

DOI: 10.37943/XMG L9885

Saule Damekova

Candidate of Pedagogical Sciences, Associate Professor, Pedagogical Institute
saule.damekova2020@gmail.com, orcid.org/0000-0001-8544-3019
Astana International University, Kazakhstan

Dmitry Mukharsky

Master's Degree, Senior Lecturer, Department of Mathematics, Physics and
Computer Science
DMyharskii@shokan.edu.kz, orcid.org/0009-0000-7448-4667
Ualikhanov Kokshetau University, Kazakhstan

Daulet Serikuly

Master's Degree, Junior Researcher, Pedagogical Institute
vwww.14732@gmail.com, orcid.org/0009-0003-5475-0739
Astana International University, Kazakhstan

Aliya Kalieva

PhD Student, Junior Researcher, Department of Mathematics, Physics and Computer
Science
kalievaaliya7@gmail.com, orcid.org/0009-0006-1196-5926
Ualikhanov Kokshetau University, Kazakhstan

Alibek Damekov

Master's Degree, Junior Researcher, Pedagogical Institute
damekov19@mail.ru, orcid.org/0000-0002-0899-4965
Astana International University, Kazakhstan

Erlan Badilov

Master's Student, Junior Researcher, Pedagogical Institute
ebadilov@gmail.com, orcid.org/0009-0001-4220-674X
Astana International University, Kazakhstan

**DESIGN AND DEVELOPMENT OF AN INTELLIGENT MULTI-AGENT
EDUCATIONAL PLATFORM FOR DEVELOPING CRITICAL THINKING AMONG
PRE-SERVICE COMPUTER SCIENCE TEACHERS**

Abstract: The development of generative artificial intelligence expands opportunities for personalised learning support but also increases the risks of cognitive dependence and weakened critical thinking. This issue is particularly relevant to pre-service computer science teachers, whose future professional practice requires logically grounded reasoning, argument evaluation, and informed decision-making. This study aimed to design, develop, and experimentally evaluate an intelligent multi-agent educational platform for fostering critical thinking among pre-service computer science teachers. The methodology included analysis of scientific approaches to critical thinking development and artificial intelligence agents in teacher education, identification of pedagogical and functional requirements, platform architecture and interaction scenario design, prototype development, and a quasi-experimental study with pre- and post-testing. The developed architecture integrates an intelligent tutoring system for argumentation, role-specific artificial intelligence agents for computer science lesson simulation, an expert retrieval-augmented generation module, and an analytics dashboard. Agent operation is organised through profiling, short- and long-term memory, pedagogically constrained planning, and action modules supporting adaptive task generation, dialogue-based guidance, formative assessment, feedback, reflection, and individual progress monitoring. The platform also enables simulation of pedagogical situations aimed at developing future teachers' professional readiness to foster school students' critical thinking. Preliminary experimental results indicate positive changes in critical thinking

performance and demonstrate the pedagogical potential of the platform. The next stage will evaluate the computer science lesson simulation mode and its contribution to professional readiness for developing school students' critical thinking. The practical significance lies in integrating the platform into both subject-specific and pedagogical-methodological components of pre-service computer science teacher education.

Keywords: critical thinking; intelligent multi-agent educational platform; artificial intelligence agent; intelligent tutoring system; artificial intelligence-enhanced educational simulation; retrieval-augmented generation.

Introduction

The widespread adoption of artificial intelligence systems has grown rapidly in recent years. Generative Artificial Intelligence (GenAI) and large language models (LLMs) are having a profound impact on various aspects of life. Students were among the first to recognize the benefits of using these technologies. However, the uncritical use of information generated by these models entails significant risks alongside the obvious advantages. UNESCO's recommendations on the use of generative AI in education emphasize the need for a human-centric approach, the development of user competencies and the maintenance of pedagogical oversight [1].

The global development of artificial intelligence systems necessitates a significant re-evaluation and updating of existing educational methods and approaches [2]. LLMs enable the retrieval, processing and presentation of any information in a concise concentrated form and in a variety of formats. Current models have been trained on the entire body of text-based information accumulated throughout human history and are capable of utilizing it. Integrating LLMs into the educational process opens up opportunities for personalised learning accompanied by robust, individualised feedback.

However, the purposeful and effective use of these powerful tools requires relevant domain knowledge and skills. Large language models are prone to hallucinations and errors [3-4]. To ensure that the information they generate is reliable and does not cause more harm than benefit, users must be able to formulate prompts competently and critically evaluate the model's outputs. Effective prompts should be tested, stored, refined and reused. This process can be automated through specialised software solutions such as artificial intelligence-enhanced educational simulations. Such simulators configure the connected AI models to perform specific roles, impose predefined constraints, adapt user queries and submit validated prompt templates to the LLM.

The OECD Future of Education and Skills 2030 project focuses on the competencies required for life and professional activity in the context of technological change and uncertainty [5]. Within the international PISA framework, creative thinking has been introduced as a distinct assessment domain, reflecting the growing attention of international research to students' ability to address non-routine problems and apply thinking strategies across different contexts [6].

This issue is also highly relevant for Kazakhstan. The need to improve the quality of workforce training, enhance the competitiveness of higher education and align educational programs with the modern demands of the economy and society is outlined in the Concept for the Development of Higher Education and Science of the Republic of Kazakhstan for 2023-2029 [7].

The Delphi Report identifies the following components within a general model of critical thinking (CT): interpretation, analysis, evaluation, inference, explanation and self-regulation [8]. Educational research emphasizes that CT is both general and domain-specific. Therefore, within a discipline-infused approach to CT development, it is appropriate to prioritize the CT skills that most closely correspond to the logic and specific characteristics of the subject being studied.

In the present study argumentation serves as such an integrative skill. Constructing an argument requires analysing a problem, evaluating evidence, drawing a conclusion, explaining one's position and reflectively examining one's own reasoning. In the studies by Osborne and Henderson, the development of argumentation is presented as a progressive transition from formulating a simple claim to using evidence, evaluating arguments and considering counterarguments [9].

S. Toulmin's model is widely used to analyse the structure of an argument and includes a claim, data, warrant, backing, qualifier, and rebuttal. Badjeber et al. [10] and Giri and Paily [11] indicate that evaluating limitations, formulating objections, and constructing counterarguments remain among the most challenging tasks for learners. In programming education, argumentation is manifested in explaining the choice of an algorithm, demonstrating the correctness of a solution, and comparing alternative approaches to solving a problem.

The development of argumentation requires a dialogic organisation of learning. Alexander [12] regards educational dialogue as a means of developing understanding, thinking, and argument evaluation, while Wegerif and Major [13] consider digital technologies as mediators of such interaction. Socratic questioning encourages students to clarify claims, examine evidence, consider alternatives, and revise their initial decisions. However, in traditional classrooms, it is difficult for teachers to sustain extended individual dialogue with every learner. Artificial intelligence agents based on large language models make it possible to scale this interaction and provide personalised dialogue-based support.

Dialogue-based guidance should be combined with formative assessment. Henderson and Lewis [14] demonstrate the role of technology-supported dialogue in facilitating classroom speaking and listening, while Shi et al. [15] show that large language models are increasingly employed in education for tasks such as generating explanations, feedback, and personalised learning support. Xi et al. [16] and Tang and Putra [17] further demonstrate the potential of LLM-based conversational and customised chatbot approaches for supporting reflective thinking, reasoning, argumentation, and personalised dialogue-based learning.

Contemporary research conceptualizes CT as an integration of cognitive skills, dispositions, metacognitive regulation, epistemic beliefs, and artificial intelligence literacy.

Analytical studies identify institutional, methodological, and cultural-pedagogical barriers within the education system that affect the development of students' CT in higher education specifically, university instructors' competence, their own level of CT and mastery of methods to foster it, and students' willingness to put in the effort to change their own mindset. Assessing CT presents an additional challenge.

A promising solution to this problem is the development and integration of artificial intelligence-enhanced educational simulations into the learning process. The development of LLMs has ushered in a new generation of AI simulators, in which virtual students and pedagogical artificial intelligence agents are implemented as multi-agent LLM systems.

Literature review and problem statement

With the emergence of LLMs, development efforts have intensified worldwide, focusing both on practical applications and on the theoretical rationale for using this new tool within educational systems. Researchers have established the theoretical basis for and implemented AI systems that function as tutors, consultants, and expert-mentors, as well as tools for automating and analysing learning activities.

Son [2] presents a systematic and comprehensive review of intelligent tutoring systems (ITSs). The authors note that numerous studies published in recent years have demonstrated the positive effects of intelligent tutoring systems on students' achievement in mathematics. However, it remains unclear whether such systems transform the nature of learning or merely automate and improve the efficiency of traditional instructional practices. Using the SAMR framework, the authors conducted a detailed review of 58 studies published between 2003 and 2023 on the use of intelligent tutoring systems in mathematics education.

The main conclusion of the review is that approximately two-thirds of contemporary intelligent tutoring systems primarily enhance the accessibility, adaptability, and personalisation of traditional instruction without substantially transforming its underlying pedagogical model. The typical functional cycle of such systems includes presenting a task, analysing the learner's response, providing hints, adapting task difficulty, and constructing an individualised learning pathway.

The second part of the article is devoted to systems that significantly rebuild the learning process. According to the authors, their main features are: students independently formulate tasks; game methods are used; research environments are created; the student teaches the virtual agent.

The authors of the article consider training systems that have one or more of the listed features to be the next step in the development of ITS.

Based on an analysis of the effectiveness of various ITS, the authors criticize the widespread approach in which such systems are designed to replace the teacher. The article argues that the highest educational outcomes are achieved in models where the teacher organizes the learning process, while the ITS functions only as a supporting tool. Activities such as discussion, inquiry, and reflection remain the responsibility of the teacher.

The platform was designed on the basis of the characteristics of the most effective intelligent tutoring systems: preserving the teacher's leading role and using artificial intelligence as a support tool. An advanced analytics dashboard enables the monitoring of students' learning activities and outcomes, thereby extending the capabilities of existing AI-based educational platforms.

