From Barriers to Breakthroughs: Formation Mechanism and Development Pathways of University Teachers' AI Literacy Under Bidirectional Regulation of Human-AI Trust

Shenghui Zhu[1, *], Shaoqi Xu[1], Zheyi Han[2], Zihan You[1]
[1]College of Education, Zhejiang Normal University, Jinhua 321004, China
[2]College of Chemistry and Materials Science, Zhejiang Normal University, Jinhua 321004, China
*Corresponding email: 3537494005@qq.com

Abstract

To address the trust barriers hindering university teachers' artificial intelligence (AI) literacy improvement, this study introduces a bidirectional human-AI trust perspective, extends the unified theory of acceptance and use of technology (UTAUT) model, and examines the formation mechanism and development pathways via partial least squares structural equation modeling (PLS-SEM) analysis of 1,519 valid questionnaires from Zhejiang Province. The findings indicate: (1) Teachers' AI literacy presents a pattern of "solid basics but weak advanced competence", with bidirectional trust imbalance as the core constraint. (2) All UTAUT core variables significantly affect AI literacy; bidirectional human-AI trust serves as a partial mediator, and AI' s adaptive trust positively moderates the link between teachers' trust in AI and literacy, forming a closed-loop effect. (3) There are notable heterogeneities across teaching experience, disciplines and institution types. This study further identifies three developmental stages and proposes a three-tiered targeted pathway, providing empirical support for advancing teachers' AI literacy from basic use to sustainable, innovative application.

Keywords

Artificial intelligence literacy, Bidirectional regulation, Human-AI trust, Double First-class universities, Unified theory of acceptance, University teachers

Introduction

The deep penetration of artificial intelligence technology is reshaping the ecological landscape of higher education, and teachers' AI literacy has become a core competence supporting the digital transformation of education and the implementation of smart education.

At the national level, China's Education Modernization 2035 explicitly identifies the development of teachers' innovative educational capabilities as a core component of high-quality educational development. The Ministry of Education released The Industry Standard for Teachers' Digital Literacy. This document builds a teachers' digital literacy framework with five dimensions. The dimensions include digital awareness, digital technology knowledge and skills, digital application, digital social responsibility, and professional development. It offers an authoritative foundation for assessing teachers' capabilities in the AI era [1].

As a national pilot province for educational digital reform, Zhejiang Province has successively released multiple policy documents. Representative examples include the Zhejiang Province's Action Plan for Promoting "AI + Education" (2025-2029) and the Zhejiang Province's Framework for Artificial Intelligence Literacy of Teachers in Higher Education Institutions (Trial). These policies regard the improvement of teachers' AI literacy as a core task for building a strong education province. Meanwhile, they offer policy guidance and practical requirements to cultivate teachers' AI literacy at the regional level.

However, there is a noticeable mismatch between the speed of technological iteration and the pace of teachers' literacy improvement. The current application of AI among university teachers generally shows a structural deviation of "high cognition but low practice". Most teachers can use AI tools to finish basic tasks. Typical examples include courseware production and information retrieval. However, they have clear deficiencies in advanced capabilities. These capabilities

cover personalized instructional design, intelligent analysis of research data, and ethical risk prevention and control. The underlying reason is not simply insufficient technical competence, but the bidirectional trust barrier between teachers and AI systems. On the one hand, teachers have doubts about the reliability, transparency, and ethics of AI technology, forming "human distrust of AI". On the other hand, AI systems show insufficient respect and adaptation to teachers' professional judgment, leading to "AI's distrust of humans".

This bidirectional trust deficit severely hinders the substantial improvement of teachers' AI literacy. Most existing improvement strategies adopt one-size-fits-all skills training. This approach neglects the core issue of trust reconstruction. It therefore leads to a persistent practical dilemma. The dilemma can be summarized as "having tools but being unwilling to use, having training but being unable to use, and having technology but being afraid to use". Against this background, this study focuses on university teachers in Zhejiang Province, takes "from barriers to breakthroughs" as the core narrative, and unfolds around three core research questions.

First, what is the overall status and core barriers of university teachers' AI literacy and bidirectional human-AI trust in Zhejiang Province? Second, what regulatory role and conduction mechanism does bidirectional human-AI trust play in the formation of teachers' AI literacy? Third, how to construct a targeted layered development pathway for teachers' AI literacy based on bidirectional trust characteristics? This study conducts large-sample empirical analysis. It reveals the internal mechanism of literacy formation and develops a phased, grouped breakthrough pathway. These findings enrich the theoretical system of educational human-computer interaction. Meanwhile, they offer replicable practical solutions. The solutions support the cultivation of university teachers' AI literacy in Zhejiang Province and nationwide.

