AI Chatbots in Education: A Systematic Review of Adoption Theories, Knowledge Outcomes, and Research Gaps (2019–2024)

Mouza Almehrzi

British University in Dubai, UAE

20001324@student.buid.ac.ae

***Abstract* -** AI-powered chatbots have become one of the most actively studied technologies in educational research, yet the field remains fragmented across adoption theories, knowledge outcome measures, and methodological approaches. This paper presents a systematic literature review of 205 peer-reviewed studies on AI chatbots in educational contexts, published between 2019 and 2024, sourced from 10 major academic databases and conducted using a PRISMA-informed protocol. The review analyses theoretical frameworks, geographical distribution, independent and dependent variables, methodological approaches, and participant demographics across the final set of studies. Technology Acceptance Model (TAM) dominates theoretical landscape (74 mentions; 36% of studies), while the Unified Theory of Acceptance and Use of Technology (UTAUT) appears in 24 studies, and Constructivist Learning Theory in 17. Quantitative surveys account for 47% of studies, qualitative methods for 45%, and mixed methods for only 7%. Student-only participant samples comprise 62% of studies, with faculty and administrative staff severely underrepresented. Ninety-six percent of studies examine chatbot adoption at the individual level, leaving organizational-level adoption research almost entirely unexplored. Three critical gaps are identified: the theoretical dominance of TAM in the absence of pedagogical integration frameworks; the near-total absence of social sustainability criteria in chatbot evaluation; and the concentration of research in five countries (China, India, Spain, Malaysia, Germany), with the UAE and broader Gulf contexts representing only 4% of the evidence base. These gaps motivate the development of integrated theoretical frameworks combining technology adoption, knowledge management, and social sustainability as an agenda for the field's next research cycle.

Keywords: *AI chatbots; educational technology; systematic literature review; TAM; UTAUT; knowledge management; social sustainability; higher education; technology adoption*

I. Introduction

Research on AI-powered chatbots in educational contexts has grown faster than almost any comparable strand of educational technology scholarship. From 18 published studies in 2019, the annual output reached 62 studies in 2023, a 244% increase in five years, with a further 38 studies published in the first half of 2024 alone. This growth is not surprising given the operational pressures it responds to, rising student enrolments, expectations of 24/7 institutional responsiveness, demand for personalized feedback at scale, and the maturation of conversational AI technologies that make deployment practically feasible for institutions with modest technical infrastructure (Dos, 2025; Crompton & Burke, 2023).

However, growth in volume does not automatically produce growth in understanding. A field that expands quickly can develop blind spots just as quickly, theoretical monocultures that crowd out alternative perspectives, methodological habits that replicate easily rather than illuminate deeply, and gaps in geographic or stakeholder representation that quietly narrow the evidence base. Prior reviews of educational chatbot research (Okonkwo & Ade-Ibijola, 2021; Smutny & Schreiberova, 2020) have identified some of these patterns, but the field has continued to expand since those reviews were completed and has not been comprehensively mapped since then.

This paper presents a systematic literature review of 205 studies on AI chatbots in education, published between 2019 and 2024. The review addresses five questions: What theoretical frameworks have been applied? Where has the research been conducted? What variables have been studied? What methodological approaches have been used? Moreover, who has been included as a research participant? Answering these questions produces a map of the field's current state and, crucially, of what is missing, the gaps that motivate the next generation of research.

The review is distinctive in three respects. First, it applies a rigorous PRISMA-informed protocol across ten databases, producing a final set of 205 studies that meet explicit inclusion and exclusion criteria. Second, it evaluates theoretical frameworks not merely by counting citations but by assessing their fit with the full range of educational chatbot outcomes, adoption, knowledge facilitation, equity, and sustainability. Third, it pays specific attention to the UAE and Gulf

Inclusion Criteria	Exclusion Criteria
Chatbot or conversational agent in education	Not related to chatbots in education
Theoretical model with hypotheses or RQs	No theoretical model or hypotheses
Written in English	Non-English language publications
Peer-reviewed: journal, conference, or book chapter	SLRs, meta-analyses, posters, theses, reports
Published 2019–2024	Published before 2019 or after March 2025
Full-text accessible	Restricted or inaccessible full text
Empirical or analytical results reported	No results reported

context, which is underrepresented in prior reviews despite being an active and growing site of educational technology adoption.

