Conversational Intelligent Tutoring Systems: Evolution, Taxonomy, and Pedagogical Frontiers

Zhengyuan Liu[❖*], **Stella Xin Yin**[♠*], **Li Zhu**[♦], **Xiaoman Wang**[✿], **Yunshi Lan**[✿], **Nancy F. Chen**[❖]
[♠] Nanyang Technological University, Singapore
[✿] East China Normal University, [♦] University of Iowa, USA
[❖] Agency for Science, Technology and Research (A*STAR), Singapore
{liu_zhengyuan,nancy_chen}@a-star.edu.sg

Abstract

Conversational Intelligent Tutoring Systems (CITS) represent a convergence of artificial intelligence, learning sciences, and human-computer interaction, aiming to provide the adaptive and dialogic learning as expert teachers. In the past decades, the advancement of machine learning and natural language processing reshaped CITS's design, implementation, and evaluation. In this survey, we analyze the evolution of computational paradigms in shifting pedagogical affordances, interaction styles, and methodological development. We frame CITS not merely as dialogue systems, but as proactive agents that reason about learner cognition, affect, and context while enacting pedagogical strategies grounded in educational theories. From domains, populations, and instructional goals, we identify key components, evaluation approaches, and unresolved challenges. Our paper highlights the promise and limitations of conversational tutoring in the generative AI era, offering a foundation for future research that seeks to maximize their pedagogical value.

1 Introduction

Intelligent Tutoring Systems (ITS) are automated educational systems designed to provide personalized instruction by modeling learner knowledge and delivering adaptive feedback, with the ambition of replicating the adaptability and efficacy of expert human tutors. This ambition is rooted in the "2 Sigma Problem" posited by Bloom (1984), which demonstrated that students receiving one-on-one tutoring performed better than those in conventional classroom settings. While traditional ITS can help narrow this gap by providing step-by-step cognitive support in structured domains like mathematics and physics, they mainly rely on menu-based inputs, drag-and-drop interfaces, or constrained text entries. These interaction types fail to capture the dialogic nature of learning central to social constructivist theories (Palincsar, 2012; Pedaste et al., 2015; Alexander, 2018).

Conversational Intelligent Tutoring Systems (CITS) represent the technological evolution designed to overcome these limitations. By integrating Natural Language Processing (NLP) with pedagogical modeling, CITS engage learners in open-ended natural language dialogue, facilitating deeper thinking, self-explanation, and reflection that characterize effective instruction (Graesser et al., 2004; Rus et al., 2013). Beyond presenting fixed response options, CITS interpret free-form student input, diagnose misconceptions, and enact instructional strategies through conversation, which mimic adaptive behavior of a human tutor. As human-machine interaction becomes increasingly mediated through sophisticated dialogue systems, conversational tutoring has become the important frontier of educational technology research.

Along with the development of ITS, many surveys review and summarize current ITS approaches, offering broad overviews of the technological landscape and pedagogical effectiveness across diverse educational settings. However, existing surveys have not addressed CITS as a distinct research area. Prior surveys of ITS focused on educational outcomes and broad application contexts, with less attention to system architecture and interaction design (Mousavinasab et al., 2021; Kulik and Fletcher, 2016; Lin et al., 2023). These surveys predate the rise of large language models (LLMs) and therefore do not treat conversation itself as a central design dimension. While more recent work (Létourneau et al., 2025; Zerkouk et al., 2025) has reviewed AI-based tutoring systems, it tends to treat such systems as a single undifferentiated category, without tracing the architectural and pedagogical evolution that has unfolded over the past decade. Appendix A summarizes the differences between our survey and prior ITS surveys.

[*] Equal contribution.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1134–1151
November 6-10, 2026 ©2026 Association for Computational Linguistics

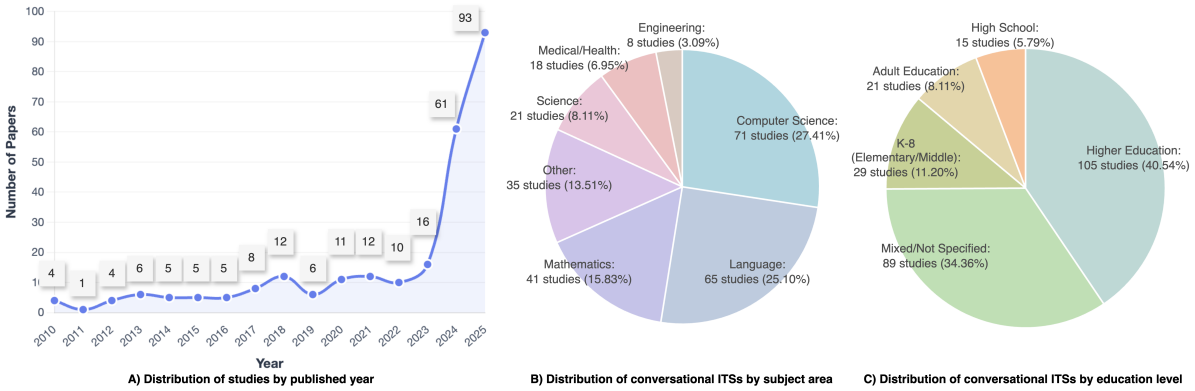


Figure 1: Distribution of the CITS related studies by published year, subject domain, and student group.

In this paper, we conduct a comprehensive review of CITS research spanning the fifteen-year period from 2010 to 2025. Through a systematic search across seven databases including ACM, IEEE, ERIC, PsycINFO, EduSource, Scopus, and Web of Science, 259 studies are included in this study (details are provided in Appendix B). Figure 1 provides an overview of these studies. The number of publication was steady before 2022, however, there has been explosive growth in recent years. Regarding the application domain, the field shows a clear preference for computer science and language education. The demographic pie chart highlights an imbalanced distribution across education levels. Many systems are designed for university students and adult learners, whereas younger students in K-12 settings are still underexplored.

Drawing on 259 papers, our key contributions are as follows. First, we provide a comprehensive review of CITS research from 2010 to 2025, tracing a paradigm shift from rule-based methods, through data-driven neural models, to LLMs, and examining roles from cognitive tutoring to affect-aware and metacognitive support. Second, we reconceptualize the ITS taxonomy for the conversational setting, arguing that dialogue restructures the functional relationships between domain knowledge, student modeling, and pedagogical strategy. Third, we provide a comparative analysis of methods within each architectural component, identifying explicit trade-offs between pedagogical control, scalability, and adaptability across paradigms. Fourth, we synthesize evaluation practices across empirical user studies and automatic metrics, illustrating gaps between current measures and effective tutoring requirements. Finally, we identify key challenges and opportunities for future research: resolving the misalignment between LLM behavior and pedagogical intent, developing long-term student models, and establishing evaluation paradigms that assess metacognitive and affective outcomes alongside cognitive gains.

2 Purpose and Role of CITS

2.1 Cognitive Support and Skill Acquisition

The primary role of CITS is to mimic human experts by scaffolding learning through natural language. Instead of delivering static content, CITS employ dialogue-based strategies to guide learners through step-by-step problem solving, and scaffolding to correct answers (Mousavinasab et al., 2021; Liu et al., 2024d). In domains like programming and mathematics, CITS serve as intelligent assistants that diagnose misconceptions and provide conceptual hints to foster active learning (Kazemitabaar et al., 2024; Daheim et al., 2024). Aside from knowledge acquisition, some systems aim to support self-regulated learning (SRL) strategies to help students monitor their own understanding and manage learning goals (Graesser and McNamara, 2010; Lintean et al., 2012).

2.2 Motivational and Affective Support

CITS increasingly serve as affective partners, with approaches ranging from sentiment analysis for identifying learner intent (Ali et al., 2025) to interfaces for empathic robotic tutoring (Bhargava et al., 2013). LLM-generated motivational feedback can support student motivation and learning (Gaeta et al., 2025), while EDEN's adaptive empathetic feedback has been linked to perceived affective support and L2 grit in a preliminary study (Siyan et al., 2024). To further personalize the experience, agents adapt to learners' learning styles (Latham et al., 2014); in EFL oral learning, per-

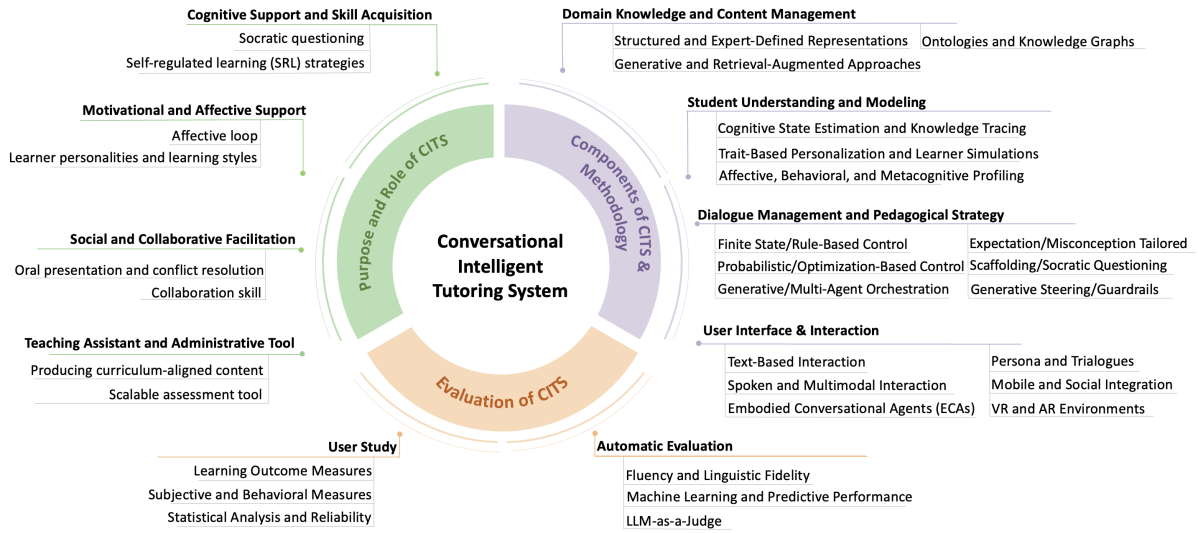


Figure 2: Overview of the systematic survey and taxonomies of conversational intelligent tutoring systems.

ceived human likeness and social presence are associated with self-efficacy and motivation (Wei et al., 2025). However, benefits vary: motivational scaffolds can enhance willingness to communicate depending on learner preferences (Ayedoun et al., 2020), whereas a trial of MathSpring found no significant overall improvements in achievement, affect, or dispositions, despite exploratory benefits among high-usage students (Feng et al., 2025).