Literature review and theoretical basis

Research progress on university teachers' AI literacy

Teachers' AI literacy is an extension and upgrade of information literacy and digital literacy in the era of artificial intelligence. Information literacy focuses on the identification, retrieval, and ethical use of information. Digital literacy focuses on the contextual application of digital tools. AI literacy further incorporates additional core connotations. These include understanding AI principles, critical usage of AI, prevention and control of ethical risks, and innovative integration of teaching and research. Such literacy represents an organic combination of technical competence and educational professional competence.

The academic community has reached a general consensus. University teachers' AI literacy is not merely technical operational skills. Instead, it refers to a comprehensive set of competencies. These competencies include understanding fundamental AI principles, assessing the applicability of AI tools, creatively leveraging AI to empower teaching and research, and abiding by relevant regulations to mitigate technical risks. This literacy covers four core dimensions: technology application, teaching innovation, research empowerment, and ethical norms.

In terms of framework construction, international organizations and scholars have developed multiple representative systems. In 2023, the United Nations Educational, Scientific and Cultural Organization (UNESCO) released the AI Competency Frameworks for Teachers (AICFT), which is based on the core concept of "digital humanism". It constructs six literacy domains: AI understanding, AI ethics, AI fundamentals, AI pedagogy, AI tools, and professional learning. It puts forward gradient requirements at three levels: understanding, application, and creation, becoming the world's first systematic AI literacy framework for teachers.

In 2020, researchers proposed a 17-item AI literacy competency framework. This framework is more aligned with the actual technical needs of frontline teachers and provides an important reference for defining basic literacy [2]. Domestically, the China Association for Educational Technology released the Grade Standard for Artificial Intelligence Literacy of University Teachers in 2025. This document classifies teachers' AI literacy into three tiers: basic, intermediate, and expert. It covers four dimensions: fundamental AI knowledge, teaching application, research application, and ethical norms. This standard provides a localized basis for evaluation.

Overall, the framework of teachers' AI literacy has undergone a gradual shift. Its focus has moved from general technological orientation toward educational

professional scenario orientation. Meanwhile, distinctive traits specific to university teachers have become more prominent. These traits include research empowerment and academic ethics.

Existing empirical studies show that the overall AI literacy of university teachers worldwide presents a structural feature of "solid basic ability but insufficient advanced competence". A survey of 2,500 university teachers in the United Arab Emirates by Muammar et al. found that only 32.00% of teachers could use AI tools for personalized instructional design, and less than 20.00% mastered intelligent analysis skills for research data [3].

Domestic studies have also verified this conclusion: A survey showed that only 23.10% of teachers could carry out in-depth processing and innovative application of digital resources, while 73.30% were satisfied with simple copying, modification, and linking. Meanwhile, notable group heterogeneities exist in teachers' AI literacy. Young teachers, science, technology, engineering, and mathematics (STEM) teachers, and teachers with higher academic degrees and professional titles generally achieve higher overall AI literacy levels. In contrast, older teachers, humanities and social science teachers, and teachers from local undergraduate universities show greater room for improvement.

Regarding influencing factors, existing studies have unfolded from two perspectives: promoting factors and barrier factors, based on the technology acceptance model (TAM) and theory of planned behavior (TPB). In terms of promoting factors, perceived ease of use, perceived usefulness, self-efficacy, organizational support, and social support have significant positive effects on teachers' AI adoption intention and behavior, among which self-efficacy is the strongest predictor. Barrier factors can be summarized into three levels: Individual, organizational, and environmental. At the individual level, weak digital awareness, technology anxiety, and insufficient endogenous motivation are major obstacles. At the organizational level, the training system suffers from "emphasizing theory over practice", lack of follow-up guidance, and imperfect incentive mechanisms. At the environmental level, insufficient digital resource supply, imperfect software and hardware facilities, and lack of interdisciplinary collaboration platforms are common issues. These studies provide a basis for understanding the formation logic of teachers' AI literacy but generally ignore the deep regulatory role of trust variables.