II. SYSTEMATIC REVIEW METHODOLOGY

A. *Search Protocol and Database Selection*

The search was conducted in three iterative rounds: an initial search on 11 June 2024, a second round on 27 October 2024, and a third round on 16 February 2025, with a final update in March 2025 to capture any publications in the period immediately preceding the review's completion. The iterative approach was chosen because the field's growth rate meant that a single search date would likely miss relevant publications within the review window.

Ten academic databases were searched: Google Scholar, ResearchGate, IEEE Xplore, Springer, ScienceDirect, Taylor & Francis, Frontiers, ProQuest, Semantic Scholar, and SAGE. These databases were selected to provide complementary coverage of engineering-oriented, education-oriented, and social science-oriented publication venues, reflecting the genuinely interdisciplinary character of educational chatbot research.

The search query applied across all databases was: "CHATBOT" AND "EDUCATION" AND ("acceptance" OR "adoption" OR "use" OR "intention" OR "behavior" OR "learning"). This query was designed to capture the dominant strand of chatbot research studies examining user acceptance and behavioral outcomes, while remaining broad enough to include studies framed in learning theory rather than in technology acceptance terms.

B. *Inclusion and Exclusion Criteria*

Studies were included if they: (1) focused on chatbots or conversational agents in educational contexts; (2) included a theoretical model with explicit hypotheses or research questions; (3) were written in English; (4) were published as peer-reviewed journal articles, conference papers, or book chapters with reported empirical or analytical results; (5) were published between 2019 and 2024; and (6) were available in full text. Studies were excluded if they were systematic reviews or meta-analyses themselves, were posters, reports, theses, or magazine articles, lacked reported results, or fell outside the review period. Sources published beyond this window are incorporated in the Introduction and Discussion as supplementary contextual evidence.

Table 1. Inclusion and Exclusion Criteria

C. *Screening and Final Sample*

The initial search returned 5,670 articles across all databases. After applying the inclusion and exclusion criteria, 226 studies were assessed for eligibility. A further 21 were excluded for insufficient alignment with the review's conceptual framework, specifically, studies that named chatbots in their titles but did not examine them in the context of educational adoption, learning outcomes, or knowledge facilitation. The final sample comprised 205 studies. Data were extracted across eleven dimensions for each study: author, publication year, database, theory or model applied, country of study, independent variables, dependent variables, moderators, methodology, participant type, and unit of analysis.

III. FINDINGS

- ***Publication Trends***

Publication volumes grew consistently across the review period, rising from 18 studies in 2019 to a peak of 62 in 2023, with 38 studies captured in the first half of 2024 alone. This trajectory places educational chatbot research among the fastest-growing strands of educational technology scholarship. It reflects both the acceleration of AI capabilities and the pandemic-driven expansion of digital learning infrastructure, which created institutional demand for scalable student support solutions.

Rank	Theory / Model	Count	% of Studies
1	Technology Acceptance Model (TAM)	74	36.1%
2	Unified Theory of Acceptance and Use of Technology (UTAUT)	24	11.7%
3	Constructivist Learning Theory	17	8.3%
4	UTAUT2	12	5.9%
5	Artificial Intelligence Theory	8	3.9%
6	Self-Regulated Learning Theory	6	2.9%
7	Self-Determination Theory (SDT)	6	2.9%
8	Cognitive Load Theory	5	2.4%
9	Task-Technology Fit (TTF)	5	2.4%
10	Diffusion of Innovation Theory	5	2.4%
11	Human-Computer Interaction (HCI) Theory	5	2.4%
12	Blended Learning Theory	4	2.0%
13	Expectation-Confirmation Model (ECM)	4	2.0%
14	Social Learning Theory	4	2.0%
15	Value-Added Model / ARCS / Chatbot-Mediated / Other (×18 theories)	26	12.7%

The most productive databases by final study count were Google Scholar (N = 55), ResearchGate (N = 55), IEEE Xplore (N = 41), Springer (N = 31), and ScienceDirect (N = 9). The concentration in Google Scholar and ResearchGate reflects the interdisciplinary and pre-print-inclusive nature of those platforms, which capture conference papers and working papers that more selective databases exclude. The strong IEEE Xplore contribution reflects the engineering-oriented character of much chatbot development research, which sits at the boundary between computer science and educational technology.