A promising contemporary approach to the design of intelligent systems is the use of agent-based and multi-agent architectures. Gao et al. [18] identify autonomy, social ability, reactivity, and pro-activeness as key capabilities of intelligent agents. These capabilities enable agents to operate independently, respond to changes in their environment, pursue goals proactively, and interact with other agents or humans. The authors further examine how these capabilities are implemented in contemporary LLM-based agents through perception, reasoning and decision-making, adaptivity, and interaction, thereby providing a conceptual basis for the design of modern artificial intelligence agents and multi-agent systems (MAS).

Li et al. [19] provide a systematic review of contemporary LLM-based multi-agent systems (MAS), focusing on their construction, applications, and current challenges. The authors propose a unified framework comprising five key components: profile, perception, self-action, mutual interaction, and evolution. The study also identifies current challenges and future research directions in the development of LLM-based multi-agent systems. However, its primary focus is on general-purpose problem-solving and world simulation rather than the application of multi-agent architectures to educational technologies and intelligent tutoring systems.

The further development of multi-agent technologies in the context of large language models is reflected in recent studies. Guo et al. [20] provide a contemporary survey of LLM-based multi-agent systems, focusing on their application domains, agent profiling and communication mechanisms, capability development, evaluation benchmarks, and current challenges. The study demonstrates the transition from individual LLM-based agents to multi-agent systems capable of collaborative problem-solving and complex environment simulation.

In the educational context, Jiang et al. [21] propose the von Neumann Multi-Agent System Framework (vNMF) for the design of LLM-based multi-agent systems. The framework represents each AI agent through four functional modules: a control unit, a logic unit, a storage unit, and input-output devices and defines four corresponding types of operations: task decomposition, self-reflection, memory processing, and tool invocation. The proposed architecture focuses primarily on the functional organisation and interaction of specialised agents to support educational processes. However, it does not provide detailed modelling of virtual students with differentiated behavioural profiles or computer science lesson situations, which are central to the educational simulation developed in the present study.

Alongside the development of multi-agent educational architectures, recent research has also examined mechanisms that enable individual LLM-based agents to monitor and regulate their own reasoning. Luo and Jho [22] investigate metacognitive self-monitoring in LLMs, particularly the ability of models to assess the reliability of their own responses. Their work focuses on the internal regulation of LLM reasoning rather than on learner analytics, longitudinal monitoring of student progress, or platform-level integration within an educational environment.

At the pedagogical level, related research has explored how LLM-based systems can support learners' critical thinking, reflection, and independent reasoning. Favero et al. [23] developed a Socratic chatbot designed to foster critical thinking through structured, thought-provoking questions rather than by providing ready-made answers. The chatbot was fine- and prompt-tuned to generate Socratic questions that encourage learners to examine assumptions, consider alternative perspectives, and reflect on their reasoning. Although the approach was evaluated using Theory of Knowledge tasks, it was primarily aimed at developing general critical thinking skills rather than subject-specific knowledge. The chatbot therefore focused mainly on reflection and the stimulation of independent reasoning. However, it was designed primarily as a dialogue-based tutoring tool and did not include an analytics dashboard, longitudinal learner progress monitoring, or integration into a comprehensive educational platform.

Xi et al. [16] developed an LLM-based Socratic intelligent conversational agent that uses structured questioning to support higher-order and reflective thinking. In an experimental study involving 94 university students, the Socratic agent demonstrated advantages over a non-Socratic conversational agent in academic achievement and reflective thinking, particularly in reflection and critical reflection. However, the study focused primarily on the effects of Socratic dialogue within a specific learning context rather than on longitudinal learner analytics or integration into a comprehensive educational platform. Krupp et al. [24] focus on preventing cognitively passive learning and on approaches to reducing students' dependence on prompts and outputs generated by large language models and artificial intelligence systems. The study involved a relatively small sample of 40 students, which limits the generalisability of the findings. In addition, the proposed approach addressed only selected elements of learner interaction with generative AI rather than a complete intelligent tutoring cycle.

Shanto et al. [25] developed and described the AI-CRITIQUE conceptual framework for fostering critical thinking through the use of large language models, particularly ChatGPT. The authors conducted extensive work on formalizing the approach to work with models of generative AI. The main focus was on the evaluation and verification of responses generated by the LLM. The authors proceed from the assumption that any AI answer may contain errors and biases. Learners receive an answer to their question from the LLM and must independently check it against additional sources. To evaluate the effectiveness of the system, the authors analysed students' responses rather than AI-generated responses. The growth of the average level of thinking from the level of simple reproduction of knowledge to the level of rationalisation was considered as an indicator of the effectiveness of the approach.

Pal Chowdhury et al. [26] developed a hybrid ITS. The training system combines both classical methods, which the authors called AutoTutor, and integration with modern LLMs. The authors paid special attention to the limitations of issuing ready-made solutions, as in the previous work. According to the authors, high controllability of the behaviour of the models was observed, and the issuance of ready-made solutions was reduced to a minimum. Pedagogical guardrails inherent in the system are described in the work. This article is relevant to the present study because the authors focused on working with mathematical problems. On the other hand, only a limited range of developed skills of CT is described.

Kahl et al. [27] describe the design of an artificial intelligence-based intelligent tutoring system for robotics. The study provides a detailed comparison of retrieval-augmented generation, prompt engineering, and fine-tuning. The results indicate that retrieval-augmented generation can improve the quality and relevance of tutor responses; however, the evaluation primarily concerns system-response quality rather than student learning outcomes.

Long et al. [28] focus entirely on the use of LLM GPT-4 for the analysis of educational dialogues. And although the system does not interact directly with students and it does not have a tutor component, the authors conducted extensive work on the automation of pedagogical analytics. A high level of agreement with expert assessments was reported. Similar problems are discussed by Lee et al. [29]. The authors developed a system for automating the evaluation of

students' written answers. The authors used the promising Chain-of-Thought method, in which the task set before the LLM is not solved immediately but is divided into successive, logically connected steps. The system neither interacts directly with students nor focuses on the development of CT skills. The work is interesting because it thoroughly examines the full and accurate assessment of answers. The ideas explored in the article are suitable for use as part of more elaborate and complete systems for control and evaluation of answers. The main methods used by the authors of the article are incorporated into the educational platform developed in the present study in a revised form.

Tang et al. [30] developed and tested an interesting idea of evaluating CT through the analysis of peer review among students. The authors did not aim to develop a fully functional educational system. In the work, the feasibility of this approach to assessing CT skills was investigated. The idea is valuable in that it automates the process of diagnosing the levels of CT and can be used in more complete educational systems as well. Future research will examine in more detail and use the proposed approach to assess the progress of the development of CT in future works.

Park and Seo [31] propose a multi-agent LLM-based debate chatbot for assessing critical thinking through argumentation and analytical reasoning. The framework compares single-agent and multi-agent assessment models and evaluates critical thinking based on argumentation skills and intellectual standards. The multi-agent model demonstrated strong agreement with human evaluators, indicating the potential of multi-agent approaches for scalable and structured assessment of critical thinking.

Ta et al. [32] developed the ExGen system for generating programming exercises in the Python language. The main result of the study is that it is not difficult to generate various exercises with the help of LLMs. LLMs perform this task effectively. The main difficulty is in the organisation of automatic selection of high-quality tasks. The central contribution of the authors' work is a multi-level system of filtering and validation of generated tasks. This publication informed the design of prompts and the selection of high-quality generated tasks in the present study.

Recent studies published in 2025–2026 demonstrate a transition toward integrated multi-agent intelligent tutoring architectures. López-Goyez et al. [33] present ELA Tutor, a self-adaptive multi-agent intelligent tutoring system integrated into the Moodle learning management system. The architecture combines specialised agents, a Meta-Agent Orchestrator, persistent experiential memory, interaction logging, databases, and real academic data to support adaptive tutoring and academic monitoring. This study is particularly relevant as an example of integrating multi-agent orchestration and learner-state information within an operational educational platform.

Dai et al. [34] investigated simulation-based teacher learning using AI-powered virtual student agents. Preservice teachers interacted with virtual students in a simulated classroom environment, enabling them to practise dialogic teaching strategies such as probing students' ideas, building on their responses, and supporting collaborative knowledge construction. The study demonstrates the potential of AI-powered virtual students for providing repeated and controlled opportunities for pedagogical practice, which is particularly relevant to the lesson simulation implemented in the present study.

Aperstein et al. [35] present a framework that advances the concept of deliberate pedagogical practice through virtual student agents and mentor agents that support the simulation of learning situations and provide feedback. Unlike the aforementioned solutions, the platform developed in this study focuses on fostering CT among pre-service computer science teachers and integrates an intelligent tutoring system, an artificial intelligence-enhanced computer science lesson simulation, a retrieval-augmented generation module, user dashboards, and learning progress analytics.

Research on closely related topics remains limited in Kazakhstan. Nevertheless, recent studies by Kabzhalelov et al. [36], Muratova and Baibekova [37], and Zhumabekova et al. [38]

demonstrate growing scholarly interest in the use of artificial intelligence in education, including its influence on critical thinking, knowledge acquisition, and the development of students' intercultural competence.

However, these studies primarily examine the effects of artificial intelligence on learning, whereas the development of a comprehensive intelligent tutoring system specifically designed for the purposeful development of CT has not yet received sufficient attention from Kazakhstani researchers.

The analysis of the above studies shows that the focus on the distribution of educational functions between interacting agents demonstrates the high efficiency of the approach. On the other hand, most of the works have either a purely conceptual orientation with no experimental verification, or are focused only on certain functions of learning. Complex platforms are rare and poorly described. The absence of such publications in Kazakhstani journals is especially striking against this background.

The aim was to develop an intelligent multi-agent educational platform, an artificial intelligence-enhanced educational simulation, which we called CritMentor, for the development of CT. The platform is based on MAS and LLMs and integrates many researchers' work in this field. The platform was also intended to support pedagogical experiments on the development of CT skills among future teachers of computer science in IT specialties.

A comparison of existing intelligent tutoring and AI-enabled educational systems with the proposed intelligent multi-agent educational platform is presented in Table 1.