Research on human-AI trust in education

Human-AI trust is a core concept in the field of artificial intelligence application. Hancock et al. proposed a multidimensional model of human-robot trust through meta-analysis, including three dimensions: performance-based trust, process-based trust, and purpose-based trust [4]. With the development of generative AI, the connotation of trust has been further expanded, with transparency, reliability, explainability, accountability, and ethical alignment becoming core components. Different from traditional interpersonal trust, human-AI trust is characterized by asymmetry, dynamism, and technology dependence. Moreover, trust and distrust are not completely opposite poles. High trust does not necessarily mean low distrust, which gives rise to the concept of "calibrated trust", emphasizing a rational and appropriate level of trust.

Research on human-AI trust in the educational field is still in its infancy. Lelescu et al. conducted a systematic review of 37 studies between 2019 and 2024, finding that educators' trust in generative AI is rooted in concerns about the "black box" nature of AI and lack of institutional support, and explainability is a key factor affecting trust [5]. Feldman-Maggor et al., taking 41 in-service chemistry teachers as samples, found that explainable AI significantly and positively affects teachers' trust in AI educational technology and adoption intention through "perceived understandability" [6].

Domestically, some scholars have proposed establishing "morality" for educational machines based on humanistic spirit, emphasizing the transparency and credibility of technical interpretation. Others have constructed a five-dimensional trust framework including privacy, security, fairness, explainability, and accountability for evaluating educational AI systems. Overall, however, existing studies mostly focus on the one-way dimension of "human trust in AI". They lack systematic discussion on "AI's trust in teachers' professional judgment" - that is, the system's respect and adaptation to teachers' autonomy. Furthermore, few studies incorporate human-AI trust as a bidirectional regulatory variable into the research framework of teachers' literacy formation.

Research critique

In summary, existing studies have made important

progress in the field of university teachers' AI literacy, forming a relatively mature conceptual framework and evaluation system, and initially exploring the role of human-AI trust in technology adoption. However, there are still three limitations.

First, in terms of research perspective, most studies focus on one-way technology acceptance logic, insufficiently analyze the dynamic regulatory mechanism of trust from "barriers" to "breakthroughs", and especially lack systematic research on bidirectional trust closed loops.

Second, in terms of theoretical integration, most domestic studies are descriptive analyses, lacking integrated theoretical frameworks and empirical tests of regulatory effects.

Third, in terms of practical relevance, existing improvement strategies are mostly macro general suggestions, lacking regional empirical studies targeting leading areas of educational digitalization and targeted pathway design with layered classification.

This study addresses the above gaps, takes "bidirectional regulation of human-AI trust" as the core entry point, uses PLS-SEM to empirically test the regulatory effect of trust in the formation of teachers' AI literacy, and extracts layered and classified development pathways "from barriers to breakthroughs".

Core theoretical foundations

Proposed by Venkatesh et al., the UTAUT model integrates eight major theories in the field of technology acceptance [7]. It includes four core variables: Performance expectancy, effort expectancy, social influence, and facilitating conditions as well as moderating variables such as gender, age, and experience, and has been widely applied in educational technology adoption research.

Based on the UTAUT model, this study adds two extended variables: Personal innovativeness and perceived risk. It also introduces bidirectional human-AI trust as a mediating and moderating variable to construct an extended model adapted to the scenario of university teachers' AI literacy.

Human-AI trust theory holds that trust is the core mediating variable in human-technology system interaction, affecting both technology adoption intention and the depth of use and innovative behavior. From the bidirectional trust perspective, trust includes two directions: One is human trust in technology, that is, users' subjective belief in the reliability, security, and ethics of the technical system; the other is technology's adaptive trust in humans, that is, the degree to which the system identifies, adapts to, and respects users' needs. The two form a dynamic closed loop that jointly affects the depth and effectiveness of technology application.

Teacher professional development theory emphasizes that the improvement of teachers' competence is a continuous, dynamic, and staged evolutionary process, jointly influenced by individual cognition, organizational environment, policies and institutions, and other factors [8]. University teachers have triple professional functions: teaching, research, and social service. Their AI literacy development presents staged and differentiated characteristics, requiring differentiated pathways designed to match the development needs of different stages.

Research model and hypotheses

Core concept definition

This study defines university teachers' AI literacy as a comprehensive competence. It refers to teachers' capacity to understand the basic principles of artificial intelligence. Teachers can reasonably and creatively deploy AI tools to accomplish work tasks in teaching, research and educational contexts. Meanwhile, they follow ethical norms and guard against technical risks. It specifically includes four dimensions: basic AI cognition, AI literacy in teaching application, AI literacy in research empowerment, and AI ethics and safety literacy.