- ***Theoretical Frameworks***

Table 2. Theoretical Frameworks in Educational Chatbot Research (N = 205). Top theories by frequency.

The Technology Acceptance Model (Davis, 1989) was by far the most frequently applied theoretical framework, appearing in 74 studies (36% of the total sample). UTAUT (Venkatesh et al., 2003) appeared in 24 studies, Constructivist Learning Theory in 17, UTAUT2 in 12, and Self-Regulated Learning Theory in six. Table 2 presents the full distribution.

TAM's dominance is consistent with prior reviews (Okonkwo & Ade-Ibijola, 2021; Smutny & Schreiberova, 2020), but the distribution in this review reveals the full extent of the theoretical narrowness: the top three theories together account for 56% of studies, while 18 additional theoretical frameworks collectively account for only 13%. This concentration has consequences for the field's ability to address questions that lie outside TAM's scope, including knowledge facilitation outcomes, institutional equity, and long-term sustainability, which require theoretical tools that TAM was not designed to provide.

The presence of Constructivist Learning Theory in 17 studies (8.3%) and SDT in six is notable: it indicates that a minority of researchers have already moved toward pedagogically integrated frameworks. Chamorro et al. (2023) provide the clearest example, demonstrating that combining TAM with constructivist learning outcomes revealed adoption-learning relationships that TAM alone could not detect. This integration approach is promising but remains at the margins of the field.

- ***Geographical Distribution***

China, India, and Spain each accounted for 16 studies in the final sample, followed by Malaysia (13), Germany (10), Indonesia (9), Taiwan (9), Turkey (9), and South Korea (9). The United Arab Emirates accounted for eight studies. Together, the top five

countries produced 74 of the 205 studies, 36% of the evidence base concentrated in five national contexts.

The Gulf region, despite being an active and well-resourced site of higher education expansion, with UAE institutions consistently ranking in global university league tables and significant government investment in AI-enabled education infrastructure, accounted for only eight studies, or 4% of the sample. This underrepresentation has direct implications for the generalizability of findings from predominantly East Asian and European contexts to Gulf institutional settings, where cultural dimensions (collectivism, power distance), language dynamics (Arabic/English bilingualism), and institutional governance structures may all moderate chatbot adoption and use in ways that existing models have not been tested against.

- ***Dependent Variables***

User satisfaction was the most frequently studied dependent variable, accounting for 29% of studies. Behavioural intention followed at 23%, learning outcomes at 19%, acceptance at 11%, effectiveness at 10%, and engagement at 9%. These proportions reveal a field that is primarily interested in whether users like and intend to use chatbots, rather than in what they learn or gain institutionally from doing so. User satisfaction and behavioral intention together account for over half of all dependent-variable measurements, a concentration that reflects the TAM-dominant theoretical landscape and, like it, limits the field's ability to address downstream knowledge and equity questions.

The 19% share attributed to learning outcomes is the most practically important finding in the distribution of dependent variables: it shows that a meaningful minority of researchers are measuring what students learn or gain from chatbot interactions, not just whether they plan to use them. This minority represents the theoretical direction that the field needs to expand toward if it is to answer the questions that educators and institutions need answered.

- ***Independent Variables***

Perceived Usefulness was the most frequently studied independent variable (25% of studies), followed by Ease of Use (22%), Social Influence (16%), Trust (15%), Performance (13%), and Facilitating Conditions (9%). This distribution maps almost exactly onto the core constructs of TAM and UTAUT. It confirms the theoretical dominance finding from Section 3.2 on the independent variable side: researchers are measuring what their theoretical frameworks direct them to measure, and the frameworks they most often use are adoption-oriented rather than outcome-oriented.