Table 1. Comparison of Existing Intelligent Tutoring and AI-Enabled Educational Systems with the Proposed Intelligent Multi-Agent Educational Platform

Study / system	Architecture type	MAS / agents	Pedagogical analytics	Empirical evaluation	Subject area	Main limitation
Socratic Chatbot [23], 2025	Monolithic chatbot	No / single agent	No	Evaluated through simulated interactions; not tested with real learners	General CT skills	Not integrated into an educational platform

S-ICA [16], 2026	LLM-based Socratic intelligent conversational agent	Single agent; Socratic questioning	Partial reflective-thinking assessment and Epistemic Network Analysis (ENA);	Yes, 94 university students ;	Higher education; Research Methods in Educational Technology	Limited to one disciplinary context and student population; no longitudinal learner monitoring or integration into a comprehensive educational platform
AI Tools against Cognitive Offloading [24], 2024	Pedagogical framework for learner interaction with GenAI aimed at reducing cognitive offloading	No	No	Yes; small sample of 40 students	Physics	Not integrated into an educational platform; CT is not the primary educational objective
AI-CRITIQUE [25], 2024	Pedagogical framework for learner interaction with ChatGPT to develop CT	No	No	Yes	Evaluation and critical analysis of ChatGPT responses in higher education	Not implemented as an educational platform
AutoTutor + LLM [26], 2024	Hybrid ITS combining an LLM with a traditional intelligent tutoring system	No	Partial	Yes	Mathematics, particularly arithmetic word problems, in higher education	Requires manual design of pedagogical scenarios; CT is not the primary objective
AI Tutor for Robotics [27], 2025	RAG, prompt engineering , and fine-tuning	No	Limited	Yes	University-level robotics course	CT is not the primary objective

Analysis of GPT-4 Learning Dialogues [28], 2024	Analytical system	No	Advanced analytics	Empirically evaluated using dialogue datasets	Mathematics and the Chinese language in school education	Does not interact directly with learners; the system analyses educational data but does not provide instruction
Automated Assessment of Written Responses [29], 2024	LLM-based assessment system using prompt engineering	No	Partial	Yes	Natural science disciplines in higher education	Lacks tutoring and simulation components; assessment quality depends on prompts and assessment criteria
CT Assessment through Multi-Agent Debate and Expert Evaluation [31], 2025	Modular multi-agent framework for CT assessment	Yes	Partial	Yes	Higher education; CT assessment	Not an instructional platform; primarily focused on assessment and discussion
ExGen [32], 2026	LLM-based educational content generation system with multistage filtering	No	Minimal	Yes	Python programming in higher education	Evaluation focuses on generation quality rather than educational effectiveness

ELA Tutor [33]	Adaptive multi-agent intelligent tutoring architecture integrated with Moodle;	Yes. Specialised agents coordinated by an RL-based Meta-Agent	Yes. Partial - academic monitoring, interaction data and learner-state information; not focused specifically on longitudinal CT analytics	Yes	Higher education; intelligent tutoring, LMS-supported learning	Focuses on adaptive tutoring and agent orchestration rather than on critical-thinking development, virtual-student simulation, or subject-specific
AI-powered virtual student simulation [34], 2026	VR-supported teaching simulation with AI-powered virtual student agents	Yes	Partial analysis of preservice teachers' emotions and dialogic teaching practices	Yes	Teacher education; dialogic teaching; virtual classroom simulation	Small sample; focus on teacher emotions and dialogic practices rather than longitudinal CT development.
Generative AI-Based Platform for Deliberate Teaching Practice [35], 2025	Conceptual platform for deliberate teaching practice supported by GenAI	Partial	Feedback is included, but advanced group analytics are not provided	No comprehensive empirical evaluation of the platform	Teacher preparation and teaching practice in higher education	Presented as a review and conceptual framework; does not include an implemented intelligent multi-agent educational platform combining an ITS for argumentation, a RAG expert module, and an analytics dashboard

Developed CritMentor Platform, 2025–2026	Intelligent multi-agent educational platform integrating an ITS for argumentation, an artificial intelligence-enhanced educational simulation, a RAG module, and pedagogical analytics	Yes	Individual - and group-level educational analytics	Yes; pilot implementation and a quasi-experimental educational study	Pre-service computer science teacher education; Python programming; methods of teaching computer science	The platform has been evaluated within one subject area. Evaluation of the artificial intelligence-enhanced computer science lesson simulation mode is planned for the next stage of the study
--	--	-----	--	--	--	--

A comparative analysis of existing intelligent educational systems reveals several key areas in which the developed CritMentor platform differs from previous solutions. The systems most closely related to the proposed platform include the multi-agent debate chatbot proposed by Park and Seo [31], which focuses on the assessment of critical thinking; AutoTutor + LLM [26] and ELA Tutor [33] as examples of intelligent tutoring systems; the AI-powered virtual student simulation investigated by Dai et al. [34] as a closely related approach to simulation-based teacher education; and the Generative AI-Based Platform for Deliberate Teaching Practice [35] as a framework supporting pre-service teachers through virtual teaching practice.

From an architectural perspective, the reviewed solutions function as separate systems, and most of their capabilities are not integrated within a single educational platform. In existing systems, CT is developed through dialogic prompting, expert evaluation, critical assessment of AI-generated responses, or reasoning-task design. In CritMentor, by contrast, CT is embedded in the interaction structure and developed through argumentation, analysis of alternative solutions, counterargumentation, reflection, and pedagogical situations related to the professional preparation of pre-service computer science teachers.

In comparable systems, pedagogical simulation is typically implemented as a standalone solution focused on classroom modelling. In CritMentor, artificial intelligence-enhanced educational simulation is integrated into the broader educational environment. Existing systems are generally limited to demonstration-level evaluation and basic analytics, whereas CritMentor provides individual and group analytics, learning-session logging, and assessment of changes in argumentation performance.

Based on the above-mentioned sources, the authors, within the framework of the grant-funded research project, developed their own educational system for the development of CT using MAS and LLM technologies. In the next section, the goals and objectives of the study of influence of AI on the possibilities of development of skills of CT are described.

The aim and objectives of the study

The aim of the study was to design, develop, and experimentally evaluate an intelligent multi-agent educational platform for fostering CT among pre-service computer science teachers.

The objectives of the study were to:

1. Analyse approaches to CT development and the use of artificial intelligence agents in teacher education.
2. Identify the pedagogical and functional requirements for the platform.
3. Develop a platform architecture that supports the development of students' argumentation and CT through subject-specific content and promotes pre-service computer science teachers' professional readiness to foster school students' CT through the simulation of pedagogical situations.
4. Design interaction scenarios between students and artificial intelligence agents.
5. Implement a platform prototype.
6. Conduct an experimental evaluation of the effectiveness of the developed platform.

Materials and methods

The platform architecture

The general architecture of the platform is presented in Figure 1. The infrastructure of the platform is organised based on the container approach using Docker and Docker Compose. The deployment method allows the main components of the system to run as interconnected but logically separate services. The platform includes the following logical services: client layer, server layer, analytics layer and database. A separate LLM service is logically part of the platform architecture but in fact it is a cloud service.

The technological stack of the platform was selected to integrate the web interface, server-side business logic, database, artificial intelligence agents, and semantic search components into a unified educational environment. The client layer is implemented using React/Vite, the server layer is built in Python with FastAPI, and data storage is provided by PostgreSQL. The expert module uses retrieval-augmented generation, supported by vector search based on pgvector. The platform infrastructure is deployed using Docker and Docker Compose, which ensures the reproducibility of the experimental environment and enables the scaling of individual services. This technological stack makes it possible to implement an intelligent multi-agent educational platform that integrates an intelligent tutoring system for argumentation, artificial intelligence-enhanced educational simulation, artificial intelligence agents, and analytics for the development of CT.

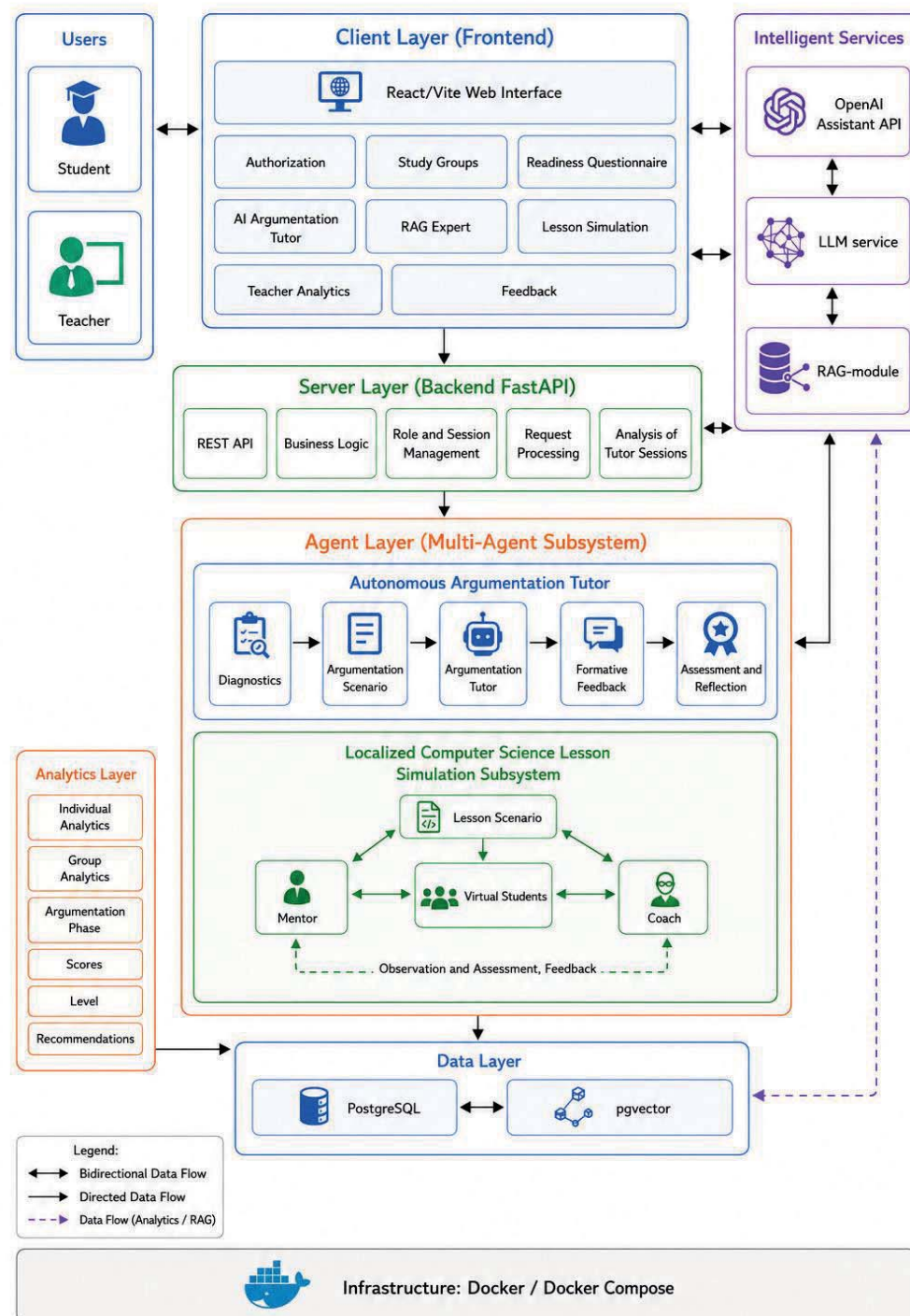


Figure 1. The architecture of the CritMentor.

