Multi-Agent Systems in Education: A Survey from the Trustworthiness Perspective

Chahana Dahal, Jinming Chen, Muchao Ye, *Member, IEEE*, and Zuobin Xiong, *Member, IEEE*

***Abstract*—Multi-agent systems are becoming an important part of educational technology by helping students learn through collaboration, feedback, and adaptive support. This paper reviews previous and recent research on multi-agent learning systems and explores the roles agents can take, such as tutors, peers, or facilitators. To the best of our knowledge, this is the first comprehensive survey of multi-agent learning systems that introduces a pedagogy-based role and interaction taxonomy, compares system architectures beyond LLMs, and links design to learning theory. The fields are grouped on the basis of system architecture, learning scenarios, and connections to learning theories such as constructivism and the Zone of Proximal Development. The survey examines how these systems are evaluated and the challenges of measuring learning, engagement, and collaboration. In addition, this work provides a comprehensive study of emerging trustworthiness frameworks that address safety, privacy, fairness, and transparency in multi-agent learning systems. Finally, the paper concludes with open challenges and future directions in using the multi-agent learning system in education scenarios, including the need for better simulation environments, stronger coordination between humans and AI agents, etc.**

***Impact Statement*—This paper examines how multi-agent educational systems are reshaping AI-supported learning. It introduces a taxonomy that links agent roles, system architectures, and learning theories. The study highlights a clear shift from rule-based tutoring to collaborative, large language model (LLM) systems that simulate classrooms, peer feedback, and research teams. These systems enable richer interaction patterns that mirror real-life learning. The paper also analyzes emerging trustworthy frameworks that address safety, privacy, fairness, and transparency in multi-agent classrooms. By comparing symbolic, hybrid, and LLM-based designs, it shows how social and cognitive theories can guide safer and more inclusive AI learning environments. The work provides educators, developers, and policymakers with a foundation for building multi-agent systems that are pedagogically sound and ethically aligned. This paper shows that educational AI is a field that is moving from single-agent instruction to collective, human-centered learning ecosystems.**

***Index Terms*—Multi-Agent Systems, AI Agents, AI in Education, Large Language Models, Trustworthy AI**

I. Introduction

In recent years, the integration of artificial intelligence (AI) into educational technologies has transformed the way learners interact with digital systems. Intelligent tutoring systems, personalized learning platforms, and AI-powered content generation tools are now widely deployed in educational settings. There is a growing shift from AI systems that act as solo tutors towards where multiple AI agents interact with each other, learners, and instructors to facilitate learning [1]. This shift toward collaborative AI systems reflects key insights from the learning sciences, which emphasize that social interaction, collaboration, and communication are essential for effective learning and cognitive development. Theories such as Vygotsky's Zone of Proximal Development [2] suggest that learners benefit the most when they receive support just beyond their current abilities. Similarly, constructivist pedagogy promotes active learning through exploration, dialogue, and problem-solving. Multi-agent educational systems apply these ideas by simulating environments where students learn through interactions with AI-powered tutors, peers, and facilitators.

Recent advances in AI have made such multi-agent systems possible. For example, multi-agent reinforcement learning (MARL) [3] allows groups of AI agents to learn to cooperate by interacting with each other in shared environments and receiving rewards for their collective performance. Symbolic AI uses logic and rule-based reasoning, which enables agents to explain their actions and follow structured problem-solving steps [4]. Recent hybrid approaches combine the strengths of both symbolic and neural methods, termed as neuro-symbolic agents, combining logical rules with deep learning and creating learning systems that are both understandable and flexible [5]. On top of these, large language models (LLMs), such as GPT or PaLM, are trained on massive text datasets and are proficient at understanding and generating human-like language [6].

This paper provides a comprehensive survey of multi-agent educational systems by introducing a novel taxonomy of agent roles (e.g., tutor, peer, evaluator, and mediator). It compares symbolic, neuro-symbolic, MARL, and LLM-based implementations, and aligns these systems with foundational learning theories. Beyond pedagogical design, the survey highlights the importance of the trustworthiness of learning agents as multi-agent systems are increasingly influencing classroom feedback and collaboration. It also outlines a roadmap for inclusive, reliable, and transparent AI learning environments that align teaching goals with safety requirements. Unlike prior surveys that focus on single-agent tutoring or coordination, this paper presents the first pedagogically and ethically grounded view of multi-agent educational systems. It includes a new role- and interaction-based taxonomy, evaluates system architectures beyond LLMs, and links design choices to both learning theory and trust/safety frameworks in AI-supported education.

Chahana Dahal and Zuobin Xiong are with the Department of Computer Science, University of Nevada Las Vegas, NV 89154 USA (e-mail: {*chahana.dahal, zuobin.xiong*}@unlv.edu)

Jinming Chen is an independent researcher. (email: u3638159@connect.hku.hk)

Muchao Ye is with the Department of Computer Science, University of Iowa, Iowa 52242 USA (e-mail: muchao-ye@uiowa.edu).

The contributions of this survey are summarized as follows:

- A novel pedagogical role taxonomy that categorizes agents based on their pedagogical roles (e.g., tutor, peer, evaluator), interaction structures (e.g., turn-taking, debate, consensus-building), and learning scenarios (e.g., collaborative coding, group discussions) is proposed.
- The symbolic, reinforcement learning, hybrid, and LLM-based multi-agent architectures used in education were compared, highlighting their strengths, limitations, and use cases.
- Multi-agent designs are compared with foundational pedagogical models such as constructivism, peer teaching, dialogic learning, and debate-based instruction, which can bring a broad impact beyond computer science.
- From the safety and trustworthiness perspective, how modern educational multi-agent systems apply safeguards for reliability, transparency, fairness, and cultural sensitivity was analyzed.
- Open problems, including agent coordination, safety assurance, and trust calibration, are analyzed, and evaluation metrics to assess learning impact, interaction quality, and ethical compliance are proposed.

Survey Methodology. A systematic literature search was conducted to map research on multi-agent educational systems. The search covered leading conferences and journals in AI (e.g., AAAI, IJCAI), educational technology (e.g., AIED, EDM, LS), natural language processing (e.g., ACL, EMNLP, NeurIPS), and human-computer interaction (e.g., CHI, ICLS), along with relevant preprints on arXiv. Keyword combinations such as *"multi-agent learning," "educational agents," "collaborative AI tutors,"* and *"AI classroom simulations"* were used. Backward reference tracing was additionally applied to capture influential works cited in key papers and surveys.

From an initial set of 60 papers, 30 were identified as relevant to multi-agent systems in educational contexts and aligned with the taxonomy of agent roles, interaction structures, and learning scenarios developed in this survey. From these, 23 primary studies were selected as full research papers. Papers were excluded if they: (a) existed only as workshop abstracts or early-stage prototypes, (b) addressed general multi-agent frameworks without direct educational application, or (c) focused solely on single-agent tutoring without multi-agent coordination. Each selected work was classified along three dimensions: Educational Role, Interaction Structure, and Learning Scenario.

For trustworthiness-related research, additional searches using terms such as *"safety of educational AI agents," "privacy in intelligent tutoring,"* and *"fairness in educational recommender systems"* were conducted, followed by reference list expansion to identify further methodologies and perspectives. The reviewed literature informed the design of the pedagogical role taxonomy (Figure 2), providing a structured overview of how current multi-agent education systems align with learning theories and practices.

Organization. The remainder of this paper is organized as follows, and an overview of the survey structure is displayed in Fig. 1. Section II introduces the background and theoretical foundations of multi-agent systems in education. It reviews prior work in AI for education and discusses key learning theories that motivate the use of multi-agent approaches. Section III delves into system architectures for multi-agent learning. Section IV examines agent roles in learning environments, including educational roles, interaction structures, and learning scenarios enabled by multi-agent coordination. Applications and case studies of multi-agent systems in education are reviewed in Section V. Section VI discusses trustworthiness in AI agents, with a focus on safety, privacy, fairness, and transparency in multi-agent educational systems. Section VII surveys evaluation and benchmarking methodologies and focuses on evaluation challenges. Finally, the research trends, open challenges, and future directions are summarized in Section VIII.

II. Background and Theoretical Foundations

A. Historical Perspective on AI in Education

The use of AI in education has evolved significantly over the past few decades. Early systems, such as Computer-Assisted Instruction (CAI) [7], used static, rule-based systems for drill-and-practice learning, yet these systems offered limited adaptability or feedback. The next major advance was Intelligent Tutoring Systems (ITS) [8]. ITS used domain models, student models, and decision rules to personalize instruction. They enabled more interactive and adaptive feedback, but remained largely single-agent systems where the collaborations in different systems are restricted. More recently, LLMs have introduced flexible, general-purpose capabilities. When used in agents, LLMs can hold conversations, give open-ended feedback, and act like peer learners. This allows for more flexible and natural interactions than what traditional ITS can offer [6].

B. Why Multi-Agent Systems for Education?

Contemporary learning theories [2], [5] emphasize that cognition is not solely an individual process but is shaped through social interaction. Frameworks like Vygotsky's Zone of Proximal Development and Bandura's Social Learning Theory highlight the role of collaboration, modeling, and feedback in learning. Multi-agent educational systems support these theories by enabling AI agents to interact not only with human learners but also with one another. These agents take on roles like peer tutors, evaluators, and facilitators, and can mirror real classroom dynamics. Through interactions such as debate, negotiation, and collaborative knowledge building, they create rich and engaging learning experiences.

Unlike single-agent systems focused on one-to-one tutoring, multi-agent systems enable distributed reasoning and group-based problem-solving. This shift enables more scalable, interactive, and pedagogically aligned AI environments, making them well-suited for virtual or hybrid classrooms. Multi-agent educational systems increasingly draw on foundational theories from the learning sciences. These theories guide how agents are designed to support cognition, motivation, and collaboration. Below, a few major frameworks in education theory are presented, and how current AI capabilities align with them are highlighted.