The relatively low share for Facilitating Conditions (9%) is noteworthy given the growing evidence from Al-Emran et al. (2020), Tarhini et al. (2017), and others that institutional infrastructure adequacy is one of the most consequential predictors of both adoption and learning outcomes in educational technology deployments. Its underrepresentation in the literature as an independent variable suggests a systematic gap between what the evidence base indicates matters and what researchers choose to study.

- ***Methodological Approaches***

Quantitative surveys dominated the methodological landscape (47%), with qualitative methods accounting for 45% and mixed methods for only 7%. This distribution is superficially balanced, but the 45% qualitative component is largely composed of small-scale case studies and interviews rather than systematic field observations or longitudinal ethnographic work. True longitudinal studies following the same participants across more than one academic semester accounted for fewer than 8% of the total sample, with an average study duration of approximately 8 weeks.

The 7% share of mixed methods is the most striking methodological gap. Mixed-methods designs that combine quantitative adoption measurement with qualitative investigation of why adoption decisions were made, or that combine survey data with LMS behavioral logs, are precisely what the field is asking for. A researcher who wants to understand not just whether students intend to use a chatbot, but also whether using it improves their ability to navigate the LMS independently, needs both a survey and behavioral records. The rarity of such designs represents a methodological opportunity as much as a gap.

- ***Participant Demographics***

Students comprised 62% of research participants across the sample, faculty 11%, administrative staff 6%, and combined or unstated samples 21%. Student concentration is expected, given that students are the primary users of chatbots in most educational deployments. However, the severe underrepresentation of faculty and administrative staff has two important consequences.

First, it limits the evidence to the adoption barriers that most directly determine whether chatbots are deployed and maintained: faculty concerns about pedagogical

fit, administrative concerns about cost-effectiveness and governance, and support staff perspectives on workload changes and professional identity. Gökçearslan et al. (2024) found that 78% of surveyed faculty believed chatbots could not provide adequate emotional support or nuanced feedback, and 45% feared chatbots might oversimplify complex subject-matter concerns that are invisible in student-only samples. Second, it means that the 40% workload reduction reported by Goel and Polepeddi (2016) for support staff following chatbot deployment has not been replicated or challenged in subsequent research, because support staff have not been included as primary participants in subsequent studies.

- ***Unit of Analysis: Individual vs. Organizational***

96% of the reviewed studies examined chatbot adoption at the individual level, measuring students' intentions, perceptions, or outcomes. Only 4% examined organizational-level impacts: cost-effectiveness, institutional adoption rates, administrative efficiency, or governance outcomes. This concentration means that the field has built extensive evidence base on whether individual students like and intend to use chatbots, while almost entirely ignoring whether institutions benefit from deploying them.

This is not merely a methodological preference; it reflects a deeper theoretical gap. The dominant theoretical frameworks (TAM, UTAUT in its original form) were designed to explain individual adoption behaviour. They do not provide constructs for measuring institutional learning, cumulative knowledge management outcomes, or the equity effects of deployment at scale. Addressing the individual–organizational gap, therefore, requires not just different research designs but different theoretical frameworks.

IV. CRITICAL GAPS AND THEORETICAL IMPLICATIONS

- ***TAM Dominance Without Pedagogical Integration***

TAM's explanatory power for initial adoption intention in educational technology contexts is well established. Meta-analyses consistently show it accounts for 30–40% of the variance in adoption intention (Šumak et al., 2011). However, TAM was designed to explain individual adoption decisions for information systems in organizational settings. Applying it to educational chatbots treats learning as incidental to adoption and casts students as tool users rather than as learners developing competencies.

Three specific limitations emerge from the review findings. First, TAM's two-construct model cannot distinguish between a chatbot that students find usefully simple and one that they find simplistically unhelpful; both might score similarly on Perceived Usefulness if the survey item asks about general utility rather than learning-specific utility. Second, TAM provides no mechanism for examining what students gain from chatbot interactions in terms of knowledge, understanding, or independence. A 74-study dominance by a framework without a learning-outcome construct is a serious structural limitation for a field that presents itself as educational research. Third, TAM does not address post-adoption outcomes, sustainability, equity, or institutional impact, which are the questions that matter most to institutional decision-makers.

The integration of TAM with Constructivist Learning Theory (17 studies), Knowledge Management frameworks, and Social Sustainability criteria represents the theoretical direction the field needs to move toward. Chamorro et al. (2023) demonstrated that blending TAM with constructivist learning outcomes revealed significant adoption-knowledge relationships that TAM alone could not detect. The finding that collaborative learning outcomes correlated with a 20% increase in peer-to-peer problem-solving, but only when teachers scaffolded chatbot use with group discussions, is the practically actionable insight that TAM alone could never produce.

- ***The Absence of Social Sustainability Criteria***

None of the 205 reviewed studies applied a social sustainability framework to evaluate chatbot deployment outcomes. The ethical concerns that the field acknowledges, including algorithmic bias (18 studies), data privacy (22 studies), lack of transparency (15 studies), equity gaps for low-income students (9 studies), and staff displacement concerns (8 studies), are identified as challenges in individual studies but have never been organized into a testable theoretical framework applied to real deployment data.

This gap is particularly significant given the distribution of equity across the dependent variables: user satisfaction (29%) and behavioral intention (23%) dominate, whereas equitable access, staff impact, and long-term viability appear as primary dependent variables in no study. A field that evaluates chatbots primarily through user satisfaction is structurally

unable to detect deployment outcomes that fall on users who are less satisfied by design students whose queries the chatbot handles poorly, staff whose work conditions are affected by automation, and institutions that allocate resources to systems they cannot sustain.

- ***Geographic Concentration and Context Specificity***

The concentration of research in five national contexts (China, India, Spain, Malaysia, Germany) raises legitimate questions about whether findings generalize across the diverse institutional, cultural, and linguistic settings in which chatbots are being deployed. UTAUT's moderating variables, gender, age, experience, and voluntariness, have been validated primarily in Western and East Asian organizational settings. The extent to which they apply in Gulf higher education contexts, where institutional authority structures, language dynamics, and cultural dimensions of technology adoption may differ systematically, has not been tested.

The eight UAE studies in the sample demonstrate that research on Gulf-context chatbots exists and is growing. However, eight studies are insufficient to identify patterns, test the framework's generalizability, or establish the contextual modifiers needed for an institution in Dubai or Abu Dhabi to make evidence-based decisions about chatbot deployment. This is a specific and addressable gap rather than a general call for more research.

- ***The Knowledge Management Gap***

Knowledge Management theory provides the conceptual language for a set of chatbot outcomes that the review finds almost entirely unstudied: knowledge acquisition (how users obtain understanding through chatbot interactions), knowledge sharing (how chatbot systems make collectively accumulated knowledge available across user communities), and knowledge application (how users act on chatbot-provided information to complete real tasks). Of the 205 studies reviewed, fewer than 12 explicitly examined knowledge outcomes using KM constructs or frameworks.

This gap is particularly stark given the deployment context. Chatbots that operate in institutional support environments, LMS help desks, academic advising services, and enrollment management are, by design, knowledge management systems. They convert tacit institutional knowledge (what experienced staff know from years of handling similar queries) into explicit, structured, scalable responses. Evaluating them without KM constructs is analogous to evaluating a library without measuring whether users find what they came for.

V. Future Research

The field's most pressing need is an integrated framework that combines technology acceptance, knowledge management, and social sustainability constructs, as none of the 205 studies reviewed bring all three together. UTAUT offers the strongest adoption foundation; its Facilitating Conditions and Social Influence constructs make it more institutionally attuned than TAM, while KM theory (Alavi & Leidner, 2001; Nonaka & Takeuchi, 1995) supplies the knowledge-outcome layer and social sustainability frameworks (Sterling, 2012; Leal Filho et al., 2019) supply the equity and long-term impact layer. This absence of integration represents the field's most significant theoretical gap.

