How Can "Digital Humans" Become "Confidants"? -The Emotional Interaction Dilemma and Breakthrough Pathways of AI Virtual Digital Humans in Psychological Counseling for Primary and Secondary School Students

Yimeng Lai*, Jiaxue Xie

Chengdu University of Information Technology, Chengdu, Sichuan Province, China

*2366767355@qq.com

Abstract. AI virtual digital humans are increasingly used in school mental health education, yet challenges such as simplified expression, emotional detachment, and lack of trust persist. This study integrates affective computing, psychological contract, situated learning, and value-sensitive design theories. Using literature review and case analysis, it examines three core dilemmas—superficial emotional interaction, trust barriers, and insufficient contextual adaptability—and proposes breakthrough pathways: shifting from recognition to empathy, building continuous and personalized trust, and establishing human-machine collaboration.

Keywords: AI virtual digital human; psychological counseling in primary and secondary schools; emotional interaction; human-machine collaboration

1 Introduction

1.1 Research Background and Problem Statement

AI virtual digital humans are increasingly deployed in school mental health education, providing emotional support to address the shortage of professional counselors. However, practical implementation reveals persistent issues: students often exhibit simplified communication and insufficient trust when interacting with AI. This raises a fundamental question: can digital humans truly become the "confidants" that students are willing to open up to?

Current research primarily focuses on optimizing AI emotion computing, with limited systematic analysis from the perspectives of educational context and trust mechanisms. Addressing this gap, this study examines emotional interaction issues between AI digital humans and students, systematically analyzes practical dilemmas, and explores breakthrough pathways.

C. F. Peng et al. (eds.), *Proceedings of the 2026 5th International Conference on Humanities, Wisdom Education and Service Management (HWESM 2026)*, Advances in Social Science, Education and Humanities Research 1024,
https://doi.org/10.2991/978-2-38476-593-5_51

1.2 Research Significance and Methods

Theoretically, this study integrates four theories to construct a multidimensional analytical framework revealing the structural causes of emotional interaction dilemmas. Practically, the proposed pathways offer guidance for designing AI products in school mental health education.

This study employs literature review and case analysis. Relevant publications were systematically retrieved from CNKI and Web of Science. Three representative AI mental health applications for primary and secondary schools in China were selected to examine their interaction performance in real educational scenarios.

2 Core Concepts and Theoretical Foundations

2.1 Definition of Core Concepts

AI virtual digital human refers to a virtual entity driven by artificial intelligence, featuring anthropomorphic appearance, language interaction capabilities, and emotional expression functions. In school psychological counseling, it typically appears as a virtual teacher, companion, or counselor.

Emotional interaction refers to the emotion-mediated interaction between humans and AI systems, encompassing emotion recognition, understanding, response, and expression. Current AI digital human emotional interaction primarily involves categorical recognition of overt emotions, without addressing the underlying individual experiential context.

Psychological counseling in primary and secondary schools refers to educational activities aimed at promoting students' mental well-being through listening, empathy, and guidance, with effectiveness grounded in trust relationships, contextual understanding, and professional boundaries.

2.2 Theoretical Basis

This study constructs an analytical framework based on four theories.

Affective computing theory (Picard, 1997) reveals the fundamental limitation of current AI emotional interaction: machines can recognize emotional labels through data training but cannot understand the personal experiences and social relationships behind them. This theory explains the technical roots of the "superficialization" dilemma[1].

Psychological contract theory (Rousseau, 1995) explains the mechanisms of implicit expectations and trust formation[2]. Students interacting with AI often hold expectations of "being understood and kept confidential." When AI responses are mechanical or privacy protections are unclear, the psychological contract is broken. This theory provides a framework for analyzing trust construction challenges.

Situated learning theory (Lave & Wenger, 1991) emphasizes that the meaning of behavior depends on specific contexts, and interactions detached from context struggle to generate genuine understanding[3]. School counseling involves multiple contexts such as classrooms, families, and peer relationships, while general AI models fail to

capture these nuances. This theory helps analyze the dilemma of insufficient educational context adaptability.

Value-sensitive design theory (Friedman et al., 2006) advocates embedding human values such as privacy, autonomy, and transparency into technical design from the outset[4]. This theory identifies ethical risks in AI digital human applications and provides methodological guidance for developing a human-machine collaborative service model.

These four theories collectively form the analytical framework of this study.