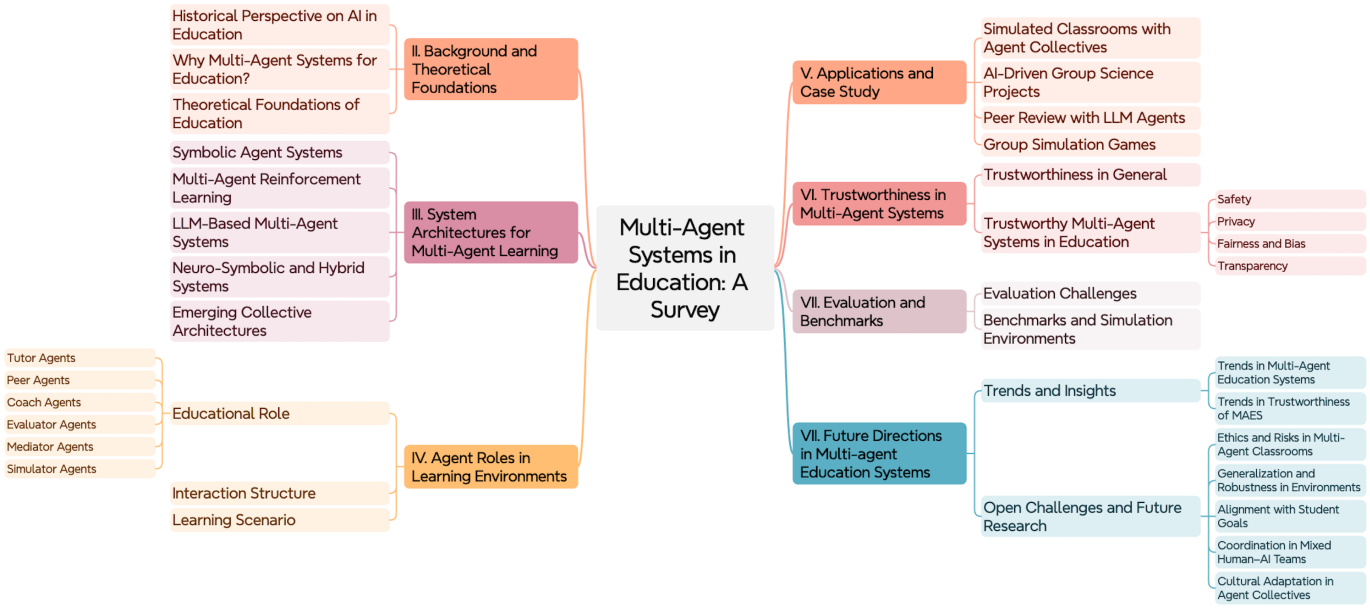

Fig. 1. The organization framework of the survey paper.

C. Theoretical Foundations of Education

1) Constructivist Learning and Distributed Cognition: Constructivist approaches suggest that learning is rarely a matter of simply absorbing facts. It happens through active engagement, experimentation, and negotiation of meaning [9]. From this angle, knowledge is not something stored in an individual's head once and for all, but something that takes shape as people interact with one another and with the tools around them. The idea of distributed cognition pushes this further: it suggests that the "unit of learning" is often not the individual at all, but the broader system that includes peers, artifacts, and environments. In multi-agent contexts, this implies that agents should not only deliver information but also participate in the co-construction of understanding. Such systems encourage learners to explore through trial and error. They promote shared responsibility in problem-solving and support reflection on outcomes with artificial peers. Instead of presenting knowledge as fixed, these agents show that learning is dynamic and shaped by the quality of interaction within the group.

2) Zone of Proximal Development: Vygotsky's concept of the Zone of Proximal Development (ZPD) captures a very practical truth. Learners often perform best when tasks are just beyond their reach, but still within range if they have the right kind of help [2]. Translating this into multi-agent systems, agents can take on the role of a slightly more capable peer. These agents do not simply give solutions, instead, they provide scaffolding that helps learners work through challenges on their own. This support can include timely hints, demonstrations of strategies, or subtle cues that keep learners engaged. Over time, the level of assistance decreases as learners gain confidence and take more responsibility: similar to how human mentors build independence. Importantly, agents designed with the ZPD in mind can also act as role models, showing ways of thinking or problem-solving that learners can adopt into their own practice.

3) Dialogic Learning and Socratic Critique: Dialogic theories suggest that understanding deepens when knowledge is worked out through dialogue. Learners deepen their knowledge when ideas are questioned, compared, and refined in conversation. Rather than viewing the learner as a passive recipient, this approach sees them as an active participant in joint exploration. AI agents based on dialogic principles act as conversational partners, where they engage learners in structured debate, prompt self-explanation, and sometimes play the role of a devil's advocate to reveal hidden assumptions. Socratic questioning is especially valuable in this process. By asking why, how, or what if at key moments, agents encourage learners to make their reasoning explicit. This dialogue strengthens content understanding and builds habits of reflection and critique that are essential to academic inquiry. From this perspective, dialogic AI agents are not just instructional tools but intellectual companions that help learners practice the forms of discourse found in scholarly communities.

4) Additional Learning Frameworks: Beyond constructivism, ZPD, and dialogic theories, several complementary frameworks guide multi-agent educational design. Self-Regulated Learning (SRL) and meta-cognitive theory emphasize how learners plan, monitor, and reflect on their learning—principles used by tutor and coach agents such as MetaTutor [10]. The ICAP framework [11] classifies learning interactions by levels of cognitive engagement—interactive, constructive, active, and passive. It provides a foundation for structured turn-taking in educational dialogue. Cognitive Apprenticeship [12] builds on constructivist theory by modeling expert reasoning and gradually reducing support, an approach used in systems like Coscientist. Coaching psychology and self-determination theory guide the design of motivational and reflective dialogue in coach agents [13]. Argumentation theory

and game-based learning inform evaluator and negotiation agents by linking debate and simulation for deeper reasoning and better decision-making.

III. System Architectures for Multi-Agent Learning

Multi-agent educational systems can be built using different AI architectures. Each architecture supports different forms of reasoning, coordination, and interaction. This section describes four major paradigms: symbolic agents, multi-agent reinforcement learning (MARL), LLM-based agents, and neuro-symbolic systems. The reference table that compares and classifies literature is provided in Table I.

A. Symbolic Agent Systems

Symbolic architectures use rule-based logic, semantic representations, and predefined inference mechanisms. These systems perform great in domains where knowledge can be clearly structured, and reasoning steps must be transparent. For instance, Betty's Brain [14] used a symbolic peer agent that learned concept maps from students, reinforcing constructivist principles of "learning by teaching." Similarly, self-explanation research [34] has shown how prompting learners to articulate their reasoning enhances metacognition and conceptual understanding. The strength of symbolic systems lies in their high explainability and pedagogical alignment, though they often struggle with adaptability in open-ended learning contexts.

B. Multi-Agent Reinforcement Learning (MARL)

In MARL-based systems, agents learn to cooperate or compete by interacting with each other and their environment. Each agent develops policies through trial-and-error reward structures, which can give rise to emergent behaviors. While fewer educational systems have fully adopted MARL compared to symbolic or LLM-based approaches, simulation-driven environments such as EcoMUVE [33] prefigured this approach by embedding agents into complex ecosystems where learners engage in inquiry-based science. MARL is especially promising for game-like or decision-making scenarios where group dynamics, negotiation, or resource management mirror real-world problem-solving. However, the scalability and interpretability of MARL remain open challenges.

C. LLM-Based Multi-Agent Systems

LLMs have transformed the design of educational agents by enabling flexible, natural language interactions. LLM-powered agents can simulate peers, coaches, or even full classrooms. For example, SimClass [18] coordinated multiple LLM agents acting as teachers, classmates, and facilitators to simulate realistic classroom interactions. AgentReview [23] has used LLM agents to replicate the peer-review process with roles such as reviewer, author, and area chair. Similarly, LLM-Powered Classrooms [19] and PairBuddy [20] showed how agents can collaborate with learners in coding or discussion tasks. These architectures are highly scalable and versatile, but their reasoning processes remain implicit and often hard to interpret.

D. Neuro-Symbolic and Hybrid Systems

Hybrid architectures combine the interpretability of symbolic reasoning with the adaptability of neural models. In education, these systems are particularly effective for tasks that require both structured guidance and flexible adaptation. Coscientist [31] illustrated this approach by coordinating multiple specialized LLM agents in scientific research pipelines while still incorporating symbolic planning to manage tasks. BioAgents [32] applied a similar idea to genomic data analysis by distributing work across LLM agents with symbolic task allocation and neural processing. Hybrid designs offer a strong balance, and they enable guided reasoning and explainability without losing the adaptability and fluency of neural systems.

E. Emerging Collective Architectures

Recent work pushes beyond individual paradigms toward collective agent ecosystems, where dozens of agents interact with each other and the learner simultaneously. MoralSim [29] demonstrated how LLM agents can simulate repeated social dilemmas and highlighted moral reasoning and social-emotional learning. Mediator systems such as Self-play negotiation with a critic/mediator model [35] and LLM-Deliberation [26] showcased how multi-agent dialogue and negotiation can emulate authentic group processes. These collective architectures move closer to simulating the social and distributed nature of real classrooms, yet they raise new challenges of coordination, consistency, and cultural adaptation.

IV. Agent Roles in Learning Environments

To understand how multi-agent educational systems operate in practice, the survey proposes a taxonomy of agent roles based on learning science and system design. This taxonomy, illustrated in Figure 2, classifies agents according to their instructional function, modes of interaction, and the learning scenarios they support. It aims to clarify how different agent roles reflect the main pedagogical strategies. Unlike traditional systems with a single tutor, multi-agent environments divide instructional tasks among various roles, such as teacher, peer, coach, evaluator, or mediator. These roles target different phases of learning, where some focus on content delivery or feedback, while others support group coordination or simulate real-world settings. Moreover, the way agents interact (e.g., turn-taking, debate) and the context of use (e.g., coding, discussion) further shape their educational impact.

A. Educational Role

In multi-agent educational systems (MAES), agents take on different educational roles that guide how they support learners and interact with each other. These roles include teacher, peer, coach, evaluator, mediator, and simulator. These reflect a different way of helping students learn. This section explains these roles and highlights systems that demonstrate them.