The platform's agent layer comprises an intelligent tutoring system for argumentation and a local multi-agent subsystem for artificial intelligence-enhanced computer science lesson simulation. The autonomous tutor operates as an independent artificial intelligence agent, whereas the artificial intelligence-enhanced educational simulation module integrates several role-specific agents, including the mentor agent, the virtual student agents, and the instructional coach agent.

The platform's agent layer was designed using a modular approach to the development of artificial intelligence agents. The underlying logic followed an autonomous agent architecture comprising profiling, memory, planning, and action modules.

The profiling module defines the role, objective, pedagogical function, and operational constraints of each artificial intelligence agent. The platform includes profiles for the intelligent tutoring system for argumentation, mentor agent, virtual student agents, and instructional coach agent. The argumentation tutor profile supports the analysis, construction, evaluation, and refinement of arguments. The mentor agent profile guides the student through the computer science lesson scenario. The virtual school-student profiles represent different types of learning responses, misconceptions, errors, and difficulties. The instructional coach profile is responsible for analysing the student's pedagogical actions and providing recommendations.

The memory module supports the storage and use of interaction context. Within the platform, it comprises short-term and long-term memory. Short-term memory includes the current dialogue, selected topic, argumentation phase, student responses, and the state of the learning session. Long-term memory is implemented through the storage of session outcomes, the student profile, assessment history, analytical logs, and materials from the retrieval-augmented generation knowledge base. This enables the artificial intelligence agents to account for the student's previous actions, retrieve relevant fragments of instructional and methodological materials, and generate more evidence-based feedback.

The planning module determines the sequence of pedagogical actions performed by each artificial intelligence agent. In the intelligent tutoring system for argumentation, planning is structured around five phases: understanding arguments, constructing arguments, evaluating arguments, developing counterarguments, and synthesis and conclusion. In the artificial intelligence-enhanced computer science lesson simulation, planning involves selecting the lesson plan, generating virtual student responses, coordinating mentor actions, and formulating assessment recommendations by the instructional coach agent. The planning process is pedagogically constrained: the agents do not act arbitrarily but follow predefined learning scenarios, assessment criteria, and methodological requirements.

The action module translates agent decisions into specific educational actions. These actions include generating questions, hints, learning tasks, virtual student responses, formative feedback, assessment comments, recommendations, and analytical records. The action module uses both the internal capabilities of the LLM and external platform services, including the retrieval-augmented generation module, database, student profile, and analytics module. Thus, the actions of the artificial intelligence agents do not constitute unrestricted text generation but represent pedagogically guided interaction aligned with the objectives of CT development.

Organisation in the form of containers ensures the reproducibility of the experimental environment and simplifies the transfer of the platform between local, test, and server environments.

Internal interaction between levels is carried out through service names in the Docker network. The server layer interacts with other layers using internal names. Thanks to this, the server part does not depend on external addresses and can work in an isolated environment. This approach simplifies the configuration of connections between components and increases the manageability of the platform.

Docker Compose enables centralised management of the platform configuration: service startup, environment variables, ports, and dependencies between components.

Access levels and internal services are differentiated depending on the role of the user (student or teacher).

The above-described organisation of the platform in the form of connected containers provides a foundation for further expansion, including the integration of monitoring, logging, and additional data-processing services.

The platform interface

The platform interface is implemented as a web application accessible to users through a web browser. The system supports two primary user roles: student and instructor. Following registration and authentication, users gain access to the platform sections corresponding to their

assigned role and educational scenario. The student workflow begins with completing a readiness questionnaire designed to assess preparedness for the development of CT.

The central instructional module of the intelligent multi-agent educational platform is the intelligent tutoring system for argumentation. Its interface is presented in Figure 2. Within this module, the student selects the interaction language, instructional module, topic, and argumentation phase. After the session is initiated, the intelligent tutoring system for argumentation presents a task, asks clarifying questions, and guides the student through the process of constructing a reasoned argument. The dialogue-based interface enables the systematic recording of the student's responses and the feedback provided by the artificial intelligence agent.

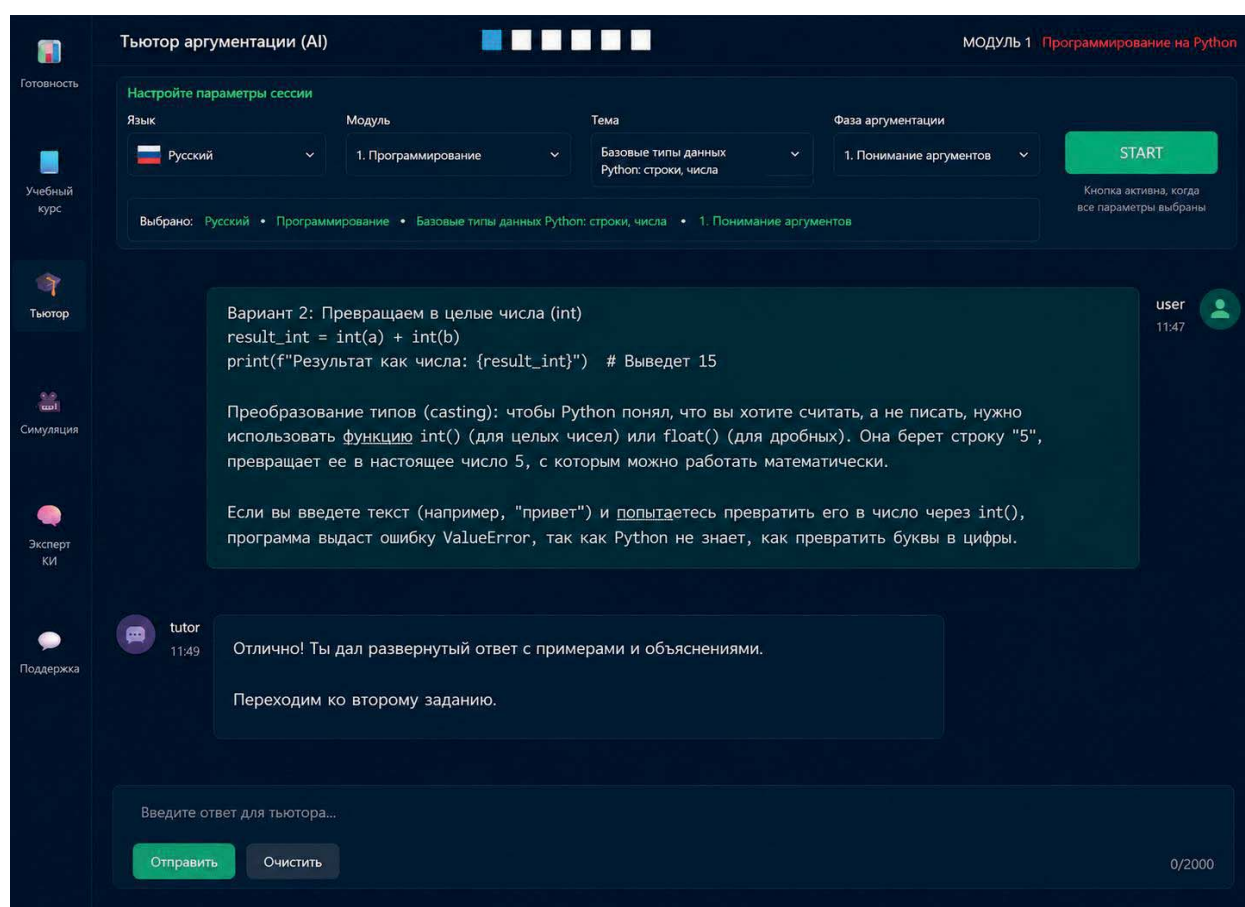


Figure 2. Interface of the Intelligent Tutoring System for Argumentation.

The operation of the intelligent tutoring system for argumentation is structured around five phases of argumentation: argument comprehension, argument construction, argument evaluation, counterargumentation, and synthesis and conclusion. In this study, argumentation is conceptualised as a teachable and assessable learning activity that includes formulating a claim, justifying a position, evaluating supporting reasons, constructing counterarguments, drawing conclusions, and engaging in reflection. This framework underlies the phase-based organisation of the LLM-supported session.

The proposed approach was implemented as a five-phase model of the intelligent tutoring system for argumentation, in which each successive phase required increasingly complex actions related to the analysis, justification, and critical interpretation of the learning task. Toulmin's model of argumentation was used to design the learning tasks, prompt structure, assessment rubric, and quality-control criteria for the responses generated by the intelligent tutoring system for argumentation in the context of Python programming instruction.