3 Real-world Challenges in Emotional Interaction

3.1 The Superficialization Dilemma

The application of AI digital humans in school psychological counseling first encounters the challenge of insufficient emotional interaction depth. Current affective computing relies primarily on recognizing and analyzing overt features such as speech, facial expressions, and text. This "recognition" essentially involves categorizing emotional representations rather than understanding emotional connotations[5]. When students confide in AI digital humans, the system can accurately identify emotional labels such as "sadness" and "anxiety," but struggles to grasp underlying factors such as personal growth experiences, family environments, and peer relationships.

In campus trials of the AI psychological companion app "Heart Language Companion," students initially provided detailed descriptions of their distress, but after several formulaic responses, they simplified their expressions to phrases like "I'm feeling down" or "It's okay," eventually abandoning their confessions. This superficial response pattern leaves students without a genuine sense of being understood, and when they realize the AI cannot truly "understand" them, they tend to simplify or even cease communication, further diminishing the quality of emotional interaction.

3.2 Trust Construction and Ethical Risk Dilemma

Trust serves as the psychological foundation for effective counseling, yet AI digital humans, as non-human entities, face structural barriers in trust-building. Primary and secondary school students are in a critical period of psychological development and are naturally sensitive and cautious about "confidants." Establishing trust with AI digital humans requires overcoming the presupposition that "machines cannot truly understand humans," a process inherently difficult, as shown in Figure 1.

More critically, ethical risks further complicate trust formation. When interacting with AI digital humans, students inevitably reveal personal emotions, psychological distress, and even private information, yet the boundaries for collecting, storing, and using such data are often unclear. The application of "AI Psychology Teacher" Xiao Zhi in a middle school illustrates this issue: many students began avoiding sensitive topics after learning that their conversation records might be reviewed by back-end systems, limiting interactions to superficial exchanges like "I'm feeling okay today." When students realize their confessions might be recorded, analyzed, or used for

algorithm optimization, defensive mechanisms activate, leading them to actively suppress genuine emotions. Trust deficiency and privacy concerns intertwine, trapping AI digital humans in a vicious cycle where "the more trust is needed, the harder it becomes to gain trust."

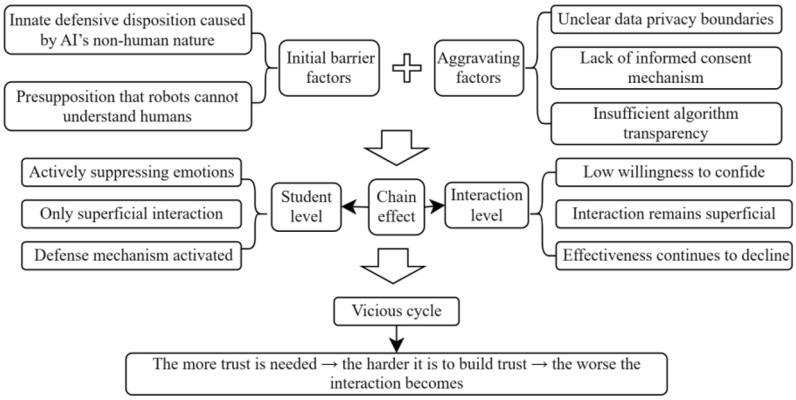

Fig. 1. Obstacles to Trust Construction and Their Chain Effects

3.3 Challenges of Insufficient Educational Context Adaptability

Psychological counseling in primary and secondary schools is highly contextualized, requiring each effective intervention to be precisely embedded in students' specific life scenarios. However, AI digital humans predominantly use generic emotional interaction models that lack deep understanding and flexible adaptation to educational contexts. The same set of response patterns, whether applied to classroom stress relief or family conflicts, inevitably proves ineffective, as seen in Table 1.

The "Sunshine Heart Talk" AI assistant used in upper elementary grades encountered such issues: it applied similar response templates to "exam anxiety" and "emotional distress caused by parental divorce," leading students to feel "the AI doesn't understand my situation" when dealing with the latter. More fundamentally, AI digital humans struggle to capture subtle interpersonal cues in educational contexts—such as teacher-student relationships, peer status, and family interaction patterns—which are critical variables affecting student mental health. When AI digital humans cannot recognize and respond to this contextual information, counseling effectiveness diminishes significantly, and irrelevant responses may even exacerbate students' frustration.