Teacher/Tutor Agents: Teacher or tutor agents are the most common in educational AI. They give direct instruction, guide students through tasks, and provide personalized feedback. For

TABLE I
THE SUMMARY OF MULTI-AGENT EDUCATIONAL SYSTEMS LITERATURE

Architecture	System / Citation	Reasoning	Theoretical Foundation(s)
Symbolic	MetaTutor [10]	Rule-based reasoning	Self-Regulated Learning; Constructivism (learning-by-teaching)
	Betty's Brain [14]	Symbolic concept reasoning	Constructivism; Zone of Proximal Development
	Self-Explaining Agents [11]	Scripted prompting	Constructivism; Metacognition
	Academically Productive Talk Agents [15]	Scripted dialogue rules	Computer-Supported Collaborative Learning; Argumentation Theory
	Peer Bots [16]	Scripted / constrained peer dialogue	Peer Learning; Collaborative Learning
	Two Heads Better [17]	Scripted multi-perspective trialogue	Social Cognition; Multi-Perspective Learning
	Tutor-Coach Hybrid [13]	Coaching dialogue policy / scripted scaffolds	Self-Regulated Learning; Coaching Psychology
LLM-Based	SimClass [18]	Implicit language-based reasoning	Social Constructivism; Dialogic Learning
	LLM-Powered Classrooms [19]	Peer-style LLM collaboration	Collaborative Learning; Social Learning Theory
	PairBuddy [20]	Conversational problem solving	Cognitive Apprenticeship; Collaborative Learning
	Proactive Conversational Coaches [21]	Goal-directed dialogue planning	Coaching Psychology; Self-Determination Theory
	Group Argumentation [22]	Debate-based reasoning	Argumentation Theory; Collaborative Learning
	AgentReview [23]	Role-based dialogic reasoning	Dialogic Theory; Social Learning Theory
	Dialogic Critique Agents [24]	Reflective critique dialogue	Dialogic Learning; Feedback Intervention Theory
	Simulation Facilitators [25]	Guided facilitation dialogue	Collaborative Learning; Guided Discovery
	LLM-Deliberation [26]	Multi-agent negotiation reasoning	Argumentation Theory; Game-Based Learning
	Post-Lecture Forum Agent [27]	Discussion facilitation	Dialogic Learning; Community of Inquiry
	Virtual Teaching Assistant [28]	Classroom orchestration	Orchestrated Classroom Discourse
	MoralSim [29]	Social dilemma reasoning with LLM agents	Social-Emotional Learning; Moral Reasoning
	Socratic LLM Debates [30]	Socratic debate / critique loops	Dialogic Learning; Socratic Questioning; Divergent Thinking
Neuro-Symbolic	Coscientist [31]	Hybrid orchestration + tool use	Cognitive Apprenticeship; Collaborative Inquiry
	BioAgents [32]	Tool-augmented multi-agent LLM workflows	Collaborative Learning; Constructivism
Simulation (MARL)	EcoMUVE [33]	Environment-based simulation reasoning	Situated Learning; Constructivism

instance, MetaTutor used multiple agents to help students learn science texts by supporting self-regulated learning [10]. The agents offer hints, ask questions, and monitor how students learn. In Betty's Brain [14], students teach a virtual peer by building concept maps, which helps learners organize their understanding and reflect on their thinking. SimClass [18] went further by simulating entire classrooms filled with LLM-based tutor agents. These agents help analyze how teaching strategies work across many students at once. Self-Explaining Agents [11] followed a similar structure, where agents ask students to explain their reasoning step-by-step. This encourages meta-cognition and helps students build a deeper understanding over time. However, overuse of tutor agents can lead to passive learning and reduce student autonomy.

Peer Agents: Peer agents act more like classmates than teachers, performing tasks such as collaboration, discussion, and shared learning. In LLM-Powered Classrooms [19], AI agents acted as peer collaborators during group learning tasks. They engaged in conversation, provided explanations, and helped learners reflect on their understanding. PairBuddy [20] consisted of conversational agents embedded within coding platforms that assisted learners by answering questions, suggesting solutions, and co-editing code.

These agents have supported a peer-based learning style by encouraging learners to explain their thinking and correct errors through guided collaboration. But these agents can provide incomplete or inaccurate explanations, especially when they are designed to behave like a learner.

Coach Agents: Coach agents support learners' motivation, focus, and self-regulation rather than delivering content. Recent systems have shown how conversational coaches can encourage meta-cognitive behaviors and persistence. Multi-turn coaching agents with different dialogue styles improve learners' sense of support and highlight evaluation gaps be-

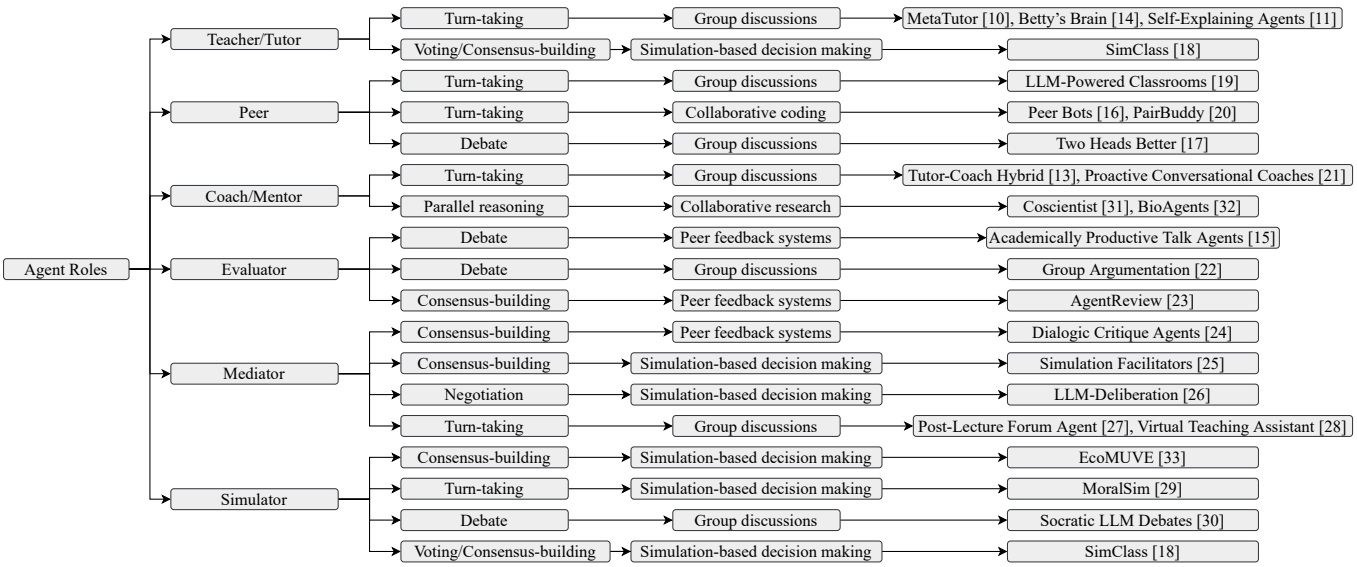

Fig. 2. Taxonomy of agent roles in multi-agent educational systems and the distribution of studies (one study may fall into multiple categories)

tween users and experts [21]. These systems used coach agents as complements to tutors by reinforcing productive learning habits and promoting meta-cognitive awareness. A key limitation is that coach agents relied on accurate estimates of learner states, and poorly timed interventions can reduce engagement.

Evaluator Agents: Evaluator agents are designed to assess performance and provide structured feedback. They can also influence decisions made by other agents in a learning ecosystem. For example, Conversational Agents for Academically Productive Talk [15] employed evaluator-style agents that assess the quality of student arguments, ask clarification questions, and prompt revisions. Such systems help learners practice defending their reasoning and engaging with opposing views through targeted feedback in collaborative discussions. However, a drawback of these systems is that they may oversimplify complex reasoning or reflect bias when evaluation criteria are unclear.

Mediator Agents: Mediator agents are designed to guide group processes rather than directly deliver content. Recent studies show that conversational agents can facilitate online peer discussions by prompting participation and increasing learner confidence in post-lecture forums [27]. At a larger scale, virtual teaching assistants (VTAs) used in classrooms can manage Q&A, coordinate turn-taking, and summarize group discussions [28]. These systems act as mediators by supporting equitable participation and productive collaboration in both online and in-person learning settings. However, they can interfere with natural group interactions if they intervene too frequently, which reduces learners' sense of ownership.

Simulator Agents: Simulator agents take on roles within interactive scenarios and help create realistic learning experiences. In SimClass [18], agents not only tutored but also simulated classroom behaviors, such as asking questions, getting distracted, or working in groups to reflect real-life classrooms. However, it may fail to capture complete real classroom complexity. This limits the transfer of learning to real-world settings.

B. Interaction Structure

Turn-taking: In turn-taking systems, agents and learners interact sequentially, which creates room for reflection between moves. MetaTutor [10] used turn-based prompts and feedback from pedagogical agents to support self-regulated learning. After each segment, learners are prompted to explain or summarize key ideas. This design aligned with evidence that structured self-explanation improves understanding, as shown in Chi's self-explanation research and the ICAP framework [11]. However, these interactions can slow learning and reduce engagement when learners prefer more flexible or exploratory interaction.

Parallel reasoning: Parallel setups run multiple agents (or multiple instances of the same agent) in parallel to generate diverse lines of reasoning. Aggregating their outputs through sampling and voting can improve reliability and limit error propagation, especially when the number of agents is increased [36], [37]. In education-related research, multi-role frameworks, such as Coscientist, distributed responsibilities across parallel agents for hypothesis generation, planning, and execution [31]. Similarly, classroom simulations, like SimClass [18], coordinated multiple roles and can incorporate agent voting to reach instructional decisions. These interactions can increase computational cost and may amplify shared errors if agent outputs lack sufficient diversity.

Debate: Debate-style systems have agents challenge one another's arguments, which can bring contradictions and sharpen reasoning. Multi-agent debate has improved factual accuracy and reasoning in mathematical or strategic tasks by running repeated argument–critique cycles under a judging agent [38], [30]. In tutoring contexts, Socratic agents guided learning through probing questions and counter-arguments instead of

direct answers [39]. However, these agents can confuse learners or increase cognitive load when arguments become overly complex or adversarial.

Negotiation and Consensus: Multi-agent negotiation models agents as stakeholders with distinct goals. In LLM-Deliberation, agents exchanged offers, made concessions, and adapted strategies to reach agreements [26]. This process creates realistic trade-off reasoning that can be applied to learning scenarios. In voting schemes, those agents independently propose solutions, then select or rank the best response according to shared criteria (e.g., correctness, clarity) to achieve a final consensus. Ensemble voting is a simple, strong baseline for reliability [36], [37]. In classroom simulations, voting among role agents helps form a consensus explanation or decide the next step. A limitation of these models is that negotiation processes may favor dominant agents or strategies, leading to suboptimal or biased outcomes.

Consensus-building: Beyond voting, reaching consensus requires agents to exchange rationales and adjust their positions until they align. AgentReview modeled this process in peer review in the field by coordinating reviewers and a chair to reach a shared decision [23]. This pattern also offers useful guidance for the building of consensus in classroom teams. However, these agents can suppress minority viewpoints and discourage creative solutions when alignment is prioritized over exploration.