Bidirectional human-AI trust in this study refers to the bidirectional trust relationship formed in the interaction between teachers and AI systems, including two dimensions and eight sub-indicators. Teachers' trust in AI covers teachers' subjective belief in AI systems, including competence trust, emotional trust, ethical trust, and risk perception (reverse scored). AI's adaptive trust in teachers covers the degree of adaptation and respect of AI systems to teachers' needs, including functional adaptation trust, personalized guidance trust, hierarchical authority trust, and risk early warning trust.

Theoretical model construction

Based on the extended UTAUT model, combined with human-AI trust theory and teacher professional development theory, this study constructs an integrated theoretical model of "antecedent variables - bidirectional human-AI trust - teachers' AI literacy". In the model,

performance expectancy, effort expectancy, social influence, facilitating conditions, personal innovativeness, and perceived risk are antecedent variables. Bidirectional human-AI trust is the mediating and moderating variable, and university teachers' AI literacy is the outcome variable.

Research hypotheses

Building on the theoretical framework above, this study proposes the following research hypotheses. Regarding the direct effects of UTAUT core variables on teachers' AI literacy: (1) Performance expectancy reflects the degree to which teachers believe AI can improve work performance, and higher perceived usefulness drives active learning and application of AI. (2) Effort expectancy refers to the perceived ease of using AI technology, and lower adoption thresholds facilitate mastery of relevant skills. (3) Social influence captures perceived support from surrounding groups, and group norms promote literacy improvement. (4) Facilitating conditions refer to organizational and technical environmental support, which provide basic conditions for literacy development. (5) Personal innovativeness reflects openness to new technologies, and innovative individuals explore AI tools more actively; perceived risk refers to concerns about potential risks of AI application, and higher risk perception inhibits AI use and literacy improvement. Accordingly:

H_1: Performance expectancy has a significant positive effect on university teachers' AI literacy.

H_2: Effort expectancy has a significant positive effect on university teachers' AI literacy.

H_3: Social influence has a significant positive effect on university teachers' AI literacy.

H_4: Facilitating conditions have a significant positive effect on university teachers' AI literacy.

H_5: Personal innovativeness has a significant positive effect on university teachers' AI literacy.

H_6: Perceived risk has a significant negative effect on university teachers' AI literacy.

Regarding the mediating effect of bidirectional human-AI trust: External factors such as performance expectancy and facilitating conditions improve teachers' trust in AI, which in turn enhances usage intention and depth, and ultimately improves AI literacy. Meanwhile, factors such as effort expectancy and personal innovativeness affect the degree of AI system adaptation to teachers, forming positive feedback that further promotes literacy improvement.

H_7: Bidirectional human-AI trust plays a mediating role between UTAUT core variables and university teachers' AI literacy.

Regarding the closed-loop regulatory effect: AI's adaptive trust in teachers positively moderates the relationship between teachers' trust in AI and AI literacy. When AI systems accurately adapt to teachers' needs and respect professional autonomy, teachers' trust in AI is more easily transformed into actual usage behavior and ability improvement; poor adaptability weakens this transformation even with high trust.

H_8: AI's adaptive trust in teachers plays a positive moderating role between teachers' trust in AI and university teachers' AI literacy.

Research design and methodology

Questionnaire design and variable measurement

This study uses a questionnaire survey to collect data. All variables adopt mature scales, which are localized and revised in combination with university teachers' professional scenarios, and are measured using a 5-point Likert scale (1= strongly disagree, 5= strongly agree). The questionnaire consists of four modules with a total of 58 items. The university teachers' AI literacy scale, revised based on provincial frameworks and the UNESCO AICFT framework, includes 16 items across four dimensions. The bidirectional human-AI trust scale, compiled based on Hancock's model and related domestic research, includes 16 items across two subscales. The UTAUT core variables scale, adapted from the classic scale by Venkatesh et al., includes 22 items covering six dimensions. Demographic variables include gender, age, teaching experience, professional title, education level, discipline category, and university type, used for group difference analysis and multi-group testing.

Pilot survey and scale validation

Before the formal survey, 3 experts in the field of educational technology were invited to complete content validity testing and optimize item wording. Subsequently, a small-sample pilot survey was conducted among teachers from 3 different types of universities in Zhejiang Province, with 112 valid questionnaires collected. Item optimization was completed through exploration factor analysis and reliability testing. The final overall Cronbach's α coefficient of the scale was 0.87, and the

reliability of each dimension was greater than 0.7, with acceptable structural validity, forming the formal survey questionnaire.