2.3 Social and Collaborative Facilitation

In social learning contexts, CITS act as both practice partners and collaboration facilitators. For social skill practice, they are designed for oral presentation, debate training, and conflict resolution through scenario-based simulation and role-play (Sydorenko et al., 2018; Othlinghaus-Wulhorst and Hoppe, 2020; Shaikh et al., 2024; Lee et al., 2024; Dell'Aquila et al., 2022). In group settings, agents facilitate Computer-Supported Collaborative Learning by monitoring peer interactions and intervening to trigger productive dialogue (Adamson et al., 2014; Tegos et al., 2014; Cai et al., 2024). Furthermore, the "learning-by-teaching" paradigm inverts the traditional role, where the CITS perform as a teachable agent, encouraging the learner to master concepts by explaining them to the agent (Jin et al., 2024; Kucharavy et al., 2025; Liu et al., 2024b; Segedy et al., 2013).

2.4 Teachers' Assistant

CITS can assist with content creation, including generating mathematics hints (Tonga et al., 2024) and personalized reading-comprehension questions (Wang et al., 2024b). Curriculum-Driven EduBot and Book2Dial generate textbook-grounded conversational data to train educational chatbots (Li et al., 2024; Wang et al., 2024a). They also function as assessment tools or intermediaries in large courses, handling routine student inquiries to allow human instructors to focus on pedagogical instructions (Wu et al., 2025a; Hobert and Berens, 2024; Sánchez-Vera, 2025). In practice, these roles are not mutually exclusive; many CITS simultaneously support both learners and teachers (Feng et al., 2025), reflecting the integrated nature of effective tutoring.

3 Taxonomy of CITS Components

The functional taxonomy of ITS, established by Woolf (2010), decomposes systems into four interdependent modules: the Domain Model, Student Model, Tutoring Model, and Interface Model. While our taxonomy preserves this organizational continuity, CITS require a reconceptualization of each component's function. In traditional ITS, these modules operate independently: domain knowledge is statically encoded, student modeling tracks skill mastery, and the tutoring model selects from predefined instructional actions. In CITS, these boundaries tightly couple that each conversational turn simultaneously updates the learner model, triggers a pedagogical decision, and shapes how domain content is expressed. Specifically, domain knowledge must be *generatable and conversationally grounded*; student modeling must capture *communicative intent and affective state*; and

pedagogical strategy must operate as a *real-time dialogic policy*. Conversation is therefore not simply a new interface layered atop the classical taxonomy but restructures the functional relationships between all four modules. As shown in Figure 2, we thus organize these components into four categories: (1) Domain Knowledge and Content Management ("*what to teach*"); (2) Student Understanding and Modeling ("*whom to teach*"); (3) Dialogue Management and Pedagogical Strategy ("*how to teach*"); and (4) User Interface & Interaction, which defines the modes of communication. We also present a matrix table (Appendix C) to show the comparison of paradigms across three components.

3.1 Domain Knowledge and Content Management

3.1.1 Structured and Expert-Defined Representations

Early CITS architectures rely on structured, expert-defined formats that trade conversational flexibility for pedagogical reliability. The dominant paradigm pairs *curriculum scripts* with Finite State Machines (FSMs), where scripts encode ideal answers, anticipated misconceptions, and hints while FSMs govern dialogue transitions based on classified student responses. AutoTutor exemplifies this approach, using LSA to match free-text inputs against predefined expectations and triggering scripted moves accordingly (D'Mello et al., 2010; Wolfe et al., 2013; Carmon et al., 2019; Jordan et al., 2015). Semantic matching can also use Word2Vec embeddings and cosine similarity to compare learner utterances with candidate sentences (Li et al., 2020). Beyond fixed scripts, BEETLE II uses symbolic language understanding to generate context-sensitive feedback during circuit experimentation (Dzikovska et al., 2014). For procedural domains requiring deeper knowledge structures, two complementary formalisms emerge: PDDL decomposes complex scenarios, such as surgical interventions, into explicit sequences of preconditions and effects, enforcing a single correct procedural path (Vannaprathip et al., 2022); whereas constraint-based models (CBM), as in SQL-Tutor, define validity through underlying domain constraints, allowing alternative solution paths and accommodating a broader range of correct student responses (Mitrovic et al., 2013; Winder et al., 2024). These approaches still depend on expert-defined domain coverage, which limits adaptability when learners move beyond modeled concepts or interaction contexts.

3.1.2 Ontologies and Knowledge Graphs

To overcome the rigidity of script-based representations, systems adopted hierarchical and semantic knowledge structures that enable inferential reasoning beyond pattern matching. These approaches represent the conceptual relationships within a domain: DeepTutor organizes knowledge through Learning Progressions (LPs) that decompose high-level concepts into subgoals, enabling the system to scaffold learning at varying levels of granularity (Rus et al., 2014; Rus and Ştefănescu, 2016); MST employs the Unified Medical Language System (UMLS) ontology to map clinical cases to learning objectives, supporting reasoning across complex, interdependent medical concepts (Afzal et al., 2020). While these symbolic structures offer richer domain coverage and inferential capacity than FSMs, they present a similar bottleneck: knowledge graphs still require substantial expert construction effort and remain brittle outside their predefined concept boundaries. More recently, WikiHowAgent combines a hierarchical knowledge graph organizing domains, topics, and articles with embedding-based semantic similarity search to support procedural learning in a multi-LLM workflow (Pei et al., 2025).

3.1.3 Generative and Retrieval-Augmented Approaches

LLMs can reduce manual content-authoring effort by generating dynamic hints and explanations, while domain-specific validation and pedagogical design remain necessary (Tithi et al., 2025; Létourneau et al., 2025). However, this scalability introduces a new and more subtle risk: unlike structured representations where errors are predictable and bounded, LLMs may generate plausible but incorrect instructional content, particularly in high-precision STEM domains (Gupta et al., 2025). Retrieval-Augmented Generation (RAG) partially mitigates this by anchoring generation in validated educational materials (e.g., textbook segments, course notes), retrieved via graph-based approaches or vector databases before each response (Chen et al., 2025; Tao et al., 2025). This grounds responses in curriculum-aligned content, with applications evaluated in computer science (Liu et al., 2024c; Frazier et al., 2024), mathematics (Levonian et al., 2025), and natural science (Liu et al., 2025c).

Nevertheless, RAG addresses factual grounding but not logical derivation errors, leaving a gap that neither generative nor retrieval-based methods currently resolve, and pointing toward neuro-symbolic integration as a key future direction (Section 5).

3.2 Student Understanding and Modeling

3.2.1 Cognitive State Estimation and Knowledge Tracing

Early systems like AutoTutor operated episodically, using LSA and regular expressions to match learner inputs against expected answers in real-time, triggering Expectation-Misconception Tailored (EMT) dialogue moves when similarity fell below thresholds (Wolfe et al., 2013; Brust-Renck et al., 2017; Carmon et al., 2019). This approach captured immediate comprehension but was blind to knowledge evolution across interactions. Probabilistic methods addressed this temporal gap. Bayesian Knowledge Tracing (BKT) models mastery as a latent variable updated over response sequences. Separately, Hidden Markov Models (HMM) and Conditional Random Fields (CRF) have been used to model student error sequences and support adaptive feedback selection (Gutierrez and Atkinson, 2011). Deep Knowledge Tracing (DKT) learns knowledge state representations directly from interaction data; in logic-heavy domains, formal verification and graph-based models further trace multi-step reasoning errors (Vannaprathip et al., 2022; Patel et al., 2025; Zeng et al., 2025). While powerful, these methods require predefined knowledge components and struggle to interpret the rich semantic content of free-form dialogue. Neural discourse analysis offers an intermediate approach, matching relational graphs of student answers and reference solutions to identify misconceptions and generate targeted feedback (Grenander et al., 2021). LLM-based approaches resolve this by processing full conversational history and natural language explanations without explicit feature engineering, operating either as direct mastery estimators or within hybrid architectures alongside BKT and DKT for error diagnosis and forgetting curve modeling (Daheim et al., 2024; Petukhova and Kochmar, 2025; Wang et al., 2025a; Bafandkar et al., 2025; Liu et al., 2025d; Kochmar et al., 2025). LLM-based tutoring platforms also incorporate student memory and adaptive feedback to support personalization (Chudziak and Kostka, 2025; Wu et al., 2025b), although these design features alone do not establish superior personalization accuracy.