Table 1. Triple Dilemma of Emotional Interaction in AI Digital Humans

Dilemma Dimension	Core Performance	Main Impact	Technical Root Cause
Superficial Emotional Recognition	Recognizes overt emotions but cannot understand underlying causes	Students struggle to gain genuine emotional resonance	Single-modal recognition, lack of situational awareness

Trust Construction Barriers	Natural defense against non-human subjects	Reduced willingness to confide, superficial interaction	Lack of continuous interaction, absence of personalization
Prominent Ethical Risks	Unclear data privacy boundaries	Students become wary, trust difficult to establish	Inadequate data governance mechanisms
Insufficient Context Adaptation	General models struggle with diverse scenarios	Reduced effectiveness, possible increased frustration	Lack of deep embedding in educational contexts

4 Breakthrough Path from "Digital Human" to "Confidant"

4.1 From Recognition to Empathy: Technological Advancements in Emotional Understanding

The fundamental pathway to overcoming the superficiality dilemma is advancing AI digital humans' emotional interaction from "recognition" to "empathy." This requires a technological leap from single-modal recognition to multi-modal integrated understanding. Integrating multi-dimensional data such as voice, facial expressions, text, and physiological signals provides AI with richer emotional judgment criteria[6]. Introducing situational awareness capabilities that contextualize emotional expressions within specific temporal-spatial backgrounds and individual growth trajectories brings AI closer to genuine "empathy."

Simultaneously, deep integration of psychological theory and educational knowledge is crucial. AI digital humans should not merely be emotion classifiers but "listeners" equipped with basic psychological literacy—capable of identifying the underlying intentions, defense mechanisms, and developmental needs behind students' words. This technological leap from recognition to empathy is a critical step in the journey from "digital human" to "confidant."

4.2 Continuity and Personalization

Establishing trust requires time accumulation and interaction continuity. In real psychological counseling, therapists gradually build stable trust through long-term interaction with clients—a logic equally applicable to AI digital humans. Transforming AI digital humans from "one-time question-and-answer tools" into "long-term companions" requires constructing student psychological growth profiles that document emotional trajectories, key events, and interaction histories, ensuring each interaction builds on sustained understanding of the individual student,as shown in Figure 2.

At the same time, endowing AI digital humans with personalization features is crucial. AI digital humans should possess stable language styles, emotional expressions, and value stances, forming a "personality image" that students can recognize and anticipate[7]. The combination of continuity and personalization helps students gradually

overcome psychological defenses against "machines," establishing stable trust in AI digital humans and making them truly worthy of being students’ "confidants."

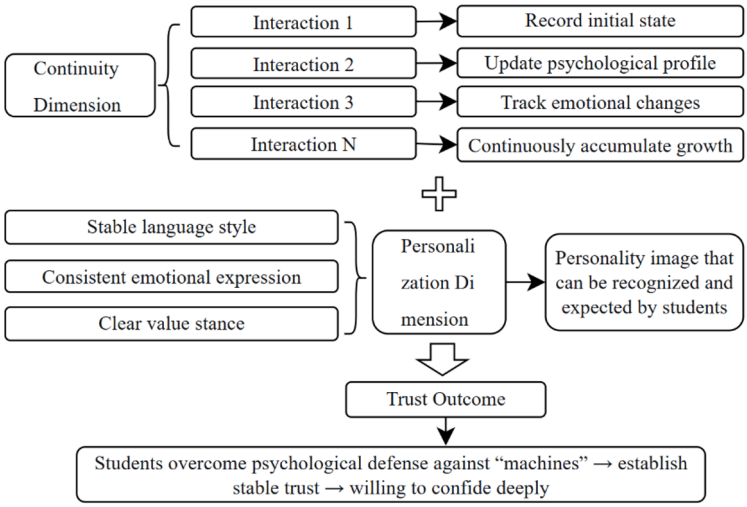

Fig. 2. Mechanism of Continuity and Personification in Constructing Trustworthy Relationships

4.3 Human-Machine Collaboration: Establishment of Ethical Governance and Service Boundaries

AI digital humans entering the field of school psychological counseling must have their role boundaries and ethical limits clearly defined. Regarding privacy protection, the principle of data minimization should be established, with students clearly informed about the scope and purpose of data collection, and granted rights to informed consent and control over personal information.