Formal survey and sample selection

A stratified sampling method was adopted for formal surveys. The surveys covered all types of universities in Zhejiang Province. These universities include Double First-class universities, provincial undergraduate universities, private undergraduate universities, and higher vocational colleges. They also span a full range of disciplines: humanities, science, engineering, agriculture, medicine, art, and physical education.

Questionnaires were distributed through a combination of online Wenjuanxing platform and offline channels in cooperation with university teacher development centers and academic departments. A total of 1,726 questionnaires were collected. Data cleaning was completed by setting logical check questions, response time threshold (less than 120 seconds considered invalid), extreme value elimination, and missing value processing. Finally, 1,519 valid samples were obtained, with an effective recovery rate of 88.00%.

The demographic characteristics of the sample are evenly distributed. Males account for 46.20% of respondents, while females account for 53.80%. In terms of teaching experience, 28.30% have 5 years or less, 41.70% have 6-15 years, and 30.00% have more than 16 years. For disciplinary backgrounds, 38.20% come from humanities and social sciences, 45.60% from science and engineering, and 16.20% from arts, physical education, agriculture and medicine. Regarding institutional types, 21.40% are from Double First-class universities, 48.70% from ordinary undergraduate universities, and 29.90% from higher vocational colleges. The sample structure is basically consistent with the overall characteristics of university teachers in Zhejiang Province, with good representativeness.

Data analysis methods

This study uses SmartPLS 4.0 software to construct PLS-SEM for data analysis. The analysis process includes seven steps: Common method bias testing via Harman's single-factor test. Measurement model testing via Cronbach's α, composite reliability (CR), average variance extracted (AVE), and heterotrait-monotrait ratio of correlations (HTMT) ratio. Structural model testing via path coefficients and significance tests. Mediating effect testing via the Bootstrap method with 5,000 resamples. Moderating effect testing via interaction terms. Multi-group analysis across teaching experience, discipline, and university type and robustness testing via variable replacement and sample winsorization.

Empirical analysis and results

Common method bias testing

Harman's single-factor test was conducted on all items with unrotated principal component analysis. The results show that the first factor explains 32.47% of the variance, which is below the critical value of 40.00%, indicating that there is no serious common method bias problem in this study.

Descriptive statistics and group difference analysis

In terms of descriptive statistics, the overall mean score of AI literacy among university teachers in Zhejiang Province is 3.42, which is at the medium to upper level. The scores of each dimension from high to low are basic AI cognition (3.78), AI ethics and safety literacy (3.51), AI literacy in teaching application (3.36), and AI literacy in research empowerment (3.03). The overall presentation shows a structural feature of "solid basic cognition but insufficient advanced application", especially the research empowerment ability has obvious shortcomings, verifying the realistic judgment of "high cognition but low practice".

For bidirectional human-AI trust, the overall mean score is 3.27, among which the mean score of teachers' trust in AI is 3.35, and the mean score of AI's adaptive trust in teachers is 3.19. The latter is significantly lower than the former, indicating that the insufficient adaptation of current AI systems to teachers' professional needs is the core pain point of bidirectional trust imbalance.

Group difference analysis via independent samples t-test and one-way analysis of variance (ANOVA) reveals clear heterogeneities. Teaching experience presents an inverted U-shaped relationship with AI literacy: teachers with 6-15 years of teaching experience have the highest scores, followed by those with 5 years or less, and those with more than 16 years have the lowest. Science, technology, engineering, and mathematics (STEM) teachers have significantly higher total AI literacy scores than humanities and social science teachers and art, physical education, agriculture and medicine teachers, especially with the largest gap in the research empowerment dimension. Teachers in Double First-class

universities score significantly higher than those in ordinary undergraduate universities and higher vocational colleges, while higher vocational teachers perform fairly well in the teaching application dimension but have obvious shortcomings in research empowerment.

Measurement model testing

The measurement model shows good psychometric properties. The Cronbach's α coefficients of all variables range from 0.78 to 0.91, and CR ranges from 0.82 to 0.93, both above the critical value of 0.7, indicating good internal consistency reliability. The average variance extracted (AVE) of all variables ranges from 0.54 to 0.68, all above 0.5, and all factor loadings are greater than 0.6 and significant at the 0.001 level, indicating good convergent validity. The HTMT values between all variables are less than 0.85, and the square root of AVE is greater than the correlation coefficient between corresponding variables, indicating good discriminant validity.

Structural model testing and hypothesis verification

The Bootstrap method (5,000 resamples) was used to test the path coefficients and significance of the structural model. The results are shown in Table 1. The R^2 of the model is 0.623, indicating that the model can explain 62.30% of the variance in teachers' AI literacy, with good explanatory power.