3.2.2 Affective, Behavioral and Metacognitive Profiling

Beyond knowledge state estimation, effective tutoring requires modeling the emotional and behavioral dimensions that mediate learning, as unaddressed frustration or disengagement can undermine cognitive scaffolding. These signals are captured at three levels of modality richness, each offering a different trade-off between *accessibility* and *signal fidelity*. In text-only settings, lexicon-based and transformer-based sentiment analysis infer emotional polarity and off-task states while behavioral logs (e.g., response latency, help-seeking frequency, and attempt counts) are analyzed to detect Self-Regulated Learning deficits and trigger metacognitive scaffolds (Ali et al., 2025; Li et al., 2025a; Graesser and McNamara, 2010; Munshi et al., 2023; Winder et al., 2024; Martins et al., 2024); these approaches are deployable without specialized hardware but are limited to surface-level proxies of affect. Multimodal systems achieve higher fidelity by fusing eye-tracking, facial action units, and physiological sensors to detect fine-grained states such as boredom, confusion, and flow (D'Mello et al., 2012; Bhargava et al., 2013; Graesser et al., 2017; Aslan et al., 2024); speech tutoring systems such as JustSpeak instead detect behavioral and performance difficulties, including stuckness, mispronunciation, and rhythm deviations (Zhang et al., 2021). Negotiated open learner models also let students discuss their assessed knowledge with the tutor, supporting self-assessment accuracy and reflection (Suleman et al., 2016). The central unresolved challenge is *closing the action loop*: detecting affect reliably is necessary but insufficient as systems must also determine when and how to intervene without disrupting cognitive engagement, a problem addressed further in Section 3.3.1.

3.2.3 Trait-Based Personalization and Learner Simulations

Long-term profiling captures stable learner traits such as learning styles, Big Five personality dimensions, and proficiency levels (e.g., CEFR), typically initialized via surveys and refined through interaction, enabling instruction to be tailored to persistent individual characteristics (Latham et al., 2012; Crockett et al., 2017; Liu et al., 2025a; Oliveira et al., 2021; Alrobai and Cristea, 2025; Maha-

jan, 2022; Li et al., 2024; Almasi and Kristensen-McLachlan, 2025). An emerging complement to real-learner profiling is LLM-based student simulation: persona-aware models that mimic specific knowledge gaps and error patterns enable scalable training and evaluation of tutor policies without human participants (Liu et al., 2024e; Weitekamp et al., 2025; Shi et al., 2025; Arana et al., 2025; Pei et al., 2025). However, the validity of these simulations remains uncertain: simulated learners may fail to capture the full cognitive and affective diversity of real students (Li et al., 2025c), meaning policies optimized on synthetic data risk overfitting to idealized rather than authentic learning behavior.

3.3 Dialogue Management and Pedagogical Strategies

3.3.1 Dialogue Management

As a core component, dialogue management can be formalized as a sequential decision-making process: given dialogue history H, student state S, and pedagogical goals G, the system must interpret the student's response to update S, select the next tutoring action a (e.g., hint, question, feedback), and generate an appropriate response, expressible as a policy $\varphi(H, S, G) \rightarrow a$. Three generations of policy design reflect a progressive trade-off between *curricular control* and *conversational adaptability*. FSM-based systems like AutoTutor execute fixed tutoring frames (e.g., pump $\rightarrow$ hint $\rightarrow$ prompt), guaranteeing curricular adherence but failing when learners deviate from predefined paths (D'Mello et al., 2010; Kopp et al., 2012). Probabilistic state-space approaches combine HMM and dynamic Bayesian network (DBN) inference, student-state estimation, and automated planning to select tutoring actions under uncertainty (Abuazizeh et al., 2020; Sosnowski and Yordanova, 2020), while retaining explicit models of states and actions. Reinforcement learning has also been used to learn culturally situated dialogue strategies for language learners (Abou-Khalil et al., 2018). LLM-based generative agents use conversational context to support flexible multi-turn interaction, with prompts specifying roles, rules, and interaction stages (Lee et al., 2023; Almasi and Kristensen-McLachlan, 2025; Liu et al., 2024d); however, their probabilistic nature introduces pedagogical unpredictability. Recent multi-agent frameworks partially address this by decomposing the policy into specialized "Planner", "Tutor", and "Critic" roles that separate instructional design from content generation, enforcing pedagogical safety while preserving generative flexibility (Zeng et al., 2025; Zakaria et al., 2025; Pei et al., 2025).

3.3.2 Incorporating Pedagogical Strategies

Dialogue management determines *when* to act; pedagogical strategy determines *how*, specifically, how to scaffold without revealing, challenge without discouraging, and adapt without losing curricular alignment. Rule-based systems address this through Expectation-Misconception Tailored (EMT) moves, comparing learner inputs against expected sub-answers to trigger targeted elicitation or error correction (Graesser et al., 2017; Pai et al., 2021; Schmucker et al., 2024). Neural systems shift toward student-led discovery via adaptive fading: scaffolding begins with high-level hints and progressively narrows to concrete prompts only as the learner continues to struggle, balancing guidance with cognitive effort (Rus et al., 2015; Albacete et al., 2019; Bafandkar et al., 2025). Alongside general reasoning techniques such as Chain-of-Thought prompting (Wei et al., 2022), tutoring-specific approaches use state-space planning and hierarchical question trees for Socratic debugging (Kargupta et al., 2024), or prompt- and fine-tuning to elicit reflective questions (Favero et al., 2025). Yet LLMs' default behavior, satisfying queries immediately, fundamentally conflicts with pedagogy; "pedagogical steering" techniques such as verify-then-generate workflows and facilitator framing are thus required to prevent premature answer revelation (Daheim et al., 2024; Wang et al., 2025a; Steindl et al., 2025).

Beyond cognitive strategies, effective CITS modulate interventions according to affective state, providing motivational support to struggling learners while increasing challenge for high-achievers (DMello and Graesser, 2012; Kar et al., 2025), and personality-aware models further tailor communication style to individual traits to foster rapport (Li et al., 2025a; Alrobai and Cristea, 2025). At the social level, trialogue architectures (Lippert et al., 2020) and learning-by-teaching scenarios extend strategy beyond dyadic interaction: the system orchestrates multi-party turn-taking, simulates peer behaviors such as deliberate errors, and inverts the tutor-learner role to deepen conceptual understanding through explanation (Yang et al., 2025; Kucharavy et al., 2025; Jin et al., 2024; Liu et al., 2024a).

3.4 User Interface & Interaction

Text remains the dominant modality in CITS due to its low deployment barrier and compatibility with existing learning environments. Beyond simple chat interfaces (Yang and Zapata-Rivera, 2010; Rajala et al., 2023), pedagogical effectiveness is enhanced by coupling conversation with visual scaffolds such as dynamic dashboards, progress indicators, and multimedia displays that support self-regulated learning (Latham et al., 2012; Afzal et al., 2020; Munshi et al., 2023), or by embedding text interfaces directly into task environments as integrated workflows in STEM contexts (Wambsganss et al., 2021; Winder et al., 2024). **Voice** interaction is essential for language learning and hands-free scenarios. JustSpeak synchronizes speech recognition with tutoring progress to provide configurable feedback during speech practice (Zhang et al., 2021). **Multimodal** tutors combine spoken explanations with visual science activities in MyST (Ward et al., 2013), and multilingual speech with picture description in SingaKids (Liu et al., 2025b). **Physical embodiment** can also influence initial enjoyment, although learning benefits are not assured (Ackermann et al., 2025).

CITS can also be integrated with **mobile applications**: CILA combines ChatGPT, Apple's Shortcuts, and LINE to answer learners' questions and record exchanges for self-reflection, with reported benefits for self-regulation and knowledge construction in blended learning (Wu et al., 2024). For spatial and psychomotor skills, CITS are increasingly deployed in **Virtual Reality (VR)** and **Augmented Reality (AR)**. These systems guide learners through immersive simulations, such as dental surgery or geometric modeling, often overlaying digital feedback onto physical objects (Ren et al., 2025; Ateş, 2025). Immersive environments also support multi-party spoken interaction for foreign-language learning (Divekar et al., 2022).

4 Evaluation of CITS Systems

4.1 Empirical Evaluation & User Study

User studies remain the gold standard for CITS evaluation, as they directly capture the learning outcomes and behavioral effects that automatic metrics cannot examine. Empirical studies usually adopt two types of measures: (1) **Learning performance** measures quantify instructional effectiveness through pre- and post-test comparisons, providing a model-agnostic indicator of knowledge gain widely adopted across ITS and automated feedback platforms (Hwang et al., 2024; Rus and Ştefănescu, 2016). Immediate post-tests primarily capture short-term cognitive outcomes; delayed assessments, such as the two-week vocabulary post-tests in chatbot-assisted dynamic assessment (Jeon, 2023), help examine retention, although longer-term and metacognitive outcomes require further measures.

(2) **Surveys and interaction logs** capture dimensions of the learning experience that performance scores alone cannot reflect, including perceived usefulness, cognitive load, confidence, and satisfaction (Yang and Zapata-Rivera, 2010), while behavioral log data, such as task completion rate and time-on-task, characterize engagement patterns and learning strategies (Nye et al., 2018; Brezack et al., 2025). Although these measures provide richer contextual evidence, self-report data is susceptible to social desirability bias. In WritingTutor, apparent benefits of learner control did not yield significant pairwise differences in reported experience or learning outcomes (Wambsganss et al., 2025).

4.2 Automatic Evaluation

Automatic evaluation in tutoring systems has evolved into a multi-layered framework that seeks to approximate pedagogical effectiveness through objective and reproducible metrics (Liu et al., 2024d; Maurya et al., 2025). For instance, MathTutorBench explicitly separates problem-solving ability from pedagogical quality, showing that subject expertise alone does not ensure effective tutoring (Macina et al., 2025).

