LLM Agents for Education: Advances and Applications

Zhendong Chu[1]*, Shen Wang[1]*, Jian Xie[2]*, Tinghui Zhu[2]*, Yibo Yan[3]*, Jinheng Ye[4]*, Aoxiao Zhong[1], Xuming Hu[3], Jing Liang[5], Philip S. Yu[6], Qingsong Wen[1]†
[1]Squirrel Ai Learning, USA [2]Fudan University, China [3]Hong Kong University of Science and Technology (Guangzhou), China [4]Tsinghua University, China [5]Squirrel Ai Learning, China [6]University of Illinois Chicago, USA

Abstract

Large Language Model (LLM) agents have demonstrated remarkable capabilities in automating tasks and driving innovation across diverse educational applications. In this survey, we provide a systematic review of state-of-the-art research on LLM agents in education, categorizing them into two broad classes: (1) *Pedagogical Agents*, which focus on automating complex pedagogical tasks to support both teachers and students; and (2) *Domain-Specific Educational Agents*, which are tailored for specialized fields such as science education, language learning, and professional development. We comprehensively examine the technological advancements underlying these LLM agents, including key datasets, benchmarks, and algorithmic frameworks that drive their effectiveness. Furthermore, we discuss critical challenges such as privacy, bias and fairness concerns, hallucination mitigation, and integration with existing educational ecosystems. This survey aims to provide a comprehensive technological overview of LLM agents for education, fostering further research and collaboration to enhance their impact for the greater good of learners and educators alike.

Keywords

Large language models, Adaptive learning, AI for education

1 Introduction

Artificial intelligence (AI) techniques are increasingly used in education to enable personalized learning and intelligent tutoring [33, 156, 243]. While traditional educational data mining approaches [3, 51, 94, 165, 176, 194], such as knowledge tracing and cognitive diagnosis, have made significant progress in reshaping the human-learning paradigm by analyzing student behaviors and assessing knowledge states, they still face major challenges in real-world applications. These challenges include shallow contextual understanding, limited interactive capabilities, and difficulties in generating adaptive, personalized learning materials, *etc* [103, 184, 256].

The strong natural language understanding of Large Language Models (LLMs) and the task automation capabilities of LLM agents make them valuable for addressing challenges in education [197, 204]. First, memory enables LLM agents to retain both long-term knowledge about students' study habits and short-term context from real-time interactions, enhancing contextual understanding and ensuring personalized learning experiences across various educational tasks [250]. Second, tool use allows LLM agents to access external resources, perform complex calculations, and retrieve real-time information, enabling them to automate intricate educational tasks such as grading, knowledge retrieval, and adaptive content generation, thereby overcoming limited interactivity and enhancing engagement [52]. Third, planning supports structured learning by decomposing complex topics, predicting optimal learning paths, and dynamically adjusting instructional strategies, allowing LLM agents to autonomously guide students through personalized learning experiences [80]. By integrating these capabilities, LLM agents not only enhance understanding, engagement, and personalization but also automate complex educational workflows, making learning more adaptive, efficient, and scalable.

In this survey, we provide a comprehensive review of LLM agents in education. We begin by introducing LLM agents, followed by a general discussion of their potential applications in educational systems. Next, we categorize them into two broad classes: (1) *Pedagogical Agents*, which automate complex pedagogical tasks such as classroom simulation and learning resource recommendation for teachers. Also, they transform student support paradigms through applications like knowledge tracing, adaptive learning, and error detection and correction. (2) *Domain-Specific Educational Agents*, which are designed to address specialized challenges in different fields, including science education, language learning, and professional development. Furthermore, we discuss the key challenges in LLM agents for education, which need to be addressed to advance research and enable effective deployment in real-world educational applications. Finally, we compile comprehensive datasets, benchmarks, and evaluation methodologies to support and encourage further research involvement in the area of LLM agents for education. We summarize our contributions as follows:

- **Comprehensive review of LLM Agents in education.** We analyze recent advancements in LLM-driven educational agents, discussing their applications, effectiveness, and challenges in automating pedagogical tasks and enhancing learning experiences.
- **Novel task-centric taxonomy.** We propose a structured classification of LLM agents, categorizing them into Pedagogical Agents and Domain-Specific Educational Agents, providing a framework for understanding their roles and capabilities.
- **Current challenges and future research directions.** We analyze critical challenges that need to be addressed for the effective deployment of LLM agents in education, including issues related to privacy, bias and fairness concerns, hallucination, and integration into real-world educational ecosystems.
- **Compilation of essential resources.** We compile comprehensive datasets and benchmarks to support future research efforts and facilitate the development of more robust and effective LLM-driven educational solutions[1].

*Equal contribution. †Corresponding author.

[1]Due to the page limit, we present more details in Appendix A.

Conference acronym 'XX, Woodstock, NY
2018. ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

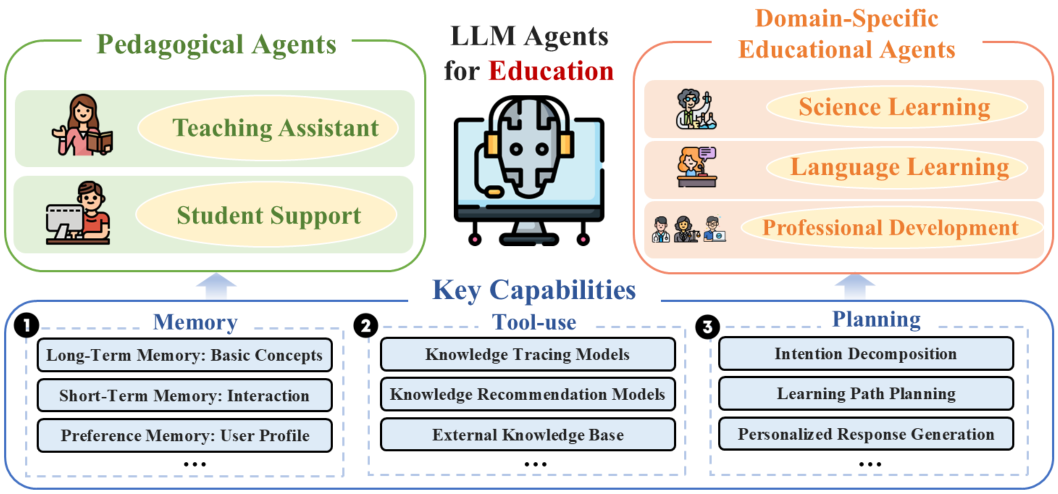

Figure 1: The overview of LLM Agents for education.

2 LLM Agents for Education

LLM agents, powered by LLMs [150, 151, 188], combine core components like memory, tool use, and planning to enhance the performance for diverse tasks [41, 124, 189, 208, 223]. They leverage long-term memory for foundational knowledge and short-term memory for real-time adaptation, enabling context-aware interactions. Active memory management, such as summarization and retrieval [31, 118], helps maintain relevant information. To overcome LLM limitations, such as knowledge cutoff and calculation inefficiency [121], LLM agents integrate external tools like search engines, calculators, and databases [157, 257], enabling real-time data access and multi-step tasks. Additionally, planning capabilities allow agents to decompose goals, simulate state transitions, and optimize actions for complex problem-solving [192, 210, 229].

Education, a field inherently focused on individualized learning and problem-solving, is one of the ideal scenarios for deploying LLM agents. These agents can significantly enhance the learning experience by providing personalized support tailored to the unique needs of each student. To achieve this, as illustrated in Figure 1, LLM agents must integrate several key capabilities, including memory, tool-use, and planning modules. The memory module enables the agent to retain an understanding of fundamental concepts, track the student's past interactions and progress, and store personalized profiles to provide accurate responses. The tool-use module allows the agent to access specialized educational tools, such as knowledge tracing and recommendation models, as well as external resources like search engines, thereby expanding its functional capabilities and ensuring up-to-date, accurate information. The planning module empowers the agent to deconstruct the student's objectives, plan tailored learning paths, and offer personalized feedback based on the student's progress and performance. Together, these modules allow LLM agents to function as dynamic, responsive tutors capable of guiding students through both simple and complex tasks. By offering context-aware feedback, adapting to evolving needs, and tailoring learning experiences, LLM agents have the potential to revolutionize education. Their ability to provide scalable, adaptive, and efficient support can transform personalized education, making it more accessible, engaging, and effective. As shown in Figure 2, we categorize LLM agents for education into two types: Pedagogical Agents and Domain-Specific Educational Agents, and outline the roadmap for discussion in the following sections.

3 Pedagogical Agents

Pedagogical agents are AI-driven systems designed to enhance both teaching and learning through automation and personalization. These agents can be broadly categorized into two types: **Agents for Teaching Assistance**, which automate key tasks for educators, and **Agents for Student Support**, which provide personalized learning guidance to students. By streamlining tasks for teachers and offering real-time, adaptive feedback to students, these agents improve instructional effectiveness, reduce workloads, and foster engaging, tailored learning experiences.

3.1 Agent for Teaching Assistance

Agents for teaching assistance are designed to support educators in the learning environment. These agents leverage LLMs to provide personalized, scalable, and efficient support across various aspects of the educational process. Their primary objectives are to enhance teaching quality, enrich student learning experiences, and reduce educators' workload. By incorporating advanced agent capabilities—such as memory management for retaining contextual information, tool integration for accessing external databases and APIs, and planning for precise student profile modeling—these agents can effectively assist in key areas, including *classroom simulation* (§3.1.1), *feedback comment generation* (§3.1.2), and *learning resource recommendation* (§3.1.3).

3.1.1 Classroom Simulation (CS). Classroom simulation refers to the ability of teaching agents to replicate and model various classroom scenarios, such as student-teacher dialogues, collaborative learning activities, and problem-solving tasks. These simulations create dynamic and interactive learning environments where educators can experiment with different teaching strategies, assess student reactions, and receive real-time feedback on how various pedagogical approaches may unfold. By simulating these classroom dynamics, educators can refine their methods, anticipate student challenges, and enhance overall instructional effectiveness, all without the constraints of a physical classroom setting.

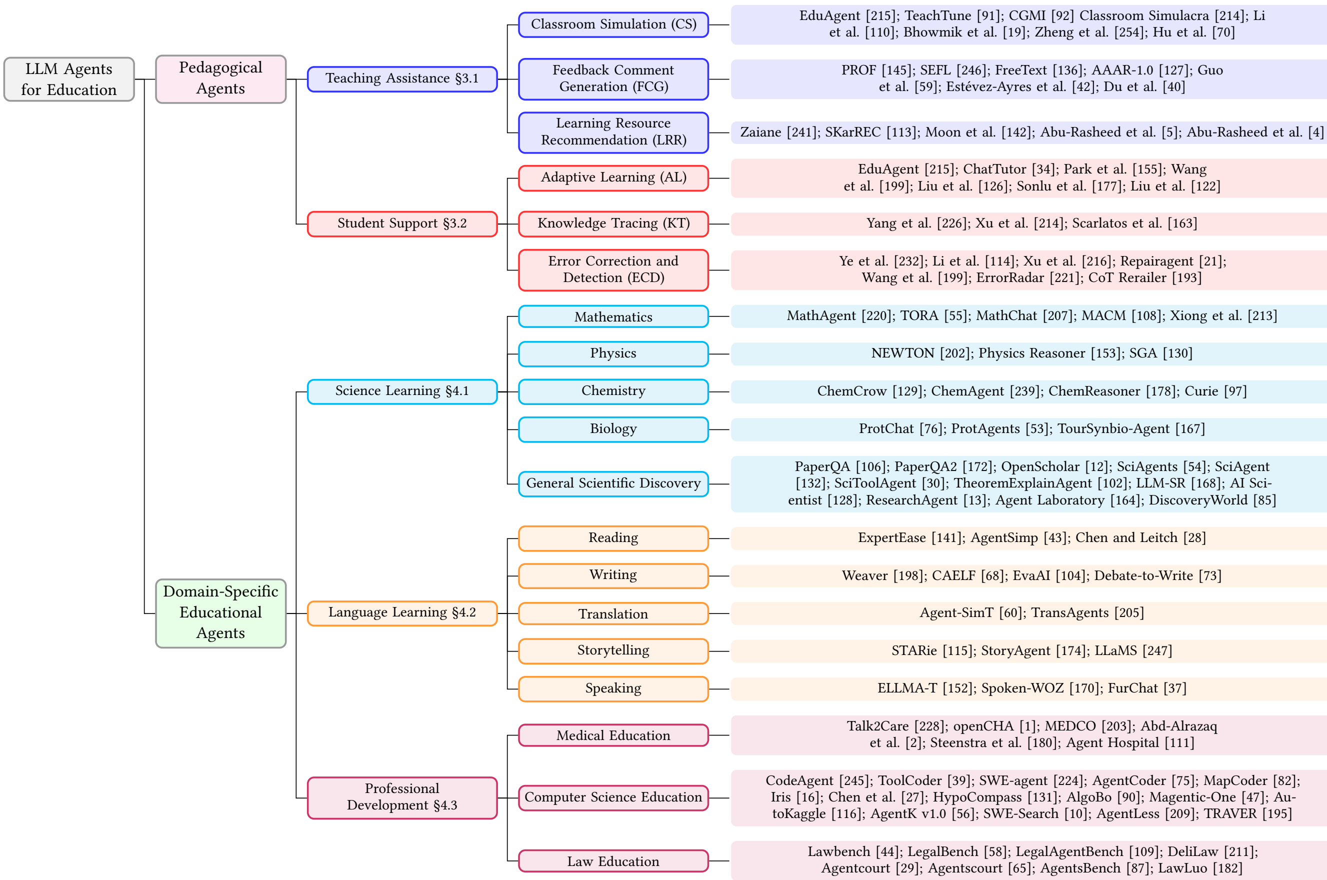

Figure 2: Taxonomy of representative research on education agents.