All six direct effect hypotheses are verified: Performance expectancy, effort expectancy, social influence, facilitating conditions, and personal innovativeness all have significant positive effects on university teachers' AI literacy, while perceived risk has a significant negative effect. Among them, performance expectancy has the largest influence coefficient, indicating that perceived usefulness is the core driving force for the improvement of teachers' AI literacy.

Table 1. Path coefficients and hypothesis testing results of the structural model.

Path	Path coefficient	t-value	p-value	Hypothesis
Performance expectancy → Teachers' AI literacy	0.187	5.243	***	Supported (H_1)
Effort expectancy → Teachers' AI literacy	0.152	4.106	***	Supported (H_2)
Social influence → Teachers' AI literacy	0.124	3.579	***	Supported (H_3)
Facilitating conditions → Teachers' AI literacy	0.168	4.872	***	Supported (H_4)
Personal innovativeness → Teachers' AI literacy	0.141	4.025	***	Supported (H_5)
Perceived risk → Teachers' AI literacy	-0.103	2.984	**	Supported (H_6)

Note: ** indicates $p<0.01$, *** indicates $p<0.001$.

Mediating and moderating effect testing

Mediating effect testing shows that the mediating effect value of bidirectional human-AI trust is 0.215, with a 95.00% confidence interval of [0.168, 0.267], which does not include 0, indicating a significant mediating effect. Among them, the mediating effect of bidirectional trust accounts for 34.20% between performance expectancy and AI literacy, and 41.70% between perceived risk and AI literacy, indicating that external factors largely affect teachers' literacy through the trust mechanism. Hypothesis H_7 is supported.

Further dimensional differentiation shows that the mediating effect of teachers' trust in AI is greater than that of AI adaptive trust, indicating that at the current stage, human trust in technology is still the main conduction path, but the chain mediating effect of AI adaptive trust is also significant.

Moderating effect testing shows that the path coefficient of the interaction term is 0.092, with a t-value of 2.763 and $p<0.01$, indicating a significant moderating effect. Simple slope analysis shows that when the level of AI adaptive trust is high, the positive effect of teachers' trust in AI-on-AI literacy is stronger (slope =0.324). When the level of AI adaptive trust is low, the positive effect weakens (slope =0.187). Hypothesis H_8 is supported, verifying the closed-loop regulatory effect of bidirectional trust.

Multi-group analysis and robustness testing

Multi-group analysis across teaching experience, discipline, and university type further reveals heterogeneous mechanisms. Among young teachers, personal innovativeness has a greater impact; among middle-aged teachers, performance expectancy and social influence play a more significant role; among older teachers, facilitating conditions and perceived risk have the largest influence on coefficients. Among STEM teachers, performance expectancy and effort expectancy have stronger effects; among humanities and social science teachers, social influence and ethical trust play a more prominent role. Among Double First-class

university teachers, personal innovativeness and research empowerment needs have greater impacts. Among higher vocational college teachers, facilitating conditions and training support have more significant effects [9].

Robustness testing was conducted in two ways: Replacing variable measurement methods with dimension means, and winsorizing the sample at the upper and lower 1.00%. The results show that there are no substantial changes in the direction and significance of core path coefficients, indicating that the conclusions of this study have good stability and reliability.

Discussion and practical implications

Discussion of main findings

The overall AI literacy of university teachers in Zhejiang Province is at the medium to upper level, but the structural imbalance problem is prominent: Basic cognition is widely popularized, but the depth of teaching application is insufficient, research empowerment ability is weak, and ethical literacy, though cognized, is insufficiently implemented.

This is consistent with relevant research conclusions at home and abroad and also fits the development stage characteristics of Zhejiang Province as a leading province in educational digital reform - technical facilities and training coverage have achieved results, but there is still much room for improvement in deep application and innovative ability.

Bidirectional human-AI trust exhibits an imbalanced characteristic: Humans hold high trust in AI yet trust in the technology itself remains low. Teachers generally acknowledge the technical performance of AI, while their concerns over risks are still prominent. Meanwhile, the poor adaptation of AI systems to teachers' professional needs constitutes the core breakdown point of the current bidirectional trust loop. This finding explains the practical dilemma of "why lots of training but poor effectiveness" - it is not enough to only improve teachers' technical cognition and trust. Technology-side adaptation optimization is equally important. Only the coordinated improvement of bidirectional trust can break through barriers.