Fluency and Linguistic Fidelity. Early evaluation drew on machine translation metrics, with n-gram overlap measures (BLEU, ROUGE, METEOR) quantifying lexical similarity to human references (Wang et al., 2023). For voice-based systems, Word Error Rate (WER), Character Error Rate (CER), Goodness of Pronunciation (GOP), and fundamental frequency (F_0) assess transcription and prosodic fidelity (Okur et al., 2022). Embedding-based metrics such as BERTScore subsequently enabled finer-grained semantic alignment beyond surface overlap (Li et al., 2020; Saha et al., 2025; Levonian et al., 2025). However, all linguistic fidelity metrics share a limitation that a response may score highly while still failing pedagogically if it lacks adaptive guidance or conceptual scaffolding.

Machine Learning and Predictive Performance. Diagnostic tasks such as grammatical-error detection are evaluated using accuracy, precision, recall, and F_1 (Jaganov et al., 2025). Pedagogy-driven evaluation also examines tutor-response quality (Maurya and Kochmar, 2025), with classifiers assessing mistake identification, mistake location, guidance, and actionability (Saha et al., 2025; Rahman et al., 2025). In clinical tutoring, Intersection-over-Union (IoU) measures overlap between candidate diagnoses extracted from student responses and a ground-truth candidate set (Wu et al., 2025a). Prediction of prior knowledge from dialogue features has been evaluated using correlation (r) and adjusted R^2 (Rus and Ştefănescu, 2016), while unsupervised learner profiling employs Silhouette Coefficient, AIC, and BIC (Brezack et al., 2025). Outcome- and risk-aware metrics include the Negative-to-Positive Impact Ratio (PNIR), which measures the ratio of negative to positive teaching-impact cases in simulated teacher–student interactions (Shi et al., 2025). Despite this breadth, predictive accuracy remains task-specific and does not guarantee improvements in long-term learning development.

LLM-as-a-Judge. LLM-based evaluation addresses the ceiling of metric-based approaches by approximating expert judgment through rubric-guided prompting (Li et al., 2025b). LLM-based checks range from answer verification and error classification in mathematics hint generation (Tonga et al., 2024) to grading logic hints for accuracy, consistency, clarity, and explanatory context, supplemented by human expert ratings (Tithi et al., 2025). Broader pedagogical assessments also report Win Rate (WR) and Desired Annotation Match Rate (DAMR) (Kar et al., 2025). While increasingly influential in large-scale ITS assessment, LLM judges may inherit foundation model biases and favor fluent but pedagogically shallow responses, underscoring the need for calibration against human pedagogical experts.

5 Challenges and Future Directions

5.1 From Episodic to Long-term Tutoring

Current CITS operate episodically: each session begins without memory of prior interactions, producing what we term *tutoring amnesia* (Baddeley and Wilson, 1994), where the system cannot reference a misconception surfaced two sessions prior or adapt to a learner's long-term forgetting curve. This undermines the personalization gains achieved by knowledge tracing and trait-based modeling, which construct rich learner profiles that are discarded at session end. Future work should develop persistent student models that compress dialogue history into memory pieces retrievable and editable across sessions, enabling tutors to reference past struggles and track conceptual development over semester-long interactions (Liu et al., 2025d).

5.2 Resolving the Pedagogy-Fluency Tension in LLMs

LLMs are optimized to satisfy queries immediately, yet effective tutoring requires strategically withholding answers to preserve productive failure (Kapur, 2008) (i.e., students attempt to solve complex, novel problems before receiving direct instruction). Over-constrained prompting produces repetitive interactions that limit the exploration needed for productive failure; under-constrained prompting produces unpredictable teaching behaviors that violate learning science principles. Current mitigations such as verify-then-generate workflows and pedagogical steering are partial and prompt-dependent: they can suppress direct answers but cannot reliably orchestrate the instruction that productive failure requires. A more principled resolution requires *intermediate reasoning layers* (Paul et al., 2024) that force the model to select a pedagogical move (e.g., withhold, hint, counter-example, decompose) before generating a response, so that instructional strategy guides response generation, reducing the model's default bias toward immediate, helpful-sounding answers

5.3 Evaluation Beyond Learning Gain

Three challenges emerge in evaluation aspects. First, most studies measure cognitive outcomes only, despite CITS increasingly targeting affective and metacognitive dimensions; this creates a blind spot where systems optimized for test scores may inadvertently suppress self-regulation. In an experiment on mathematics feedback, hallucinated feedback increased confusion and reduced perceived usefulness, while learning gains under fully erroneous and accurate feedback were not statistically separable (Steinbach et al., 2025). Second, the field is heavily skewed toward computer science and language learning at university level, with K-12 and domains such as social science and engineering underrepresented, limiting generalizability of current findings. Third, while LLM-as-a-Judge shows

promise for scalable pedagogical assessment, it currently lacks calibration against human expert judgment. Future benchmarks should shift from measuring *response correctness* to *pedagogical efficacy*: whether a dialogue maximized the student's cognitive effort, produced conceptual change, and supported metacognitive development.

5.4 Equity, Privacy, and Human-AI Orchestration

The reliance on cloud-based LLMs for personalization creates a tension with student data privacy; Small Language Models deployable on-device offer a privacy-preserving alternative, but currently sacrifice the pedagogical sophistication that makes LLM-based CITS effective. Besides, recent research suggests that LLM tutors may exhibit implicit algorithmic bias, offering differential levels of encouragement or scaffolding complexity based on inferred dialect or demographics (Chinta et al., 2024). Beyond data governance, the most challenging problem remains human-AI orchestration (Cukurova et al., 2025): rather than positioning AI as a replacement for human instructors, the next-generation CITS should be conceived as collaborative infrastructures in which human teachers and AI agents each operate within their respective zones of competence, with explicit mechanisms for handoff and shared oversight.

6 Conclusion

Intelligent tutoring systems have evolved from script-based instructors into dialogic agents. This evolution has altered not only system architectures, and the epistemic assumptions underlying tutoring itself. Conversation is beyond a peripheral interface layered atop instruction; it is the primary medium through which knowledge is constructed, negotiated, and regulated. Therefore, the core challenge facing the field is no longer how to enable dialogue, but how to make dialogue pedagogically accountable. Looking ahead, progress in CITS research will depend on tighter integration between learning sciences and generative AI techniques. Advances in student modeling, scaffolding planning, and multimodal learning analytics must co-evolve with models. Future systems will become proactive learning companions that can model long-term learner trajectories and regulate affect in real-time, rather than treating every session in isolation. Moreover, the field must redefine effectiveness, shifting evaluation from task completion to measuring knowledge construction and metacognitive development. Ultimately, CITS should advance a vision of education that places human agency and intellectual growth at its center.

Limitations

Several limitations should be acknowledged. First, our taxonomy, while reconceptualized for the conversational setting, focuses more on system-side design dimensions than learner-side experience, potentially underrepresenting how students actively co-construct knowledge through interaction. Second, the included studies are predominantly from computer science and language learning contexts, reflecting a limitation in content coverage and evaluation methodology. For example, domains requiring critical thinking or creative writing lack comparable evaluation frameworks. Furthermore, framing CITS evolution as a paradigm shift from rule-based to LLM-based systems risks overstating the obsolescence of earlier approaches, which retain practical advantages in high-stakes or resource-constrained settings where LLM deployment is impractical. Finally, the long-term pedagogical effects of sustained LLM tutoring on learner autonomy, metacognitive development, and potential over-reliance remain empirically open questions that the current literature, and by extension this survey, cannot yet resolve.

Ethics and Impact Statement

All authors affirm adherence to the ACL Code of Ethics. This survey synthesizes publicly published research; no new experiments involving human participants or proprietary models were conducted. We do not release any dataset or code. The intended impact of this paper is positive: to provide a structured, critical foundation for developing conversational tutoring systems that are pedagogically sound, transparent, and equitable. However, we caution that existing CITS—particularly those built on large language models—may propagate or amplify biases related to dialect, socioeconomic status, or educational resources if deployed without careful fairness auditing. Researchers and practitioners should prioritize inclusive evaluation across diverse learner populations and languages. This work is not intended for commercial surveillance, high-stakes assessment without human oversight, or any application that diminishes the role of human educators.

Information About Use Of AI Assistants

We used AI tools (e.g., ChatGPT) only for grammar checking and proofreading. No AI was used for raw content generation, literature synthesis, analysis, or insight development. All intellectual contributions, claims, and interpretations are the authors' own, and all information has been manually verified.

Acknowledgements

This research is supported by the Agency for Science, Technology and Research (A*STAR), Singapore, the Japan-Singapore Joint Call: Japan Science and Technology Agency (JST) (No. 251043539) and A*STAR 2024 (R24I6IR136), and the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-018). We thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

Victoria Abou-Khalil, Toru Ishida, Masayuki Otani, Brendan Flanagan, Hiroaki Ogata, and Donghui Lin. 2018. Learning culturally situated dialogue strategies to support language learners. *Research and Practice in Technology Enhanced Learning*, 13.

Moh'd Abuazizeh, Thomas Kirste, and Kristina Yordanova. 2020. Computational state space model for intelligent tutoring of students in nursing subjects. In *ACM International Conference Proceeding Series*, pages 324–330. Association for Computing Machinery.

Helene Ackermann, Anja Henke, Johann Chevalère, Hae Seon Yun, Verena V. Hafner, Niels Pinkwart, and Rebecca Lazarides. 2025. Physical embodiment and anthropomorphism of ai tutors and their role in student enjoyment and performance. *npj Science of Learning*, 10.