A separate module of the platform is the artificial intelligence-enhanced educational simulation. Its interface is presented in Figure 3. In this module, the student assumes the role of a prospective computer science teacher. The user specifies the lesson parameters, including the topic, subject, grade level, lesson type, lesson aim, and learning objectives. During the simulation, the student formulates the teacher's responses, receives replies from virtual students, and may request mentor feedback. The simulation interface enables the modelling of pedagogical situations and supports the development of the professional practices of prospective teachers.

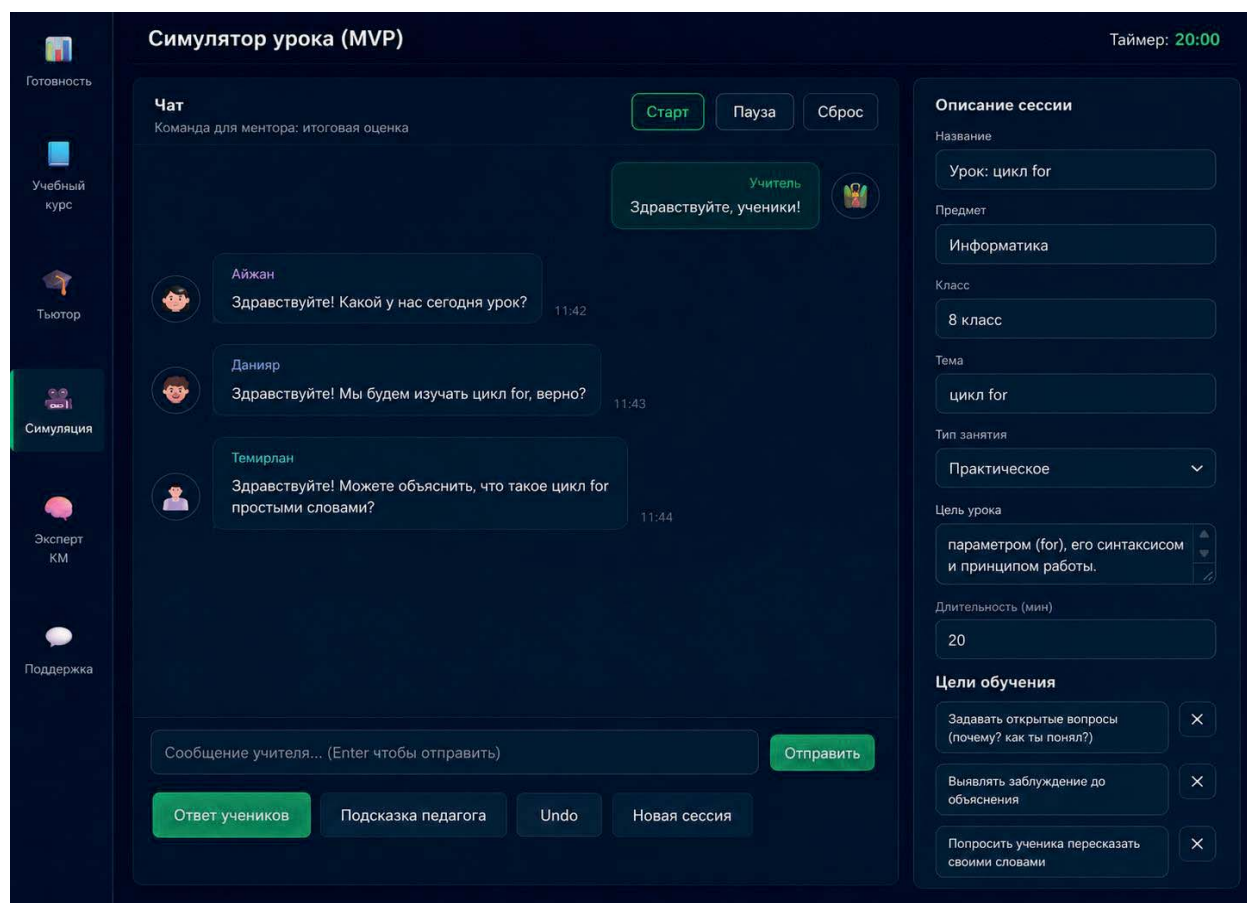


Figure 3. Interface of the Artificial Intelligence-Enhanced Educational Simulation.

For instructors, the key interface component is the analytics dashboard. The analytics dashboard supports two modes of operation: group-level analysis and individual student analysis. In the group analytics mode, the instructor selects an academic group and obtains average performance indicators across the phases of argumentation. In the individual analytics mode, the instructor can locate a specific student, review their average score, phase-specific performance, levels, and recommendations, and access the complete dialogue from the ITS session.

The expert module for CT is implemented as a consultation interface that performs both reference and methodological functions. Users may submit questions related to the theoretical foundations of CT, argumentation, logic, cognitive biases, self-regulation, and metacognitive development.

The Server Layer and Platform Database

The server layer of the platform is implemented in Python using the FastAPI framework. It functions as the central backend component that integrates the user interface, database, intelligent modules, and instructor analytics dashboard. The primary responsibilities of the server layer include processing requests from the client application, managing user sessions, enforcing

role-based access control, administering academic groups, initiating ITS sessions, communicating with LLM services, performing RAG-based semantic retrieval, and preparing analytical data.

PostgreSQL is used as the database management system. The database serves as the central data repository of the platform and supports the storage of user data, academic groups, questionnaire results, ITS sessions, dialogue records, lesson simulations, feedback submissions, analytics indicators, and knowledge-base fragments used by the retrieval-augmented generation system.

User authentication is implemented through server-side sessions. After a user signs in, the backend identifies the user's role and grants access to the corresponding sections of the platform. Students are provided with access to assigned courses, the intelligent tutoring system for argumentation, the artificial intelligence-enhanced educational simulation, and the expert module. Instructors are granted access to academic group management tools and the analytics dashboard. User passwords are stored in the database exclusively as cryptographic hashes, thereby providing a basic level of credential protection.

Academic groups establish relationships among instructors, students, and academic disciplines. These relationships are used not only to organize the learning process but also to support subsequent individual- and group-level analytics. The system enables instructors to create courses, assign students to groups, and retrieve data for a specific academic group or an individual student.

A separate database component is dedicated to diagnosing students' readiness for the development of CT. It stores students' questionnaire responses, total scores, and final readiness levels. These data make it possible to establish an initial baseline of each learner's readiness and to use it as an additional diagnostic indicator in the pedagogical experiment.

The intelligent tutoring system for argumentation sessions occupy a central position in both the server-side logic and the database structure. When a session is initiated, the backend associates it with the user, academic group, topic, and argumentation phase. The student's message is transmitted to the LLM service, while the response generated by the intelligent tutoring system for argumentation is stored in the database together with the original message. As a result, each tutoring session is recorded as a complete dialogue history that can subsequently be accessed for pedagogical analysis.

To improve the quality of the generated tasks, a preliminary LLM-based validation stage is incorporated into the workflow of the intelligent tutoring system for argumentation. Before a task is presented to the student, the system verifies its alignment with the selected topic, difficulty level, argumentation phase, and clarity of formulation. It also evaluates whether the task provides sufficient conditions for constructing a reasoned response and whether its wording avoids directing the student immediately toward a predetermined solution.

Each ITS session stores dialogue history, phase-specific scores, proficiency level, recommendations, and outcome summaries, enabling both individual and aggregated learning analytics.

The expert module for CT is implemented using a retrieval-augmented generation approach. When a user submits a query, the server layer performs a semantic search for relevant fragments within a curated knowledge base, constructs the query context, and transmits it to the language model. The knowledge base contains fragments of instructional and methodological content, including their titles, textual data, and vector representations. Semantic similarity search is performed using pgvector, which enables the system to retrieve context not only through keyword matching but also according to the conceptual relevance of the content to the user's query.

To improve the reliability of the retrieval-augmented generation component, the knowledge base is organised as a curated repository that integrates several types of materials, including theoretical sources on CT and methodological guidelines. Each knowledge-base fragment is supplemented with metadata, such as the source title, material type, thematic category, key concepts, and date of the most recent update. These metadata make it possible to trace the

origin of the retrieved context and facilitate the subsequent pedagogical verification of generated responses.

Based on the metadata, the system automatically generates references to the knowledge-base fragments used in producing the response, thereby enabling the expert module to display the sources on which its output is based. The relationship between the generated response and the retrieved knowledge-base fragments is preserved throughout the response-generation process. This linkage enhances the transparency of the expert module by allowing students to identify the instructional and methodological materials underlying the generated answer.

The artificial intelligence-enhanced educational simulation module is also managed by the server layer. The backend receives the parameters of the pedagogical situation, the message history, and the selected simulation mode, after which it constructs a request to the language model. In response, the system generates either the virtual students' utterances or mentor feedback. The database stores the simulation parameters, including the subject, grade level, topic, lesson type, lesson aim, lesson plan, and the messages produced during the simulated pedagogical interaction.

The analytics component of the server layer is designed to process the outcomes of intelligent tutoring system sessions. The backend extracts the argumentation phase, score, proficiency level, recommendations, and a concise summary of the outcome from the stored dialogues. These data are used in the instructor analytics dashboard to display individual- and group-level dynamics in the development of argumentation skills.

Educational Methodology for the Use of the CritMentor Platform

The educational methodology underlying the use of the CritMentor platform is based on a discipline-infused approach to CT development, in which the skills of analysis, argumentation, evidence evaluation, counterargumentation, and reflection are developed not in isolation but within the subject-specific preparation of pre-service computer science teachers. Python programming tasks and pedagogical situations related to computer science teaching methodology were used as the disciplinary context.

The platform's instructional logic comprises two interconnected learning pathways. The first is implemented through an intelligent tutoring system for argumentation, which supports students in analysing task conditions, formulating a position, selecting evidence, evaluating alternative solutions, and constructing a well-reasoned conclusion. The second is implemented through an artificial intelligence-enhanced computer science lesson simulation, in which students interact with virtual student agents, analyse their responses, identify errors and learning difficulties, select appropriate pedagogical actions, and receive feedback from an instructional coach agent.