Effective classroom simulation depends on accurately modeling student behavior. Previous studies [91, 110, 215] demonstrate that LLM-based agents can predict fine-grained student behaviors across diverse personas and past learning patterns, aligning closely with real teachers' expectations. To enhance simulation, the CGMI framework [92] uses a tree-based cognitive architecture with memory, reflection, and planning modules to simulate roles like teacher, student, and supervisor, improving realism. Similarly, Classroom Simulacra [214] incorporates a transferable iterative reflection module for more accurate behavior simulation. These systems enable automated interactions that reduce educators' task loads while broadening the exploration of student profiles. Simulations can also test educational strategies tailored to different profiles, enhancing teaching quality, as shown in studies by Bhowmik et al. [19] and Zheng et al. [254]. Additionally, Hu et al. [70] demonstrate how LLMs can refine teaching plans through integrated simulations. To sum up, classroom simulations can be leveraged to test different educational strategies tailored to diverse student profiles, ultimately enhancing the quality of education.

3.1.2 Feedback Comment Generation (FCG). Providing timely, relevant, and constructive feedback is a cornerstone of effective education. Teaching assistant agents can generate automated feedback comments on students' assignments, quizzes, and projects. For example, Guo et al. [59] presents a training-free system for providing accurate feedback to students, which introduces a two-agent system. Specifically, Agent 1 generates initial feedback based on the students' responses, while Agent 2 evaluates and refines this feedback to prevent over-praise and excessive inferences. Furthermore, Nair et al. [145] designs a training strategy called PROF, which trains an automated LLM-based writing comment generator through reinforcement learning. This system adopts an iterative pipeline to simulate various student writing styles and incorporates a more advanced revision model (e.g., GPT-4) to provide the quality of the feedback as rewards. Similarly, SEFL [246] enhances feedback generation by having LLM agents role-play both students and teachers to generate data, which is then used to fine-tune models and improve feedback capabilities. These systems have also been deployed in real-world applications, such as FreeText [136], which pairs student responses with teacher-provided criteria, enabling the agent to identify strengths and weaknesses and provide targeted feedback for improvement. Beyond traditional feedback, advanced LLM agents are now capable of handling more complex, expertise-intensive tasks. For example, Du et al. [40] explores the potential of LLM agents as assistants for natural language processing (NLP) paper reviewing tasks, while AAAR-1.0 [127] evaluates agents' capabilities in areas such as equation inference, experiment design, paper weakness analysis, and review critique, revealing their potential in conducting sophisticated research tasks.

However, another line of research highlights that agent-generated feedback still faces challenges in handling complex tasks, such as programming and the review of professional academic papers [42, 127]. For instance, these agents may struggle with concepts like starvation and deadlocks, leading to inaccurate or incorrect feedback. Future work could focus on integrating external tools (e.g., search engines) and enhancing memory mechanisms to better address complex problem-solving scenarios, as well as refining the personalization of feedback across diverse learning contexts.

3.1.3 *Learning Resource Recommendation (LRR).*

To ensure that students access the most suitable learning resources aligned with their learning path and knowledge domain, it is crucial to develop an effective recommendation system. Zaiane [241] first introduced recommendation systems into e-learning, utilizing web mining techniques to suggest online learning activities or shortcuts on course websites. Since then, the integration of LLM-based agents has opened new avenues for learning resource recommendations. These recommendations can be generated using both retrieval-based and generation-based methods. Retrieval-based methods involve agents accessing a database or their own memory to suggest existing resources—such as textbooks, research papers, or online content—based on student queries, past behavior, or content similarities [113, 166]. In contrast, generative methods create new learning content tailored to an individual student's learning style, knowledge gaps, and interests [142]. Moreover, to enhance the understanding and acceptance of recommended content, these agents need to provide reasons for their recommendations. Explaining the rationale behind suggestions fosters trust and enables students to make more informed decisions about the resources they engage with. For example, Abu-Rasheed et al. [4, 5] incorporate knowledge graphs to recommend human-curated sources of information. This approach not only enhances the interpretability of recommendations but also reduces the risk of generating misinformation, thereby improving the quality of the learning experience.

Future work should focus on integrating agents with adaptive learning systems to dynamically adjust recommendations based on real-time student performance. Also, a hybrid approach combining generative and retrieval-based methods could enhance both accuracy and diversity. Furthermore, incorporating multi-modal content—including interactive media, video, and immersive simulations—could further enrich the learning experience.

3.2 Agent for Student Support

LLM agent-based student support systems enable independent, personalized learning by providing real-time academic assistance without requiring direct teacher involvement. Unlike traditional rule-based learning systems, they deliver interactive, tailored feedback, allowing students to learn at their own pace. Broadly, their core functionalities include *adaptive learning* (§3.2.1), *knowledge tracing* (§3.2.2), and *error correction and detection* (§3.2.3).

3.2.1 *Adaptive Learning (AL).*

LLM agents can automate tasks, making it possible to build self-sustaining adaptive learning systems without teacher involvement. These systems operate within an action-memory framework, continuously refining instruction based on student profiling, content selection, and difficulty adaptation. The agent maintains a structured student profile, dynamically selecting materials based on past interactions. After each task, it analyzes student behavior and adjusts teaching strategies accordingly, enhancing personalization and responsiveness.

Several implementations exemplify this adaptive approach. EduAgent [215] introduces a structured profiling mechanism comprising four distinct cognitive patterns: gaze behavior linked to physiological memory, motor behavior mapped to motor memory, cognitive state associated with cognitive memory, and post-course assessments contributing to knowledge memory. This structured representation enhances adaptive decision-making by providing a multi-faceted view of student learning states. Chen et al. [34] propose a system consisting of interaction, reflection, and reaction, with each component composed of specific LLM tools and memory modules. Furthermore, a meta-agent is introduced to control the information flow through these agents.

Building on these adaptive mechanisms, LLM agents operate within a state-action framework, allowing them to dynamically adjust instructional strategies in response to a student's evolving cognitive and affective state. The state space in these systems encodes the student's cognitive and affective profile, while the action space governs how the agent adjusts its teaching approach. A typical cognitive state representation includes tracking a student's knowledge proficiency, comprehension levels, and misconceptions [122, 126, 155], allowing the agent to tailor explanations, adjust difficulty levels, and reinforce concepts dynamically. Recent studies highlight the importance of affective state modeling, as emotional factors such as motivation, interest, and self-efficacy significantly influence learning outcomes. For instance, Park et al. [155] propose an affective state model that enables agents to adjust feedback tone, provide encouragement, and regulate pacing to maintain engagement. Another crucial dimension of adaptation involves learning preferences and personality traits. These studies integrate personality with memory design, tracking the psychological state of the students. Wang et al. [199] integrate learning preferences into state modeling, recognizing that students process information differently depending on instructional format and modality. Adapting content to these preferences enhances retention and learning efficiency. Moreover, Liu et al. [126] apply the Big Five personality model [160] to personalize tutoring strategies, acknowledging that individual differences shape learning experiences [177].

Recent approaches explore multi-agent systems for adaptive learning, where specialized agents collaborate to enhance personalization. For example, Wang et al. [199] design five agents—Gap Identifier, Learner Profiler, Dynamic Learner Simulator, Learning Path Scheduler, and Content Creator—to deliver goal-oriented, personalized instruction. Similarly, OATutor [154] provides an experimental platform for modular adaptive learning, allowing researchers to design scalable, domain-general tutoring agents.

3.2.2 *Knowledge Tracing (KT).*

Knowledge tracing is a fundamental component of intelligent tutoring systems, enabling the tracking of a learner's evolving understanding over time to predict future performance and knowledge retention. Traditional knowledge tracing methods often rely on statistical models or deep learning approaches to estimate student mastery levels. In contrast, LLM agents introduce a more dynamic and personalized approach, leveraging their ability to process natural language interactions, infer conceptual mastery, and adapt instructional strategies accordingly.

Recent advancements have explored multi-agent frameworks for knowledge tracing. For instance, Yang et al. [226] propose a multi-agent system with three specialized agent roles: administrator, judger, and critic. In this framework, the administrator delegates knowledge tracing tasks to judgers, who collaborate through discussions to assess the student's cognitive state. The critic agent then evaluates the outcome and determines whether the assessment criteria are met, ensuring a structured yet flexible knowledge tracing process. Other agent-based approaches explore alternative strategies for modeling student knowledge. Xu et al. [215] propose simulating students as different personas, allowing agents to adaptively trace knowledge progression based on varied learning profiles. Meanwhile, Scarlatos et al. [163] employ dialogue-driven interactions to probe students' conceptual boundaries, using conversational exchanges to refine knowledge estimation dynamically.

3.2.3 *Error Correction and Detection (ECD).*

Error detection and correction are critical components of intelligent tutoring systems, enabling students to refine their understanding by receiving real-time, context-aware feedback. LLM agents can identify errors in student responses across various domains, including academic writing, programming, and mathematical reasoning. By dynamically adapting feedback to the learner's proficiency level, these agents serve as intelligent reviewers, debuggers, and writing assistants [114, 232].

Agent-based systems leverage state representations and adaptive inference mechanisms to track error patterns and misconceptions dynamically [21, 126, 155, 199]. Recent advancements extend this capability into the multi-modal domain, incorporating direct analysis of student-generated drafts. Xu et al. [216] propose a multimodal LLM framework that processes handwritten or digitally drafted student work. The system first extracts and converts draft content into natural language, enabling the LLM-based agent to interpret and analyze handwritten responses. The agent then provides indirect yet effective instructional feedback, guiding students toward self-correction and deeper comprehension. Moreover, CoT Rerailer [193] designs a derailment identification process and a rerailment process to conduct error detection. Zhang et al. [249] propose the MathCCS benchmark and develop a sequence error analysis framework, adopting multi-agent collaboration. As the first benchmark for multimodal error detection, ErrorRadar [221] can provide a data foundation for multimodal agents in error detection task.

4 Domain-Specific Educational Agents

Recent research on LLM agents in education has also shown growing interest in domain-specific applications. We explore their use in *science learning*, *language learning*, and *professional development*, focusing on their algorithmic frameworks, agentic designs, and relevant datasets and benchmarks.

4.1 Agent for Science Learning

An agent for science learning is an intelligent system powered by LLMs, designed to assist students in acquiring and applying scientific knowledge through personalized, interactive experiences [22, 149, 158, 222]. The significance of these agents in education lies in their ability to offer tailored feedback, enhance conceptual understanding, and promote active engagement with complex scientific ideas. In the following sections, we explore the impact of LLM agents in four key scientific disciplines: *mathematics* (§4.1.1), *physics* (§4.1.2), *chemistry* (§4.1.3), and *biology* (§4.1.4), as well as their broader contributions to *general scientific discovery* (§4.1.5).

4.1.1 *Mathematics.*

In mathematics, LLM agents provide substantial support by helping students navigate complex problems and reinforcing their understanding of abstract concepts [140, 183, 206, 213, 219, 221]. For instance, Gou et al. [55] introduce TORA (Tool-integrated Reasoning Agents), a framework that integrates natural language reasoning and program-based tool use to handle mathematical reasoning. MathAgent [220] similarly proposes Mixture-of-Math-Agent framework to address multimodal error detection in real-world K-12 scenarios, and flexibly transform the visual information of different types of questions into forms that are more easily understood by LLMs (*e.g.*, converting plane geometry images into formalized expression). Additionally, MathChat [207] serves as a conversational mathematical problem-solving agent, which consists of a chat-based LLM agent and a tool-based user agent. Furthermore, Xiong et al. [213] propose to use reinforcement learning from human feedback (RLHF) to further improve tool-integrated agents for mathematical problem-solving, and formulate this method as a Markov decision process, distinguishing it from the typical contextual bandit approach used in RLHF. Besides, MACM [108] discuss the limitations of LLMs in handling complex mathematical logical deduction, thus introducing a multi-agent system, which comprises three interactive agents: Thinker, Judge, and Executor.