David Adamson, Gregory Dyke, Hyeju Jang, and Carolyn Penstein Rosé. 2014. Towards an agile approach to adapting dynamic collaboration support to student needs. *International Journal of Artificial Intelligence in Education*, 24:92–124.

Shazia Afzal, Tejas Indulal Dhamecha, Paul Gagnon, Akash Nayak, Ayush Shah, Jan Carlstedt-Duke, Smriti Pathak, Sneha Mondal, Akshay Gugnani, Nabil Zary, and Malolan Chetlur. 2020. Ai medical school tutor: Modelling and implementation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12299 LNAI, pages 133–145.

Patricia Albacete, Pamela Jordan, Sandra Katz, Irene-Angelica Chounta, and Bruce M. McLaren. 2019. The impact of student model updates on contingent scaffolding in a natural-language tutoring system. In *International Conference on Artificial Intelligence in Education (AIED 2019)*, pages 43–55.

Robin Alexander. 2018. Developing dialogic teaching: Genesis, process, trial. *Research papers in education*, 33(5):561–598.

Muhammad Hamza Ali, Mahbuba Afrin, Redowan Mahmud, and Aneesh Krishna. 2025. Cross-layered sentiment analysis for identifying learner intent in ai chatbots. *IEEE Transactions on Technology and Society*, pages 1–10.

Mina Almasi and Ross Deans Kristensen-McLachlan. 2025. Alignment drift in CEFR-prompted LLMs for interactive Spanish tutoring. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 70–88.

Amani Alrobai and Alexandra I. Cristea. 2025. *Personality-Aware Conversational Intelligent Tutoring System with GenAI: Studying the Effect on Learners in Introductory Programming*, pages 322–327. Springer Nature Switzerland, Cham.

Jasper Meynard Arana, Kristine Ann M. Carandang, Ethan Robert Casin, Christian Alis, Daniel Stanley Tan, Erika Fille Legara, and Christopher Monterola. 2025. Foundations of peers: Assessing llm role performance in educational simulations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, volume 4, pages 908–918.

Sinem Aslan, Lenitra M. Durham, Nese Alyuz, Rebecca Chierichetti, Pete A. Denman, Eda Okur, David I.Gonzalez Aguirre, Julio C.Zamora Esquivel, Hector A. Cordourier Maruri, Sangita Sharma, Giuseppe Raffa, Richard E. Mayer, and Lama Nachman. 2024. What is the impact of a multi-modal pedagogical conversational ai system on parents' concerns about technology use by young children? *British Journal of Educational Technology*, 55:1625–1650.

Hüseyin Ateş. 2025. Integrating augmented reality into intelligent tutoring systems to enhance science education outcomes. *Education and Information Technologies*, 30:4435–4470.

Emmanuel Ayedoun, Yuki Hayashi, and Kazuhisa Seta. 2020. Toward personalized scaffolding and fading of motivational support in l2 learner-dialogue agent interactions: An exploratory study. *IEEE Transactions on Learning Technologies*, 13:604–616.

Alan Baddeley and Barbara A Wilson. 1994. When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia*, 32(1):53–68.

Shayan Bafandkar, Sungyong Chung, Homa Khosravian, and Alireza Talebpour. 2025. Pappl: Personalized ai-powered progressive learning platform. *arXiv preprint arXiv:2508.14109*.

Shweta Bhargava, Srinivasan Janarthanam, Helen Hastie, Amol Deshmukh, Ruth Aylett, Lee Corrigan, and Ginevra Castellano. 2013. Demonstration of the emote wizard of oz interface for empathic robotic tutors. In *SIGDIAL 2013 Conference*, pages 22–24. Association for Computational Linguistics.

Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, 13(6):4–16.

Natalie Brezack, Melissa Lee, Kelly Collins, Wynnie Chan, and Mingyu Feng. 2025. Uncovering student profiles with problem solving and effort data from a tutoring system: Insights from hierarchical cluster heatmap and latent profile analyses. In *L@S 2025 - Proceedings of the 12th ACM Conference on Learning @ Scale*, pages 336–340.

Priscila G. Brust-Renck, Valerie F. Reyna, Evan A. Wilhelms, Christopher R. Wolfe, Colin L. Widmer, Elizabeth M. Cedillos-Whynott, and A. Kate Morant. 2017. Active engagement in a web-based tutorial to prevent obesity grounded in fuzzy-trace theory predicts higher knowledge and gist comprehension. *Behavior Research Methods*, 49:1386–1398.

Zhenyao Cai, Seehee Park, Nia Nixon, and Shayan Doroudi. 2024. Advancing knowledge together: Integrating large language model-based conversational ai in small group collaborative learning. In *Conference on Human Factors in Computing Systems - Proceedings*.

Colin M. Carmon, Andrew J. Hampton, Brent Morgan, Zhiqiang Cai, Lijia Wang, and Arthur C. Graesser. 2019. Semantic matching evaluation of user responses to electronics questions in autotutor. In *Proceedings of the 6th 2019 ACM Conference on Learning at Scale, L@S 2019*.

Penghe Chen, Zhilin Fan, and Yu Lu. 2025. Knowstu: Diagnosing students' problem behaviors using fine-tuned llm and rag. *IEEE Transactions on Learning Technologies*, 18:962–975.

Sribala Vidyadhari Chinta, Zichong Wang, Zhipeng Yin, Nhat Hoang, Matthew Gonzalez, T Le Quy, and Wenbin Zhang. 2024. Fairaied: Navigating fairness, bias, and ethics in educational ai applications. *arXiv preprint arXiv:2407.18745*.

Jarosław A. Chudziak and Adam Kostka. 2025. Ai-powered math tutoring: Platform for personalized and adaptive education. *arXiv preprint arXiv:2507.12484*.

Keeley Crockett, Annabel Latham, and Nicola Whitton. 2017. On predicting learning styles in conversational intelligent tutoring systems using fuzzy decision trees. *International Journal of Human Computer Studies*, 97:98–115.

Mutlu Cukurova, Wannapon Suraworachet, Qi Zhou, and Sahan Bulathwela. 2025. Towards synergistic teacher-ai interactions with generative artificial intelligence. *Preprint*, arXiv:2511.19580.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise verification and remediation of student reasoning errors with large language model tutors. *arXiv preprint arXiv:2407.09136*.

Elena Dell'Aquila, Federica Vallone, Maria Clelia Zurlo, and Davide Marocco. 2022. Sg-accord: Designing virtual agents for soft skills training in the school context. *Education Sciences*, 12.

Rahul R. Divekar, Jaimie Drozdal, Samuel Chabot, Yalun Zhou, Hui Su, Yue Chen, Houming Zhu, James A. Hendler, and Jonas Braasch. 2022. Foreign language acquisition via artificial intelligence and extended reality: design and evaluation. *Computer Assisted Language Learning*, 35:2332–2360.

Sidney DMello and Art Graesser. 2012. Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems*, 2.

Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human Computer Studies*, 70:377–398.

Sidney K. D'Mello, Art Graesser, and Brandon King. 2010. Toward spoken human-computer tutorial dialogues. *Human-Computer Interaction*, 25:289–323.

Myroslava Dzikovska, Natalie Steinhauser, Elaine Farrow, Johanna Moore, and Gwendolyn Campbell. 2014. Beetle ii: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics. *International Journal of Artificial Intelligence in Education*, 24:284–332.

Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2025. Enhancing critical thinking in education by means of a socratic chatbot. In *Communications in Computer and Information Science*, volume 2519 CCIS, pages 17–32.

Mingyu Feng, Natalie Brezack, Chunwei Huang, Yvonne Kao, Kelly Collins, Melissa Lee, Megan Schneider, and Wynnie Chan. 2025. Evaluating the efficacy of an intelligent tutoring system that integrates affective supports into math learning. *Journal of Computer Assisted Learning*, 41.

Matthew Frazier, Kostadin Damevski, and Lori Pollock. 2024. Customizing chatgpt to help computer science principles students learn through conversation. In *Annual Conference on Innovation and Technology*

in Computer Science Education, ITiCSE, volume 1, pages 633–639. Association for Computing Machinery.

Angelo Gaeta, Francesco Orciuoli, Antonella Pascuzzo, and Angela Peduto. 2025. Enhancing traditional its architectures with large language models for generating motivational feedback. *Computers and Education: Artificial Intelligence*, 9.

Arthur Graesser and Danielle McNamara. 2010. Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist*, 45:234–244.

Arthur C. Graesser, Zhiqiang Cai, Brent Morgan, and Lijia Wang. 2017. Assessment with computer agents that engage in conversational dialogues and trialogues with learners. *Computers in Human Behavior*, 76:607–616.

Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M Louwerse. 2004. Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36(2):180–192.

Matt Grenander, Robert Belfer, Ekaterina Kochmar, Iulian V. Serban, François St-Hilaire, and Jackie C. K. Cheung. 2021. Deep discourse analysis for generating personalized feedback in intelligent tutor systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:15534–15544.

Adit Gupta, Jennifer Reddig, Tommaso Calo, Daniel Weitekamp, and Christopher J. MacLellan. 2025. Beyond final answers: Evaluating large language models for math tutoring.

Fernando Gutierrez and John Atkinson. 2011. Adaptive feedback selection for intelligent tutoring systems. *Expert Systems with Applications*, 38:6146–6152.

Sebastian Hobert and Florian Berens. 2024. Developing a digital tutor as an intermediary between students, teaching assistants, and lecturers. *Educational Technology Research and Development*, 72:797–818.

Wu Yuin Hwang, Bo Chen Guo, Anh Hoang, Ching Chun Chang, and Nien Tsu Wu. 2024. Facilitating authentic contextual efl speaking and conversation with smart mechanisms and investigating its influence on learning achievements. *Computer Assisted Language Learning*, 37:1632–1658.