CT development is organised progressively across five phases: understanding arguments, constructing arguments, evaluating arguments, developing counterarguments, and synthesis and conclusion. At each phase, the student performs a learning task, receives formative feedback from an artificial intelligence agent, and has the outcome recorded in the LLM-supported session logs. This structure integrates subject-specific learning in programming, argumentation development, and the formation of professional pedagogical skills required by pre-service computer science teachers.

Within this methodology, the instructor is responsible for designing learning scenarios, selecting topics and tasks, interpreting the platform's analytical data, and organising subsequent discussion of the results. The artificial intelligence agents do not replace the instructor; instead, they provide automated formative support by asking clarifying questions, offering hints, modelling learner responses, helping students reflect on their own solutions, and generating recommendations for further development.

Learning outcomes were assessed through a combination of external CT assessment, the platform's internal rubric, and analysis of learning-session logs. The internal rubric was used to evaluate the quality of argumentation, the validity of conclusions, the ability to consider alternative

perspectives, and the level of reflection. Thus, the educational methodology of the CritMentor platform integrates argumentation training, artificial intelligence-enhanced educational simulation for pedagogical practice, formative assessment, and learning analytics.

To enhance the transparency of automated feedback, the CritMentor platform incorporates elements of pedagogically interpretable explainability for AI-generated feedback, recommendations, and assessments. Explainability is implemented not as a separate technical module but as a methodological principle governing feedback generation: each AI-generated comment or assessment conclusion is accompanied by the relevant rubric criterion, the specific fragment of the student's response on which the artificial intelligence agent relies, a brief justification for the assessment, and a recommendation for improvement. When retrieval-augmented generation is used, the relevant fragment of the instructional or methodological material that informed the feedback is also taken into account.

This approach enables students to identify which elements of their argumentation were assessed as strengths or weaknesses, while allowing teachers to verify the justification of automated assessment. Within the CritMentor platform, explainability serves a pedagogical function by increasing the transparency of formative support, reducing risks associated with AI-based assessment, and promoting students' conscious reflection.

A teacher's readiness to foster CT is conceptualised as an integrative construct comprising cognitive and methodological components, dispositions, value orientations, and reflexivity. Since existing instruments do not sufficiently account for the specific characteristics of pre-service computer science teacher education, an original diagnostic model is being developed and is currently undergoing empirical validation.

The pedagogical requirements of the platform include a discipline-infused approach, the progressive development of CT and argumentation, adaptive support, formative assessment, and reflection. The platform functions include diagnostic assessment, task generation and validation, dialogue-based guidance, personalisation of task difficulty, storage of interaction history, recording of learning outcomes, and visualisation of individual and group progress.

The study is being conducted as a two-stage quasi-experiment with pre- and post-testing at Sh. Ualikhanov Kokshetau University, where the control group was formed, and Astana International University, where the experimental group was established. During the first stage, implemented in the second semester of the 2025–2026 academic year, an instructional and methodological unit aimed at developing the basic CT, argumentation, and reflection skills of pre-service computer science teachers was piloted. In the control group, instruction was delivered in a traditional format using instructional and methodological materials, whereas the experimental group additionally used an intelligent multi-agent educational platform with an artificial intelligence-enhanced educational simulation that provided dialogue-based support, task generation, formative feedback, and individual progress tracking.

During the second stage, planned for the first semester of the 2026–2027 academic year, the experimental work will be expanded within the course “Methods of Teaching Computer Science” and other subject-specific disciplines. This stage will focus on developing more advanced skills in critical analysis, pedagogical design, argumentation in professional pedagogical situations, and the solution of interdisciplinary educational and professional tasks.

Data were collected through pre- and post-assessment of students' CT, platform logs, the internal assessment rubric, and expert analysis of learning sessions. The platform recorded student responses across the phases of argumentation, hints, AI-generated feedback, recommendations, and progress indicators. Data analysis included descriptive statistics, comparative and correlational analyses, Student's t-test, and analysis of covariance (ANCOVA). Statistical significance was set at $p < 0.05$, and data processing followed the analytical conventions of SPSS Statistics.

Results

In accordance with the stated objectives, the following results were obtained.

The analysis of scientific approaches showed that CT is most effectively developed through a combination of a discipline-infused approach, dialogic interaction, argumentation, formative assessment, and reflection. The review of contemporary intelligent tutoring systems revealed a growing use of multi-agent architectures, personalised support, artificial intelligence-enhanced educational simulations, and learning analytics. However, most existing solutions implement these functions separately and do not integrate CT development, subject-specific learning, and the professional pedagogical preparation of pre-service teachers within a single platform.

Based on this analysis, the pedagogical and functional requirements for the platform were identified. The pedagogical requirements included a discipline-infused approach, the progressive development of argumentation, the diagnosis of students' readiness, adaptive support, formative assessment, reflection, the simulation of pedagogical situations, and the preservation of the instructor's leading role. The functional requirements included user registration and authentication, role-based interfaces, learning-group management, diagnostic assessment, adaptive task generation, dialogue-based guidance, response assessment, interaction-history storage, learning-session logging, and the visualisation of individual and group progress.

The architecture of the CritMentor intelligent multi-agent educational platform was developed. It integrates an intelligent tutoring system for argumentation, artificial intelligence-enhanced computer science lesson simulation, an expert retrieval-augmented generation module, a database, and an instructor analytics dashboard. The agent layer comprises profiling, short-term and long-term memory, pedagogically constrained planning, and action modules. The autonomous tutor supports argumentation development, whereas the artificial intelligence-enhanced educational simulation integrates role-specific artificial intelligence agents representing a mentor, virtual students, and an instructional coach. This architecture supports CT development through subject-specific content and contributes to pre-service computer science teachers' readiness to foster CT among school students.

Two interconnected scenarios of student interaction with artificial intelligence agents were designed. The first is implemented through the intelligent tutoring system for argumentation and consists of five phases: understanding arguments, constructing arguments, evaluating arguments, developing counterarguments, and synthesis and conclusion. The second scenario is implemented through the artificial intelligence-enhanced computer science lesson simulation, in which the student analyses virtual student responses, identifies errors and learning difficulties, selects pedagogical actions, and receives recommendations from the instructional coach agent. Interaction outcomes are stored in learning-session logs and used for formative assessment and progress monitoring.

A platform prototype was implemented, including modules for user registration and authentication, user and learning-group management, a questionnaire assessing students' readiness for CT development, an intelligent tutoring system for argumentation, an artificial intelligence-enhanced computer science lesson simulation, an expert retrieval-augmented generation module, and an instructor analytics dashboard. The developed software product was entered into the State Register of Rights to Copyright-Protected Objects of the Republic of Kazakhstan, as confirmed by Certificate No. 76272 dated July 8, 2026.

A quasi-experimental study with pre- and post-testing was conducted at two universities: Sh. Ualikhanov Kokshetau University ($n = 60$) and Astana International University ($n = 60$). The first stage of the experiment involved 120 second-year pre-service computer science teachers enrolled in the following educational programmes: Computer Science, Computer Science and Robotics, Mathematics and Computer Science, and Physics and Computer Science. The study was conducted during the second semester of the 2025–2026 academic year and lasted 15 weeks.

The experiment employed a discipline-infused model in which CT development methods and relevant learning content were integrated into the course "Python Programming." At Sh.

Ualikhanov Kokshetau University, CT was developed through traditional instructional methods, including lectures, practical tasks, problem-based learning, dialogue, reasoned justification of solutions, and reflection. At Astana International University, a blended approach was used: traditional instruction was supplemented by the intelligent multi-agent educational platform incorporating an intelligent tutoring system for argumentation. The characteristics of the control and experimental groups are presented in Table 2.

Table 2. Characteristics of the Study Groups

Indicator	Control Group	Experimental Group
University	Sh. Ualikhanov Kokshetau University	Astana International University
Number of students	60	60
Kazakh-medium instruction	55	49
Russian-medium instruction	5	11
Male students	36	24
Female students	24	36

Two freely available practice tests based on the Watson–Glaser format and hosted on the AssessmentDay platform were used to assess CT. The pre-test consisted of 17 items, whereas the post-test included 50 items. Both forms covered the same components of CT and contained items of comparable difficulty. To improve measurement precision and provide a more differentiated assessment of performance, the number of items in the post-test form was increased. To ensure comparability, the scores were converted into percentages of the maximum possible score. The comparability of the two forms in terms of content and expected level of difficulty was confirmed through expert evaluation by three instructors.

As shown in Table 3, both groups demonstrated positive changes, while the experimental group showed a greater mean gain than the control group.

Table 3. Descriptive statistics for CT scores

Group	Pre-test, M $\pm$ SD, percentage points	Post-test, M $\pm$ SD, percentage points	Mean gain, M $\pm$ SD, percentage points
Control group, n = 60	53.14 $\pm$ 13.57	57.53 $\pm$ 11.83	4.40 $\pm$ 18.37
Experimental group, n = 60	61.67 $\pm$ 18.36	68.77 $\pm$ 13.27	7.10 $\pm$ 21.00

In the experimental group, the mean score increased from 61.67% to 68.77%, representing a gain of 7.10 percentage points. A paired-samples Student's t-test showed that this improvement was statistically significant, $t(59) = 2.62$, $p = 0.011$. The within-group effect size was Cohen's $d_z = 0.34$, indicating a small effect. The 95% confidence interval for the mean gain ranged from 1.68 to 12.52 percentage points.

In the control group, the mean score increased from 53.14% to 57.53%, corresponding to a gain of 4.40 percentage points. However, this change did not reach the conventional level of statistical significance, $t(59) = 1.85$, $p = 0.069$, Cohen's $d_z = 0.24$. The 95% confidence interval included zero and ranged from -0.35 to 9.14 percentage points.

Before the intervention, the experimental group demonstrated a higher mean level of CT than the control group (61.67% vs. 53.14%). This difference was statistically significant, $t = 2.89$, $p = 0.005$. Therefore, the pre-test score was included as a covariate in the ANCOVA model when comparing the post-test results.

The results of the ANCOVA comparing post-test CT scores while controlling for pre-test scores are presented in Table 4.