This study verifies the applicability of the extended UTAUT model in the scenario of university teachers' AI literacy. All six core variables significantly affect teachers' AI literacy, among which performance expectancy is the strongest driving factor, indicating that "usefulness" is still the core motivation for teachers to actively improve AI literacy.

More importantly, this study reveals the dual role of bidirectional human-AI trust. On the one hand, trust plays an important mediating role between external factors and literacy. External support needs to be transformed into internal motivation through trust, which explains "why many training investments cannot be converted into literacy improvement" - training that does not touch trust reconstruction can only stay at the cognitive level and is difficult to transform into deep application behavior. On the other hand, AI adaptive trust plays a positive moderating role, forming a bidirectional closed loop. The higher the technology adaptability, the higher the conversion efficiency of teachers' trust into literacy. This finding breaks through the one-way trust research paradigm and proves the importance of bidirectional interaction.

Multi-group analysis reveals differences in the literacy formation mechanism among different groups, rooted in different groups' professional needs, technical foundations, and resource acquisition capabilities. Young teachers have high technology acceptance but insufficient professional experience, relying more on individual exploration. Middle-aged teachers are in the golden period of career development, value the effect of technology on performance improvement more, and are more susceptible to group norms. Older teachers have weak technical foundations, are more sensitive to risks, and rely more on external environmental support [10]. STEM teachers have more research scenarios and high requirements for technical performance. Humanities and social science teachers pay more attention to social identity and ethical norms. Differences among university types of stems from different school-running orientations and resource endowments. These heterogeneous characteristics provide a basis for targeted and classified policy implementation.

Theoretical contributions

Theoretically, this study makes three main contributions. First, this research breaks the traditional one-way trust research paradigm. It innovatively adopts the perspective of "bidirectional human-AI trust". The study integrates two components into one unified analytical framework: teachers' trust in AI and AI's adaptive trust in teachers. It identifies the bidirectional closed-loop regulatory

mechanism. Moreover, it extends the research boundary of educational human-computer interaction. Second, this study extends the classic UTAUT model. On this basis, it builds an integrated analytical framework. This improvement enhances the explanatory power of technology acceptance theory within smart education scenarios. It also offers a new analytical paradigm for future research. Third, this study builds a five-level empirical chain based on large-sample regional data. The chain consists of descriptive statistics, difference test, path analysis, moderating effect analysis, and robustness test. It achieves in-depth exploration shifting from "correlation" to "causal mechanism". This also makes the research conclusions more scientific and generalizable.

Practical implications and development pathways

Practically, based on empirical conclusions, this study proposes the core logic of "breaking trust barriers first, then improving literacy ability", and constructs a complete practical system of "stage identification - targeted breakthrough - three-dimensional guarantee".

This study takes "bidirectional human-AI trust level" and "overall AI literacy level" as core indicators. Combined with K-means clustering results and verification from qualitative interviews, it divides the evolution of university teachers' AI literacy into three stages. These stages are the cognitive awakening stage, the trial application stage, and the integrated innovation stage. The trial application stage features basic cognition and preliminary trust but single application scenarios, with poor tool adaptability and lack of scenario-based guidance as main obstacles. The integrated innovation stage features high literacy and high trust, with ethical risk management and innovation ability improvement as main challenges.

Targeted breakthrough pathways are designed for each stage. For the cognitive awakening stage, a trust-breaking path focuses on reducing technology anxiety and building basic trust through AI general education, typical case demonstration, and low-threshold experiential training. For the trial application stage, an adaptive empowerment path focuses on improving tool adaptability and strengthening scenario-based application through personalized tool customization, scenario-based practical training, and optimized explainability of AI systems. For the integrated innovation stage, a capability advancement path focuses on improving advanced innovation ability and improving ethical norms through interdisciplinary innovation communities, ethical guideline construction, and calibrated trust cultivation.

A three-dimensional linked practical guaranteed system supports the pathways above. At the individual level, staged self-learning resource packages, teacher AI literacy growth profiles, and self-efficacy enhancement strategies guide teachers to establish correct AI cognition and trust views. At the university level, a five-in-one cultivation system of "trust first - layered training - scenario adaptation - incentive guarantee - dynamic evaluation" is built, supported by improved technical infrastructure and positive incentive mechanisms. At the government level, special policies, provincial training bases, product standardization, and a provincial monitoring and evaluation system jointly promote the sustainable improvement of teachers' AI literacy.

Conclusion

Taking 1,519 university teachers in Zhejiang Province as samples, this study empirically examines the formation mechanism of university teachers' AI literacy under the bidirectional regulation of human-AI trust and draws the following main conclusions.