C. Learning Scenarios

Beyond roles and coordination, the deployment of multi-agent systems varies by learning scenario. These scenarios define the learning context, task type, and the nature of agent-learner or agent-agent interactions.

Collaborative Coding: In collaborative coding settings, agents act as real-time partners that help learners write, debug, and understand code. PairBuddy was a conversational agent designed for pair programming. It offered suggestions, explained code segments, and asked guiding questions to support problem-solving. By simulating a coding partner, the agent encouraged learners to verbally express reasoning and strengthen debugging skills through dialogue.

Group Discussions: In group discussions, agents manage dialogue, balance participation, and foster critical reasoning. In Two Heads May Be Better Than One [17], multiple pedagogical agents simulated peer perspectives to help students compare arguments and deepen understanding. More recent LLM-based facilitators [25] guided group learning by prompting elaboration, evaluating arguments, and maintaining equitable participation. These systems train learners to defend their reasoning and engage constructively with opposing views.

Simulation-based Decision Making: Simulation-based environments allow learners to make strategic decisions in dynamic, interactive settings, in which the agents take on different roles or model complex systems. SimClass [18] simulated a full classroom where multiple tutor and peer agents interact with each other and with human learners, where they reproduce dynamics such as confusion, disagreement, and distraction.

Collaborative Research: Multi-agent systems like Co-scientist and BioAgents modeled scientific collaboration by assigning research roles to specialized LLM agents. In Co-scientist [31], some agents jointly designed and conducted chemistry experiments, whereas the others can handle literature review and lab automation. BioAgents [32] coordinated agents for genomic data analysis, with each responsible for a specific subtask.

Peer Feedback Systems: Peer feedback systems emphasize reflection, explanation, and critique. Agents support learners in evaluating their own and others' work more effectively. Critique Agents scaffold the feedback by prompting constructive comments, clarifications, and argument refinement. This interaction promotes meta-cognitive awareness and higher-level thinking.

Comparative Analysis. Across the reviewed architectures, a clear trade-off emerges between interpretability and scalability. Symbolic systems (e.g., MetaTutor, Betty's Brain) demonstrated higher interpretability and evaluation rigor, supported by pre/post test designs, but remain limited in scalability and open-ended adaptability [14], [10]. LLM-based systems (e.g., SimClass, AgentReview) offered greater scalability and interaction flexibility, but their reasoning processes remain implicit and longitudinal learning outcome evidence is largely absent [18], [23]. Neuro-symbolic systems such as Coscientist and BioAgents balanced these trade-offs but have not yet been evaluated in an authentic classroom settings [31], [32]. These gaps represent key open challenges addressed in Section VIII-B.

V. Applications and Case Study

Multi-agent systems in education are moving beyond individual tutoring to simulate real-world learning settings. In this section, the applications are grouped into four major areas: simulated classrooms, collaborative science projects, peer review, and educational games. For each application, its key system, architectures, agent roles, and how they connect to learning theories are described.

A. Simulated Classrooms with Agent Collectives

SimClass [18] is a large-scale classroom simulation using LLM agents. Agents have roles like teacher, TA, and classmates. They interacted with each other and human learners to teach lessons, ask questions, and lead discussions. The architecture combined LLM prompting with a coordination layer that manages classroom dynamics. SimClass was grounded in social constructivism and dialogic learning and applies frameworks like Flanders Interaction Analysis and the Community of Inquiry model. The results from simulated classrooms have shown that multi-agent interaction increases student engagement and immersion. While SimClass is an advanced example, AI-based classroom collectives are becoming more common. These systems demonstrate how multi-agent LLMs can simulate diverse classroom interactions to support both instruction and peer dialogue.

B. AI-Driven Group Science Projects

Earlier systems like EcoMUVE [33] created virtual ecosystems for agent-guided learning. These agents led students through inquiry-based investigations based on situated and constructivist learning theory. Although not LLM-based, EcoMUVE pioneered collaborative, agent-supported science learning.

Modern systems use LLMs for real-world research support. Coscientist [31] connected multiple LLMs (e.g., GPT-4 and Claude) to plan and execute chemistry experiments. Its agents can retrieve the literature, select protocols, and control laboratory robots. BioAgents [32] used a similar structure for genomic data analysis. Here, specialized agents handled subtasks like tool selection and workflow generation, supervised by a reasoning agent. These systems applied cognitive apprenticeship and collaborative learning principles, so that the agents can model expert reasoning, divide complex tasks, and guide learners through scientific processes.

C. Peer Review with LLM Agents

AgentReview [23] simulated the academic peer review process with LLM agents. Agents took on roles like reviewer, author, and area chair and are configured with traits such as expertise or bias. The system combined LLMs with agent-based modeling to simulate realistic reviewer behaviors. Because real reviews are confidential, AgentReview used synthetic papers to study group decision-making. The system is based on social learning and dialogic theories. It explores how agents influence each other and revise feedback. Experiments showed that agents change their scores after discussion, often aligning with others, which mirrors conformity in real peer review. Specifically, AgentReview [23] is both a research tool and a learning environment where students can understand peer feedback, critique, and academic discourse.

D. Group Simulation Games (Negotiation and Ethical Dilemmas)

Some systems use agents in game-like environments to teach negotiation or ethics. LLM-Deliberation [26], for example, modeled multi-issue negotiations with six LLM agents. Each agent represented a stakeholder with private preferences. The agents negotiated over time using zero-shot prompting strategies and simulated real-world negotiation. Based on argumentation theory and game-based learning, this system allowed students to observe agents displaying strategies like concession and goal balancing.

In ethical scenarios, MoralSim [29] repeatedly presented LLM agents with moral dilemmas, such as the Prisoner's Dilemma or Public Goods Game. Each agent retained a memory of previous decisions and can be prompted for moral reasoning. The results indicated LLMs do not always prioritize ethical behavior; they may act in self-interest depending on the context. This framework is linked to social-emotional learning and moral reasoning, which can help students explore ethical decisions, group cooperation, and fairness through simulations.

Across all of these areas, multi-agent systems are evolving from simple tutor models to rich ecosystems of intelligent agents. Most of these systems use LLMs and are shaped by learning theories like constructivism, dialogic learning, ZPD, and collaborative cognition.

VI. Trustworthiness in Educational AI Agents

Trustworthiness is a key feature that multi-agent systems should have in education, which can ensure that educators and students trust the use of AI technology and improve the results of education. The trustworthiness of autonomous AI agents has emerged as a core topic in both academia and industry. This section examines four key dimensions, safety, privacy, fairness, and transparency, and reviews the mitigation methods commonly presented in current literature.

A prevailing approach in AI agent safety is value alignment, which strives to align agent behaviors with human values and ethical norms. Reinforcement Learning from Human Feedback (RLHF) is a widespread adoption of this, leveraging human evaluations to iteratively refine agent behavior [40]. Complementing this, safe reinforcement learning frameworks integrate explicit safety constraints into policy optimization. This process effectively limits risks during the agent's exploratory learning phase. Critical pre-deployment practices also include red-teaming and adversarial testing, in which domain experts design adversarial or high-risk prompts to identify latent unsafe interaction patterns or vulnerabilities, enabling targeted corrections [41].

Privacy preservation ensures AI agents do not leak or misuse personal information when processing sensitive data, such as demographics, medical records, or educational transcripts. A foundational technique is differential privacy (DP), which injects controlled statistical noise into the training process (e.g., Differentially Private Stochastic Gradient Descent, DP-SGD) or model outputs to minimize re-identification risk [42]. Another widely employed method is federated learning (FL). FL enables decentralized model training where raw user data remains localized on devices [43]. Supplementary privacy-enhancing technologies include secure multi-party computation (SMPC) and data anonymization. SMPC facilitates collaborative training without exposing raw data, and data anonymization removes or encrypts personally identifiable information (PII) to prevent unauthorized disclosure in collaborative computing or model-sharing scenarios [44].

Fairness and bias mitigation ensure AI agents do not perpetuate systemic discrimination or unequal treatment across different demographic subgroups. Bias in AI agents can stem from unrepresentative training data, historical societal biases, or model architecture flaws. Bias mitigation methods are typically categorized into three types. First, pre-processing techniques address bias at the data level, including re-sampling to balance demographic groups and re-weighting to amplify the influence of marginalized samples during training [45]. Second, in-processing methods integrate fairness constraints directly into model training. Examples include fairness-aware regularization, which adds penalties for disparate outcomes, and constraint optimization, which enforces metrics like demographic parity or equalized odds [46]. Third, post-processing adjustments refine model outputs to enhance equity. This can

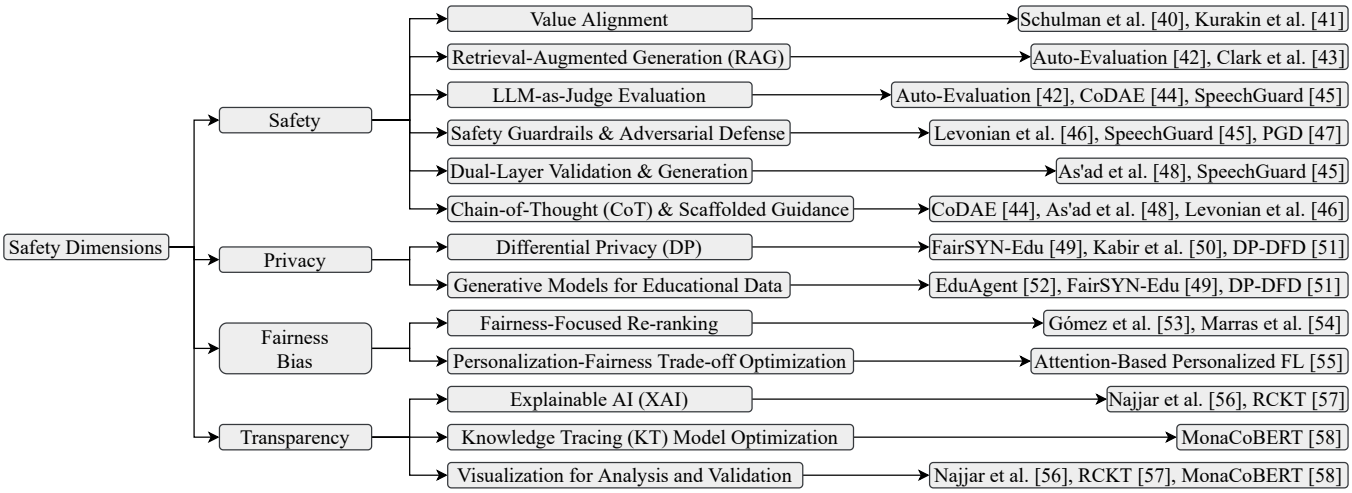

Fig. 3. The taxonomy of trustworthy research in multi-agent education systems along with their methods

involve calibrating prediction probabilities or adjusting decision thresholds to balance error rates across subgroups [47]. Additionally, bias detection tools and algorithmic auditing frameworks are deployed to continuously monitor bias. These tools help ensure that emerging biases are identified and addressed promptly [48].