4.1.2 *Physics.*

In the field of physics, LLM agents help students make sense of challenging concepts and offer interactive tools to simulate physical phenomena [15, 46, 88, 144, 153, 218]. Wang et al. [202] introduce NEWTON, the first pipeline and benchmark to explore the physical reasoning abilities of LLMs. Furthermore, Kortemeyer [98] describe a case study exploring if an LLM agent can pass an introductory calculus-based physics course. In addition, Physics Reasoner [153], a novel knowledge-augmented framework for physics problem-solving, leverages a comprehensive formula set and detailed checklists to ensure accuracy and completeness. It can serve as an agent consisting of three stages - problem analysis, formula retrieval, and guided reasoning. Besides, Ma et al. [130] describe the Scientific Generative Agent (SGA), a bilevel optimization framework designed for physical scientific discovery, and highlight the use of LLMs for generating and revising scientific hypotheses and implementing an exploit-and-explore strategy.

4.1.3 *Chemistry.*

Chemistry education also benefits greatly from LLM agents, which can explain molecular structures, chemical reactions, and experimental processes in an engaging and interactive way [62, 129, 159, 190, 239]. For example, ChemCrow [129] is the first LLM chemistry agent capable of autonomous planning and execution of chemical syntheses, including an insect repellent and three organocatalysts. Yu et al. [239] further present ChemAgent, an enhanced chemistry agent improved over ChemCrow, with a focus on two essential cognitive abilities of chemistry problem-solving: reasoning and grounding. Besides, Curie [97] is an agent framework aimed at incorporating rigor into the experimentation process via three core elements: an intra-agent rigor module to boost reliability, an inter-agent rigor module to ensure systematic control, and an experiment knowledge module to improve interoperability. Recent studies have explored the capabilities of LLMs in complex chemical discovery [84, 143, 161, 178, 227], and their potential can be advanced by leveraging the interactivity of agent-based tool use and the flexibility of planning strategies [159, 175].

4.1.4 Biology. In biology, LLM agents enhance learning by offering detailed explanations of biological processes and providing interactive experiences to explore living systems [18, 50, 179, 222, 252]. For example, ProtChat [76] is a multi-agent tool leveraging GPT-4 and Protein Language Models for seamless protein analysis automation, thus evolutionizing the complexities of protein sequence interpretation. ProtAgents [53] is introduced as a multi-agent modeling framework that combines state-of-the-art LLMs with diverse tools to tackle protein design and analysis. It consists of a team of agents: User, Planner, Assistant, Critic, and Group Chat Manager. Besides, Shen et al. [167] present TourSynbio-Agent, an innovative agent framework that leverages TourSynbio-7B's protein understanding ability to perform various protein engineering tasks, such as mutation analysis, inverse folding, and visualization.

4.1.5 General Scientific Discovery. LLM agents support general scientific discovery by assisting students in data interpretation, hypothesis testing, and creative problem-solving [30, 35, 54, 164, 222]. These LLM agents, such as PaperQA [106], PaperQA2 [172], OpenScholar [12], SciAgents [54], TheoremExplainAgent [102], and LLM-SR [168], can analyze complex scientific datasets, helping students uncover patterns and trends that may not be immediately apparent. In addition, Narayanan et al. [146] present Aviary, an extensible gymnasium for language agents for three challenging scientific tasks: manipulating DNA constructs for molecular cloning, answering research questions by accessing scientific literature, and engineering protein stability. Furthermore, both SciAgent [132] and SciToolAgent [30] extend to a tool-augmented scientific reasoning setting with the help of domain-specific tools. Besides, Agent Laboratory [164] emerges as an agent framework that automates the research process of three phases (Literature Review, Experimentation, and Report Writing) via various LLM agents (PhD, Postdoc, ML Engineer, *etc.*).

4.2 Agent for Language Learning

The integration of LLM agents into language learning is revolutionizing how core competencies—reading, writing, listening, and speaking—are taught and practiced [237]. These skills form the foundation of effective communication and language acquisition, and recent advancements in LLM-based agents have significantly enhanced how learners interact with and acquire these skills. Below, we introduce recent advancements in each subdomain, highlighting the role of LLM agents in enhancing pedagogical outcomes in language acquisition for students and second language (L2) speakers through engaging and adaptive approaches [77, 231, 233].

4.2.1 Reading. Reading comprehension is a vital component of language learning, and LLM agents are playing an increasingly important role in enhancing students' reading abilities. For instance, ExpertEase [141] employs a multi-agent framework to adapt documents for grade-specific audiences, simulating expert-teacher-student collaboration to enhance comprehension. AgentSimp [43] tackles document-level simplification by leveraging multiple agents with distinct roles to ensure coherence and accessibility. Additionally, various LLMs [6] have been used as academic reading companions, demonstrating improved engagement and understanding of complex qualitative texts in educational settings [28].

4.2.2 Writing. The development of writing skills has benefited significantly from NLP tasks like explainable grammatical error correction (EXGEC) [235, 236, 258] and automatic essay scoring (AES) systems [181]. Weaver [198], a family of LLMs fine-tuned for writing tasks [234, 238], outperforms generalist LLMs like GPT-4 in generating human-like narratives. Moreover, Weaver natively supports retrieval-augmented generation (RAG) and function calling, serving as a qualified foundational model for LLM agents. For interactive feedback on student essays, CAELF [68] introduces a multi-agent framework that enables interactive essay feedback. By combining Teaching-Assistant agents' evaluations with teacher-agent arbitration, students can contest grades and engage with the feedback, addressing the "black box" limitations of traditional automated scoring. Inspired by the process of human debate, Debate-to-Write [73] construct a persona-based multi-agent framework that can enable agents to collaboratively debate, discuss ideas, and form a comprehensive plan for argument writing.

4.2.3 Translation. LLM agents demonstrate remarkable advancements in both simultaneous [61, 100] and literary translation [36, 173]. Translation tasks benefit from LLM agents through their ability to integrate specialized tools and orchestrate multi-agent collaboration. Agent-SiMT [60] combines the decision-making capabilities of a Simultaneous Machine Translation (SiMT) policy agent with the generative power of a translation agent, achieving state-of-the-art performance in simultaneous translation by dynamically balancing reading and generation actions. For literary translation, TransAgents [205] employs a multi-agent framework to replicate the complex workflows of human translation teams, addressing cultural nuances and stylistic challenges through collaborative reasoning. This approach not only improves translation quality but also extends LLM applications to linguistically and culturally rich domains. These contributions underscore the importance of tool use and agent collaboration in advancing translation education [253].

4.2.4 Storytelling. Storytelling applications leverage LLM agents to create immersive and interactive learning experiences [171]. STARie [115], a peer-like embodied conversational agent, integrates multimodal tools such as speech synthesis and facial animation to scaffold children's storytelling, fostering narrative creativity and oral communication skills [17, 25]. StoryAgent [174] combines top-down story drafting with bottom-up asset generation to transform simple prompts into coherent, multi-modal digital narratives. By automating complex storytelling workflows [119], it democratizes content creation and enhances engagement in language learning. LLaMS [247], a multi-modal agent framework, is designed to generate multi-modal human-level stories characterized by expressiveness and consistency, incorporating the Story-Adapter module for long image sequence illustration. These systems demonstrate the potential of LLM agents to support both cognitive and creative aspects of language education for children.

4.2.5 Speaking. LLM agents are revolutionizing spoken language education by integrating reasoning and multi-agent collaboration to build adaptive dialogue systems [14, 125]. ELLMA-T [152] employs contextual reasoning and role-playing in social VR environments to provide personalized feedback and language assessments, enabling learners to practice speaking in realistic scenarios [112, 120]. SpokenWOZ [170] introduces a large-scale benchmark for task-oriented spoken dialogue, highlighting the importance of reasoning and multi-turn interaction in addressing real-world conversational challenges. FurChat [37], an embodied conversational agent, combines verbal and non-verbal communication cues to simulate natural interactions, making it a valuable tool for improving speaking skills through immersive and realistic practice. By employing multimodal signals such as speech and gestures, SpeechAgents [244] enhances the authenticity of dialogue simulations, capturing consistent content, natural rhythm, and rich emotional expression. Through Multi-Agent Tuning [117], it optimizes LLM capabilities for large-scale simulations involving up to 25 agents, enabling applications like drama creation and audio novel generation.

4.3 Agent for Professional Development

Agents for professional development harness the capabilities of LLMs to offer scalable, adaptive, and context-aware learning experiences tailored to domain-specific needs. This section summarizes how recent studies develop agents to revolutionize professional training in fields including *medical* (§4.3.1), *computer science* (§4.3.2), and *law education* (§4.3.3).

4.3.1 Medical Education. The deployment of LLM agents in healthcare has created new opportunities for personalized, interactive, and scalable systems, with several health agents introduced [169] such as Talk2Care [228] and openCHA [1]. Additionally, Abd-Alrazaq et al. [2] highlight the educational potentials of LLMs in crafting personalized curricula, adaptive learning plans, and dynamic assessment tools for medical education, while concurrently addressing challenges including algorithmic bias, misinformation, and privacy issues. MEDCO [203], a multi-agent system, has the capacity to replicate real-world medical training environments through agent collaboration with virtual patients, expert physicians, and radiologists, enhancing interdisciplinary learning and peer interaction. Furthermore, Abbasian et al. [1] introduce openCHA as a personalized LLM-powered framework that integrates external resources and orchestrates multi-step problem-solving for complex healthcare queries [230], emphasizing tool use and action planning. Beyond traditional education, Steenstra et al. [180] explore LLMs in creating fantasy narrative games for adolescent health education, demonstrating the agents' ability to generate engaging, doctor-validated content that enhances knowledge retention through gamification. Li et al. [111] present Agent Hospital, a simulation environment where LLM-driven agents evolve through autonomous interactions, demonstrating significant improvements in medical reasoning and performance on benchmarks like MedQA [89] after treating thousands of simulated patients. Collectively, these investigations highlight the versatility of LLM agents within medical education, demonstrating their abilities in reasoning, collaboration, tool integration, and adaptive learning to effectively address a broad spectrum of educational and clinical challenges [93, 187, 191].

4.3.2 Computer Science Education. An agent for computer science (CS) education greatly enhances learning by providing personalized guidance on coding, debugging, and understanding CS principles [99, 107, 123, 131]. For example, CodeAgent [245] serves as an LLM agent framework for repo-level code generation, incorporating external tools such as WebSearch and DocSearch. Recent studies have demonstrated the potential of agent-based code generation systems such as ToolCoder [39], SWE-agent [224], AgentCoder [75], and MapCoder [82], which can significantly enhance students' coding efficiency [48, 90, 195]. Furthermore, Bassner et al. [16] introduce Iris, an LLM-driven virtual tutor designed to offer personalized, context-aware assistance to CS students within the interactive learning platform Artemis. Besides, Chen et al. [27] propose Learning-by-Teaching (LBT) as an effective pedagogical strategy for CS education, and leverage the advantages of LLM agents (*e.g.*, contextual conversation & learning from demonstrations).

4.3.3 Law Education. LLM agents leverage pre-trained legal knowledge, interactive capabilities, and reasoning skills to support law education through judicial interpretation, moot court simulation, and case analysis [29, 105, 148, 240]. However, evaluations from LawBench [44] and LegalBench [58] reveal that LLMs struggle with legal knowledge application and judicial aid. LegalAgentBench [109] further highlights their limitations in multi-hop reasoning and defense statement writing, showing that LLM agents require significant improvements to effectively assist in complex legal tasks. Despite these challenges, LLM agents are emerging as valuable tools for moot court simulations, a crucial component of legal reasoning and advocacy training. DeliLaw [211] enhances law education by integrating legal and case retrieval modules, enabling students to practice legal research, case analysis, statutory interpretation, and mock consultations. LawLuo [182] applies a multi-agent framework with retrieval-augmented generation to simulate multi-turn legal consultations, improving personalization and ambiguity handling. Similarly, AgentCourt [29] and AgentsCourt [65] simulate courtroom interactions and judicial decision-making, providing a realistic training ground for law students. AgentsBench [87] extends this by offering multi-agent legal reasoning and case analysis, further advancing AI-driven legal education.