Table 4. Results of the Covariance Analysis of Post-Test Scores

Indicator	Value
Adjusted mean, control group, %	57.78
Adjusted mean, experimental group, %	68.52
Adjusted between-group difference, percentage points	10.75
t for group effect	4.52
p	< 0.001
Partial η^2	0.149
Model R^2	0.173
Model F	12.25
Model p	< 0.001

As shown in Table 4, after controlling for pre-test scores, the adjusted mean in the experimental group was significantly higher than that in the control group, $t = 4.52$, $p < 0.001$, partial $\eta^2 = 0.149$.

The findings indicate that the experimental group demonstrated a statistically significant improvement in CT scores, whereas the positive change observed in the control group did not reach statistical significance. At the same time, a direct comparison of the mean gains did not reveal a statistically significant between-group effect. However, after controlling for baseline differences using ANCOVA, the adjusted post-test score of the experimental group was significantly higher than that of the control group.

The analysis of changes across the five phases of interaction with the intelligent tutoring system for argumentation revealed an overall upward trend. The mean score increased from 3.26 during the argument-understanding phase to 3.60 during the counterargumentation, synthesis, and conclusion-formulation phases. The most pronounced change occurred between the first and fourth phases, after which the scores stabilised.

A repeated-measures analysis of variance revealed statistically significant differences across the phases, $F(4, 184) = 3.42$, $p = 0.010$, partial $\eta^2 = 0.069$. This result was confirmed by the Friedman test, $\chi^2(4) = 13.50$, $p = 0.009$, Kendall's $W = 0.072$. A planned comparison between the

first and fifth phases also demonstrated a statistically significant increase, $t(46) = 2.86$, $p = 0.006$, Cohen's $d_z = 0.42$. The Wilcoxon signed-rank test produced a comparable result ($p = 0.008$).

The fifth-phase score was positively correlated with the Watson–Glaser post-test result ($r = 0.435$, $p < 0.001$), indicating consistency between the platform's internal assessment of argumentation performance and the external assessment of CT.

Discussion

The designed CritMentor architecture demonstrates that the development of students' CT is more appropriately supported not by a standalone AI tool, but by an integrated intelligent multi-agent educational platform. The integration of an intelligent tutoring system for argumentation, an artificial intelligence-enhanced educational simulation, an expert retrieval-augmented generation module, and educational analytics tools makes it possible to connect subject-specific learning, formative support, the modelling of pedagogical activity, and the monitoring of individual progress.

A distinctive feature of CritMentor is its dual-mode architecture. In student mode, the intelligent tutoring system for argumentation supports the development of pre-service teachers' own CT through subject-specific learning materials. In pedagogical simulation mode, artificial intelligence-enhanced educational simulations reproduce school lesson situations that require the analysis of virtual students' responses, the selection of appropriate pedagogical interventions, the formulation of cognitively stimulating questions, and reflection on the instructional decision made. Thus, the platform architecture links the subject-specific preparation of pre-service computer science teachers with the development of their methodological readiness to foster school students' CT. However, during the first stage of the study, experimental evaluation focused primarily on the student mode; therefore, the reported statistical results relate specifically to the operation of the intelligent tutoring system for argumentation.

The functional differentiation of intelligent agents is based on the classical understanding of an agent as a system capable of perceiving the state of its environment, performing actions autonomously, and exhibiting goal-directed behaviour. The works of Wooldridge and Jennings established the theoretical foundations for the development of autonomous and multi-agent systems. Contemporary research on LLM-based multi-agent systems identifies such key components as agent profiling, memory, planning, interaction, and action execution. Within the CritMentor architecture, these principles are operationalised through modules for profiling, short-term and long-term memory, planning, action execution, diagnosis, and formative feedback.

Agent autonomy in CritMentor is pedagogically constrained. System behaviour is determined by the role profiles of the tutor, mentor, virtual student, and instructional coach, as well as by the subject domain, the sequence of the learning scenario, assessment rubrics, and feedback provision rules. These constraints reduce the likelihood of unstructured and pedagogically unguided interaction between the learner and the language model. The platform also incorporates mechanisms for verifying the alignment of generated tasks with the subject content, using external knowledge sources, preserving interaction histories, and enabling the instructor to review the complete dialogue between the student and the artificial intelligence agent.

Unlike systems primarily focused on assessing explanations or responses, CritMentor is designed to organise learning activities that involve analysis, argumentation, evaluation of evidence, consideration of alternative perspectives, counterargumentation, and reflection. The intelligent tutoring system for argumentation facilitates instructional dialogue and performs a formative function. At the same time, the formulation of learning objectives, the selection of content, and final pedagogical decision-making remain the responsibility of the instructor. This distribution of functions is consistent with a human-centred approach to the use of artificial intelligence in education.

The quasi-experimental study revealed positive changes in the CT of pre-service computer science teachers who used the intelligent multi-agent educational platform. In the experimental group, the mean score increased from 61.67% to 68.77%, and the difference between the pre-test

and post-test scores was statistically significant ($p = 0.011$; Cohen's $d_z = 0.34$). In the control group, the mean score also increased; however, the change did not reach the conventional level of statistical significance ($p = 0.069$).

A direct between-group comparison of the mean gains did not reveal a statistically significant difference ($p = 0.454$). At the same time, the experimental group had a higher baseline level of CT. Therefore, analysis of covariance was applied to control for the initial non-equivalence of the groups. After controlling for the pre-test score, the adjusted post-test score of the experimental group was statistically significantly higher than that of the control group ($p < 0.001$; partial $\eta^2 = 0.149$).

The difference between the results of the gain-score analysis and ANCOVA does not constitute a statistical contradiction. Gain-score analysis is based on individual differences between two measurements, which showed substantial variability in the present study. By contrast, ANCOVA estimates the between-group difference in post-test scores after statistically adjusting for baseline performance and generally provides more precise control of initial group non-equivalence.

Additional insight into the mechanism underlying the possible effect is provided by the analysis of the five phases of argumentation. The later and more cognitively demanding phases of interaction with the intelligent tutoring system for argumentation were associated with higher levels of students' argumentation performance. The mean score increased from the argument-understanding phase to the phases of counterargumentation, synthesis, and conclusion formulation. The most pronounced change occurred between the first and fourth phases, after which the scores stabilised.

This pattern may be associated with the staged organisation of learning activities, dialogic support, formative feedback, and mandatory reflection. At the same time, the stabilisation of scores during the fourth and fifth phases indicates not continuous linear growth, but rather the maintenance of the achieved level as students progressed towards integrating arguments, considering alternative perspectives, and formulating a final conclusion. The phase-based dynamics may also be regarded as evidence of the internal consistency of the pedagogical approach.

The positive association between the fifth-phase score obtained during interaction with the intelligent tutoring system for argumentation and the Watson–Glaser post-test result indicates convergent alignment between the platform's internal assessment rubric and the external assessment of CT. Future research will compare LLM-based assessment results with independent evaluations conducted by instructors.

The findings are consistent with the broader development of intelligent tutoring systems. The systematic review by Son [2] indicates that the pedagogical effectiveness of intelligent tutors depends not only on the technical characteristics of the system, but also on how the technology is integrated into instructional content and teaching methods. In the present study, the intelligent tutoring system for argumentation was embedded in a discipline-infused model in which CT was developed directly through the subject content of the “Python Programming” course.

The positive changes observed in the experimental group are also consistent with the findings of Shanto et al. [25], who argue that CT is promoted not by learners' access to ChatGPT itself, but by a structured sequence of questioning, consideration of multiple perspectives, evaluation of responses, synthesis of information, and reflection. A similar pedagogical logic was implemented in the developed AI tutor through the sequential phases of argument understanding and analysis, argument construction, evaluation, counterargumentation, synthesis, and reflection.

The importance of structured dialogue is supported by the study of Favero et al. [23], in which a Socratic chatbot does not provide ready-made answers but instead encourages learners to consider multiple perspectives and engage in reflective reasoning through a sequence of questions. In the present study, a similar principle was implemented in an authentic educational setting with

student participation, thereby extending the empirical evidence on the potential of dialogic support for CT development.

The use of diagnostic criteria and assessment rubrics is consistent with the findings of Lee et al. [29]. The authors demonstrated that the quality of automated assessment of extended responses improves when the language model receives not only the learner's response but also the task instructions, contextual information, and an assessment rubric. In CritMentor, the assessment covers the clarity of the claim, the relevance and sufficiency of evidence, the logical connection between data and conclusions, and the quality of counterargumentation, synthesis, and reflection. The results of automated assessment are used as formative learning analytics and are made available to the instructor through the analytics dashboard for additional pedagogical review.

The study has several limitations. The groups were formed at two different universities without random assignment of participants. Therefore, the possible effect of the platform cannot be fully separated from institutional differences, variations in teaching practices, educational programmes, and student characteristics. The groups also differed significantly in their baseline levels of CT, although this initial non-equivalence was partially controlled for using ANCOVA. The intelligent tutoring system for argumentation was evaluated primarily within a single course, "Python Programming," which limits the generalisability of the findings to other subject areas. The artificial intelligence-enhanced computer science lesson simulation mode has not yet undergone comprehensive experimental evaluation. In addition, the internal assessment of argumentation was partly based on LLM-generated evaluations and therefore requires further validation through independent expert assessment.

The next stage of the study will involve expanding the subject-specific content of the intelligent tutoring system while preserving the overall agent-based architecture. The system is planned to be integrated into the courses "Artificial Intelligence," "Computer Networks," "Computer Architecture," and "Methods of Teaching Computer Science." A separate research direction will focus on the experimental evaluation of the pedagogical simulation mode, including the analysis of pre-service teachers' actions when interacting with virtual students and the assessment of their methodological readiness to foster school students' CT.

Future research will also increase the sample size, incorporate learning analytics indicators, and employ a longitudinal design. This will make it possible to assess the sustainability of the observed effect, its transferability across disciplines, and the contribution of the intelligent multi-agent educational platform to the development of pre-service computer science teachers' methodological readiness.

Conclusion

In this study, the authors developed the comprehensive CritMentor architecture, which integrates an intelligent tutoring system for argumentation, an artificial intelligence-enhanced computer science lesson simulation, a retrieval-augmented generation module, and learning analytics tools. The preliminary findings indicate positive changes in students' CT and suggest the pedagogical potential of the platform for the subject-specific and methodological preparation of pre-service computer science teachers.

