

PAPER

Design and Optimization of an Intelligent Mobile Interaction System for Adult Education

Lin Chen ,Linping Han  (✉)Baoding Open University,
Baoding, Chinahanlinping66666688@163.com**ABSTRACT**

Mobile learning scenarios in adult education impose mutually constraining technical requirements regarding real-time interaction performance, recommendation accuracy, and data privacy protection. Neither cloud-centric centralized inference nor device-only local computation can adequately satisfy these demands. In this study, a terminal-edge-cloud collaborative intelligent mobile interaction system was proposed. A fuzzy-inference-based context-awareness engine was deployed on the terminal device, performing local fused inference on temporal, spatial, and device-status information to dynamically generate micro-interaction strategies, with inference latency kept below 5 ms. The recommendation model was compressed to 20 MB through knowledge distillation and 8-bit integer (INT8) quantization, enabling a single inference to be completed within 50 ms. At the edge nodes, user short-term interests were maintained via differentially private gradients, while at the cloud side, model parameters were aggregated through federated averaging to achieve global evolution without raw data leaving the devices. This study provides a feasible technical pathway for the design of mobile interaction systems in resource-constrained and privacy-sensitive scenarios.

KEYWORDS

intelligent mobile interaction, adult education, context awareness, federated learning, edge computing, model lightweighting

1 INTRODUCTION

The global market for adult education has experienced sustained expansion, with mobile devices having become the predominant terminal through which adult learners acquire knowledge. Unlike full-time students, adult learners exhibit educational behaviors that are strongly characterized by contextual specificity: individual learning sessions are largely concentrated within durations of five to 20 minutes, learning activities occur frequently across commuting, work-break, and home settings, and the content of learning is often associated with highly sensitive information such as career development, salary levels, and knowledge deficiencies [1, 2].

Chen, L., Han, L. (2026). Design and Optimization of an Intelligent Mobile Interaction System for Adult Education. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(18), pp. 71–85. <https://doi.org/10.3991/ijim.v20i18.63064>

Article submitted 2026-04-28. Revision uploaded 2026-07-25. Final acceptance 2026-07-28.

© 2026 by the authors of this article. Published under CC-BY.

These characteristics impose technical demands on mobile learning systems that substantially exceed those of traditional educational scenarios. Real-time responsiveness is required to be maintained on mobile terminals where computational power, battery life, and network conditions are all constrained, while accurate personalized content adaptation must be achieved without compromising privacy [3, 4]. The construction of a mobile interaction system that is both real-time and intelligent under the dual pressures of resource constraints and privacy preservation has therefore become a core issue that urgently demands attention within the field of educational technology.

Recent advances in mobile computing have provided new pathways toward addressing the aforementioned challenges. Computational capacity has been off-loaded to the network edge through edge computing, by which service latency is effectively reduced [5]; collaborative multi-party training has been enabled via federated learning without exposure of raw data [6]; and efficient deployment of deep models on mobile terminals has been progressively realized through model compression techniques such as knowledge distillation and quantization [7, 8]. However, these technologies have largely been treated as independent optimization measures in existing studies, lacking systematic integration specifically tailored for adult education scenarios. How these three paradigms might be organically combined—such that responsiveness is guaranteed at the terminal side, personalization is enhanced at the edge side, global evolution is accomplished at the cloud side, and user privacy is locked throughout the entire process—remains an open technical problem [9, 10].

Three principal deficiencies have been identified in existing intelligent mobile learning systems. The mechanisms of context awareness are generally shallow: in most systems, contextual information is exploited only along a single dimension, without multidimensional fused modeling of temporal, spatial, and device-status factors, and the inference process relies almost entirely on the cloud, such that once network connectivity weakens or is interrupted, the perception capability ceases to function—a reality that severely diverges from the frequently mobile state of adult learners [3, 4]. The deployment paradigm for recommendation systems confronts inherent dilemmas. The cloud-inference-with-terminal-collection pattern requires continuous uploading of user behavioral data, which proves particularly problematic in adult education scenarios, where learners are highly vigilant about privacy leakage concerning their career trajectories and knowledge weaknesses and where the round-trip latency of cloud-based inference can reach several seconds under poor network conditions, substantially degrading the interactive experience [5, 6]. Although federated recommendation has been explored in general domains, it has not been specifically designed for the content structure of mobile learning or the behavioral rhythms of adult learners [11, 12]. At the system architecture level, a top-level design for terminal-edge-cloud collaboration is lacking: pure terminal-side solutions are constrained by computational power and cannot evolve continuously, while pure cloud-side solutions sacrifice real-time performance and privacy. The functional division and dynamic collaboration mechanisms among the three tiers remain without systematic technical investigation or engineering validation [9, 10].

In response to the above deficiencies, an intelligent mobile interaction system for adult education is proposed and implemented. The main contributions are as follows. A terminal-side context-awareness engine based on fuzzy inference is designed, by which multidimensional contextual fusion and micro-interaction strategy output are accomplished locally, without requiring network communication throughout the entire process. A three-tier collaborative recommendation architecture is constructed, comprising lightweight terminal-side inference, short-term

edge-side modeling, and cloud-side federated aggregation. The terminal model, after compression through distillation and quantization, enables inference within fifty milliseconds; short-term interests are maintained at edge nodes via differentially private gradient updates; and global parameters are aggregated at the cloud side through federated averaging, thereby balancing system evolution with data privacy. A complete prototype system is developed and systematically evaluated through experiments across five dimensions: response latency, resource consumption, recommendation accuracy, privacy protection, and cold-start performance. The remainder of this study is organized as follows: The overall system architecture and design principles are presented in Section 2. The context-awareness engine is detailed in Section 3. The terminal-edge-cloud collaborative recommendation architecture is introduced in Section 4. The experimental setup and results are reported in Section 5. Conclusions and future work are discussed in Section 6.

2 OVERALL SYSTEM ARCHITECTURE

The system adopts a three-layer physical topology consisting of the terminal layer, the edge layer, and the cloud layer. The terminal layer is deployed on the mobile devices of adult learners, hosting the context-awareness module and the lightweight recommendation model. All sensor data acquisition, multidimensional contextual fusion inference, and recommendation computation are performed locally in a closed loop, without incurring any network dependency. The edge layer is composed of computing nodes deployed at the network edge, each serving a geographically defined group of terminals. These nodes are responsible for receiving model gradients or embedding vectors that have been noised with differential privacy from each terminal and for maintaining a time-rolling short-term interest vector for each learner. The cloud layer is equipped with a federated learning server, which does not store any raw user data but solely undertakes the aggregation of local model parameters uploaded from each edge node. Global recommendation models are generated via the federated averaging algorithm and then redistributed to the edge layer. The collaboration among the three layers is governed by the principles of terminal-side real-time responsiveness as the highest priority, edge-side on-demand coordination, and cloud-side periodic aggregation, seeking an optimal balance among response speed, personalization capability, and system evolution. The overall architecture of the terminal-edge-cloud collaborative mobile interaction system for adult education is illustrated in Figure 1.

The data flow of the system is manifested as three closed loops operating at distinct temporal scales. The local inference closed loop runs entirely within the terminal. Sensor data flow sequentially through the context-awareness engine and the lightweight recommendation model, ultimately outputting interaction strategies and recommendation results. The entire loop cycle is controlled within one hundred milliseconds and is completely independent of network communication, constituting the cornerstone of the system's responsiveness. The edge-enhanced closed loop is triggered on demand: when the confidence of the terminal-side model for a given recommendation item falls below a preset threshold, a collaborative inference request is initiated from the terminal to the edge node. The edge node performs supplementary computation based on the user's short-term interest vector and returns the result. The communication frequency of this loop is strictly regulated by the threshold so that unnecessary network overhead is avoided. The global evolution closed loop operates on an hourly or daily cycle. Model updates from the terminals

under its jurisdiction are periodically aggregated by each edge node and uploaded to the cloud. After federated averaging is completed at the cloud, the global model is disseminated downward layer by layer, driving continuous evolution of the system's intelligence, while the raw user data never leaves the terminal devices throughout the entire process. The three closed loops are decoupled from one another in temporal scale while being organically interconnected at the functional level, jointly underpinning the complete operation of the system.

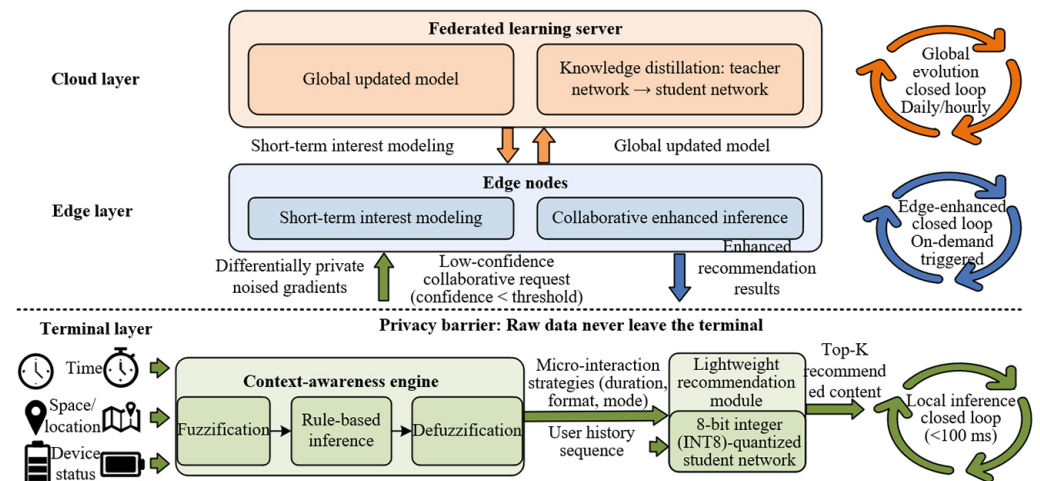


Fig. 1. Overall architecture of the terminal-edge-cloud collaborative mobile interaction system for adult education

3 TERMINAL-SIDE CONTEXT-AWARENESS MICRO-INTERACTION ENGINE BASED ON FUZZY INFERENCE

The learning behavior of adult learners is highly dependent on the real-time context in which they are situated. Significant differences are often observed in content preferences, available time investment, and interaction willingness for the same user between commuting and office scenarios, yet the response of existing mobile learning systems to such contextual changes is typically delayed and coarse-grained. To endow the system with real-time perception and adaptive interaction capabilities, a lightweight context-awareness engine has been designed locally on the terminal device. This engine takes multidimensional contextual information as input and outputs micro-interaction strategies adapted to the current scenario through fuzzy inference, with all computation performed on the terminal side and without reliance on any network communication. The contextual state of an adult learner at time t is denoted as the triplet $C_t = (T_t, S_t, D_t)$. Among the temporal context $T_t = (t_{period}, t_{duration}, t_{frequency})$, t_{period} is the time-period encoding, taking values from {morning, daytime, evening, late night}; $t_{duration}$ is the current continuous learning duration; and $t_{frequency}$ is the cumulative learning frequency for the current day. In the spatial context $S_t = (s_{loc}, s_{noise})$, s_{loc} is the geographic location category, inferred in a lightweight manner through the global positioning system and Wi-Fi fingerprinting, and $s_{noise} \in [0, 1]$ represents the ambient noise level. In the device context $D_t = (d_{battery}, d_{network}, d_{screen})$, $d_{battery} \in [0, 1]$ is the current battery percentage, $d_{network}$ is the network status, and d_{screen} is the screen size. All of the above features are acquired through native sensors and system application programming interfaces on the mobile device, and all raw data are locked locally on the terminal from the moment of acquisition.

Multidimensional contextual information exhibits significant fuzziness and uncertainty. Whether a battery level below 30% qualifies as “low battery” depends on the specific usage scenario, and the noise level during commuting is also a continuously varying quantity rather than a dichotomous variable. To characterize such uncertainty, fuzzy sets are employed to model the membership of continuous contextual variables. Taking battery level $d_{battery}$ as an example, three fuzzy sets—low, medium, and high—are defined, with their membership functions given respectively as:

$$\mu_{low}(x) = \begin{cases} 1, & x \leq 0.15 \\ (0.35 - x)/0.20, & 0.15 < x < 0.35 \\ 0, & x \geq 0.35 \end{cases} \quad (1)$$

$$\mu_{medium}(x) = \begin{cases} 0, & x \leq 0.15 \\ (x - 0.15)/0.20, & 0.15 < x < 0.35 \\ 1, & 0.35 \leq x \leq 0.65 \\ (0.85 - x)/0.20, & 0.65 < x < 0.85 \\ 0, & x \geq 0.85 \end{cases} \quad (2)$$

$$\mu_{high}(x) = \begin{cases} 0, & x \leq 0.65 \\ (x - 0.65)/0.20, & 0.65 < x < 0.85 \\ 1, & x \geq 0.85 \end{cases} \quad (3)$$

The above trapezoidal membership functions smoothly map continuous battery level values onto three fuzzy sets, with the linear transition intervals between 0.15 and 0.35, and between 0.65 and 0.85, reflecting the uncertain boundaries of gradual battery-state transitions. For continuous variables such as $t_{duration}$ in the temporal context and s_{noise} in the spatial context, membership functions of identical structure are employed, with only the threshold parameters adjusted according to the respective physical meanings. For discrete variables such as t_{period} , s_{loc} , and $d_{network}$, singleton membership functions are adopted, whereby the membership degree is assigned a value of 1 when the variable takes the corresponding category value and 0 otherwise. The parameters of all membership functions are pre-stored on the terminal in the form of lookup tables and are directly retrieved during inference without incurring additional computational overhead from real-time calculation.

On the basis of fuzzification, the mapping from context to strategy is accomplished through a set of fuzzy rules. The interaction strategy parameter vector is defined as $P = (p_{duration}, p_{format}, p_{mode})$, where $p_{duration}$ is the adapted duration of the recommended learning unit, p_{format} is the content presentation format, and p_{mode} is the interaction mode. The rule base covers typical combinations across the temporal, spatial, and device dimensions, with each rule taking the following form:

$$R_{ijk} : \text{IF } T_t \in A_i \text{ AND } S_t \in B_j \text{ AND } D_t \in C_k, \text{ THEN } P = (v_{duration}^{ijk}, v_{format}^{ijk}, v_{mode}^{ijk}) \quad (4)$$

where, A_i , B_j , and C_k are the fuzzy sets for each contextual dimension, and $v_{duration}^{ijk}$, v_{format}^{ijk} , and v_{mode}^{ijk} in the consequent part of the rule are the optimal strategy parameters under the corresponding contextual combination. The rule consequents are set according to prior knowledge from studies on adult learning behavior, with the core logic summarized as follows: short-form video or audio formats are assigned to commuting scenarios; longer durations and text-image reading modes are supported in

home scenarios; and offline caching mode is automatically switched under low-battery and poor-network conditions. With three fuzzy sets each for time, space, and device, the rule base comprises a total of 27 rules, covering the primary contextual combinations faced by adult learners. The Mamdani model is adopted for inference. For a given current context C_t , the activation strength of the r -th rule is determined by the minimum of the membership degrees across all dimensions:

$$w_r = \min(\mu_{A_i}(T_t), \mu_{B_j}(S_t), \mu_{C_k}(D_t)) \quad (5)$$

This *min* operator reflects the logic and relationship among the antecedents of a rule, which must be satisfied simultaneously. When w_r takes a value of zero, the rule is considered not activated; the larger the value, the higher the confidence assigned to the consequent of that rule.

The consequents of all activated rules are required to be fused into a final strategy output through a defuzzification step. The weighted average method is adopted, in which the activation strength of each rule is used as the weight to perform a weighted summation of the strategy parameters in the consequents:

$$p = \frac{\sum_{r=1}^R w_r \cdot v_r}{\sum_{r=1}^R w_r} \quad (6)$$

where, R is the total number of rules with activation strength greater than zero, and v_r is the specific value of the corresponding strategy parameter in the consequent of the r -th rule. The weighted average method offers the advantages of simple computation and smooth output, making it suitable for resource-constrained terminal deployment environments. $P_{duration}$ is rounded to discrete levels of 5, 10, 15, or 20 minutes, while p_{format} and p_{mode} are selected from the consequent sets according to the maximum membership principle, by which the category with the highest accumulated membership is chosen. The defuzzified strategy parameter P constitutes the context-adapted interaction strategy for that moment. The entire inference process involves only lookup-table operations, minimum computation, and weighted averaging. The rule base and membership parameters are pre-stored on the terminal in JSON format. Single-inference latency can be controlled within 5 milliseconds, imposing no perceptible impact on the user interaction experience.

The strategy parameters output by the context-awareness engine are fed directly into the recommendation module, such that the subsequent content filtering is constrained not only by user preference modeling but also by the current context. Specifically, the preference score assigned by the recommendation model f_θ to a candidate learning unit $item_j$ is extended as:

$$\hat{y}_{ij} = f_\theta(C_t, H_t, item_j, P_t) \quad (7)$$

where, H_t is the short-term historical interaction sequence of the learner, and P_t is the strategy parameter vector output by the engine at the current moment. This coupling between the context-awareness and recommendation modules ensures that the recommendation results are not only aligned with the user's long-term interest distribution but also matched to the cognitive load capacity and interaction conditions under the current scenario. For example, when the engine detects that the user is commuting and the network is in 4G status, the strategy parameter p_{format}

is biased toward short-form video formats, and $p_{duration}$ is set to five minutes; the recommendation model accordingly prioritizes content items in the candidate set that are compatible with these constraints, thereby achieving a complete closed loop from scene perception to content adaptation. Because the engine operates entirely locally on the terminal, the context-reasoning capability remains fully intact even in environments where network connectivity is unavailable, providing a fundamental guarantee for the sustained learning experience of adult learners in mobile states.

4 TERMINAL-EDGE-CLOUD COLLABORATIVE LIGHTWEIGHT ADAPTIVE RECOMMENDATION ARCHITECTURE

The context-awareness engine addresses the problem of scenario adaptation for interaction strategies, yet the core value of the system in adult education scenarios ultimately depends on the accuracy of content recommendation. The recommendation model is required to perform real-time inference under limited terminal resources while also possessing generalization capability for new users and new content and continuously evolving with usage data. