privacy and intellectual property, evaluating AI-generated content, and solving digital problems.

This development can be linked to the practical nature of the intervention. Students did not study digital competence solely through descriptive content. They repeatedly selected tools, constructed resources, solved technical problems, evaluated digital information, and revised a final product. Technology-integration research has emphasized that pre-service teachers need opportunities to connect technical knowledge with pedagogical decisions and authentic classroom purposes. Operational familiarity alone is unlikely to produce meaningful instructional integration (Ertmer & Ottenbreit-Leftwich, 2010).

The digital competence agent may have reduced delays caused by minor technical obstacles while preserving the need for students to implement the suggested procedures. Immediate support enabled students to continue working within the task rather than postpone progress until the next instructor interaction. At the same time, the verification and ethics agent directed attention toward privacy, source credibility, intellectual property, accessibility, and responsible AI use. Digital competence was therefore developed as a combination of practical performance and critical responsibility.

The significantly higher digital product scores provide complementary evidence. The product rubric evaluated applied performance beyond the digital-competence test and was scored by independent raters. The large standardized difference indicates that the experimental group's advantage was visible in a complex product developed over time, not only in short practical tasks.

However, because the product did not have an equivalent pretest, the product comparison cannot control directly for prior product-development ability. Random allocation and baseline equivalence on digital competence reduce this concern but do not remove it completely.

5.3 Development of Professional Readiness for Teaching

The strongest adjusted difference was observed for professional readiness measured by the scale. The corresponding situational judgment result also favored the experimental group. Together, these findings support the third and fourth hypotheses and indicate that the difference was present in both perceived readiness and responses to realistic professional situations.

Professional readiness is unlikely to develop through information delivery alone. It requires repeated opportunities to interpret situations, anticipate consequences, make decisions, receive feedback, and reconsider professional action. The environment provided such opportunities through lesson-planning tasks, classroom cases, simulated decisions, assessment problems, and reflective prompts. These activities connected abstract pedagogical knowledge with situations resembling those encountered in teaching.

The professional simulation agent required students to evaluate several possible courses of action rather than identify a rule in isolation. Students were prompted to consider learner needs, classroom constraints, professional responsibilities, ethical implications, and the likely consequences of each response. The reflection component further encouraged them to examine why an initial decision might require revision.

The use of both a scale and a situational judgment test is methodologically important. A self-report gain could reflect increased confidence without corresponding development in professional judgment. In the present study, improvement was also evident in the situational test, reducing—although not eliminating—the possibility that the result represented favorable self-perception alone.

The Field Training context probably contributed to the effect because the tasks had direct relevance to participants' approaching professional roles. Research conducted during the rapid transition to online education demonstrated that teachers' ability to adapt depended not only on technological access but also on pedagogical knowledge, confidence, and capacity to respond to unfamiliar professional demands (König et al., 2020). The present intervention integrated these elements rather than treating professional readiness and digital competence as unrelated outcomes.

5.4 Quality of the Digital Instructional Products

Experimental-group students produced substantially higher-quality digital instructional products. The difference was evident across criteria concerning instructional alignment, organization, digital design, interactivity, assessment, technical functionality, accessibility, usability, documentation, revision, and responsible AI use.

Three mechanisms may explain this result. First, students received feedback while products were still under development, allowing them to correct weaknesses before final submission. Second, the specialized agents provided different forms of review rather than one general response. Third, the environment encouraged iterative revision and retained interaction histories that students could consult while improving their products.

The finding is consistent with the broader evidence that well-designed intelligent tutoring systems can produce meaningful learning benefits, although effect sizes vary with implementation quality, outcome alignment, and the nature of the comparison condition. Kulik and Fletcher's (2016) meta-analysis of controlled evaluations found positive effects of intelligent tutoring while also demonstrating that outcomes depended considerably on instructional and methodological conditions. The present result similarly should be attributed to the alignment among support, tasks, and assessment rather than to intelligent technology in the abstract.

It is also important that the official product ratings were assigned by independent human raters rather than the product evaluation agent. The excellent inter-rater reliability supports the reproducibility of the scores and avoids circular evaluation in which the intervention system alone judges the products it helped create.

5.5 Agent-Interaction Patterns

Experimental-group students obtained relatively high interaction-log scores, and these scores were positively associated with all four primary posttest outcomes and product quality. The strongest association was observed with critical thinking, followed by professional readiness measured by the scale. This pattern is consistent with the design of the rubric, which emphasized evidence seeking, comparison of agent outputs, human control, revision, and reflection rather than counting messages or time spent online.

The result supports a distinction between exposure and productive use. Frequent access to a platform does not necessarily indicate meaningful engagement. A student could generate many requests without evaluating or applying the responses. Conversely, a smaller number of purposeful interactions could lead to substantial revision. The interaction rubric therefore considered how students used the agents and what they did with the resulting feedback.

This approach addresses an important concern in learning analytics. Platform metrics may oversimplify learning if observable behavior is treated as equivalent to cognitive development. Guzmán-Valenzuela et al. (2021) cautioned that learning analytics research may emphasize data generation and prediction without adequately representing the complexity of learning. In the present study, the logs were consequently used as complementary behavioral evidence and were interpreted alongside performance and outcome measures.

The correlations remain exploratory. They do not establish that stronger agent interaction caused higher posttest performance. Students who were more motivated, digitally capable, or professionally engaged may have used the environment more effectively. A future study could investigate these relationships through longitudinal process analysis or experimental manipulation of agent-support patterns.

5.6 Usability, Trust, and Satisfaction

Students evaluated the environment favorably, with an average response above four on the five-point scale. This suggests that most participants perceived the environment as understandable, useful, appropriately coordinated, and satisfactory. Positive evaluation is practically important because a technically capable system may have limited educational value if students cannot identify agent roles, navigate tasks, or interpret recommendations.

The questionnaire explicitly distinguished appropriate trust from uncritical acceptance. High trust was not defined as believing every generated response. Instead, students were expected to understand the limitations of AI, verify important claims, and retain control over final decisions. This conception is particularly important in teacher education because professional responsibility cannot be transferred to an automated system.

The moderate positive association between satisfaction and interaction-log quality suggests that students who used the agents more purposefully also tended to evaluate the environment more favorably. Satisfaction was also associated with product quality. These findings are informative but should not be interpreted as evidence that satisfaction caused learning. Perceptions were collected only after the intervention, and favorable responses may partly reflect students' awareness of their successful task completion.

5.7 Integration of the Findings

Taken together, the results support the conceptual model proposed for the study. The environment appears to have influenced the outcomes through several interacting mechanisms:

1. specialized scaffolding matched assistance to the task;
2. critical dialogue required evidence and justification;
3. immediate feedback enabled revision during performance;
4. authentic production connected digital skills with pedagogical purposes;
5. simulated situations supported professional decision-making;
6. interaction histories made revision and reflection visible; and
7. human oversight preserved learner and instructor responsibility.

The differentiated architecture may have been particularly valuable because the dependent variables were interrelated. Critical thinking

supported the evaluation of digital information and AI-generated suggestions. Digital competence enabled students to translate pedagogical decisions into instructional products. Professional readiness required the integration of reasoning, technology use, ethical responsibility, and contextual judgment. The environment addressed these relationships through coordinated support rather than independent instructional units.

5.8 Theoretical Contributions

The study makes three principal theoretical contributions.

First, it extends the concept of instructional scaffolding from a single intelligent tutor to a coordinated system of functionally differentiated agents. The results suggest that specialization may be valuable when learning tasks require several qualitatively different forms of reasoning and performance.

Second, the findings support an integrated understanding of pre-service teacher development. Critical thinking, digital competence, and professional readiness were measured separately but developed within the same authentic Field Training activities. This supports the view that contemporary teacher competence cannot be reduced to isolated technical or pedagogical components.

Third, the study demonstrates the importance of learner control in educational AI. The agents were designed to support reasoning and revision, while students retained responsibility for evaluation and final decisions. The positive outcomes therefore provide evidence for a human-centered instructional model rather than for the replacement of professional judgment by automated output.

5.9 Practical Implications

Teacher-education programs considering similar environments should begin with pedagogical functions rather than the number or novelty of agents. Each agent should have a clearly defined role, explicit boundaries, and prompts aligned with course outcomes. Overlapping roles may confuse students and produce repetitive feedback.

AI support should be embedded in authentic tasks. Lesson planning, classroom-case analysis, digital product development, assessment design, and reflective practice provide more meaningful contexts than isolated demonstrations of AI functionality.

Students require explicit preparation for responsible use. Orientation should address verification, privacy, intellectual property, bias, appropriate trust, and the continuing responsibility of the human decision-maker. AI literacy should be integrated into professional practice rather than treated exclusively as technical training.

Learning analytics should focus on meaningful behavior. Counts of logins, messages, or time online should not be interpreted as direct measures of learning. Analysis should consider evidence seeking, comparison, revision, decision ownership, and transfer to professional performance.

Finally, AI-generated formative feedback should remain separate from official high-stakes assessment. Human raters or instructors should retain responsibility for consequential judgments, particularly when products or performance contribute to course certification.

6. Limitations

Several limitations should be considered when interpreting the findings.

First, the study was conducted at one Egyptian faculty of education with fourth-year students enrolled in the General Education Program. The sample represented several academic specializations, but the results may not generalize directly to other institutions, teacher-education systems, academic levels, or cultural contexts.

Second, the intervention lasted 10 weeks. The study demonstrated immediate post-intervention effects but did not include delayed measurement. It therefore cannot determine whether the observed development was retained or transferred to later independent classroom practice.