Transparency and accountability require that AI agent decision-making processes are explainable and traceable. System developers and deployers must also be responsible for agent behaviors. Local Interpretable Model-agnostic Explanations (LIME) generates simplified local surrogate models to explain individual predictions. SHapley Additive exPlanations (SHAP) quantifies each input feature's contribution to model outputs. Attention visualization highlights key input components (e.g., student interaction patterns in educational AI) that drive decisions. All these methods demystify the "black-box" nature of complex models [49].

Additionally, model and data documentation frameworks enhance accountability. Model Cards document a model's performance, limitations, and intended uses [50]. Data sheets for datasets detail data collection methods, sources, and potential biases [50]. These frameworks facilitate peer review and research reproducibility. From a governance perspective, accountability mechanisms are essential, including audit trails (logging input-output pairs and decision timestamps), decision logs (recording high-stakes decision rationales), and Human-in-the-Loop (HITL) monitoring (enabling human intervention for erroneous or harmful behaviors) [51]. Such mechanisms ensure clear attribution of responsibility and actionable remedies for unintended or harmful AI outcomes.

A. *Safety in Educational AI Agents*

Due to the uniqueness of different educational multi-agent systems applied in different applications, researchers have proposed different methodologies to solve safety problems.

Value Alignment: As a prevailing approach in AI agent safety, value alignment strives to align agent behaviors with human values and ethical norms. Reinforcement Learning from Human Feedback (RLHF) is a widespread adoption of this approach, leveraging human evaluations to iteratively refine agent behavior [40]. Complementing this, safe reinforcement learning frameworks integrate explicit safety constraints into policy optimization. This process effectively limits risks during the agent's exploratory learning phase. Critical pre-deployment practices also include red-teaming and adversarial testing, in which domain experts design adversarial or high-risk prompts to identify latent unsafe interaction patterns or vulnerabilities, enabling targeted corrections [41].

Retrieval-Augmented Generation (RAG): RAG applies a two-stage (retrieval + generation) framework, which balances LLM flexibility with knowledge base accuracy, a critical priority for education where factual correctness directly enhances the safety of the models. For example, Auto-Evaluation [52] built an AI-powered auto-evaluation agent to assess the output of Aila at scale, which is an AI lesson-planning tool anchored in 13,000 human-vetted open educational resources (OER), leveraging RAG to ground Aila in OER aligned with England's national curriculum, securing the outcome of the agent. Another related work is conducted by Clark et al. [53], who used the OER corpus as the foundational context for Aila, ensuring generated lessons align with national curriculum standards and avoid age-inappropriate or factually flawed content.

RAG consistently improves LLM output accuracy by grounding generation in domain-specific, vetted data, which effectively solves safety problems of multi-agent systems in education.

LLM-as-Judge Evaluation: This method delegates the assessment of AI-generated content to a powerful LLM, which is prompted to rate outputs against predefined criteria, with scores validated against small human-annotated gold standards to ensure alignment with expert judgment. To exemplify, the aforementioned Auto-Evaluation [52] also used LLM-as-Judge methodology to measure 24 quality and accuracy benchmarks (e.g., cultural bias, quiz question progression). Latest works such as CODAE [54] and SpeechGuard [55] used LLMs such as LLaMA-3.3-70B-Instruct and Claude 2.1 as an

automated judge to enable scalable, standardized evaluation of AI-generated educational content without over-reliance on human annotators, which massively enhances the safety of the outputs.

Safety Guardrails and Adversarial Defense: Safety guardrails encompass a spectrum of proactive and reactive measures to prevent AI education tools from generating or accepting harmful content, while adversarial defense focuses on mitigating targeted attacks that bypass these guardrails. A representative effort is made by Levonian et al. [56] focused on content moderation for student-facing tools with a WhatsApp math tutor (Rori) targeting African middle-schoolers, which was paired with safety guardrails: a curse-word filter and OpenAI's moderation API. Meanwhile, a new framework named SpeechGuard [55] addressed adversarial perturbations in multi-model (speech-text) systems by applying white-box (PGD [57]) and transfer adversarial attacks to SpeechVerse, a conformer-audio-encoder and LLM architecture. Overall, safety guardrails and adversarial defense can thus prioritize proactive safety.

Dual-Layer Validation and Generation: This framework involves two sequential LLM steps: a first layer generates an initial response or output, and a second layer validates, refines, or corrects the first layer's output against predefined criteria. The work of As'ad et al. [58] focused on content quality validation. They developed a GenAI-enhanced intelligent tutoring system (ITS) for healthcare root cause analysis (RCA) that used a dual-layer AI validation, role-specific AI agents, an AI mentor, and a GenAI-powered scoring mechanism. Such a design is similarly used in the aforementioned framework SpeechGuard [55], where the first layer of ASR pre-adaptation enables the model to process speech inputs, and a second layer of cross-modal instruction fine-tuning aligns the output with educational and safety objectives. In education, this two-stage process reduces errors and ensures outputs meet strict quality standards, as the second layer acts as a "checker" to catch flaws the first layer missed.

Chain-of-Thought (CoT) and Scaffolded Guidance: CoT prompting encourages LLMs to generate step-by-step reasoning for their outputs, while scaffolded guidance structures learner interactions to build knowledge incrementally. To detail, the CODAE [54] framework augmented real-world student-tutor dialogues with CoT prompting, adding targeted adversarial cases. As'ad et al. [58] and Levonian et al. [56] integrated scaffolded guidance to ensure students master foundational concepts before advancing to complex analysis and prevent digressions. This method emphasizes step-by-step reasoning over direct answer provision, aligning with pedagogical principles of guided learning.

B. Privacy in Educational AI Agents

To reduce privacy risks, mainstream methodologies and structures, including differential privacy (DP) and generative models, are applied in multi-agent educational systems.

Differential Privacy: DP provides a rigorous statistical framework to protect individual privacy by limiting the inference of personal information from datasets or models. For example, FairSYN-Edu [59] proposed a diffusion-based framework designed for educational synthetic data generation. It integrated differential privacy stochastic gradient descent [60] and adversarial debiasing to optimize utility, fairness, and privacy of the generated data. Another exemplary work was conducted by Kabir et al. [61], who applied DP to three knowledge tracing (KT) models, including BKT, DKT, Mona-CoBERT, with user-level privacy guarantees and protected a student's entire interaction sequence. Other representative pipelines included DP-DFD [62], protecting sensitive labels from the teacher and avoiding direct access to private data. Overall, DP is tailored to sensitive student data and adapts to task-specific needs, from data synthesis to model training and label protection.

Generative Models for Educational Data: Generative models synthesize realistic educational data to address data scarcity, privacy constraints, or the need for diverse training samples. For example, EduAgent [63] used LLM-powered generative agents to simulate fine-grained student behaviors, avoiding collecting sensitive real student data and reducing privacy exposure. Diffusion-based generative models and GAN-based generators are commonly used in this domain, which include representative works such as FairSYN-Edu [59] and DP-DFD [62]. In education, generative models can mimic real student behaviors, interaction patterns, or academic records, enabling downstream tasks like model training, simulation, and policy testing without relying on sensitive real data.

Utility and Privacy Trade-off Optimization: This framework jointly optimizes utility and privacy. FairSYN-Edu [59] explicitly optimized three objectives via a unified loss function, which achieves state-of-the-art fairness with competitive utility and privacy. While Kabir et al. [61] and DP-DFD [62] prioritize privacy and utility, implicitly aligning with this multi-objective paradigm. In education, this method addresses ethical risks and legal constraints (e.g., GDPR and FERPA) while maintaining tool effectiveness for teachers and students.

C. Fairness and Bias

To ensure fairness and reduce bias of educational multi-agent systems, the following methods and structures are applied to directly or indirectly mitigate bias of the outcome.

Fairness-focused Re-ranking: This method adjusts the output of pre-trained recommendation or prediction models without modifying the core architecture or training process. It prioritizes fairness objectives by reordering, weighting, or calibrating outputs, making it flexible and compatible with any underlying model. For example, Gómez et al. [64] proposed a multi-class re-ranking algorithm that optimizes both visibility and exposure for teacher groups across continents to mitigate bias. In addition, Marras et al. [65] used a maximizing marginal relevance (MMR) re-ranking approach to balance learner preferences and consistency with educational principles, balancing personalization and equality. Attention-Based Personalized Federated Learning [66] implicitly leveraged post-hoc adaptation of global models to subpopulations, where local meta-learning updates acted as a form of personalized post-processing, which ensured fairness across demographic subgroups while retaining personalized prediction

performance. In education, such re-ranking addresses biases in final deliverables while preserving the model's original predictive power for personalization.

Multi-stage Bias Mitigation: This framework integrates bias mitigation at multiple stages of the machine learning pipeline, which are pre-processing (data adjustment), in-processing (fairness-constrained training), and post-processing (output calibration), to address biases holistically. Raftopoulos et al. [67] compared 10 bias mitigation techniques across all three stages: pre-processing (via reweighting, learning fair representations, and disparate impact remover), in-processing (via adversarial debiasing and prejudice remover), and post-processing (via equalized odds and calibrated equalized odds). Attention-Based Personalized FL [66] combined pre-processing and in-processing. The pre-training step improved feature representation of underrepresented learners, while meta-learning in federated training ensures personalized models retain fairness. Though Gómez et al. [64] and Marras et al. [65] focused on post-processing and re-ranking, both rely on pre-processing steps to enable effective fairness adjustments. In education, the above methods tackle bias from data collection, model learning, and deployment in a coordinated manner.

Personalization vs. Fairness Optimization: This method balances two competing objectives, personalization and fairness. It uses weighted loss functions, adaptive hyperparameters, or modular architectures to avoid prioritizing one goal at the expense of the other. For example, Attention-Based Personalized FL [66] developed subgroup-specific models via meta-learning, balancing personalization and fairness to reduce bias across subgroups. The trade-off optimization addressed the tension between personalized learning and systemic equity in multi-agent education systems.