5 Challenges & Future Directions

In this section, we discuss key challenges that must be addressed to ensure the effectiveness, reliability, and ethical deployment of LLM agents for education. We focus on three critical areas: privacy, bias, and fairness concerns; hallucination and overreliance; and integration with existing educational ecosystems. For each challenge, we outline potential research directions to enhance the robustness and applicability of LLM agents in learning environments.

5.1 Privacy, Bias and Fairness Considerations

The integration of LLM agents into educational applications offers transformative potential but also raises significant ethical concerns, particularly regarding privacy and bias.

LLM agents process vast datasets, often containing sensitive personal information, leading to potential privacy risks. Studies highlight low technological readiness and insufficient privacy measures in educational contexts [217]. Emerging research [32, 49, 64, 74, 81, 251] underscores new privacy and security concerns, emphasizing the need for stronger data protection mechanisms [78, 83, 95]. Additionally, bias in LLMs remains a pressing concern, as models trained on large datasets can inadvertently reinforce stereotypes and disparities, affecting educational fairness. Recent work calls for bias mitigation strategies to promote equitable learning experiences [8, 9, 137]. Addressing these biases is essential to ensuring inclusive, unbiased educational outcomes.

To overcome the above issues, a number of future directions can be explored: (i) *Privacy-preserving memory management:* develop context-aware memory mechanisms that allow agents to retain useful learning progress while forgetting sensitive user data when necessary. (ii) *Bias detection and mitigation:* develop automated fairness-checking models that evaluate real-time content generated by LLM agents to detect biased explanations, language, or examples. (iii) *Culturally adaptive LLM Agents for global education*: train multilingual, culturally aware educational agents that dynamically adjust explanations based on regional learning norms, historical perspectives, and diverse curricula.

5.2 Hallucination and Overreliance

The phenomenon of "hallucinations" [248], where LLMs generate plausible-sounding but incorrect or nonsensical information, poses challenges to their reliability in educational contexts.

In educational settings, hallucinations can mislead learners by presenting plausible but inaccurate information as factual, leading to misconceptions [38, 67, 86]. For example, AI-generated content may fabricate historical events or scientific data, which students may unknowingly accept as true. This issue is particularly concerning given the authoritative tone of AI-generated responses, making errors harder to detect and correct. Also, recent researches [7, 101, 138] show that overreliance on AI-generated content may hinder genuine skill acquisition and impede in-depth learning.

Some directions can be explored to mitigate hallucinations in LLM agents for education: (i) *Self-correcting AI tutors:* develop LLM agents with self-reflection capabilities, where models review, verify, and refine their own generated content before presenting it to students. (ii) *Hybrid Human-AI feedback loops for educational content verification:* develop teacher-in-the-loop AI systems where educators can review and correct AI-generated responses, refining LLM performance over time. (iii) *Pedagogical-aware educational agents:* design agentic frameworks that align with human expert pedagogical practices.

5.3 Integration of LLM agents into Existing Educational Ecosystems

The integration of LLM agents into education presents opportunities for personalized learning, curriculum development, and project-based learning (PBL), but also raises challenges related to structured deployment, equitable access, and privacy.

One major challenge is the lack of structured frameworks for integrating LLM agents into educational systems. While models like the FOKE framework [71] combine foundation models, knowledge graphs, and prompt engineering to provide interactive and explainable learning services, broader adoption requires scalable models that can be validated in diverse real-world educational settings. Additionally, LLMs have been explored as tools to enhance creativity and collaboration in PBL, supporting students through brainstorming, problem-solving, and project execution. However, studies indicate that their effectiveness is limited by the absence of structured guidance frameworks that help educators and students seamlessly incorporate LLM agents into PBL workflows [242]. Another critical challenge is ensuring equitable access to LLM-powered educational tools, particularly in underfunded schools and institutions with limited AI infrastructure. Platforms such as AI-VERDE [139] aim to democratize access by providing a unified *LLM-as-a-platform* service with built-in access control, privacy-preserving mechanisms, and budget management. However, achieving widespread adoption still depends on scalable and cost-effective deployment strategies that can support educational institutions at different resource levels.

Future research should focus on developing standardized frameworks to guide the structured deployment of LLMs in personalized learning, PBL, and assessment. Expanding models like FOKE with adaptive learning strategies, multimodal content, and real-time feedback could enhance instructional effectiveness. Additionally, integrating interactive AI tutors that support student collaboration, project tracking, and contextual guidance would further improve PBL applications. Ensuring cost-effective AI deployment through cloud-based and decentralized models would make LLM-powered learning tools more accessible to a wider range of institutions. Furthermore, providing educators with AI literacy training and oversight tools is essential for responsible and effective integration.

6 Conclusion

In this survey, we provided a comprehensive review of LLM agents for education, highlighting their potential to revolutionize personalized learning, intelligent tutoring, and pedagogical automation. We introduced a task-centric taxonomy, categorizing LLM agents into Pedagogical Agents and Domain-Specific Educational Agents, and analyzed their applications across various educational domains. Furthermore, we discussed key challenges and future research directions, including ethical considerations, bias mitigation, and real-world integration. To support further advancements, we compiled essential datasets and benchmarks to facilitate research in this emerging field. As LLM agents continue to evolve, their impact on engagement, automation, and personalized learning will grow, but ensuring their effectiveness, reliability, and ethical deployment remains a key challenge. We hope this survey serves as a foundation for future research, driving innovations in AI-driven education and advancing educational equity.

References

[1] Mahyar Abbasian, Iman Azimi, Amir M Rahmani, and Ramesh Jain. 2023. Conversational health agents: A personalized llm-powered agent framework. *arXiv preprint arXiv:2310.02374* (2023).

[2] Alaa Abd-Alrazaq, Rawan AlSaad, Dari Alhuwail, Arfan Ahmed, Padraig Mark Healy, Syed Latifi, Sarah Aziz, Rafat Damseh, Sadam Alabed Alrazak, Javaid Sheikh, et al. 2023. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Medical Education* 9, 1 (2023), e48291.

[3] Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes. 2023. Knowledge tracing: A survey. *Comput. Surveys* 55, 11 (2023), 1–37.

[4] Hasan Abu-Rasheed, Mohamad Hussam Abdulsalam, Christian Weber, and Madjid Fathi. 2024. Supporting student decisions on learning recommendations: An llm-based chatbot with knowledge graph contextualization for conversational explainability and mentoring. *arXiv preprint arXiv:2401.08517* (2024).

[5] Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi. 2024. Knowledge graphs as context sources for llm-based explanations of learning recommendations. In *2024 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 1–5.

[6] Adebowale Jeremy Adetayo, Mariam Oyinda Aborisade, and Basheer Abiodun Sanni. 2024. Microsoft Copilot and Anthropic Claude AI in education and library service. *Library Hi Tech News* (2024).

[7] Tosin Adewumi, Lama Alkhaled, Claudia Buck, Sergio Hernandez, Saga Brilioth, Mkpe Kekung, Yelvin Ragimov, and Elisa Barney. 2023. Procot: Stimulating critical thinking and writing of students through engagement with large language models (llms). *arXiv preprint arXiv:2312.09801* (2023).

[8] Tosin Adewumi, Lama Alkhaled, Namrata Gurung, Goya van Boven, and Irene Pagliai. 2024. Fairness and bias in multimodal ai: A survey. *arXiv preprint arXiv:2406.19097* (2024).

[9] Amanda Aird, Paresha Farastu, Joshua Sun, Elena Stefancová, Cassidy All, Amy Voida, Nicholas Mattei, and Robin Burke. 2024. Dynamic fairness-aware recommendation through multi-agent social choice. *ACM Transactions on Recommender Systems* 3, 2 (2024), 1–35.

[10] Antonis Antoniades, Albert Örwall, Kexun Zhang, Yuxi Xie, Anirudh Goyal, and William Wang. 2024. SWE-Search: Enhancing Software Agents with Monte Carlo Tree Search and Iterative Refinement. *arXiv preprint arXiv:2410.20285* (2024).

[11] Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. SD-Eval: A Benchmark Dataset for Spoken Dialogue Understanding Beyond Words. (2024). arXiv:2406.13340 [cs.CL]

[12] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199* (2024).

[13] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738* (2024).

[14] Raluca Balan, Anca Dobrean, and Costina R Poetar. 2024. Use of automated conversational agents in improving young population mental health: a scoping review. *NPJ Digital Medicine* 7, 1 (2024), 75.

[15] Kristian G Barman, Sascha Caron, Emily Sullivan, Henk W de Regt, Roberto Ruiz de Austri, Mieke Boon, Michael Färber, Stefan Fröse, Faegheh Hasibi, Andreas Ipp, et al. 2025. Large Physics Models: Towards a collaborative approach with Large Language Models and Foundation Models. *arXiv preprint arXiv:2501.05382* (2025).

[16] Patrick Bassner, Eduard Frankford, and Stephan Krusche. 2024. Iris: An ai-driven virtual tutor for computer science education. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*. 394–400.

[17] Jackylyn Beredo and Ethel Ong. 2021. Beyond the Scene: A Comparative Analysis of Two Storytelling-Based Conversational Agents. In *Proceedings of the Asian CHI Symposium 2021*. 189–195.

[18] Manojit Bhattacharya, Soumen Pal, Srijan Chatterjee, Sang-Soo Lee, and Chiranjib Chakraborty. 2024. Large language model to multimodal large language model: A journey to shape the biological macromolecules to biological sciences and medicine. *Molecular Therapy-Nucleic Acids* 35, 3 (2024).

[19] Saptarshi Bhowmik, Luke West, Alex Barrett, Nuodi Zhang, Chih-Pu Dai, Zlatko Sokolikj, Sherry Southerland, Xin Yuan, and Fengfeng Ke. 2024. Evaluation of an LLM-Powered Student Agent for Teacher Training. In *European Conference on Technology Enhanced Learning*. Springer, 68–74.

[20] Ben Bogin, Kejuan Yang, Shashank Gupta, Kyle Richardson, Erin Bransom, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Super: Evaluating agents on setting up and executing tasks from research repositories. *arXiv preprint arXiv:2409.07440* (2024).

[21] Islem Bouzenia, Premkumar Devanbu, and Michael Pradel. 2024. Repairagent: An autonomous, llm-based agent for program repair. *arXiv preprint arXiv:2403.17134* (2024).

[22] Cameron Brown and Laura Cruz Castro. 2025. Coordinate: A Virtual Classroom Management Tool For Large Computer Science Courses Using Discord. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*. 165–171.

[23] Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. Medbench: A large-scale chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17709–17717.

[24] Ruisheng Cao, Fangyu Lei, Haoyuan Wu, Jixuan Chen, Yeqiao Fu, Hongcheng Gao, Xinzhuang Xiong, Hanchong Zhang, Wenjing Hu, Yuchen Mao, et al. 2024. Spider2-v: How far are multimodal agents from automating data science and engineering workflows? *Advances in Neural Information Processing Systems* 37 (2024), 107703–107744.

[25] Justine Cassell. 2022. Socially interactive agents as peers. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*. 331–366.

[26] Haw-Shiuan Chang, Hwai-Jung Hsu, Kuan-Ta Chen, et al. 2015. Modeling exercise relationships in E-learning: A unified approach.. In *EDM*. 532–535.

[27] Angxuan Chen, Yuang Wei, Huixiao Le, and Yan Zhang. 2024. Learning-by-teaching with ChatGPT: The effect of teachable ChatGPT agent on programming education. *arXiv preprint arXiv:2412.15226* (2024).

[28] Celia Chen and Alex Leitch. 2024. LLMs as Academic Reading Companions: Extending HCI Through Synthetic Personae. *arXiv preprint arXiv:2403.19506* (2024).

[29] Guhong Chen, Liyang Fan, Zihan Gong, Nan Xie, Zixuan Li, Ziqiang Liu, Chengming Li, Qiang Qu, Shiwen Ni, and Min Yang. 2024. Agentcourt: Simulating court with adversarial evolvable lawyer agents. *arXiv preprint arXiv:2408.08089* (2024).

[30] Huajun Chen, Keyan Ding, Jing Yu, Junjie Huang, Yuchen Yang, and Qiang Zhang. 2025. SciToolAgent: A Knowledge Graph-Driven Scientific Agent for Multi-Tool Integration. (2025).

[31] Howard Chen, Ramakanth Pasunuru, Jason Weston, and Asli Celikyilmaz. 2023. Walking down the memory maze: Beyond context limit through interactive reading. *arXiv preprint arXiv:2310.05029* (2023).