Timur Jaganov, John Blake, Julián Villegas, and Nicholas Carr. 2025. Large language model-driven dynamic assessment of grammatical accuracy in english language learner writing. *IEEE Access*, 13:151538–151550.

Jaeho Jeon. 2023. Chatbot-assisted dynamic assessment (ca-da) for l2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, 36:1338–1364.

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 *Conference on Human Factors in Computing Systems - Proceedings*.

Pamela Jordan, Patricia Albacete, and Sandra Katz. 2015. Exploring the effects of redundancy within a tutorial dialogue system: Restating students' responses. In *SIGDIAL 2015 Conference*, pages 51–59. Association for Computational Linguistics.

Manu Kapur. 2008. Productive failure. *Cognition and instruction*, 26(3):379–424.

Debanjana Kar, Leopold Böss, Dacia Braca, Sebastian Maximilian Dennerlein, Nina Christine Hubig, Philipp Wintersberger, and Yufang Hou. 2025. Mathbuddy: A multimodal system for affective math tutoring. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 821–838. Association for Computational Linguistics.

Priyanka Kargupta, Ishika Agarwal, Dilek Hakkani Tur, and Jiawei Han. 2024. Instruct, not assist: Llm-based multi-turn planning and hierarchical questioning for socratic code debugging. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9475–9495. Association for Computational Linguistics.

Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Z. Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Conference on Human Factors in Computing Systems - Proceedings*.

Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the bea 2025 shared task on pedagogical ability assessment of ai-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033. Association for Computational Linguistics.

Kristopher J. Kopp, M. Anne Britt, Keith Millis, and Arthur C. Graesser. 2012. Improving the efficiency of dialogue in tutoring. *Learning and Instruction*, 22:320–330.

Andrei Kucharavy, Cyril Vallez, and Dimitri Percia David. 2025. Llms protégés: Tutoring llms with knowledge gaps improves student learning outcome. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 248–257. Association for Computational Linguistics.

James A Kulik and John D Fletcher. 2016. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research*, 86(1):42–78.

Annabel Latham, Keeley Crockett, and David McLean. 2014. An adaptation algorithm for an intelligent natural language tutoring system. *Computers and Education*, 71:97–110.

Annabel Latham, Keeley Crockett, David McLean, and Bruce Edmonds. 2012. A conversational intelligent tutoring system to automatically predict learning styles. In *Computers and Education*, volume 59, pages 95–109.

Eun Young Lee, Ngagaba Gogo Dae Il, Gi Hong An, Sungchul Lee, and Kiho Lim. 2023. Chatgpt-based debate game application utilizing prompt engineering. In *2023 Research in Adaptive and Convergent Systems RACS 2023*.

Unggi Lee, Yeil Jeong, Junbo Koh, Gyuri Byun, Yunseo Lee, Youngsun Hwang, Hyeoncheol Kim, and Cheolil Lim. 2024. Can ChatGPT be a debate partner? Developing ChatGPT-based application 'DEBO' for debate education, findings and limitations. *Educational Technology & Society*, 27(2):321–346.

Zachary Levonian, Owen Henkel, Chenglu Li, and Millie-Ellen Postle. 2025. Designing safe and relevant generative chats for math learning in intelligent tu-toring systems. *Journal of Educational Data Mining*, 17:66–97.

Hai Li, Wanli Xing, Chenglu Li, Wangda Zhu, Bailing Lyu, Fan Zhang, and Zifeng Liu. 2025a. Who should be my tutor? analyzing the interactive effects of automated text personality styles between middle school students and a mathematics chatbot. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pages 910–917. ACM.

Kuo Chen Li, Maiga Chang, and Kuan Hsing Wu. 2020. Developing a task-based dialogue system for english language learning. *Education Sciences*, 10:1–20.

Minzhi Li, Zhengyuan Liu, Shumin Deng, Shafiq Joty, Nancy Chen, and Min-Yen Kan. 2025b. DnA-eval: Enhancing large language model evaluation through decomposition and aggregation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2277–2290.

Ruijia Li, Yuan-Hao Jiang, Jiatong Wang, and Bo Jiang. 2025c. How real is ai tutoring? comparing simulated and human dialogues in one-on-one instruction. *arXiv preprint arXiv:2509.01914*.

Yu Li, Shang Qu, Jili Shen, Shangchao Min, and Zhou Yu. 2024. Curriculum-driven edubot: A framework for developing language learning chatbots through synthesizing conversational data. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 400–419.

Ching-Chung Lin, Anna Y. Q. Huang, and Owen H. T. Lu. 2023. Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments*, 10(1):41.

Mihai Lintean, Vasile Rus, Zhiqiang Cai, Amy Witherspoon-Johnson, Arthur C. Graesser, and Roger Azevedo. 2012. *Computational Aspects of the Intelligent Tutoring System MetaTutor*, pages 247–260. IGI Global.

Anne Lippert, Keith Shubeck, Brent Morgan, Andrew Hampton, and Arthur Graesser. 2020. Multiple agent designs in conversational intelligent tutoring systems. *Technology, Knowledge and Learning*, 25:443–463.

Ben Liu, Jihai Zhang, Fangquan Lin, Xu Jia, and Min Peng. 2025a. One size doesn't fit all: A personalized conversational tutoring agent for mathematics instruction. In *WWW Companion 2025 - Companion Proceedings of the ACM Web Conference 2025*, pages 2401–2409.

Chen Chung Liu, Wan Jun Chen, Fang Ying Lo, Chia Hui Chang, and Hung Ming Lin. 2024a. Teachable q&a agent: The effect of chatbot training by students on reading interest and engagement. *Journal of Educational Computing Research*, 62:1122–1154.

Chen-Chung Liu, Wan-Jun Chen, Fang ying Lo, Chia-Hui Chang, and Hung-Ming Lin. 2024b. Teachable q&a agent: The effect of chatbot training by students on reading interest and engagement. *Journal of Educational Computing Research*, 62(4):902–934.

Suqing Liu, Zezhu Yu, Feiran Huang, Yousef Bulbulia, Andreas Bergen, and Michael Liut. 2024c. Can small language models with retrieval-augmented generation replace large language models when learning computer science? In *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, volume 1, pages 388–393.

Zhengyuan Liu, Geyu Lin, Hui Li Tan, Huayun Zhang, Yanfeng Lu, Xiaoxue Gao, Stella Xin Yin, Sun He, Hock Huan Goh, Lung Hsiang Wong, and Nancy F. Chen. 2025b. Singakids: A multilingual multimodal dialogic tutor for language learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, volume 6, pages 1244–1253. Association for Computational Linguistics.

Zhengyuan Liu, Stella Xin Yin, Dion Hoe-Lian Goh, and Nancy Chen. 2025c. COGENT: A curriculum-oriented framework for generating grade-appropriate educational content. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 129–143.

Zhengyuan Liu, Stella Xin Yin, Carolyn Lee, and Nancy F. Chen. 2024d. Scaffolding language learning via multi-modal tutoring systems with pedagogical instructions. In *Proceedings - 2024 IEEE Conference on Artificial Intelligence, CAI 2024*, pages 1258–1265.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024e. Personality-aware student simulation for conversational intelligent tutoring systems. In *Proceedings of the 2024 Conference on Empirical*

Methods in Natural Language Processing, pages 626–642.

Zhengyuan Liu, Stella Xin Yin, Bryan Chen Zhengyu Tan, Roy Ka-Wei Lee, Guimei Liu, Dion Hoe-Lian Goh, Wenya Wang, and Nancy F Chen. 2025d. The imperfect learner: Incorporating developmental trajectories in memory-based student simulation. *arXiv preprint arXiv:2511.05903*.

Angélique Létourneau, Marion Deslandes Martineau, Patrick Charland, John Alexander Karran, Jared Boasen, and Pierre-Majorique Léger. 2025. A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning*, 10(1):29.

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.

Muskan Mahajan. 2022. Bela: Bot for english language acquisition. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 142–148. Association for Computational Linguistics.

Ludmila Martins, Maite Fernández-Ferrer, and Eloi Puertas. 2024. Analysing conversation pathways with a chatbot tutor to enhance self-regulation in higher education. *Education Sciences*, 14.

Kaushal Kumar Maurya and Ekaterina Kochmar. 2025. Pedagogy-driven evaluation of generative ai-powered intelligent tutoring systems. *arXiv preprint arXiv:2510.22581*.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, volume 1, pages 1234–1251.

Antonija Mitrovic, Stellan Ohlsson, and Devon K. Barrow. 2013. The effect of positive feedback in a constraint-based intelligent tutoring system. *Computers and Education*, 60:264–272.

Elham Mousavinasab, Nahid Zarifsanaiey, Sharareh R. Niakan Kalhori, Mahnaz Rakhshan, Leila Keikha, and Marjan Ghazi Saeedi. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1):142–163.

Anabil Munshi, Gautam Biswas, Ryan Baker, Jaclyn Ocumpaugh, Stephen Hutt, and Luc Paquette. 2023. Analysing adaptive scaffolds that help students develop self-regulated learning behaviours. *Journal of Computer Assisted Learning*, 39:351–368.

Benjamin D. Nye, Philip I. Pavlik, Alistair Windsor, Andrew M. Olney, Mustafa Hajeer, and Xiangen Hu. 2018. Skope-it (shareable knowledge objects as portable intelligent tutors): overlaying natural language tutoring on an adaptive learning system for mathematics. *International Journal of STEM Education*, 5.

Eda Okur, Saurav Sahay, Roddy Fuentes Alba, and Lama Nachman. 2022. End-to-end evaluation of a spoken dialogue system for learning basic mathematics. *arXiv preprint arXiv:2211.03511*.