Fairness Metrics for Educational Contexts: Context-specific metrics are needed to quantify fairness in educational scenarios, moving beyond generic machine learning fairness measures to align with domain goals. Existing literature defined various metrics from different perspectives. For example, in the work [64], authors defined two key metrics, visibility (share of recommendations) and exposure (ranking position), to assess disparities favoring overrepresented continents, aligning recommendations with platform principles. Marras et al. introduced a metric combining consistency (alignment between recommended courses and educational principles via Manhattan distance) and equality (1-Gini index of consistency across learners) [65]. This metric identified systemic inequalities, providing a reference for mitigating bias. Another exemplary work is conducted in [67], who adopted 3 key metrics, including statistical parity (equal positive prediction rates across groups), disparate impact (ratio of favorable outcomes for unprivileged/privileged groups), and consistency (similar predictions for similar learners). Moreover, in [66], authors used subgroup-specific AUC to measure fairness, evaluating model performance for demographic subgroups separately. Metrics focus on actionable, education-relevant outcomes and are validated against real-world educational constraints, effectively revealing the fairness of models from different aspects.

D. Transparency

Transparency is essential for educational AI agents because learning is not just about getting the right answer, but understanding why an answer is correct. For educators, transparent educational AI supports trust, accountability, and effective oversight, enabling them to diagnose misconceptions, assess learning progress, and integrate AI feedback meaningfully into instruction. The related works are summarized as follows.

Explainable AI (XAI): XAI techniques demystify the decision-making processes of AI models by generating human-understandable explanations, bridging the "black-box" gap in educational AI. It tailors explanations to diverse stakeholders and aligns with ethical goals by clarifying how models arrive at outcomes. For instance, Transparency Index Framework [68] explicitly mandated XAI as a component of algorithmic transparency, required ed-tech companies to share XAI tools and their limitations, and ensured explanations are understandable for non-technical stakeholders like teachers. Najjar et al. [69] used LIME to identify discriminative features between human and AI-generated text, which enhance accountability of educational content. RCKT [70] employed counterfactual reasoning as an ante-hoc XAI method, quantifying how individual student responses influence future performance predictions, which allowed educators to see which past answers drove a model's forecast.

Knowledge Tracing (KT) Model Optimization: KT models optimize performance and interpretability by integrating advanced architectures and domain-specific embedding strategies. These innovations address limitations of traditional KT models. For example, MonaCoBERT [71] presented a BERT-based KT model with monotonic convolutional multi-head attention, which combines monotonic attention for forgetting behavior and ConvBERT attention for representation power. By analyzing attention weights, the link between absolute value and subtraction allowed educators to trace knowledge dependencies, ensuring accountable knowledge tracing practices.

Visualization for Analysis and Validation: Visualization tools translate abstract model behaviors, data patterns, or concept relationships into intuitive formats, enabling researchers and educators to validate model efficacy, identify biases, or interpret outcomes without deep technical expertise. To exemplify, Najjar et al. [69] applied word clouds to highlight vocabulary differences between human and AI text, and TF-IDF weight plots to visualize key discriminative terms. These visualizations confirmed that human text prioritizes practical language while AI text favors formal terms, improving model accountability. Other visualization tools, such as knowledge proficiency visualization [70] and Grad-CAM and t-SNE [71], have been used to achieve state-of-the-art performance on benchmark KT datasets.

E. Summary

To summarize, from the reviewed studies, a unified pipeline that organizes trustworthiness in multi-agent educational systems was derived. As shown in Figure 4, the pipeline links input, reasoning, and output safeguards into a continuous trust

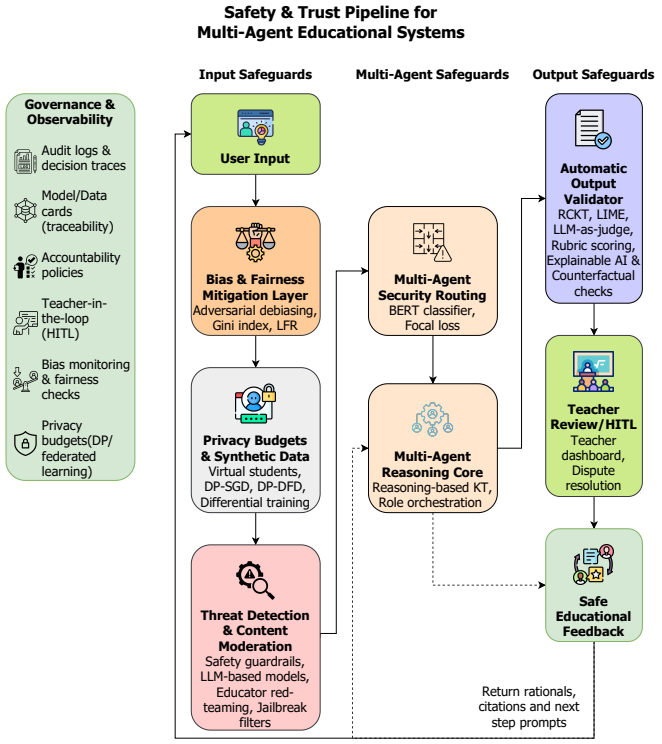

Fig. 4. The pipeline of building trustworthy multi-agent education systems in practice.

loop, which maps concrete methods, such as threat detection, privacy control, and bias mitigation, across the agent lifecycle. These safeguards are directly related to pedagogical processes such as reasoning, evaluation, and human-in-the-loop oversight. The framework captures consistent patterns across symbolic, hybrid, and LLM-based systems, offering a practical guide for building transparent and reliable educational AI agents.

VII. Evaluation and Benchmarks

A. Evaluation Challenges

Evaluating multi-agent educational systems is more complex than assessing single-agent tools. The key challenge is measuring collective learning, which requires examining how agents interact, provide mutual support, and contribute to learner understanding over time, rather than assessing an isolated agent's actions. Similarly, causal attribution is another caveat. Because many interactions are concurrent, tracing a specific learning outcome to an individual agent's action is challenging. Other important factors include learner engagement and trust, since learners may disengage if they do not feel comfortable or supported. All of these factors, such as the feelings of learners and the trust in agents, are difficult to measure using traditional metrics, which deserve more thorough investigation in future work.

B. Benchmarks and Simulation Environments

Current multi-agent benchmarks focus on general AI tasks and do not fully address educational needs. For example, ALFWorld [72] evaluated agents in text-based, embodied task environments. MultiAgentBench [73] tested LLM agents across six interactive scenarios spanning collaboration (e.g., co-authoring, coding) and competition (e.g., Werewolf, bargaining). It also compared coordination protocols (e.g., star, chain, graph) and used milestone-based KPIs with planning scores to measure both task success and coordination quality. These general-purpose setups can inform educational agent design but leave gaps for classroom-specific skills and assessments. Other frameworks test coordination or planning, but not learning outcomes. They do not measure how well agents support learning, explain their reasoning, or build user trust [74]. To close this gap, we suggest metrics and evaluation methods for multi-agent educational systems in Table II, which includes multiple dimensions, such as learning progress, interaction quality, and user perception.

VIII. Future Directions in Multi-agent Education Systems

A. Trends and Insights

Trends in Multi-Agent Education Systems. Across these selected studies, there are several clear trends in the research domain. Over the past decade, multi-agent educational systems have evolved from early, tightly scripted designs into far more dynamic and adaptive learning environments. Early systems, such as EcoMUVE, Betty's Brain, and MetaTutor, used symbolic or rule-based architectures that emphasized explainability and structured guidance. They were grounded in constructivist and self-regulated learning theories. These systems focused on clear feedback loops and scaffolded reasoning. While effective for targeted skill development, these were often limited to narrow domains and lacked the flexibility for open-ended social learning.

From around 2020 onward, the field shifted toward hybrid and language model–driven architectures. Systems like SimClass, AgentReview, Coscientist, and BioAgents used LLMs, which allow agents to communicate, debate, and coordinate naturally and closely simulate real classroom or research interactions. These newer systems blurred the line between tutoring and collaboration. Instead of one AI teacher, multiple agents now worked together as peers, evaluators, or mediators to create a shared learning space. This move has expanded the focus of educational AI from delivering instruction to orchestrating dialogue, collaboration, and reflection. The publication landscape mirrors this evolution as shown in Figure 5. Before 2018, few multi-agent education papers appeared, mostly in learning sciences or educational technology venues. By 2023–2025, publications rose sharply in major AI and NLP conferences, which signals recognition beyond education circles. The use of LLMs, in particular, has encouraged researchers to revisit pedagogical questions about social interaction, argumentation, and moral reasoning in digital learning. This convergence of AI research and pedagogy has produced systems that are not just technically impressive but also grounded in theories of collaborative learning. In the aspect of learning theories, most designs are grounded in social and constructivist learning theories, but interaction

TABLE II
PROPOSED METRICS AND EVALUATION METHODS FOR MULTI-AGENT EDUCATIONAL SYSTEMS.

Metric Category	Example Metric	Evaluation Method
Learning Gains	Pre/post test scores	Compare learner scores before and after interaction
Engagement	Dialogue depth, turn count	Track number and depth of dialogue turns
Collaboration Quality	Diversity of ideas, peer support	Analyze variety of ideas and helpfulness in feedback
Trust and Explainability	User surveys, recall tests	Ask users about trust, and test understanding of agent reasoning
Attribution	Agent influence tracing	Use ablation or logging to see which agents contributed most
Robustness	Handling of incorrect answers	Insert errors and measure how agents recover or correct them

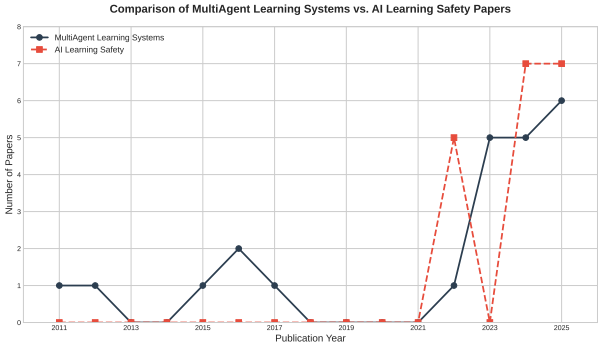

Fig. 5. Paper publication comparison for multi-agent education systems and related trustworthy topics.

patterns have grown more sophisticated, including debate, consensus-building, and negotiation, expanding beyond simple turn-taking dialogue. Learning scenarios have also diversified. Beyond tutoring, systems now support simulation-based decision-making, collaborative science projects, peer review simulations, and ethical dilemma games. These shifts suggest a deeper understanding of teaching/learning as a social, context-dependent, and complex process, which constitutes the field with a gradual shift toward designing agent ecologies that enrich human learning rather than automate it.

Trends in Trustworthiness of MAES. Safety in multi-agent education systems has evolved from an afterthought to a central design principle. Early agents relied on simple content filters or manual oversight. In contrast, modern systems treat safety as a multi-layer, socio-technical process. In conversational environments, "safe chat" mechanisms combined lexical filters, API-level classifiers, and educator-led red-teaming to monitor sensitive exchanges. In multimodal and speech-based tutoring, methods like time-domain noise flooding (TDNF) defended against adversarial attacks, while automated safety checks flag jailbreak attempts. Multi-agent architectures increasingly embed dedicated safety or intent agents to detect, route, and manage unsafe inputs, which makes safety an active, distributed capability rather than a passive guardrail.

Alongside these technical advances, privacy by design is now a non-negotiable requirement for education-focused AI systems. Classroom tools handled deeply personal and longitudinal data, making them more sensitive to misuse than generic conversational agents. New systems adopted three recurring strategies: generating synthetic or differentially private datasets to de-risk training; implementing user-level differential privacy to keep student histories confidential; and using proxy or virtual learners to replace real student data. These approaches signal a shift from reactive data protection to proactive privacy engineering, where the model never accesses identifiable data. The field recognizes that genuine educational privacy requires both content control and ethical data stewardship at every design layer.

The third trend is pedagogically aligned resilience. Instead of only training agents to block unsafe prompts, researchers now fine-tune models to behave like responsible educators. This includes refusing premature answers, expressing uncertainty, or redirecting emotional and coercive inputs. Such alignment ensures safety mechanisms reinforce, rather than hinder, the learning goals. Many existing works combine these behavioral refinements with structured documentation and accountability practices, including model cards, data sheets, decision logs, and teacher-in-the-loop auditing. Together, these tools create traceable systems where errors can be investigated and corrected without undermining trust or transparency.

More broadly, the conversation around the trustworthiness of educational AI is diverging from that of general AI safety. General frameworks emphasize robustness, fairness, and content moderation. Multi-agent education environments, however, demand an additional layer of pedagogical and social safety. Agents must not only avoid harm but also actively support equitable, culturally sensitive, and age-appropriate learning. Educational safety, therefore, extends beyond preventing model misbehavior, which requires aligning agents with the specific values and responsibilities of education.

B. Open Challenges and Future Directions

Ethics and Risks in Multi-Agent Classrooms. The growing adoption of multi-agent educational AI systems introduces ethical and pedagogical risks that demand careful attention in future research and system design. Agents that converge too easily on shared conclusions may amplify common errors or biases, leading to groupthink, reduced perspective diversity, and diminished creativity—outcomes that can disproportionately affect underrepresented learners. Conversely, overly authoritative agents may misinterpret cultural, linguistic, or communication differences, resulting in unfair or misleading feedback. Establishing clear boundaries on agent authority is therefore critical for maintaining trust and equity. At the same time, productive disagreement among agents offers opportunities to expose learners to diverse viewpoints, but unexplained inconsistencies in feedback can confuse students and undermine learning. Future systems must prioritize transparency by clearly articulating the rationale behind differing responses, helping learners make sense of conflicting perspec-

tives, fostering critical self-reflection, and encouraging deeper engagement.

Generalization and Robustness in Environments. Many existing simulation environments for training and evaluating multi-agent educational systems are limited to narrow, task-specific scenarios and fail to capture the complexity of real classroom interactions. As a result, agent behaviors learned in these settings often do not generalize well across subjects, learner populations, or instructional styles. Future research should focus on building more robust and generalizable environments that support open-ended interaction among multiple AI agents and human learners, adapt to diverse educational contexts, and remain effective under varying pedagogical goals. Meanwhile, robustness must extend beyond environments to learning dynamics themselves. When multiple AI agents interact, there is a risk of premature consensus or uncritical agreement with student responses, which can undermine cognitive challenge and lead to superficial learning. Robust multi-agent educational systems should deliberately introduce constructive tension, encourage explanation and justification, and promote independent reasoning.

Alignment with Student Goals and Learning Trajectories. Learners differ widely in their goals, motivations, prior knowledge, and learning preferences, making effective feedback alignment a central challenge for multi-agent educational systems. If agent responses are overly generic or misaligned with a learner's intentions, they risk being perceived as irrelevant and may fail to support meaningful progress. Future systems should be capable of modeling individual learning trajectories by recognizing a student's goals, current understanding, and evolving preferences, and dynamically tailoring feedback to balance personalized support with overarching instructional objectives.

Coordination in Mixed Human-AI Teams. Modern educational settings increasingly involve collaboration among students, human instructors, and AI agents, giving rise to complex team dynamics that require careful role definition. Key questions, such as when AI agents should take initiative, how authority should be shared between humans and AI, and how to prevent overreliance on automated support, remain open challenges. Future research is needed to design multi-agent educational systems that enable effective coordination, clearly communicate roles and responsibilities, and foster mutual understanding and trust between human participants and AI collaborators.

Cultural Adaptation in Agent Collectives. AI agents often reflect the data and cultural norms they were trained on. However, educational values, communication styles, and social expectations vary greatly across cultures. For example, direct feedback may be welcomed in one culture, but considered rude in another. Future multi-agent systems must be able to recognize and adapt to cultural differences in real time. This includes adjusting language, tone, feedback style, and even the way learning activities are structured. Culturally adaptive AI can help ensure that students from all backgrounds feel respected and supported. This is especially important in global or multilingual education platforms, where the same system can be used by various groups of learners [75].

IX. Conclusion

Multi-agent educational systems are expanding how we design collaborative and adaptive learning. This survey outlined key system architectures, agent roles, learning scenarios, and theoretical foundations of current multi-agent education systems with a new taxonomy. It highlights open problems in the trustworthiness of the systems from safety, privacy, fairness, and transparency, and corresponding threats and defenses. Going forward, these systems should be designed with the goal of achieving trustworthiness and utility, following the insights and suggestions derived from the key observations and propositions. The authors hope this work supports future efforts to develop trustworthy and inclusive multi-agent education systems and environments.

Acknowledgment

This work was supported by the National Science Foundation under grants No. 2429960, 2434899, and 2548041.

References

[1] A. Lippert, K. Shubeck, B. Morgan, A. Hampton, and A. Graesser, "Multiple agent designs in conversational intelligent tutoring systems," *Technology, Knowledge and Learning*, vol. 25, pp. 443–463, 2020.

[2] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, 1978.

[3] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," in *Handbook of Reinforcement Learning and Control*, C.-Y. Cheng and X. Lu, Eds. Springer, 2021, pp. 321–384.

[4] J. E. Laird, A. Newell, and P. S. Rosenbloom, "Soar: An architecture for general intelligence," *Artificial Intelligence*, vol. 33, no. 1, pp. 1–64, 1987.

[5] T. R. Besold, A. S. d'Avila Garcez, S. Bader, H. Bowman, P. Domingos *et al.*, "Neural-symbolic learning and reasoning: A survey and interpretation," *arXiv preprint arXiv:1711.03902*, 2017. [Online]. Available: https://arxiv.org/abs/1711.03902

[6] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022. [Online]. Available: https://arxiv.org/abs/2204.02311

[7] C.-L. C. Kulik and J. A. Kulik, "Effectiveness of computer-based instruction: An updated analysis," *Computers in Human Behavior*, vol. 7, no. 1-2, pp. 75–94, 1991.

[8] R. Nkambou, J. Bourdeau, and R. Mizoguchi, *Advances in Intelligent Tutoring Systems*. Springer, 2010.

[9] E. Hutchins, *Cognition in the Wild*. Cambridge, MA: MIT Press, 1995.

[10] R. Azevedo *et al.*, "Lessons learned and future directions of metatutor," *Frontiers in Psychology*, vol. 13, p. 813632, 2022.

[11] M. T. H. Chi and M. Wylie, "The icap framework: Linking cognitive engagement to active learning outcomes," *Educational Psychologist*, vol. 51, no. 2, pp. 219–243, 2016.

[12] A. Collins, J. S. Brown, and S. E. Newman, "Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics," in *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*, L. B. Resnick, Ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1989, pp. 453–494.

[13] T. Nguyen, R. Patel, and R. Simmons, "Adaptive rl coaching agents for self-regulated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[14] G. Biswas, K. Leelawong, N. J. Vye, D. L. Schwartz, J. D. Davis, J. D. Bransford, and the Teachable Agents Group at Vanderbilt, "From design to implementation to practice: A learning by teaching system – Betty's Brain," *International Journal of Artificial Intelligence in Education*, vol. 26, no. 1, pp. 350–364, 2015.

[15] S. Tegos, S. Demetriadis, P. M. Papadopoulos, and A. Weinberger, "Conversational agents for academically productive talk: A comparison of directed and undirected agent interventions," *International Journal of Computer-Supported Collaborative Learning*, vol. 11, no. 4, pp. 417–440, 2016.

[16] J. Johnson, S. Lee, and V. Kumar, "Peer bots: Conversational agents for collaborative coding," in *Proceedings of the Conference on Artificial Intelligence in Education (AIED)*, 2022.

[17] A. C. Graesser, C. M. Forsyth, and B. A. Lehman, "Two heads may be better than one: Learning from computer agents in conversational trialogues," *Teachers College Record: The Voice of Scholarship in Education*, vol. 119, no. 3, pp. 1–20, 2017, (Original work published 2017). [Online]. Available: https://doi.org/10.1177/016146811711900309

[18] Z. Zhang, D. Zhang-Li, J. Yu, L. Gong, J. Zhou, Z. Hao, J. Jiang, J. Cao, H. Liu, Z. Liu, L. Hou, and J. Li, "Simulating classroom education with LLM-empowered agents," in *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 10 364–10 379. [Online]. Available: https://aclanthology.org/2025.naacl-long.520/

[19] Z. Hao, F. Qin, J. Jiang, J. Cao *et al.*, "Ai as learning partners: Students' interactions and perceptions in a simulated classroom with multiple LLM-powered agents," in *Proceedings of the 19th International Conference of the Learning Sciences (ICLS 2025)*. Helsinki, Finland: International Society of the Learning Sciences, 2025, pp. 1789–1793.

[20] P. Robe and S. K. Kuttal, "Designing pairbuddy—a conversational agent for pair programming," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 2023.