[32] Junkai Chen, Zhijie Deng, Kening Zheng, Yibo Yan, Shuliang Liu, PeiJun Wu, Peijie Jiang, Jia Liu, and Xuming Hu. 2025. SAFEERASER: Enhancing Safety in Multimodal Large Language Models through Multimodal Machine Unlearning. *arXiv preprint arXiv:2502.12520* (2025).

[33] Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial intelligence in education: A review. *Ieee Access* 8 (2020), 75264–75278.

[34] Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2024. Empowering private tutoring by chaining large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 354–364.

[35] Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. 2024. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080* (2024).

[36] Shanbo Cheng, Zhichao Huang, Tom Ko, Hang Li, Ningxin Peng, Lu Xu, and Qini Zhang. 2024. Towards Achieving Human Parity on End-to-end Simultaneous Speech Translation via LLM Agent. *arXiv preprint arXiv:2407.21646* (2024).

[37] Neeraj Cherakara, Finny Varghese, Sheena Shabana, Nivan Nelson, Abhiram Karukayil, Rohith Kulothungan, Mohammed Afil Farhan, Birthe Nesset, Meriam Moujahid, Tanvi Dinkar, Verena Rieser, and Oliver Lemon. 2023. FurChat: An Embodied Conversational Agent using LLMs, Combining Open and Closed-Domain Dialogue with Facial Expressions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani (Eds.). Association for Computational Linguistics, Prague, Czechia, 588–592. doi:10.18653/v1/2023.sigdial-1.55

[38] Wildemarkes de Almeida da Silva, Luis Carlos Costa Fonseca, Sofiane Labidi, and José Chrystian Lima Pacheco. 2024. Mitigation of Hallucinations in Language Models in Education: A New Approach of Comparative and Cross-Verification. In *2024 IEEE International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 207–209.

[39] Hanxing Ding, Shuchang Tao, Liang Pang, Zihao Wei, Jinyang Gao, Bolin Ding, Huawei Shen, and Xueqi Chen. 2025. ToolCoder: A Systematic Code-Empowered Tool Learning Framework for Large Language Models. *arXiv preprint arXiv:2502.11404* (2025).

[40] Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. 2024. LLMs Assist NLP Researchers: Critique Paper (Meta-) Reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 5081–5099.

[41] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint*

arXiv:2401.03568 (2024).

[42] Iria Estévez-Ayres, Patricia Callejo, Miguel Ángel Hombrados-Herrera, Carlos Alario-Hoyos, and Carlos Delgado Kloos. 2024. Evaluation of LLM tools for feedback generation in a course on concurrent programming. *International Journal of Artificial Intelligence in Education* (2024), 1–17.

[43] Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative Document Simplification Using Multi-Agent Systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 897–912. https://aclanthology.org/2025.coling-main.60/

[44] Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289* (2023).

[45] Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction* 19 (2009), 243–266.

[46] Mingquan Feng, Yixin Huang, Yizhou Liu, Bofang Jiang, and Junchi Yan. [n. d.]. PhysPDE: Rethinking PDE Discovery and a Physical HYpothesis Selection Benchmark. In *The Thirteenth International Conference on Learning Representations*.

[47] Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, et al. 2024. Magentic-one: A generalist multi-agent system for solving complex tasks. *arXiv preprint arXiv:2411.04468* (2024).

[48] Eduard Frankford, Ingo Höhn, Clemens Sauerwein, and Ruth Breu. 2024. A Survey Study on the State of the Art of Programming Exercise Generation using Large Language Models. In *2024 36th International Conference on Software Engineering Education and Training (CSEE&T)*. IEEE, 1–5.

[49] Yuyou Gan, Yong Yang, Zhe Ma, Ping He, Rui Zeng, Yiming Wang, Qingming Li, Chunyi Zhou, Songze Li, Ting Wang, et al. 2024. Navigating the risks: A survey of security, privacy, and ethics threats in llm-based agents. *arXiv preprint arXiv:2411.09523* (2024).

[50] Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering biomedical discovery with AI agents. *Cell* 187, 22 (2024), 6125–6151.

[51] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. 2021. RCD: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 501–510.

[52] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* 2 (2023).

[53] Alireza Ghafarollahi and Markus J Buehler. 2024. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery* 3, 7 (2024), 1389–1409.

[54] Alireza Ghafarollahi and Markus J Buehler. 2024. SciAgents: Automating Scientific Discovery Through Bioinspired Multi-Agent Intelligent Graph Reasoning. *Advanced Materials* (2024), 2413523.

[55] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452* (2023).

[56] Antoine Grosnit, Alexandre Maraval, James Doran, Giuseppe Paolo, Albert Thomas, Refinath Shahul Hameed Nabeezath Beevi, Jonas Gonzalez, Khyati Khandelwal, Ignacio Iacobacci, Abdelhakim Benechehab, et al. 2024. Large language models orchestrating structured reasoning achieve kaggle grandmaster level. *arXiv preprint arXiv:2411.03562* (2024).

[57] Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, et al. 2024. Blade: Benchmarking language model agents for data-driven science. *arXiv preprint arXiv:2408.09667* (2024).

[58] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems* 36 (2023), 44123–44279.

[59] Shuchen Guo, Ehsan Latif, Yifan Zhou, Xuan Huang, and Xiaoming Zhai. 2024. Using Generative AI and Multi-Agents to Provide Automatic Feedback. *arXiv preprint arXiv:2411.07407* (2024).

[60] Shoutao Guo, Shaolei Zhang, Zhengrui Ma, Min Zhang, and Yang Feng. 2024. Agent-simt: Agent-assisted simultaneous machine translation with large language models. *arXiv preprint arXiv:2406.06910* (2024).

[61] Shoutao Guo, Shaolei Zhang, Zhengrui Ma, Min Zhang, and Yang Feng. 2024. Sillm: Large language models for simultaneous machine translation. *arXiv preprint arXiv:2402.13036* (2024).

[62] Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems* 36 (2023), 59662–59688.

[63] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, et al. 2023. Fabric: Automated scoring and feedback generation for essays. *arXiv preprint arXiv:2310.05191* (2023).

[64] Feng He, Tianqing Zhu, Dayong Ye, Bo Liu, Wanlei Zhou, and Philip S Yu. 2024. The emerged security and privacy of llm agent: A survey with case studies. *arXiv preprint arXiv:2407.19354* (2024).

[65] Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. AgentsCourt: Building Judicial Decision-Making Agents with Court Debate Simulation and Legal Knowledge Augmentation. *arXiv preprint arXiv:2403.02959* (2024).

[66] Zexue He, Yu Wang, An Yan, Yao Liu, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2023. MedEval: A Multi-Level, Multi-Task, and Multi-Domain Medical Benchmark for Language Model Evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8725–8744. doi:10.18653/v1/2023.emnlp-main.540

[67] Huu-Tuong Ho, Duc-Tin Ly, and Luong Vuong Nguyen. 2024. Mitigating Hallucinations in Large Language Models for Educational Application. In *2024 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. IEEE, 1–4.

[68] Shengxin Hong, Chang Cai, Sixuan Du, Haiyue Feng, Siyuan Liu, and Xiuyi Fan. 2024. " My Grade is Wrong!": A Contestable AI Framework for Interactive Feedback in Evaluating Student Essays. *arXiv preprint arXiv:2409.07453* (2024).

[69] Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, et al. 2024. E-eval: a comprehensive Chinese k-12 education evaluation benchmark for large language models. *arXiv preprint arXiv:2401.15927* (2024).

[70] Bihao Hu, Jiayi Zhu, Yiying Pei, and Xiaoqing Gu. 2025. Exploring the potential of LLM to enhance teaching plans through teaching simulation. *npj Science of Learning* 10, 1 (2025), 7.

[71] Silan Hu and Xiaoning Wang. 2024. Foke: A personalized and explainable education framework integrating foundation models, knowledge graphs, and prompt engineering. In *China National Conference on Big Data and Social Computing*. Springer, 399–411.

[72] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22170–22183.

[73] Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025. Debate-to-Write: A Persona-Driven Multi-Agent Framework for Diverse Argument Generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 4689–4703. https://aclanthology.org/2025.coling-main.314/

[74] Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. 2024. Trustagent: Towards safe and trustworthy llm-based agents through agent constitution. In *Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*.

[75] Dong Huang, Jie M Zhang, Michael Luck, Qingwen Bu, Yuhao Qing, and Heming Cui. 2023. Agentcoder: Multi-agent-based code generation with iterative testing and optimisation. *arXiv preprint arXiv:2312.13010* (2023).

[76] Huazhen Huang, Xianguo Shi, Hongyang Lei, Fan Hu, and Yunpeng Cai. 2024. ProtChat: An AI Multi-Agent for Automated Protein Analysis Leveraging GPT-4 and Protein Language Model. *Journal of Chemical Information and Modeling* 65, 1 (2024), 62–70.

[77] Haojing Huang, Jingheng Ye, Qingyu Zhou, Yinghui Li, Yangning Li, Feng Zhou, and Hai-Tao Zheng. 2023. A frustratingly easy plug-and-play detection-and-reasoning module for chinese spelling check. *arXiv preprint arXiv:2310.09119* (2023).

[78] Lan Huang. 2023. Ethics of artificial intelligence in education: Student privacy and data protection. *Science Insights Education Frontiers* 16, 2 (2023), 2577–2587.

[79] Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2023. Mlagentbench: Evaluating language agents on machine learning experimentation. *arXiv preprint arXiv:2310.03302* (2023).

[80] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of LLM agents: A survey. *arXiv preprint arXiv:2402.02716* (2024).

[81] Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. 2025. MMUNLEARNER: Reformulating Multimodal Machine Unlearning in the Era of Multimodal Large Language Models. *arXiv preprint arXiv:2502.11051* (2025).

[82] Md Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. 2024. Mapcoder: Multi-agent code generation for competitive problem solving. *arXiv*