The authors acknowledge that CritMentor is currently an experimental research prototype. The effectiveness of the platform in developing CT requires further longitudinal evaluation. Future research will therefore focus on testing the pedagogical simulation mode and conducting longitudinal educational experiments across courses included in pre-service computer science teacher education programmes.

Acknowledgments

This publication was prepared within the framework of the research project IRN AP26104087, "Developing Critical Thinking Skills of Pre-Service Computer Science Teachers through Artificial Intelligence Tools," funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan.

References

- [1] Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. <https://doi.org/10.54675/EWZM9535>
- [2] Son, T. (2024). Intelligent tutoring systems in mathematics education: A systematic literature review using the substitution, augmentation, modification, redefinition model. *Computers*, 13(10), Article 270. <https://doi.org/10.3390/computers13100270>
- [3] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), Article 42. <https://doi.org/10.1145/3703155>
- [4] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631, 755–759. <https://doi.org/10.1038/s41586-024-07566-y>
- [5] OECD. (2018). *The future of education and skills: Education 2030*. OECD Publishing. <https://doi.org/10.1787/54ac7020-en>
- [6] OECD. (2024). *PISA 2022 results: Creative minds, creative schools (Vol. 3)*. OECD Publishing. <https://doi.org/10.1787/765ee8c2-en>
- [7] Pravitel'stvo Respubliki Kazakhstan [Government of the Republic of Kazakhstan]. (2023). *Kontseptsiya razvitiya vysshego obrazovaniya i nauki Respubliki Kazakhstan na 2023–2029 gody* [Concept for the development of higher education and science in the Republic of Kazakhstan for 2023–2029]. <https://adilet.zan.kz/rus/docs/P2300000248>
- [8] Facione, P. A. (2023). *Critical thinking: What it is and why it counts (2023 update)*. Insight Assessment. <https://insightassessment.com/wp-content/uploads/2023/12/Critical-Thinking-What-It-Is-and-Why-It-Counts.pdf>
- [9] Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S.-Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, 53(6), 821–846. <https://doi.org/10.1002/tea.21316>
- [10] Badjeber, R., Supriyatman, S., Afadil, A., & Hadi, W. (2024). A review of Toulmin's argumentation in mathematics and science learning: Implementation, impact, and the role of the teacher. *Jurnal Pendidikan MIPA*, 25(3), 1199–1213. <https://doi.org/10.23960/jpmipa/v25i3.pp1199-1213>
- [11] Giri, V., & Paily, M. U. (2020). Effect of scientific argumentation on the development of critical thinking. *Science & Education*, 29(3), 673–690. <https://doi.org/10.1007/s11191-020-00120-y>
- [12] Alexander, R. (2018). Developing dialogic teaching: Genesis, process, trial. *Research Papers in Education*, 33(5), 561–598. <https://doi.org/10.1080/02671522.2018.1481140>
- [13] Wegerif, R., & Major, L. (2024). *The theory of educational technology: Towards a dialogic foundation for design*. Routledge. <https://doi.org/10.4324/9781003198499>
- [14] Henderson, J. B., & Lewis, A. (2022). Utilizing technology to DiALoG classroom speaking and listening. *Science Scope*, 45(3), 14–18. <https://doi.org/10.1080/08872376.2022.12291453>
- [15] Shi, Y., Yu, K., Dong, Y., & Chen, F. (2026). Large language models in education: A systematic review of empirical applications, benefits, and challenges. *Computers and Education: Artificial Intelligence*, 10, Article 100529. <https://doi.org/10.1016/j.caeai.2025.100529>
- [16] Xi, L., Zhang, Y., & Wang, Q. (2026). Investigating the effects of an LLM-based Socratic conversational agent on students' academic performance and reflective thinking in higher education. *Computers & Education*, 241, Article 105494. <https://doi.org/10.1016/j.compedu.2025.105494>
- [17] Tang, K.-S., & Putra, G. B. S. (2026). Generative AI as a dialogic partner: Enhancing multiple perspectives, reasoning, and argumentation in science education with customized

chatbots. *Journal of Science Education and Technology*. Advance online publication. <https://doi.org/10.1007/s10956-025-10240-1>

[18] Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11, Article 1259. <https://doi.org/10.1057/s41599-024-03611-3>

[19] Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. *Vicinagearth*, 1, Article 9. <https://doi.org/10.1007/s44336-024-00009-2>

[20] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, 8048–8057. <https://doi.org/10.24963/ijcai.2024/890>

[21] Jiang, Y.-H., Li, R., Zhou, Y., Qi, C., Hu, H., Wei, Y., Jiang, B., & Wu, Y. (2024). AI agent for education: Von Neumann multi-agent system framework. In *Proceedings of the 28th Global Chinese Conference on Computers in Education* (pp. 77–84). Global Chinese Conference on Computers in Education. <https://arxiv.org/abs/2501.00083>

[22] Luo, W., & Jho, H. (2026). Predictive metacognition: A neuro-computational framework for self-monitoring in large language models. *Scientific Reports*, 16, Article 23962. <https://doi.org/10.1038/s41598-026-54840-2>

[23] Favero, L., Pérez-Ortiz, J. A., Käser, T., & Oliver, N. (2025). Enhancing critical thinking in education by means of a Socratic chatbot. In *AI in Education and Educational Research (CCIS, Vol. 2519)*, pp. 17–32). Springer. https://doi.org/10.1007/978-3-031-93409-4_2

[24] Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2024). Challenges and opportunities of moderating usage of large language models in education. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia* (pp. 249–254). Association for Computing Machinery. <https://doi.org/10.1145/3701571.3701590>

[25] Shanto, S. S., Ahmed, Z., & Jony, A. I. (2024). Enriching the learning process with generative AI: A proposed framework to cultivate critical thinking in higher education using ChatGPT. *Tuijin Jishu/Journal of Propulsion Technology*, 45(1), 3019–3029. <https://doi.org/10.52783/tjjpt.v45.i01.4680>

[26] Pal Chowdhury, S., Zouhar, V., & Sachan, M. (2024). AutoTutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (pp. 5–15). ACM. <https://doi.org/10.1145/3657604.3662041>

[27] Kahl, S., Löffler, F., Maciol, M., Ridder, F., Schmitz, M., Spanagel, J., Wienkamp, J., Burgahn, C., & Schilling, M. (2025). Evaluating the impact of advanced LLM techniques on AI lecture tutors for a robotics course. In *AI in Education and Educational Research (CCIS, Vol. 2519)*, pp. 149–160). Springer. https://doi.org/10.1007/978-3-031-93409-4_14

[28] Long, Y., Luo, H., & Zhang, Y. (2024). Evaluating large language models in analysing classroom dialogue. *npj Science of Learning*, 9, Article 60. <https://doi.org/10.1038/s41539-024-00273-3>

[29] Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, 6, Article 100213. <https://doi.org/10.1016/j.caeai.2024.100213>

[30] Tang, T., Sha, J., Zhao, Y., Wang, S., Wang, Z., & Shen, S. (2024). Unveiling the efficacy of ChatGPT in evaluating critical thinking skills through peer feedback analysis: Leveraging existing classification criteria. *Thinking Skills and Creativity*, 53, Article 101607. <https://doi.org/10.1016/j.tsc.2024.101607>

[31] Park, B., & Seo, K. (2025). Assessing critical thinking through a multi-agent LLM-based debate chatbot. In *Extended Abstracts of the 2025 CHI Conference on Human Factors in Computing Systems* (Article 80, pp. 1–13). Association for Computing Machinery. <https://doi.org/10.1145/3706599.3721207>

[32] Ta, N. B. D., Nguyen, H. G. P., & Gottipati, S. (2026). Automatic generation of introductory programming exercises with large language models. *Research and Practice in Technology Enhanced Learning*, 21, Article 025. <https://doi.org/10.58459/rptel.2026.21025>

[33] López-Goyez, J. P., González-Briones, A., & Demazeau, Y. (2026). An adaptive multi-agent architecture with reinforcement learning and generative AI for intelligent tutoring systems: A Moodle-based case study. *Applied Sciences*, 16(3), 1323. <https://doi.org/10.3390/app16031323>

[34] Dai, C.-P., Ke, F., Zhang, N., Barrett, A., West, L., Bhowmik, S., Southerland, S. A., & Yuan, X. (2026). Investigating preservice teachers' emotions and dialogic teaching practices in simulation-based learning with artificial intelligence-powered virtual student agents. *Learning and Instruction*, 104, 102381. <https://doi.org/10.1016/j.learninstruc.2026.102381>

[35] Aperstein, Y., Cohen, Y., & Apartsin, A. (2025). Generative AI-based platform for deliberate teaching practice: A review and a suggested framework. *Education Sciences*, 15(4), Article 405. <https://doi.org/10.3390/educsci15040405>

[36] Kabzhalelov, K. R., Korganbaeva, Zh. K., Nurakhmetova, A. R., Meirmanova, A. A., & Sagimbayeva, A. E. (2026). The impact of AI tools on critical thinking in science education: A systematic review of theoretical advancements and practical outcomes. *Bulletin of the Karaganda University. Pedagogy Series*, 31(1), 234–247. <https://doi.org/10.31489/2026ped1/234-247>

[37] Muratova, K., & Baibekova, M. (2025). Vliyanie iskusstvennogo intellekta na vospriyatie i usvoenie znanii studentami v protsesse samostoyatel'noi raboty [The influence of artificial intelligence on students' perception and acquisition of knowledge during independent work]. *Bulletin of the Karaganda University. Pedagogy Series*. <https://doi.org/10.31489/2025ped4/106-117>

[38] Zhumabekova, G. B., Zhusupova, R. F., Islam, A., & Kemaldan, E. G. (2024). Fostering intercultural competence: Unleashing the power of AI-based tools in shaping student critical thinking skills. *Bulletin of Ablai Khan KazUIRandWL. Pedagogical Sciences Series*, 73(2), 191–202. <https://doi.org/10.48371/PEDS.2024.73.2.012>