First, the overall AI literacy of university teachers in Zhejiang Province presents a structural feature of "solid foundation but insufficient advanced competence". The imbalance of bidirectional human-AI trust is the core barrier restricting literacy improvement, among which the insufficient adaptation of AI systems to teachers' professional needs is the key shortcoming.

Second, performance expectancy, effort expectancy, social influence, facilitating conditions, and personal innovativeness all have significant positive effects on teachers' AI literacy, while perceived risk has a significant negative effect. Bidirectional human-AI trust plays a partial mediating role in the above paths, and AI's adaptive trust in teachers exerts a positive moderating effect on the relationship between teachers' trust in AI and AI literacy, forming a bidirectional closed-loop regulatory effect.

Third, there are significant group heterogeneities in the formation mechanism of teachers' AI literacy. Teacher groups with different teaching experience, disciplines, and university types show differences in core influence

paths, providing a basis for layered and classified policy implementation.

Fourth, the evolution of university teachers' AI literacy can be divided into three stages: cognitive awakening stage, trial application stage, and integrated innovation stage. Designing targeted pathways for the core barriers of each stage and building a three-dimensional linked guaranteed system of individual - university - government can effectively achieve the leap from barriers to breakthroughs.

This study still has certain limitations. First, it adopts cross-sectional survey data, which can reveal the mechanism between variables but is difficult to fully verify causal relationships and dynamic evolution laws. Second, it uses self-reported scales for measurement, which may have social desirability bias. Third, the sample focuses on Zhejiang Province, so the generalizability of research conclusions needs further verification. Fourth, intervention experiments have not been carried out to verify the actual effect of the improvement pathways. Future research can be expanded into three aspects. First, carry out longitudinal tracking studies to reveal the dynamic evolution law of teachers' AI literacy and bidirectional trust. Second, adopt quasi-experimental research methods to verify the effectiveness of staged improvement pathways through intervention experiments. Third, expand the sample scope to different regions across the country to enhance the generalizability of conclusions. Fourth, further explore in-depth issues such as AI ethics and academic integrity to improve the theoretical system of teacher professional development in the intelligent era.

Funding

This work was supported by the Provincial Undergraduate Training Program on Innovation and Entrepreneurship (Grant No: S202610345***).

Acknowledgments

The authors would like to thank the experts who provided guidance for the scale design, the universities that supported the survey distribution, and all the teachers who participated in the survey.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Wu, Y., Huang, Y. (2026) Faculty AI literacy in higher education systems: a systematic review and future-adaptive framework. *Interactive Learning Environments*, 1-31.

[2] Tully, S. M., Longoni, C., Appel, G. (2025) Lower artificial intelligence literacy predicts greater AI receptivity. *Journal of Marketing*, 89(5), 1-20.

[3] Muammar, M., Ruqoiyyah, S., Ningsih, N. S. (2023) Implementing the teaching at the right level (TaRL) approach to improve elementary students' initial reading skills. *JOLLT Journal of Languages and Language Teaching*, 11(4), 610-625.

[4] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., Parasuraman, R. (2011) A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5), 517-527.

[5] Lelescu, A., Sava, S., Grosseck, G., Malita, L. (2025) Exploring trust in generative AI for higher education institutions: a systematic literature review focused on educators. *Humanities and Social Sciences Communications*, 12(1), 1961.

[6] Feldman-Maggor, Y., Cukurova, M., Kent, C., Alexandron, G. (2025) The impact of explainable AI on teachers' trust and acceptance of AI EdTech recommendations: the power of domain-specific explanations. *International Journal of Artificial Intelligence in Education*, 35(5), 2889-2922.

[7] Venkatesh, V., Morris, M. G., Davis, G. B., Davis, F. D. (2003) User acceptance of information technology: toward a unified view1. *MIS Quarterly*, 27(3), 425-478.

[8] Edu, N. (2025) Enhancing teachers' professional development for more productive society. *International Journal of Innovative Development and Policy Studies*, 13(1), 197-203.

[9] Yang, C., Guo, R., Cui, Y. (2023) What affects vocational teachers' acceptance and use of ICT in teaching? A large-scale survey of higher vocational college teachers in China. *Behavioral Sciences*, 13(1), 77.

[10] Wood, E., Mueller, J., Willoughby, T., Specht, J., Deyoung, T. (2005) Teachers' perceptions: barriers and supports to using technology in the classroom. *Education, Communication & Information*, 5(2), 183-206.