preprint arXiv:2405.11403 (2024).
[83] Islam Asim Ismail. 2025. Protecting Privacy in AI-Enhanced Education: A Comprehensive Examination of Data Privacy Concerns and Solutions in AI-Based Learning. *Impacts of Generative AI on the Future of Research and Education* (2025), 117–142.
[84] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2024. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* 6, 2 (2024), 161–169.
[85] Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. 2024. DISCOVERYWORLD: A virtual environment for developing and evaluating automated scientific discovery agents. *Advances in Neural Information Processing Systems* 37 (2024), 10088–10116.
[86] Hunkoog Jho. 2024. Leveraging generative AI in physics education: Addressing hallucination issues in large language models. (2024).
[87] Cong Jiang and Xiaolei Yang. 2024. Agents on the Bench: Large Language Model Based Multi Agent Framework for Trustworthy Digital Justice. *arXiv preprint arXiv:2412.18697* (2024).
[88] Zhoumingju Jiang and Mengjun Jiang. 2024. Beyond answers: Large language model-powered tutoring system in physics education for deep learning and precise understanding. *arXiv preprint arXiv:2406.10934* (2024).
[89] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11, 14 (2021), 6421.
[90] Hyoungwook Jin, Seonghee Lee, Hyungyu Shin, and Juho Kim. 2024. Teach ai how to code: Using large language models as teachable agents for programming education. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–28.
[91] Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. 2024. TeachTune: Reviewing Pedagogical Agents Against Diverse Student Profiles with Simulated Students. *arXiv preprint arXiv:2410.04078* (2024).
[92] Shi Jinxin, Zhao Jiabao, Wang Yilei, Wu Xingjiao, Li Jiawen, and He Liang. 2023. Cgmi: Configurable general multi-agent interaction framework. *arXiv preprint arXiv:2308.12503* (2023).
[93] Mert Karabacak and Konstantinos Margetis. 2023. Embracing large language models for medical applications: opportunities and challenges. *Cureus* 15, 5 (2023).
[94] Anupam Khan and Soumya K Ghosh. 2021. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and information technologies* 26, 1 (2021), 205–240.
[95] Wajahat Naseeb Khan. 2024. Ethical Challenges of AI in Education: Balancing Innovation with Data Privacy. *Journal of AI in Education: Innovations, Opportunities, Challenges, and Future Directions* 1, 1 (2024), 1–13.
[96] Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649* (2024).
[97] Patrick Tser Jern Kon, Jiachen Liu, Qiuyi Ding, Yiming Qiu, Zhenning Yang, Yibo Huang, Jayanth Srinivasa, Myungjin Lee, Mosharaf Chowdhury, and Ang Chen. 2025. Curie: Toward Rigorous and Automated Scientific Experimentation with AI Agents. *arXiv preprint arXiv:2502.16069* (2025).
[98] Gerd Kortemeyer. 2023. Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research* 19, 1 (2023), 010132.
[99] Tomaž Kosar, Dragana Ostojić, Yu David Liu, and Marjan Mernik. 2024. Computer science education in chatgpt era: Experiences from an experiment in a programming course for novice programmers. *Mathematics* 12, 5 (2024), 629.
[100] Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024. Transllama: Llm-based simultaneous translation system. *arXiv preprint arXiv:2402.04636* (2024).
[101] Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2024. Challenges and opportunities of moderating usage of large language models in education. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*. 249–254.
[102] Max Ku, Thomas Chong, Jonathan Leung, Krish Shah, Alvin Yu, and Wenhu Chen. 2025. TheoremExplainAgent: Towards Multimodal Explanations for LLM Theorem Understanding. arXiv:2502.19400 [cs.AI] https://arxiv.org/abs/2502.19400
[103] Kristjan-Julius Laak and Jaan Aru. 2024. AI and personalized learning: bridging the gap with modern educational goals. *arXiv preprint arXiv:2404.02798* (2024).
[104] Paraskevas Lagakis and Stavros Demetriadis. 2024. EvaAI: a multi-agent framework leveraging large language models for enhanced automated grading. In *International Conference on Intelligent Tutoring Systems*. Springer, 378–385.
[105] Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and S Yu Philip. 2024. Large language models in law: A survey. *AI Open* (2024).
[106] Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559* (2023).
[107] Soohwan Lee and Ki-Sang Song. 2024. Teachers' and students' perceptions of AI-generated concept explanations: Implications for integrating generative AI in computer science education. *Computers and Education: Artificial Intelligence* 7 (2024), 100283.
[108] Bin Lei, Yi Zhang, Shan Zuo, Ali Payani, and Caiwen Ding. 2025. Macm: Utilizing a multi-agent system for condition mining in solving complex mathematical problems. *Advances in Neural Information Processing Systems* 37 (2025), 53418–53437.
[109] Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, et al. 2024. LegalAgentBench: Evaluating LLM Agents in Legal Domain. *arXiv preprint arXiv:2412.17259* (2024).
[110] Haoxuan Li, Jifan Yu, Xin Cong, Yang Dang, Yisi Zhan, Huiqin Liu, and Zhiyuan Liu. 2025. Exploring LLM-based Student Simulation for Metacognitive Cultivation. *arXiv preprint arXiv:2502.11678* (2025).
[111] Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957* (2024).
[112] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. 2025. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems* 37 (2025), 100428–100534.
[113] Qingyao Li, Wei Xia, Kounianhua Du, Qiji Zhang, Weinan Zhang, Ruiming Tang, and Yong Yu. 2024. Learning Structure and Knowledge Aware Representation with Large Language Models for Concept Recommendation. *arXiv preprint arXiv:2405.12442* (2024).
[114] Yinghui Li, Shang Qin, Haojing Huang, Jingheng Ye, Yangning Li, Libo Qin, Xuming Hu, Wenhao Jiang, Hai-Tao Zheng, and Philip S Yu. 2024. Rethinking the roles of large language models in chinese grammatical error correction. *arXiv preprint arXiv:2402.11420* (2024).
[115] Zhixin Li and Ying Xu. 2023. Designing a realistic peer-like embodied conversational agent for supporting children's storytelling. *arXiv preprint arXiv:2304.09399* (2023).
[116] Ziming Li, Qianbo Zang, David Ma, Jiawei Guo, Tuney Zheng, Minghao Liu, Xinyao Niu, Yue Wang, Jian Yang, Jiaheng Liu, et al. 2024. Autokaggle: A multi-agent framework for autonomous data science competitions. *arXiv preprint arXiv:2410.20424* (2024).
[117] Xuechen Liang, Meiling Tao, Yinghui Xia, Tianyu Shi, Jun Wang, and JingSong Yang. 2024. Cmat: A multi-agent collaboration tuning framework for enhancing small language models. *arXiv preprint arXiv:2404.01663* (2024).
[118] Xinnian Liang, Bing Wang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. 2023. Unleashing Infinite-Length Input Capacity for Large-scale Language Models with Self-Controlled Memory System. *arXiv preprint arXiv:2304.13343* (2023).
[119] Johannes Liem, Jakob Kusnick, Samuel Beck, Florian Windhager, and Eva Mayr. 2023. A Workflow Approach to Visualization-Based Storytelling with Cultural Heritage Data. In *2023 IEEE 8th Workshop on Visualization for the Digital Humanities (VIS4DH)*. IEEE, 13–17.
[120] Sue Lim, Ralf Schmälzle, and Gary Bente. 2024. Artificial social influence via human-embodied AI agent interaction in immersive virtual reality (VR): Effects of similarity-matching during health conversations. *arXiv preprint arXiv:2406.05486* (2024).
[121] Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*. PMLR, 13604–13622.
[122] Dejian Liu, Ronghuai Huang, Ying Chen, Michael Agyemang Adarkwah, Xiangling Zhang, Xin Li, Junjie Zhang, and Ting Da. 2024. Personalized Tutoring Through Conversational Agents. In *Using Educational Robots to Enhance Learning: An Analysis of 100 Academic Articles*. Springer, 59–85.
[123] Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J Malan. 2024. Teaching CS50 with AI: leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM technical symposium on computer science education V. 1*. 750–756.
[124] William Liu, Liang Liu, Yaxuan Guo, Han Xiao, Weifeng Lin, Yuxiang Chai, Shuai Ren, Xiaoyu Liang, Linghao Li, Wenhao Wang, et al. 2025. Llm-powered gui agents in phone automation: Surveying progress and prospects. (2025).
[125] Zihan Liu, Han Li, Anfan Chen, Renwen Zhang, and Yi-Chieh Lee. 2024. Understanding public perceptions of AI conversational agents: A cross-cultural analysis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
[126] Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024. Personality-aware Student Simulation for Conversational Intelligent Tutoring Systems.

In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 626–642. doi:10.18653/v1/2024.emnlp-main.37

[127] Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, et al. 2024. AAAR-1.0: Assessing AI's Potential to Assist Research. *arXiv preprint arXiv:2410.22394* (2024).

[128] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292* (2024).

[129] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence* 6, 5 (2024), 525–535.

[130] Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. 2024. Llm and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. *arXiv preprint arXiv:2405.09783* (2024).

[131] Qianou Ma, Hua Shen, Kenneth Koedinger, and Sherry Tongshuang Wu. 2024. How to teach programming in the ai era? using llms as a teachable agent for debugging. In *International Conference on Artificial Intelligence in Education*. Springer, 265–279.

[132] Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, et al. 2024. Sciagent: Tool-augmented language models for scientific reasoning. *arXiv preprint arXiv:2402.11451* (2024).

[133] Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536* (2023).

[134] Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. MathTutorBench: A Benchmark for Measuring Open-ended Pedagogical Capabilities of LLM Tutors. *arXiv preprint arXiv:2502.18940* (2025).

[135] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2024. Discoverybench: Towards data-driven discovery with large language models. *arXiv preprint arXiv:2407.01725* (2024).

[136] Jordan K Matelsky, Felipe Parodi, Tony Liu, Richard D Lange, and Konrad P Kording. 2023. A large language model-assisted education tool to provide feedback on open-ended responses. *arXiv preprint arXiv:2308.02439* (2023).

[137] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. 2024. Integrity-based explanations for fostering appropriate trust in AI agents. *ACM Transactions on Interactive Intelligent Systems* 14, 1 (2024), 1–36.

[138] Silvia Milano, Joshua A McGrane, and Sabina Leonelli. 2023. Large language models challenge the future of higher education. *Nature Machine Intelligence* 5, 4 (2023), 333–334.

[139] Paul Mithun, Enrique Noriega-Atala, Nirav Merchant, and Edwin Skidmore. 2025. AI-VERDE: A Gateway for Egalitarian Access to Large Language Model-Based Resources For Educational Institutions. *arXiv preprint arXiv:2502.09651* (2025).

[140] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830* (2024).

[141] Kaijie Mo and Renfen Hu. 2024. ExpertEase: A Multi-Agent Framework for Grade-Specific Document Simplification with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 9080–9099. doi:10.18653/v1/2024.findings-emnlp.530

[142] Hyeonseok Moon, Jaewook Lee, Sugyeong Eo, Chanjun Park, Jaehyung Seo, and Heui-Seok Lim. 2024. Generative Interpretation: Toward Human-Like Evaluation for Educational Question-Answer Pair Generation. In *Findings of the Association for Computational Linguistics: EACL 2024*. 2185–2196.

[143] Michael Moret, Irene Pachon Angona, Leandro Cotos, Shen Yan, Kenneth Atz, Cyrill Brunner, Martin Baumgartner, Francesca Grisoni, and Gisbert Schneider. 2023. Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. *Nature Communications* 14, 1 (2023), 114.

[144] Christopher E Mower and Haitham Bou-Ammar. 2025. Al-Khwarizmi: Discovering Physical Laws with Foundation Models. *arXiv preprint arXiv:2502.01702* (2025).

[145] Inderjeet Nair, Jiaye Tan, Xiaotian Su, Anne Gere, Xu Wang, and Lu Wang. 2024. Closing the Loop: Learning to Generate Writing Feedback via Language Model Simulated Student Revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 16636–16657.

[146] Siddharth Narayanan, James D Braza, Ryan-Rhys Griffiths, Manu Ponnapati, Albert Bou, Jon Laurent, Ori Kabeli, Geemi Wellawatte, Sam Cox, Samuel G Rodriques, et al. 2024. Aviary: training language agents on challenging scientific tasks. *arXiv preprint arXiv:2412.21154* (2024).

[147] Deepak Nathani, Lovish Madaan, Nicholas Roberts, Nikolay Bashlykov, Ajay Menon, Vincent Moens, Amar Budhiraja, Despoina Magka, Vladislav Vorotilov, Gaurav Chaurasia, et al. 2025. MLGym: A New Framework and Benchmark for Advancing AI Research Agents. *arXiv preprint arXiv:2502.14499* (2025).

[148] Jack Nelson. 2024. The Other'LLM': Large Language Models and the Future of Legal Education. *European Journal of Legal Education* 5, 1 (2024), 127–155.

[149] Davy Tsz Kit Ng, Chee Wei Tan, and Jac Ka Lok Leung. 2024. Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *British Journal of Educational Technology* 55, 4 (2024), 1328–1353.

[150] OpenAI. 2022. ChatGPT.

[151] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[152] Mengxu Pan, Alexandra Kitson, Hongyu Wan, and Mirjana Prpa. 2024. ELLMA-T: an Embodied LLM-agent for Supporting English Language Learning in Social VR. *arXiv preprint arXiv:2410.02406* (2024).

[153] Xinyu Pang, Ruixin Hong, Zhanke Zhou, Fangrui Lv, Xinwei Yang, Zhilong Liang, Bo Han, and Changshui Zhang. 2024. Physics Reasoner: Knowledge-Augmented Reasoning for Solving Physics Problems with Large Language Models. *arXiv preprint arXiv:2412.13791* (2024).

[154] Zachary A. Pardos, Matthew Tang, Ioannis Anastasopoulos, Shreya K. Sheel, and Ethan Zhang. 2023. OATutor: An Open-source Adaptive Tutoring System and Curated Content Library for Learning Sciences Research. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 416, 17 pages. doi:10.1145/3544548.3581574

[155] Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. 2024. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–10.

[156] Francesc Pedro, Miguel Subosa, Axel Rivas, and Paula Valverde. 2019. Artificial intelligence in education: Challenges and opportunities for sustainable development. (2019).

[157] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789* (2023).

[158] Nishat Raihan, Mohammed Latif Siddiq, Joanna CS Santos, and Marcos Zampieri. 2025. Large language models in computer science education: A systematic literature review. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*. 938–944.

[159] Mayk Caldas Ramos, Christopher J Collison, and Andrew D White. 2025. A review of large language models and autonomous agents in chemistry. *Chemical Science* (2025).

[160] Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin* 28, 6 (2002), 789–801.

[161] Yixiang Ruan, Chenyin Lu, Ning Xu, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, Xiaodong Shen, Ning Ye, et al. 2024. Accelerated end-to-end chemical synthesis development with large language models. (2024).

[162] Michael J Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English Text Simplification: A Unified Multilingual Benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4898–4927. doi:10.18653/v1/2023.acl-long.269

[163] Alexander Scarlatos, Ryan S Baker, and Andrew Lan. 2025. Exploring Knowledge Tracing in Tutor-Student Dialogues using LLMs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*. 249–259.