Regarding responsibility attribution, AI digital humans should be positioned as "assistants" rather than "substitutes." Psychological counseling is fundamentally a form of deep interpersonal interaction involving complex emotional connections and value judgments—areas that current and foreseeable AI cannot replace[8]. Therefore, AI digital humans should undertake auxiliary functions such as initial screening, daily companionship, and emotional guidance, while complex psychological issues must be identified and addressed by professional psychological teachers or counselors.

5 Conclusion and Prospects

5.1 Research Conclusions

AI digital humans face three structural dilemmas in emotional interaction within primary and secondary school psychological counseling. The superficiality dilemma lies

in technology's ability to recognize only emotional representations while failing to understand individual experiential contexts; the trust construction and ethical risk dilemma stems from the fragility of human-machine psychological contracts and unclear data privacy boundaries; the educational context adaptability dilemma arises from the difficulty of embedding general models into diverse educational settings. These three dilemmas intertwine, collectively constraining the effectiveness of AI digital humans in psychological counseling.

To bridge the gap from "digital human" to "confidant," three breakthrough pathways are proposed: advancing emotional interaction from "recognition" to "empathy" by contextualizing emotional expressions within individual growth trajectories, constructing stable psychological contracts through interaction continuity and personalization features, and establishing a "human-machine collaborative service model" with ethical governance in privacy protection and transparency.

This study constructs an integrated analytical framework combining affective computing theory, psychological contract theory, situated learning theory, and value-sensitive design theory, providing systematic theoretical tools for understanding and addressing the emotional interaction dilemmas of AI digital humans.

5.2 Future Outlook

Technologically, multi-modal emotional understanding and situational awareness should be core indicators. In schools, AI should be positioned as auxiliary with "AI alert + teacher intervention" mechanisms. Ethically, guidelines for AI psychological services for minors should be developed.

This study is based on literature and case analysis. Future research could validate pathways through longitudinal experiments. As affective computing evolves, AI digital humans may truly become a "warm" auxiliary force in school mental health education.

Acknowledgments.

I would like to express my deepest gratitude to my supervisor for their invaluable guidance, rigorous academic attitude, and patient encouragement throughout this research. I am also sincerely grateful to the teachers and students who participated in the case studies for their cooperation and openness, which greatly enriched the empirical foundation of this study. My appreciation extends to the scholars whose works I have cited, as their research provided important theoretical support. Finally, I wish to thank my family and friends for their unwavering understanding and encouragement during my academic journey.

Fundings. This paper serves the National College Student Innovation and Entrepreneurship Training Program project "Feasibility Analysis of AI Digital Humans Applied to Mental Health Practice in Primary and Secondary Schools" (Grant No. 202510621013), which has been approved for national-level funding.

References

1. Hong, Y., & Huang, Y.Hong, Y., & Huang, Y. (2024). Social and ethical impact of emotional AI advancement: the rise of pseudo-intimacy relationships and challenges in human interactions. Frontiers in Psychology, 15, 1410462.
2. Farrukh Moin, M., Behl, A., Zuopeng Zhang, J., & Shankar, A. Farrukh Moin, M., Behl, A., Zuopeng Zhang, J., & Shankar, A. (2024). AI in the Organizational Nexus: Building Trust, Cementing Commitment, and Evolving Psychological Contracts. Information Systems Frontiers.
3. Vargas, E. G., Chiappe, A., & Durand, J. Vargas, E. G., Chiappe, A., & Durand, J. (2024). Reshaping Education in the Era of Artificial Intelligence: Insights from Situated Learning Related Literature. Journal of Social Studies Education Research, 15(2), 1–28.
4. Kieslich, K., et al. (2025). AI ethics unwrapped: an empirical investigation of ethical principles in collaborative ideation processes. AI and Ethics, 5, 3159–3172.
5. Shi Dandan. Application Strategies of AI in Mental Health Education [J]. Student·Parent·Society, 2025, (34):108-110.
6. Ma Tianyu, Bai Wenli, Tian Jiaxing. Optimization Pathways of AI-Enabled Adolescent Mental Health Education Courses [J]. Science Education Journal, 2026, (02):28-30.
7. Huang Xinyu. Exploring the Application of Generative AI in Mental Health Education for Junior High School Students [J]. Educational Observation, 2025,14(29):50-52.
8. Han Qing, Zang Peng, Chang Sheng. AI-enabled mental health education in primary and secondary schools: Rational data and emotional warmth [J]. Information Technology Education for Primary and Secondary Schools, 2025, (07):29-30.