[21] V. Srinivas, X. Xu, X. Liu, K. Ayush, I. Galatzer-Levy, S. Patel, D. McDuff, and T. Althoff, "Substance over style: Evaluating proactive conversational coaching agents," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 2025, p. to appear. [Online]. Available: https://aclanthology.org/2025.acl-long.1017/

[22] M. Lee and W. Tan, "Ai-guided group argumentation for critical writing," in *Proceedings of the ACM Conference on Learning at Scale*, 2023.

[23] Y. Jin, Q. Zhao, Y. Wang, H. Chen, K. Zhu, Y. Xiao, and J. Wang, "Agentreview: Exploring peer review dynamics with LLM agents," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 1208–1226. [Online]. Available: https://aclanthology.org/2024.emnlp-main.70/

[24] H. Brown, M. Zhou, and S. Carter, "Dialogic critique agents for classroom debate," in *Proceedings of the International Conference on Artificial Intelligence in Education*, 2023.

[25] J. Lee, K. Tan, and et al., "Integrating large language model-based conversational facilitation into collaborative learning environments," in *Proceedings of the 11th ACM Conference on Learning at Scale*, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3613905.3650868

[26] S. Abdelnabi, A. Gomaa, S. Sivaprasad, L. Schönherr, and M. Fritz, "Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games," *arXiv preprint arXiv:2309.17234*, 2023, neurIPS 2024 (poster). [Online]. Available: https://arxiv.org/abs/2309.17234

[27] J. A. Haqbeen and coauthors, "Llm-based conversational agent enhances post-lecture discussions in online courses," in *Proceedings of ACM Collective Intelligence (CI 2025)*, 2025, pp. 70:1–70:10. [Online]. Available: https://ci.acm.org/2025/wp-content/uploads/CI2025_paper_70.pdf

[28] S. Kweon, S. Nam, H. Lim, H. Hong, and E. Choi, "A large-scale real-world evaluation of an llm-based virtual teaching assistant," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*. Vienna, Austria: Association for Computational Linguistics, 2025, pp. 850–864. [Online]. Available: https://aclanthology.org/2025.acl-industry.60/

[29] S. Backmann, D. G. Piedrahita, E. Tewolde, R. Mihalcea, B. Schölkopf, and Z. Jin, "When ethics and payoffs diverge: Llm agents in morally charged social dilemmas," https://arxiv.org/abs/2505.19212, 2025, arXiv preprint.

[30] T. Liang *et al.*, "Encouraging divergent thinking in large language models via multi-agent debate," in *EMNLP 2024*, 2024. [Online]. Available: https://aclanthology.org/2024.emnlp-main.992.pdf

[31] D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes, "Autonomous chemical research with large language models," *Nature*, vol. 624, no. 7992, pp. 570–578, 2023.

[32] N. Mehandru, A. K. Hall, O. Melnichenko, Y. Dubinina, D. Tsirulnikov, D. Bamman, A. Alaa, S. Saponas, and V. S. Malladi, "Bioagents: Democratizing bioinformatics analysis with multi-agent systems," *arXiv preprint arXiv:2501.06314*, 2025. [Online]. Available: https://arxiv.org/abs/2501.06314

[33] C. Dede, T. Grotzer, A. Kamarainen, and S. Metcalf, "Ecomuve: A case study of design-based research in virtual worlds for learning complex causal relationships," in *Proceedings of the 10th International Conference of the Learning Sciences (ICLS 2012)*. Sydney, Australia: International Society of the Learning Sciences, 2012, pp. 290–297.

[34] M. T. H. Chi, "Self-explaining: The dual process of generating inferences and repairing mental models," in *Advances in Instructional Psychology: Educational Design and Cognitive Science*, R. Glaser, Ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000, vol. 5, pp. 161–238.

[35] Y. Fu, H. Peng, T. Khot, and M. Lapata, "Improving language model negotiation with self-play and in-context learning from ai feedback," *arXiv preprint arXiv:2305.10142*, 2023. [Online]. Available: https://arxiv.org/abs/2305.10142

[36] J. Li, Q. Zhang, Y. Yu, Q. Fu, and D. Ye, "More agents is all you need," *Transactions on Machine Learning Research*, 2024. [Online]. Available: https://arxiv.org/abs/2402.05120

[37] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain-of-thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022. [Online]. Available: https://arxiv.org/abs/2203.11171

[38] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," *arXiv preprint arXiv:2305.14325*, 2023. [Online]. Available: https://arxiv.org/abs/2305.14325

[39] J. Liu *et al.*, "Socraticlm: Exploring socratic personalized teaching with llms," in *NeurIPS 2024*, 2024. [Online]. Available: https://neurips.cc/virtual/2024/poster/93477

[40] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[41] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[42] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[43] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[44] C. Zhao, S. Zhao, M. Zhao, Z. Chen, C.-Z. Gao, H. Li, and Y.-a. Tan, "Secure multi-party computation: theory, practice and applications," *Information Sciences*, vol. 476, pp. 357–372, 2019.

[45] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE international conference on data mining workshops*. IEEE, 2009, pp. 13–18.

[46] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2012, pp. 35–50.

[47] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[48] L. Oneto and S. Chiappa, "Fairness in machine learning," in *Recent trends in learning from data: Tutorials from the inns big data and deep learning conference (innsbddl2019)*. Springer, 2020, pp. 155–196.

[49] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[50] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[51] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[52] H.-B. Clark, M. Dowland, L. Benton, R. Budai, I. K. Keskin, E. Searle, M. Gregory, M. Hodierne, W. Gayne, and J. Roberts, "Auto-evaluation: A critical measure in driving improvements in quality and safety of ai-generated lesson resources," *arXiv preprint arXiv:2502.10410*, 2025.

[53] H.-B. Clark, L. Benton, E. Searle, M. Dowland, M. Gregory, W. Gayne, and J. Roberts, "Building effective safety guardrails in ai education tools," in *International Conference on Artificial Intelligence in Education*. Springer, 2025, pp. 129–136.

[54] S. Yuan, W. LaCroix, H. Ghoshal, E. Nie, and M. Färber, "Codae: Adapting large language models for education via chain-of-thought data augmentation," *arXiv preprint arXiv:2508.08386*, 2025.

[55] R. Peri, S. M. Jayanthi, S. Ronanki, A. Bhatia, K. Mundnich, S. Dingliwal, N. Das, Z. Hou, G. Huybrechts, S. Vishnubhotla *et al.*, "Speechguard: Exploring the adversarial robustness of multimodal large language models," *arXiv preprint arXiv:2405.08317*, 2024.

[56] Z. Levonian and O. Henkel, "Safe generative chats in a whatsapp intelligent tutoring system," *arXiv preprint arXiv:2407.04915*, 2024.

[57] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[58] M. As' ad, "Intelligent tutoring systems, generative artificial intelligence (ai), and healthcare agents: A proof of concept and dual-layer approach," *Cureus*, vol. 16, no. 9, 2024.

[59] K. Kesgin, "Fairsyn-edu a diffusion-based model for fair and private educational data synthesis," *Discover Education*, vol. 4, no. 1, pp. 1–18, 2025.

[60] Z. Xiong, Z. Cai, D. Takabi, and W. Li, "Privacy threat and defense for federated learning with non-iid data in aiot," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1310–1321, 2021.

[61] A. Kabir, C. Tankala, and D. Lowd, "On the practicality of differential privacy for knowledge tracing," 2025.

[62] B. Liu, J. Lu, P. Wang, J. Zhang, D. Zeng, Z. Qian, and S. Ge, "Privacy-preserving student learning with differentially private data-free distillation," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2022, pp. 01–06.

[63] S. Xu, X. Zhang, and L. Qin, "Eduagent: Generative student agents in learning," *arXiv preprint arXiv:2404.07963*, 2024.

[64] E. Gómez, C. S. Zhang, L. Boratto, M. Salamó, and G. Ramos, "Enabling cross-continent provider fairness in educational recommender systems," *Future Generation Computer Systems*, vol. 127, pp. 435–447, 2022.

[65] M. Marras, L. Boratto, G. Ramos, and G. Fenu, "Equality of learning opportunity via individual fairness in personalized recommendations," *International Journal of Artificial Intelligence in Education*, vol. 32, no. 3, pp. 636–684, 2022.

[66] Y.-W. Chu, S. Hosseinalipour, E. Tenorio, L. Cruz, K. Douglas, A. Lan, and C. Brinton, "Mitigating biases in student performance prediction via attention-based personalized federated learning," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 3033–3042.

[67] G. Raftopoulos, G. Davrazos, and S. Kotsiantis, "Evaluating fairness strategies in educational data mining: A comparative study of bias mitigation techniques," *Electronics*, vol. 14, no. 9, p. 1856, 2025.

[68] M. Chaudhry, M. Cukurova, and R. Luckin, "A transparency index framework for ai in education. arxiv," 2022.

[69] A. A. Najjar, H. I. Ashqar, O. A. Darwish, and E. Hammad, "Detecting ai-generated text in educational content: Leveraging machine learning and explainable ai for academic integrity," *arXiv preprint arXiv:2501.03203*, 2025.

[70] J. Cui, M. Yu, B. Jiang, A. Zhou, J. Wang, and W. Zhang, "Interpretable knowledge tracing via response influence-based counterfactual reasoning," in *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 2024, pp. 1103–1116.

[71] U. Lee, Y. Park, Y. Kim, S. Choi, and H. Kim, "Monacobert: Monotonic attention based convbert for knowledge tracing," in *International Conference on Intelligent Tutoring Systems*. Springer, 2024, pp. 107–123.

[72] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, and L. Zettlemoyer, "Alfworld: Aligning text and embodied environments for interactive learning," https://arxiv.org/abs/2010.03768, 2020.

[73] K. Zhu, H. Du, Z. Hong, X. Yang, S. Guo, Z. Wang, Z. Wang, C. Qian, X. Tang, H. Ji, and J. You, "Multiagentbench: Evaluating the collaboration and competition of LLM agents," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Long Papers)*. Association for Computational Linguistics, 2025, pp. 8580–8622. [Online]. Available: https://aclanthology.org/2025.acl-long.421/

[74] R. M. Aratchige and W. M. K. S. Ilmini, "Llms working in harmony: A survey on the technological aspects of building effective llm-based multi-agent systems," https://www.arxiv.org/abs/2504.01963, 2025.

[75] C. Dahal, "Revolutionizing education through ai-powered inclusive learning systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 21, March 2024, pp. 23 736–23 737.