A second priority is longitudinal and multi-site research. Because most existing studies are cross-sectional and confined to a single semester, little is known about how adoption dynamics and knowledge outcomes evolve as chatbot use matures. Key open questions include whether Performance Expectancy dominates early adoption while Facilitating Conditions matter more later, whether knowledge gains persist over time, and whether skills transfer across LMS contexts. Answering these will require cohort studies tracked over three to six semesters alongside multi-site comparisons across institutions using different LMS platforms.

The field also needs more faculty, staff, and organizational-level research. With only 11% of faculty and 6% of administrative staff represented across the reviewed literature, these groups should be treated as priorities rather than as incidental gaps: faculty decisions shape whether chatbots are integrated into courses or remain peripheral, while administrative decisions determine ongoing maintenance and resourcing. Both groups are well-suited to UTAUT-based designs and represent accessible opportunities for institutions with existing deployments.

Gulf and broader MENA contexts remain underrepresented despite their strategic importance combining strong institutional resources, robust national AI policies, multilingual student populations, and cultural dimensions (collectivism, pronounced Social Influence, high Power Distance) likely to

moderate UTAUT's predictions differently than East Asian or European settings. Studies conducted across institutions in the UAE, Saudi Arabia, Qatar, and Bahrain would add regional evidence while refining the theory cross-culturally.

REFERENCES

Al-Emran, M., Mezhuyev, V. & Kamaludin, A. (2020). 'Technology acceptance model in m-learning context: A systematic review', Computers & Education, 144, p. 103638.

Alavi, M. and Leidner, D. E. (2001). 'Knowledge management and knowledge management systems: Conceptual foundations and research issues', MIS Quarterly, 25(1), pp. 107–136.

Chamorro, O., Gonzalez, P. and Herrera, M. (2023). 'Chatbot adoption in higher education: Blending TAM with constructivist learning outcomes', Computers & Education, 192, pp. 104–118.

Crompton, H. and Burke, D. (2023). 'Artificial intelligence in higher education: The state of the field', International Journal of Educational Technology in Higher Education, 20(1), p. 22.

Davis, F. D. (1989). 'Perceived usefulness, perceived ease of use, and user acceptance of information technology', MIS Quarterly, 13(3), pp. 319–340.

Dos, B. (2025). 'A systematic review of chatbot research in education: A bibliometric analysis', Education and Information Technologies, 30(1), pp. 1–28.

Goel, A. K. and Polepeddi, L. (2016). Jill Watson: A virtual teaching assistant for online education. Georgia Institute of Technology.

Gokcearslan, S., Tosun, C. and Erdemir, Z. (2024). 'Faculty perceptions of AI chatbots in higher education', Computers & Education, 192, p. 104642.

Leal Filho, W., Azul, A. M., Brandli, L., Ozuyar, P. G. and Wall, T. (eds.) (2019). Sustainable cities and communities. Springer.

Nonaka, I. and Takeuchi, H. (1995). The knowledge-creating company. Oxford University Press.

Okonkwo, C. W. and Ade-Ibijola, A. (2021). 'Chatbots applications in education: A systematic review', Computers and Education: Artificial Intelligence, 2, p. 100033.

Rudolph, J., Tan, S. and Tan, S. (2023). 'ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?', Journal of Applied Learning and Teaching, 6(1).

Smutny, P. and Schreiberova, P. (2020). 'Chatbots for learning: A review of educational chatbots for the Facebook Messenger', Computers & Education, 151, p. 103862.

Sterling, S. (2012). The future fit framework: An introductory guide to teaching and learning for sustainability in HE. Higher Education Academy.

Sumak, B., Hericko, M. and Pusnik, M. (2011). 'A meta-analysis of e-learning technology acceptance: The role of user types and e-learning technology types', Computers in Human Behavior, 27(6), pp. 2067–2077.

Tarhini, A., Hone, K., Liu, X. and Tarhini, T. (2017). 'Examining the moderating effect of individual-level cultural values on users' acceptance of E-learning in developing countries', Interactive Learning Environments, 25(3), pp. 306–328.

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). 'User acceptance of information technology: Toward a unified view', MIS Quarterly, 27(3), pp. 425–478.