Eduardo Oliveira, Paula Galvao de Barba, and Linda Corrin. 2021. Enabling adaptive, personalised and context-aware interaction in a smart learning environment: Piloting the icollab system. *Australasian Journal of Educational Technology*, 37:1–23.

Julia Othlinghaus-Wulhorst and H. Ulrich Hoppe. 2020. A technical and conceptual framework for serious role-playing games in the area of social skill training. *Frontiers in Computer Science*, 2.

Kai Chih Pai, Bor Chen Kuo, Chen Huei Liao, and Yin Mei Liu. 2021. An application of chinese dialogue-based intelligent tutoring system in remedial instruction for mathematics learning. *Educational Psychology*, 41:137–152.

Annemarie Sullivan Palincsar. 2012. Social constructivist perspectives on teaching and learning. *An introduction to Vygotsky*, pages 290–319.

Manooshree Patel, Rayna Bhattacharyya, Thomas Lu, Arnav Mehta, Niels Voss, Narges Norouzi, and Gireeja Ranade. 2025. Leantutor: A formally-verified ai tutor for mathematical proofs. *arXiv preprint arXiv:2506.08321*.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2024. Refiner: Reasoning feedback on intermediate representations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1100–1126.

Margus Pedaste, Mario Mäeots, Leo A Siiman, Ton De Jong, Siswa AN Van Riesen, Ellen T Kamp, Constantinos C Manoli, Zacharias C Zacharia, and Eleftheria Tsourlidaki. 2015. Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational research review*, 14:47–61.

Jiahuan Pei, Fanghua Ye, Xin Sun, Wentao Deng, Koen Hindriks, and Junxiao Wang. 2025. Conversational education at scale: A multi-llm agent workflow for procedural learning and pedagogic quality assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2984–2997.

Kseniia Petukhova and Ekaterina Kochmar. 2025. Intent matters: Enhancing ai tutoring with fine-grained pedagogical intent annotation. *arXiv preprint arXiv:2506.07626*.

Md. Abdur Rahman, Md Al Amin, Sabik Aftahee, Muhammad Junayed, and Md Ashiqur Rahman. 2025. SmolLab_SEU at BEA 2025 shared task: A transformer-based framework for multi-track pedagogical evaluation of AI-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1127–1134.

Jaakko Rajala, Jenni Hukkanen, Maria Hartikainen, and Pia Niemelä. 2023. "call me kiran" chatgpt as a tutoring chatbot in a computer science course. In *ACM International Conference Proceeding Series*, pages 83–94.

Haocheng Ren, Muzhe Wu, Gregory Thomas Croisdale, Anhong Guo, and Xu Wang. 2025. Rubikon: Intelligent tutoring for rubik's cube learning through ar-enabled physical task reconfiguration. In *DIS 2025 - Proceedings of the 2025 ACM Designing Interactive Systems Conference*, pages 3549–3562.

Vasile Rus, Sidney D'Mello, Xiangen Hu, and Arthur Graesser. 2013. Recent advances in conversational intelligent tutoring systems. *AI magazine*, 34(3):42–54.

Vasile Rus, Nobal Niraula, and Rajendra Banjade. 2015. Deeptutor: An effective, online intelligent tutoring system that promotes deep learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29.

Vasile Rus, Dan Stefanescu, Nobal Niraula, and Arthur C. Graesser. 2014. Deeptutor: Towards macro- and micro-adaptive conversational intelligent tutoring at scale. In *L@S 2014 - Proceedings of the 1st ACM Conference on Learning at Scale*, pages 209–210. Association for Computing Machinery.

Vasile Rus and Dan Ştefănescu. 2016. Non-intrusive assessment of learners' prior knowledge in dialogue-based intelligent tutoring systems. *Smart Learning Environments*, 3.

Trishita Saha, Shrenik Ganguli, and Maunendra Sankar Desarkar. 2025. Nlip at bea 2025 shared task: Evaluation of pedagogical ability of ai tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1242–1253. Association for Computational Linguistics.

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. Ruffle&riley: From lesson text to conversational tutoring. In *L@S 2024 - Proceedings of the 11th ACM Conference on Learning @ Scale*, pages 547–549. Association for Computing Machinery, Inc.

James R. Segedy, John S. Kinnebrew, and Gautam Biswas. 2013. The effect of contextualized conversational feedback in a complex open-ended learning environment. *Educational Technology Research and Development*, 61:71–89.

Omar Shaikh, Valentino Chai, Michele J. Gelfand, Diyi Yang, and Michael S. Bernstein. 2024. Rehearsal: Simulating conflict to teach conflict resolution. In *Conference on Human Factors in Computing Systems - Proceedings*.

Yao Shi, Rongkeng Liang, and Yong Xu. 2025. Educationq: Evaluating llms' teaching capabilities through multi-agent dialogue framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32799–32828.

Li Siyan, Teresa Shao, Zhou Yu, and Julia Hirschberg. 2024. Eden: Empathetic dialogues for english learning. *arXiv preprint arXiv:2406.17982*.

Tomasz Sosnowski and Kristina Yordanova. 2020. A probabilistic conversational agent for intelligent tutoring systems. In *ACM International Conference Proceeding Series*, pages 317–323. Association for Computing Machinery.

Marlene Steinbach, Shreya Bhandari, Jennifer Meyer, and Zachary A. Pardos. 2025. When llms hallucinate: Examining the effects of erroneous feedback in math tutoring systems. In *L@S 2025 - Proceedings of the 12th ACM Conference on Learning @ Scale*, pages 139–150.

Sebastian Steindl, Fabian Brunner, Nada Sissouno, Dominik Schwagerl, Florian Schöler-Niewiera, and Ulrich Schäfer. 2025. On the effectiveness of prompt-moderated llms for math tutoring at the tertiary level. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11310–11323.

Raja M. Suleman, Riichiro Mizoguchi, and Mitsuru Ikeda. 2016. A new perspective of negotiation-based dialog to enhance metacognitive skills in the context of open learner models. *International Journal of Artificial Intelligence in Education*, 26:1069–1115.

Tetyana Sydorenko, Phoebe Daurio, and Steven L. Thorne. 2018. Refining pragmatically-appropriate oral communication via computer-simulated conversations. *Computer Assisted Language Learning*, 31:157–180.

Fulgencio Sánchez-Vera. 2025. Subject-specialized chatbot in higher education as a tutor for autonomous exam preparation: Analysis of the impact on academic performance and students' perception of its usefulness. *Education Sciences*, 15.

Mingxu Tao, Bowen Tang, Mingxuan Ma, Yining Zhang, Hourun Li, Feifan Wen, Ma Hao, and Jia Yang. 2025. Aloha: Empowering multilingual agent for university orientation with hierarchical retrieval. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 383–392.

Stergios Tegos, Stavros Demetriadis, and Thrasyvoulos Tsiatsos. 2014. A configurable conversational agent

to trigger students' productive dialogue: A pilot study in the call domain. *International Journal of Artificial Intelligence in Education*, 24:62–91.

Sutapa Dey Tithi, Arun Kumar Ramesh, Clara DiMarco, Xiaoyi Tian, Nazia Alam, Kimia Fazeli, and Tiffany Barnes. 2025. The promise and limits of llms in constructing proofs and hints for logic problems in intelligent tutoring systems. *Computers and Education: Artificial Intelligence*, 9.

Junior Cedric Tonga, Benjamin Clement, and Pierre-Yves Oudeyer. 2024. Automatic generation of question hints for mathematics problems using large language models in educational technology. *arXiv preprint arXiv:2411.03495*.

Narumol Vannaprathip, Peter Haddawy, Holger Schultheis, and Siriwan Suebnukarn. 2022. Intelligent tutoring for surgical decision making: a planning-based approach. *International Journal of Artificial Intelligence in Education*, 32:350–381.

Thiemo Wambsganss, Ivo Benke, Alexander Maedche, Kenneth Koedinger, and Tanja Käser. 2025. Evaluating the impact of learner control and interactivity in conversational tutoring systems for persuasive writing. *International Journal of Artificial Intelligence in Education*, 35:791–822.

Thiemo Wambsganss, Tobias Kung, Matthias Sollner, and Jan Marco Leimeister. 2021. Arguetutor: An adaptive dialog-based learning system for argumentation skills. In *Conference on Human Factors in Computing Systems - Proceedings*.

Jian Wang, Yinpei Dai, Yichi Zhang, Ziqiao Ma, Wenjie Li, and Joyce Chai. 2025a. Training turn-by-turn verifiers for dialogue tutoring agents: The curious case of llms as your coding tutors. *arXiv preprint arXiv:2502.13311*.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024a. Book2dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9707–9731. Association for Computational Linguistics.

Lingzhi Wang, Mrinmaya Sachan, Xingshan Zeng, and Kam-Fai Wong. 2023. Strategize before teaching: A conversational tutoring system with pedagogy self-distillation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2268–2274.

Tianfu Wang, Yi Zhan, Jianxun Lian, Zhengyu Hu, Nicholas Jing Yuan, Qi Zhang, Xing Xie, and Hui Xiong. 2025b. Llm-powered multi-agent framework for goal-oriented learning in intelligent tutoring system. In *WWW Companion 2025 - Companion Proceedings of the ACM Web Conference 2025*, pages 510–519. Association for Computing Machinery, Inc.

Xizhe Wang, Yihua Zhong, Changqin Huang, and Xiaodi Huang. 2024b. Chatprcs: A personalized support system for english reading comprehension based on chatgpt. *IEEE Transactions on Learning Technologies*, 17:1762–1776.

Wayne Ward, Ron Cole, Daniel Bolaños, Cindy Buchenroth-Martin, Edward Svirsky, and Tim Weston. 2013. My science tutor: A conversational multimedia virtual tutor. *Journal of Educational Psychology*, 105:1115–1125.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wenbo Wei, An Zhao, and Hongning Ma. 2025. Understanding how ai chatbots influence efl learners' oral english learning motivation and outcomes: Evidence from chinese learners. *IEEE Access*, 13:56699–56716.

Daniel Weitekamp, Momin N. Siddiqui, and Christopher J. MacLellan. 2025. Tutorgym: A testbed for evaluating ai agents as tutors and students. *arXiv preprint arXiv:2505.01563*.

Jaxton Winder, Erik Falor, Seth Poulsen, and John Edwards. 2024. The shell tutor: An intelligent tutoring system for the unix command shell and git. In *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE*, volume 1, pages 548–554. Association for Computing Machinery.

Christopher R. Wolfe, Colin L. Widmer, Valerie F. Reyna, Xiangen Hu, Elizabeth M. Cedillos, Christopher R. Fisher, Priscilla G. Brust-Renck, Triana C. Williams, Isabella Damas Vannucchi, and Audrey M. Weil. 2013. The development and analysis of tutorial dialogues in autotutor lite. *Behavior Research Methods*, 45:623–636.

Beverly Park Woolf. 2010. *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.

Qian Wu, Zheyao Gao, Longfei Gou, and Qi Dou. 2025a. Ddxtutor: Clinical reasoning tutoring system with differential diagnosis-based structured reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 30934–30957. Association for Computational Linguistics.

Ting Ting Wu, Hsin Yu Lee, Pin Hui Li, Chia Nan Huang, and Yueh Min Huang. 2024. Promoting self-regulation progress and knowledge construction in blended learning via chatgpt-based learning aid. *Journal of Educational Computing Research*, 61:3–31.

Yang Wu, Rujing Yao, Tong Zhang, Yufei Shi, Zhuoren Jiang, Zhushan Li, and Xiaozhong Liu.

2025b. Teaching according to students' aptitude: Personalized mathematics tutoring via persona-, memory-, and forgetting-aware llms. *arXiv preprint arXiv:2511.15163*.

Hui Chun Yang and Diego Zapata-Rivera. 2010. Interlanguage pragmatics with a pedagogical agent: The request game. *Computer Assisted Language Learning*, 23:395–412.

Xinming Yang, Haasil Pujara, and Jun Li. 2025. Learning by teaching: Engaging students as instructors of large language models in computer science education. *arXiv preprint arXiv:2508.05979*.

Rida Zakaria, Boukachour Hadhoum, Ennaji Mourad, and Machkour Mustapha. 2025. Dynamic adaptation in an intelligent multi-tutoring system: A multi-agent approach. *IEEE Access*, 13:124587–124601.

Biqing Zeng, Mengquan Liu, and Zongwei Zhen. 2025. Graphmasal: A graph-based multi-agent system for adaptive learning. *arXiv preprint arXiv:2511.11035*.

Meriem Zerkouk, Miloud Mihoubi, and Belkacem Chikhaoui. 2025. A comprehensive review of ai-based intelligent tutoring systems: Applications and challenges. *Preprint*, arXiv:2507.18882.

Xinlei Zhang, Takashi Miyaki, and Jun Rekimoto. 2021. Justspeak: Automated, user-configurable, interactive agents for speech tutoring. *Proceedings of the ACM on Human-Computer Interaction*, 5.

A Comparison of Prior Surveys

Table 1 compares the present survey with prior reviews of intelligent tutoring systems in terms of temporal coverage, number of studies, scope, and analytical dimensions. The comparison indicates differences in the extent to which these surveys address conversational design, system architecture, interaction design, LLM-era developments, and evaluation methodology.

B Survey Methodology

We conducted a systematic review following PRISMA guidelines. In June 2025, we searched seven databases (ACM Digital Library, IEEE Xplore, ERIC, PsycINFO, EduSource, Scopus, and Web of Science), using the Boolean query: ["intelligent tutoring system*" OR "ITS"] AND ["conversation*" OR "dialog*" OR "chat*" OR "interact*"] AND ["teach*" OR "instruction*" OR "pedag*" OR "learn*"], applied to titles, abstracts, and keywords. The search covered peer-reviewed publications and preprints written in English and published between January 2010 and June 2025. The initial search returned 9,190 records; after removing 4,378 duplicates, 4,812 records proceeded to

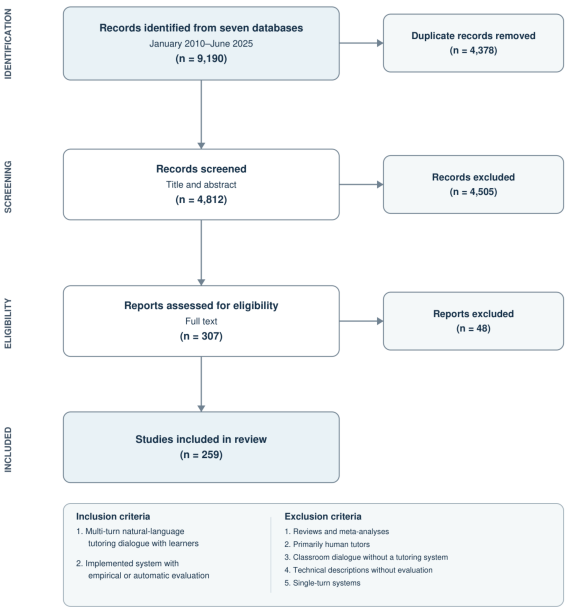


Figure 3: PRISMA Diagram

title-and-abstract screening. Studies were included if they (a) presented a tutoring system that interacts with learners through multi-turn natural language dialogue, and (b) reported an implementation together with an empirical or automatic evaluation. We excluded: (1) systematic reviews, literature reviews, and meta-analyses; (2) papers focusing primarily on human tutors; (3) papers describing classroom dialogue without a tutoring system; (4) technical descriptions without any evaluation; and (5) systems supporting only single-turn interaction. Title-and-abstract screening removed 4,505 records, and full-text eligibility assessment removed a further 48, yielding 259 included studies. Screening was performed independently by two annotators; disagreements were resolved through discussion, with inter-annotator agreement of Cohen's $\kappa = .81$ on a shared sample of 40 papers. The PRISMA flow diagram is presented below:

C Trade-off Matrix Table

Table 2 compares representative approaches to domain knowledge, student modeling, and dialogue management. The ratings summarize the pedagogical control, scalability, adaptability, and interpretability associated with each paradigm; these qualitative assessments are not based on a common benchmark.

Survey	Span	#Studies	Scope	Conversation as design dimension	Architectural analysis	Interaction design	LLM-era coverage	Evaluation methodology
Kulik & Fletcher (2016)	–2013	50	ITS (all)	×	×	×	×	✓[a]
Mousavinasab et al. (2021)	2007–2017	53	ITS (all)	×	(✓)[b]	×	×	✓
Lin et al. (2023)	2012–2022	125	AI-based ITS	×	(✓)[b]	×	×	✓
Létourneau et al. (2025)	2009–2025	28	AI-driven ITS (K–12)	×	×	×	×	✓
Zerkouk et al. (2025)	2010–2025	127	AI-based ITS	(✓)[c]	(✓)[b]	(✓)[d]	✓	✓
This survey	**2010–2025**	**259**	**CITS**	✓	✓	✓	✓	✓

✓ = central focus; (✓) = partial treatment; × = not addressed.
[a] Meta-analysis of learning outcomes; does not synthesize evaluation methodologies.
[b] Describes system types or AI techniques without comparative component-level analysis.
[c] Dialogue systems are covered as one NLP application area, not as an organizing design dimension.
[d] Usability and AR/VR/IoT interfaces are discussed under evaluation and emerging trends, not as an analytical category.

Table 1: Comparison of prior surveys with the present survey. CITS = conversational intelligent tutoring systems.

Component	Paradigm	Pedagogical control	Scalability	Adaptability	Interpretability	Representative Studies
Domain knowledge (§3.1)	Scripts + FSM / CBM	High	Low	Low	High	AutoTutor (D'Mello et al., 2010); SQL-Tutor (Mitrovic et al., 2013)
	Ontologies / KG	High	Low–Med	Med	High	MST (Afzal et al., 2020)
	Generative / RAG	Low–Med	High	High	Low	Math Tutor (Levonian et al., 2025)
Student modeling (§3.2)	Episodic matching (LSA/regex)	Med	Med	Low	Med	AutoTutor Lite (Wolfe et al., 2013)
	Sequential error models (CRF)	Med	Med	Med	Med–High	Adaptive Feedback Tutor (Gutierrez and Atkinson, 2011)
	LLM-based error diagnosis	Low	High	High	Low	Reasoning Tutor (Daheim et al., 2024)
Dialogue management (§3.3)	FSM frames	High	Low	Low	High	AutoTutor (D'Mello et al., 2010)
	RL / probabilistic	Med	Med	Med	Med	Language Tutor (Abou-Khalil et al., 2018); Nursing Tutor (Abuazizeh et al., 2020)
	Generative / multi-agent	Low–Med	High	High	Low	GenMentor (Wang et al., 2025b)

FSM = finite-state machine; CBM = constraint-based modeling; KG = knowledge graph; RAG = retrieval-augmented generation; LSA = latent semantic analysis; CRF = conditional random field; LLM = large language model; RL = reinforcement learning; MST = medical school tutor.

Table 2: Qualitative comparison of architectural paradigms across three components. Ratings summarize relative pedagogical control, scalability, adaptability, and interpretability.