Third, the same instructor taught both groups. This controlled instructor-related variation but created the possibility of instructor expectancy effects. Complete instructor blinding was impossible

because the instructor knew which Moodle course included the AI agents.

Fourth, the study compared a comprehensive Multi-Agent AI-based environment with conventional Moodle. It did not isolate the effect of each agent or distinguish the influence of specialization, immediate feedback, prompt design, interaction history, or increased support availability. A factorial or dismantling study would be required to identify the active contribution of individual components.

Fifth, the digital product was assessed only after the intervention. Although random allocation and baseline digital-competence equivalence support the comparison, the absence of a product pretest prevents direct adjustment for prior product-development performance.

Sixth, professional readiness was assessed through a scale and a situational judgment test rather than direct observation of sustained teaching in school classrooms. The instruments captured perceived preparedness and professional judgment, but they cannot fully represent actual teaching performance under authentic classroom conditions.

Seventh, the interaction logs were available only for the experimental group. Their associations with outcomes cannot be compared with equivalent behavioral processes in the control condition.

Eighth, language-model outputs are probabilistic and may vary across requests. Although generation parameters and prompt templates were standardized, exact response replication cannot be guaranteed. The findings therefore concern the implemented system and period rather than all models or later AI systems.

Ninth, the pilot sample included 40 students. Although the instruments produced acceptable overall reliability estimates and were reviewed by 11 specialists, larger independent samples would permit more rigorous examination of dimensionality and measurement invariance.

Tenth, no formal institutional ethics committee approval was obtained. The study applied informed consent, voluntary participation, withdrawal rights, grade separation, pseudonymization, and restricted data access, but the absence of documented prospective review remains an administrative limitation that should be disclosed transparently.

7. Conclusion

This randomized controlled educational study found that a Multi-Agent AI-based learning environment embedded in an online Field Training course was associated with substantially stronger critical thinking, digital competence, professional readiness for teaching, situational professional judgment, and digital instructional product quality than conventional Moodle instruction. These differences remained significant after adjustment for baseline performance and robustness analysis.

The findings suggest that the educational value of AI depends less on unrestricted access to a language model than on the instructional architecture surrounding its use. In this study, specialized roles, critical questioning, immediate formative feedback, authentic production, professional simulation, and structured reflection created conditions in which students were required to evaluate and apply AI-generated support rather than merely receive it.

The study consequently supports a human-centered model of Multi-Agent AI in teacher education. AI agents can extend the availability and differentiation of instructional support, but professional responsibility, ethical judgment, and final decision-making must remain with students and educators. Within these boundaries, a coordinated Multi-Agent AI environment offers a promising approach to preparing future teachers for digitally complex and professionally demanding educational contexts.

Declarations

Ethics Statement

This study was conducted in accordance with recognized ethical principles governing educational research involving human participants. Before participation, students received clear

written information explaining the purpose of the study, its procedures, expected duration, voluntary nature, the collection of learning-platform and AI-interaction data, and the measures adopted to protect confidentiality.

Written informed consent was obtained from all participants before data collection. Students were informed that participation was entirely voluntary and that they could decline participation or withdraw at any stage without academic penalty or any effect on their course grades. Scores obtained from the research instruments were stored separately from the official assessment of the Field. Personal identifiers were replaced with pseudonymous study codes before data analysis. Access to raw Moodle records and AI-agent interactions was restricted to the researcher, and the findings were reported only in aggregated form. No participant withdrew after random allocation.

Informed Consent

Written informed consent was obtained from all participants before their enrollment in the study. Participants also consented to the collection and research analysis of their pseudonymized Moodle activity records and interactions with the Multi-Agent AI system.

Data Availability

The anonymized data supporting the findings of this study are available from the author upon reasonable request, subject to participant-confidentiality requirements. Raw Moodle records and AI-agent conversations are not publicly available because they may contain contextually identifiable educational information. Anonymized derived data, scoring documentation, and statistical analysis procedures may be made available for research verification where ethically and legally permissible.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The author declares that she has no competing financial or non-financial interests that could have influenced the design, implementation, analysis, interpretation, or reporting of this study.

Acknowledgments

The author gratefully acknowledges the pre-service teachers who participated in the study. She also thanks the specialists who reviewed the educational content, learning environment, and research instruments, as well as the independent raters who evaluated the digital instructional products and analyzed the pseudonymized Multi-Agent AI interaction logs.

References

1. Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. Review of Educational Research, 85(2), 275–314. https://doi.org/10.3102/0034654314551063
2. Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 32(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9
3. Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. Topics in Cognitive Science, 1(1), 73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x
4. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.
5. Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), Knowing, learning, and instruction: Essays in honor of Robert Glaser (pp. 453–494). Lawrence Erlbaum Associates.

6. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/BF02310555
7. Darling-Hammond, L. (2006). Constructing 21st-century teacher education. Journal of Teacher Education, 57(3), 300–314. https://doi.org/10.1177/0022487105285962
8. Ertmer, P. A., & Ottenbreit-Leftwich, A. T. (2010). Teacher technology change: How knowledge, confidence, beliefs, and culture intersect. Journal of Research on Technology in Education, 42(3), 255–284. https://doi.org/10.1080/15391523.2010.10782551
9. Facione, P. A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction—The Delphi report. The California Academic Press.
10. Falloon, G. (2020). From digital literacy to digital competence: The teacher digital competency framework. Educational Technology Research and Development, 68(5), 2449–2472. https://doi.org/10.1007/s11423-020-09767-4
11. Guzmán-Valenzuela, C., Gómez-González, C., Rojas-Murphy Tagle, A., & Lorca-Vyhmeister, A. (2021). Learning analytics in higher education: A preponderance of analytics but very little learning? International Journal of Educational Technology in Higher Education, 18, Article 23. https://doi.org/10.1186/s41239-021-00258-x
12. Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81–112. https://doi.org/10.3102/003465430298487
13. Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. Behavior Research Methods, 39(4), 709–722. https://doi.org/10.3758/BF03192961
14. Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 6(2), 107–128. https://doi.org/10.3102/10769986006002107
15. Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70. https://doi.org/10.2307/4615733
16. Instefjord, E. J., & Munthe, E. (2017). Educating digitally competent teachers: A study of integration of professional digital competence in teacher education. Teaching and Teacher Education, 67, 37–45. https://doi.org/10.1016/j.tate.2017.05.016
17. König, J., Jäger-Biela, D. J., & Glutsch, N. (2020). Adapting to online teaching during COVID-19 school closure: Teacher education and teacher competence effects among early career teachers in Germany. European Journal of Teacher Education, 43(4), 608–622. https://doi.org/10.1080/02619768.2020.1809650
18. Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012
19. Korthagen, F. A. J. (2010). Situated learning theory and the pedagogy of teacher education: Towards an integrative view of teacher behavior and teacher learning. Teaching and Teacher Education, 26(1), 98–106. https://doi.org/10.1016/j.tate.2009.05.001
20. Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2(3), 151–160. https://doi.org/10.1007/BF02288391
21. Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. Review of Educational Research, 86(1), 42–78.

https://doi.org/10.3102/0034654315581420
22. Lai, E. R. (2011). Critical thinking: A literature review. Pearson.
23. Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. Journal of Educational Psychology, 106(4), 901–918. https://doi.org/10.1037/a0037123
24. Miao, F., Holmes, W., Huang, R., & Zhang, H. (2021). AI and education: Guidance for policy-makers. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000376709
25. Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. Teachers College Record, 108(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x
26. Moodle Pty Ltd. (2022). Moodle (Version 4.0) [Computer software]. https://download.moodle.org/releases/
27. OpenAI. (2022). OpenAI API [Computer software]. https://openai.com/api/
28. Ramdani, D., Susilo, H., Suhadi, & Sueb. (2022). The effectiveness of collaborative learning on critical thinking, creative thinking, and metacognitive skill ability: Meta-analysis on biological learning. European Journal of Educational Research, 11(3), 1607–1628. https://doi.org/10.12973/eu-jer.11.3.1607
29. Redecker, C. (2017). European framework for the digital competence of educators: DigCompEdu (Y. Punie, Ed.). Publications Office of the European Union. https://doi.org/10.2760/159770
30. Schulz, K. F., Altman, D. G., & Moher, D., for the CONSORT Group. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. BMJ, 340, Article c332. https://doi.org/10.1136/bmj.c332
31. Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153–189. https://doi.org/10.3102/0034654307313795
32. Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive load theory. Springer. https://doi.org/10.1007/978-1-4419-8126-4
33. Tondeur, J., van Braak, J., Sang, G., Voogt, J., Fisser, P., & Ottenbreit-Leftwich, A. (2012). Preparing pre-service teachers to integrate technology in education: A synthesis of qualitative evidence. Computers & Education, 59(1), 134–144. https://doi.org/10.1016/j.compedu.2011.10.009
34. Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.
35. Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. Biometrika, 34(1–2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28
36. Willis, L.-D., Shaukat, S., & Low-Choy, S. (2022). Preservice teacher perceptions of preparedness for teaching: Insights from survey research exploring the links between teacher professional standards and agency. British Educational Research Journal, 48(2), 228–252. https://doi.org/10.1002/berj.3761
37. Wooldridge, M. (2009). An introduction to multiagent systems (2nd ed.). Wiley.
38. Zawacki-Richter, O., Marín, V. I., Bond, M., & 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, Article 39. https://doi.org/10.1186/s41239-019-0171-0