[164] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227* (2025).

[165] Dalia Abdulkareem Shafiq, Mohsen Marjani, Riyaz Ahamed Ariyaluran Habeeb, and David Asirvatham. 2022. Student retention using educational data mining and predictive analytics: a systematic literature review. *IEEE Access* 10 (2022), 72480–72503.

[166] Tariq Shahzad, Tehseen Mazhar, Muhammad Usman Tariq, Wasim Ahmad, Khmaies Ouahada, and Habib Hamam. 2025. A comprehensive review of large language models: issues and solutions in learning environments. *Discover Sustainability* 6, 1 (2025), 27.

[167] Yiqing Shen, Zan Chen, Michail Mamalakis, Yungeng Liu, Tianbin Li, Yanzhou Su, Junjun He, Pietro Liò, and Yu Guang Wang. 2024. Toursynbio: A multi-modal large model and agent framework to bridge text and protein sequences for protein engineering. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2382–2389.

[168] Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. 2024. Llm-sr: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400* (2024).

[169] Roma Shusterman, Allison C Waters, Shannon O'Neill, Marshall Bangs, Phan Luu, and Don M Tucker. 2025. An active inference strategy for prompting

reliable responses from large language models in medical practice. *npj Digital Medicine* 8, 1 (2025), 119.
[170] Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *Advances in Neural Information Processing Systems* 36 (2023), 39088–39118.
[171] Nisha Simon and Christian Muise. 2022. TattleTale: storytelling with planning and large language models. In *ICAPS Workshop on Scheduling and Planning Applications*.
[172] Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740* (2024).
[173] Milena Škobo and Vedran Petričević. 2023. Navigating the challenges and opportunities of literary translation in the age of AI: Striking a balance between human expertise and machine power. *Društvene i humanističke studije* 8, 2 (23) (2023), 317–336.
[174] Samuel S Sohn, Danrui Li, Sen Zhang, Che-Jui Chang, and Mubbasir Kapadia. 2024. From Words to Worlds: Transforming One-line Prompt into Immersive Multi-modal Digital Stories with Communicative LLM Agent. *arXiv preprint arXiv:2406.10478* (2024).
[175] Tao Song, Man Luo, Linjiang Chen, Yan Huang, Qing Zhu, Daobin Liu, Baicheng Zhang, Gang Zou, Fei Zhang, Weiwei Shang, et al. 2024. A multi-agent-driven robotic AI chemist enabling autonomous chemical research on demand. (2024).
[176] Xiangyu Song, Jianxin Li, Taotao Cai, Shuiqiao Yang, Tingting Yang, and Chengfei Liu. 2022. A survey on deep learning based knowledge tracing. *Knowledge-Based Systems* 258 (2022), 110036.
[177] Sinan Sonlu, Bennie Bendiksen, Funda Durupinar, and Uğur Güdükbay. 2024. The effects of embodiment and personality expression on learning in llm-based educational agents. *arXiv preprint arXiv:2407.10993* (2024).
[178] Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. 2024. ChemReasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *arXiv preprint arXiv:2402.10980* (2024).
[179] Kamali N Sripathi, Rosa A Moscarella, Matthew Steele, Rachel Yoho, Hyesun You, Luanna B Prevost, Mark Urban-Lurain, John Merrill, and Kevin C Haudek. 2024. Machine learning mixed methods text analysis: An illustration from automated scoring models of student writing in biology education. *Journal of mixed methods research* 18, 1 (2024), 48–70.
[180] Ian Steenstra, Prasanth Murali, Rebecca B Perkins, Natalie Joseph, Michael K Paasche-Orlow, and Timothy Bickmore. 2024. Engaging and Entertaining Adolescents in Health Education Using LLM-Generated Fantasy Narrative Games and Virtual Agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
[181] Jiamin Su, Yibo Yan, Fangteng Fu, Han Zhang, Jingheng Ye, Xiang Liu, Jiahao Huo, Huiyu Zhou, and Xuming Hu. 2025. EssayJudge: A Multi-Granular Benchmark for Assessing Automated Essay Scoring Capabilities of Multimodal Large Language Models. *arXiv preprint arXiv:2502.11916* (2025).
[182] Jingyun Sun, Chengxiao Dai, Zhongze Luo, Yangbo Chang, and Yang Li. 2024. Lawluo: A chinese law firm co-run by llm agents. *arXiv preprint arXiv:2407.16252* (2024).
[183] Melanie Swan, Takashi Kido, Eric Roland, and Renato P dos Santos. 2023. Math agents: Computational infrastructure, mathematical embedding, and genomics. *arXiv preprint arXiv:2307.02502* (2023).
[184] Kehui Tan, Tianqi Pang, Chenyou Fan, and Song Yu. 2023. Towards applying powerful large ai models in classroom teaching: Opportunities, challenges and prospects. *arXiv preprint arXiv:2305.03433* (2023).
[185] Xiangru Tang, Yuliang Liu, Zefan Cai, Yanjun Shao, Junjie Lu, Yichi Zhang, Zexuan Deng, Helan Hu, Kaikai An, Ruijun Huang, et al. 2023. ML-Bench: Evaluating Large Language Models and Agents for Machine Learning Tasks on Repository-Level Code. *arXiv preprint arXiv:2311.09835* (2023).
[186] Minyang Tian, Luyu Gao, Shizhuo Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. 2024. Scicode: A research coding benchmark curated by scientists. *Advances in Neural Information Processing Systems* 37 (2024), 30624–30650.
[187] Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. 2024. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics* 25, 1 (2024), bbad493.
[188] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[189] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *arXiv preprint arXiv:2501.06322* (2025).
[190] Meng-Lin Tsai, Chong Wei Ong, and Cheng-Liang Chen. 2023. Exploring the use of large language models (LLMs) in chemical engineering education: Building core course problem models with Chat-GPT. *Education for Chemical Engineers* 44 (2023), 71–95.
[191] Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. 2024. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology* 19, 1 (2024), 43.
[192] Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems* 36 (2023), 38975–38987.
[193] Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. 2024. CoT Rerailer: Enhancing the Reliability of Large Language Models in Complex Reasoning Tasks through Error Detection and Correction. *arXiv preprint arXiv:2408.13940* (2024).
[194] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. 2022. NeuralCD: a general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering* 35, 8 (2022), 8312–8327.
[195] Jian Wang, Yinpei Dai, Yichi Zhang, Ziqiao Ma, Wenjie Li, and Joyce Chai. 2025. Training Turn-by-Turn Verifiers for Dialogue Tutoring Agents: The Curious Case of LLMs as Your Coding Tutors. *arXiv preprint arXiv:2502.13311* (2025).
[196] Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking and Improving Long-Text Translation with Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 7175–7187. doi:10.18653/v1/2024.findings-acl.428
[197] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (2024), 186345.
[198] Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, et al. 2024. Weaver: Foundation models for creative writing. *arXiv preprint arXiv:2401.17268* (2024).
[199] Tianfu Wang, Yi Zhan, Jianxun Lian, Zhengyu Hu, Nicholas Jing Yuan, Qi Zhang, Xing Xie, and Hui Xiong. 2025. LLM-powered Multi-agent Framework for Goal-oriented Learning in Intelligent Tutoring System. *arXiv preprint arXiv:2501.15749* (2025).
[200] Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. CMB: A Comprehensive Medical Benchmark in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6184–6205. doi:10.18653/v1/2024.naacl-long.343
[201] Xiyu Wang, Yufei Wang, Satoshi Tsutsui, Weisi Lin, Bihan Wen, and Alex Kot. 2024. Evolving storytelling: benchmarks and methods for new character customization with diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3751–3760.
[202] Yi Ru Wang, Jiafei Duan, Dieter Fox, and Siddhartha Srinivasa. 2023. NEWTON: Are large language models capable of physical reasoning? *arXiv preprint arXiv:2310.07018* (2023).
[203] Hao Wei, Jianing Qiu, Haibao Yu, and Wu Yuan. 2024. Medco: Medical education copilots based on a multi-agent framework. *arXiv preprint arXiv:2408.12496* (2024).
[204] Lilian Weng. 2023. LLM-powered Autonomous Agents. *lilianweng.github.io* (Jun 2023).
[205] Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *arXiv preprint arXiv:2405.11804* (2024).
[206] Yiran Wu. 2025. An Empirical Study on Challenging Math Problem Solving with LLM-based Conversational Agents. (2025).
[207] Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. 2023. Mathchat: Converse to tackle challenging math problems with llm agents. *arXiv preprint arXiv:2306.01337* (2023).
[208] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences* 68, 2 (2025), 121101.
[209] Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. 2024. Agentless: Demystifying llm-based software engineering agents. *arXiv preprint arXiv:2407.01489* (2024).
[210] Jian Xie, Kexun Zhang, Jiangjie Chen, Siyu Yuan, Kai Zhang, Yikai Zhang, Lei Li, and Yanghua Xiao. 2024. Revealing the barriers of language agents in planning. *arXiv preprint arXiv:2410.12409* (2024).

[211] Nan Xie, Yuelin Bai, Hengyuan Gao, Ziqiang Xue, Feiteng Fang, Qixuan Zhao, Zhijian Li, Liang Zhu, Shiwen Ni, and Min Yang. 2024. DeliLaw: A Chinese Legal Counselling System Based on a Large Language Model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 5299–5303.
[212] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Jing Hua Toh, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems* 37 (2024), 52040–52094.
[213] Wei Xiong, Chengshuai Shi, Jiaming Shen, Aviv Rosenberg, Zhen Qin, Daniele Calandriello, Misha Khalman, Rishabh Joshi, Bilal Piot, Mohammad Saleh, et al. 2024. Building math agents with multi-turn iterative preference learning. *arXiv preprint arXiv:2409.02392* (2024).
[214] Songlin Xu, Hao-Ning Wen, Hongyi Pan, Dallas Dominguez, Dongyin Hu, and Xinyu Zhang. 2025. Classroom Simulacra: Building Contextual Student Generative Agents in Online Education for Learning Behavioral Simulation. *arXiv preprint arXiv:2502.02780* (2025).
[215] Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2024. Eduagent: Generative student agents in learning. *arXiv preprint arXiv:2404.07963* (2024).
[216] Tianlong Xu, Yi-Fan Zhang, Zhendong Chu, Shen Wang, and Qingsong Wen. 2024. Ai-driven virtual teacher for enhanced educational efficiency: Leveraging large pretrain models for autonomous error analysis and correction. *arXiv preprint arXiv:2409.09403* (2024).
[217] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* 55, 1 (2024), 90–112.
[218] Yibo Yan and Joey Lee. 2024. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4163–4167.
[219] Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2024. A Survey of Mathematical Reasoning in the Era of Multimodal Large Language Model: Benchmark, Method & Challenges. *arXiv preprint arXiv:2412.11936* (2024).
[220] Yibo Yan, Shen Wang, Jiahao Huo, Xuming Hu, and Qingsong Wen. 2025. MathAgent: Leveraging a Mixture-of-Math-Agent Framework for Real-World Multimodal Mathematical Error Detection. *arXiv* (2025).
[221] Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, et al. 2024. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509* (2024).
[222] Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. 2025. Position: Multimodal Large Language Models Can Significantly Advance Scientific Reasoning. *arXiv preprint arXiv:2502.02871* (2025).
[223] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM Web Conference 2024*. 4006–4017.
[224] John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2025. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems* 37 (2025), 50528–50652.
[225] John Yang, Carlos E Jimenez, Alex L Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R Narasimhan, et al. 2024. SWE-bench Multimodal: Do AI Systems Generalize to Visual Software Domains? *arXiv preprint arXiv:2410.03859* (2024).
[226] Kaiqi Yang, Yucheng Chu, Taylor Darwin, Ahreum Han, Hang Li, Hongzhi Wen, Yasemin Copur-Gencturk, Jiliang Tang, and Hui Liu. 2024. Content knowledge identification with multi-agent large language models (LLMs). In *International Conference on Artificial Intelligence in Education*. Springer, 284–292.
[227] Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2024. MOOSE-Chem: Large Language Models for Rediscovering Unseen Chemistry Scientific Hypotheses. *arXiv preprint arXiv:2410.07076* (2024).
[228] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–35.
[229] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.
[230] Jingheng Ye, Yong Jiang, Xiaobin Wang, Yinghui Li, Yangning Li, Hai-Tao Zheng, Pengjun Xie, and Fei Huang. 2024. ProductAgent: Benchmarking Conversational Product Search Agent with Asking Clarification Questions. *arXiv preprint arXiv:2407.00942* (2024).
[231] Jingheng Ye, Yinghui Li, Yangning Li, and Hai-Tao Zheng. 2023. Mixedit: Revisiting data augmentation and beyond for grammatical error correction. *arXiv preprint arXiv:2310.11671* (2023).
[232] Jingheng Ye, Yinghui Li, Shirong Ma, Rui Xie, Wei Wu, and Hai-Tao Zheng. 2022. Focus is what you need for chinese grammatical error correction. *arXiv preprint arXiv:2210.12692* (2022).
[233] Jingheng Ye, Yinghui Li, and Haitao Zheng. 2023. System report for CCL23-eval task 7: THU KELab (sz)-exploring data augmentation and denoising for Chinese grammatical error correction. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*. 262–270.
[234] Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023. CLEME: debiasing multi-reference evaluation for grammatical error correction. *arXiv preprint arXiv:2305.10819* (2023).
[235] Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Peng Xing, Zishan Xu, Guo Cheng, and Zhao Wei. 2024. EXCGEC: A Benchmark of Edit-wise Explainable Chinese Grammatical Error Correction. *arXiv preprint arXiv:2407.00924* (2024).
[236] Jingheng Ye, Shang Qin, Yinghui Li, Hai-Tao Zheng, Shen Wang, and Qingsong Wen. 2025. Corrections Meet Explanations: A Unified Framework for Explainable Grammatical Error Correction. *arXiv preprint arXiv:2502.15261* (2025).
[237] Jingheng Ye, Shen Wang, Deqing Zou, Yibo Yan, Kun Wang, Hai-Tao Zheng, Zenglin Xu, Irwin King, Philip S Yu, and Qingsong Wen. 2025. Position: LLMs Can be Good Tutors in Foreign Language Education. *arXiv preprint arXiv:2502.05467* (2025).
[238] Jingheng Ye, Zishan Xu, Yinghui Li, Xuxin Cheng, Linlin Song, Qingyu Zhou, Hai-Tao Zheng, Ying Shen, and Xin Su. 2024. CLEME2. 0: Towards More Interpretable Evaluation by Disentangling Edits for Grammatical Error Correction. *arXiv preprint arXiv:2407.00934* (2024).
[239] Botao Yu, Frazier N Baker, Ziru Chen, Garrett Herb, Boyu Gou, Daniel Adu-Ampratwum, Xia Ning, and Huan Sun. 2024. Tooling or Not Tooling? The Impact of Tools on Language Agents for Chemistry Problem Solving. *arXiv preprint arXiv:2411.07228* (2024).
[240] Weikang Yuan, Junjie Cao, Zhuoren Jiang, Yangyang Kang, Jun Lin, Kaisong Song, Pengwei Yan, Changlong Sun, Xiaozhong Liu, et al. 2024. Can Large Language Models Grasp Legal Theories? Enhance Legal Reasoning with Insights from Multi-Agent Collaboration. *arXiv preprint arXiv:2410.02507* (2024).
[241] Osmar R Zaiane. 2002. Building a recommender agent for e-learning systems. In *International Conference on Computers in Education, 2002. Proceedings.* IEEE, 55–59.
[242] Siyu Zha, Yuehan Qiao, Qingyu Hu, Zhongsheng Li, Jiangtao Gong, and Yingqing Xu. 2024. Designing child-centric AI learning environments: Insights from LLM-enhanced creative project-based learning. *arXiv preprint arXiv:2403.16159* (2024).
[243] Xuesong Zhai, Xiaoyan Chu, Ching Sing Chai, Morris Siu Yung Jong, Andreja Istenic, Michael Spector, Jia-Bao Liu, Jing Yuan, and Yan Li. 2021. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity* 2021, 1 (2021), 8812542.
[244] Dong Zhang, Zhaowei Li, Pengyu Wang, Xin Zhang, Yaqian Zhou, and Xipeng Qiu. 2024. Speechagents: Human-communication simulation with multi-modal multi-agent systems. *arXiv preprint arXiv:2401.03945* (2024).
[245] Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. 2024. Codeagent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. *arXiv preprint arXiv:2401.07339* (2024).
[246] Mike Zhang, Amalie Pernille Dilling, Léon Gondelman, Niels Erik Ruan Lyngdorf, Euan D Lindsay, and Johannes Bjerva. 2025. SEFL: Harnessing Large Language Model Agents to Improve Educational Feedback Systems. *arXiv preprint arXiv:2502.12927* (2025).
[247] Rongsheng Zhang, Jiji Tang, Chuanqi Zang, Mingtao Pei, Wei Liang, Zeng Zhao, and Zhou Zhao. 2025. Let storytelling tell vivid stories: A multi-modal-agent-based unified storytelling framework. *Neurocomputing* 622 (2025), 129316.
[248] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* 2, 5 (2023).
[249] Yi-Fan Zhang, Hang Li, Dingjie Song, Lichao Sun, Tianlong Xu, and Qingsong Wen. 2025. From Correctness to Comprehension: AI Agents for Personalized Error Diagnosis in Education. *arXiv preprint arXiv:2502.13789* (2025).
[250] Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501* (2024).
[251] Zaibin Zhang, Yongting Zhang, Lijun Li, Hongzhi Gao, Lijun Wang, Huchuan Lu, Feng Zhao, Yu Qiao, and Jing Shao. 2024. Psysafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety. *arXiv preprint arXiv:2401.11880* (2024).
[252] Haiteng Zhao, Chang Ma, FangZhi Xu, Lingpeng Kong, and Zhi-Hong Deng. 2025. BioMaze: Benchmarking and Enhancing Large Language Models for Biological Pathway Reasoning. *arXiv preprint arXiv:2502.16660* (2025).

[253] Jiawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *arXiv preprint arXiv:2402.15061* (2024).

[254] Longwei Zheng, Fei Jiang, Xiaoqing Gu, Yuanyuan Li, Gong Wang, and Haomin Zhang. 2025. Teaching via LLM-enhanced simulations: Authenticity and barriers to suspension of disbelief. *The Internet and Higher Education* 65 (2025), 100990.

[255] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854* (2023).

[256] Xi Zhu, Yu Wang, Hang Gao, Wujiang Xu, Chen Wang, Zhiwei Liu, Kun Wang, Mingyu Jin, Linsey Pang, Qingsong Wen, et al. 2024. Recommender Systems Meet Large Language Model Agents: A Survey. *Available at SSRN 5062105* (2024).

[257] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems* 36 (2023), 50117–50143.

[258] Deqing Zou, Jingheng Ye, Yulu Liu, Yu Wu, Zishan Xu, Yinghui Li, Hai-Tao Zheng, Bingxu An, Zhao Wei, and Yong Xu. 2025. Revisiting Classification Taxonomy for Grammatical Errors. *arXiv preprint arXiv:2502.11890* (2025).

A Datasets & Benchmarks

In Table 1, we provide a comprehensive summary of publicly available datasets and benchmarks designed to evaluate LLM agents for education across various domains. It categorizes resources based on their primary goal, target users, subject domain, education level, language, modality and dataset size. We hope this collection can support and advance research on LLM agents for education.

Several datasets are designed to evaluate the pedagogical agents, such as ASSIST09 [45] and Junyi [26], which support knowledge tracing (KT) in K-12 math education, while others like EduAgent [215] facilitate adaptive learning (AL) by dynamically adjusting content based on student profiles. In addition, error correction and detection (ECD) datasets, such as Virtual Teacher [216] and Math-CCS [249], assess LLM agents' ability to identify and rectify student mistakes in math learning. Other datasets cater to writing, reading, and language learning, including FABRIC [63], EssayJudge [181], and EXCGEC [235], which focus on feedback and generation (FCG) for student essays. MultiSim [162] and Wang et al. [196] provide multi-lingual translation and storytelling benchmarks, expanding LLM capabilities beyond English-language education.

Several datasets support domain-specific educational agents across science, law, medicine, and computer science. Beyond their primary goal of evaluating pedagogical ability, these datasets assess LLM agents in domain-specific applications. They provide insights into how LLM agents can be adapted for specialized instruction, evaluating their ability to deliver subject-specific knowledge, facilitate problem-solving, and enhance interactive learning experiences across diverse educational fields. ScienceAgentBench [35] and TheoremExplainBench [102] assess scientific reasoning and theorem explanation, while ML-Bench [185] and MLAgentBench [79] focus on machine learning education. In law education, datasets like LawBench [44], LegalBench [58], and AgentCourt [29] evaluate legal knowledge application, case analysis, and court simulations. Medical education datasets, including MedBench [23] and OmniMedVQA [72], test clinical reasoning and medical knowledge retrieval. For computer science, SWE-Bench [224] and Programming Feedback [42] assess code generation, debugging, and software engineering instruction. These benchmarks help refine LLM agents for specialized tutoring, enhancing AI-driven learning in professional fields.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Table 1: Summary of existing datasets and benchmarks of LLM agents for education.

Dataset&Benchmark	Goal	User	Domain	Level	Language	Modality	Amount	Source
ASSIST09	KT	Student	Math	K12	EN	text	227k	[45]
Junyi	KT	Student	Math	K12	ZH	text	2.5M	[26]
EduAgent	AL & CS	Student	-	Graduate	EN	text & image	1,015	[215]
MathDial	AL & KT	Student	Math	K12	EN	text	45	[133]
MultiArith	AL	Student	Math	K12	EN	text	180	[215]
CoMTA	KT	Student	Math	K12	EN	text	153	[163]
MaCKT	KT	Student	Math	K12	EN	text	452	[226]
Virtual Teacher	ECD	Student	Math	K12	ZH	text & image	420	[216]
MathCCS	ECD	Student	Math	K12	ZH	text & image	420	[249]
MathTutorBench	ECD & FCG	Student	Math	K12	EN	text	7 tasks	[134]
MultiSim	-	Student	Reading	-	Multi-lingual	text	1.7M	[162]
FABRIC	FCG	Student	Writing	-	EN	text	1,782	[63]
EssayJudge	FCG	Student	Writing	-	EN	text & image	1,054	[181]
PROF	FCG	Teacher	Writing	-	EN	text	363	[145]
EXCGEC	FCG	Student	Writing	-	ZH	text	8,216	[235]
Wang et al. [196]	-	All	Translation	-	Multi-lingual	text	70K	[196]
NewEpisode	-	Student	Storytelling	-	EN	text & image	24.5K	[201]
SD-Eval	ECD	Student	Speaking	-	EN	speech	7,303	[11]
Programming Feedback	FCG	Teacher	Computer Science	-	Code	text	52	[42]
Review Critique	FCG	All	Computer Science	-	EN	text	440	[40]
AAAR-1.0	FCG	All	Computer Science	-	EN	text & image	1,000	[127]
ScienceAgentBench	-	All	General Science	-	EN	text	102	[35]
TheoremExplainBench	-	All	General Science	-	EN	text	240	[102]
MLGym-Bench	-	All	General Science	-	EN	text	13 tasks	[147]
ML-Bench	-	All	Machine Learning	-	EN	text & image	9,641	[185]
MLAgentBench	-	All	Machine Learning	-	EN	text	13 tasks	[79]
SciCode	-	All	General Science	-	EN	text	338	[186]
BLADE	-	All	General Science	-	EN	text	12 tasks	[57]
DiscoveryBench	-	All	General Science	-	EN	text	1,167	[135]
SUPER	-	All	General Science	-	EN	text	796	[20]
E-EVAL	-	All	Math & Language & General Science	K12	ZH	text	4,351	[69]
MedBench	-	All	Medical	-	ZH	text	40,041	[23]
CMB	-	All	Medical	-	ZH	text	280,839	[200]
OmniMedVQA	-	All	Medical	-	ZH	text & image	127,995	[72]
MedEval	-	All	Medical	-	EN	text & image	22,779	[66]
OSWorld	-	All	Computer Science	-	EN	text & image	369	[212]
Spider2-V	-	All	Computer Science	-	EN	text & image	812	[24]
VisualWebArena	-	All	Computer Science	-	EN	text & image	910	[96]
WebArena	-	All	Computer Science	-	EN	text & image	812	[255]
SWE-Bench	-	All	Computer Science	-	EN	text	2,294	[224]
SWE-Bench M	-	All	Computer Science	-	EN	text & image	617	[225]
Magentic-One	-	All	Computer Science	-	EN	text & image	617	[47]
Lawbench	-	All	Law	-	ZH	text	20 tasks	[44]
LegalBench	-	All	Law	-	EN	text	162 tasks	[58]
LegalAgentBench	-	All	Law	-	ZH	text	300 tasks	[109]
Agentcourt	-	All	Law	-	ZH	text	550	[29]
SimuCourt	-	All	Law	-	ZH	text	420	[65]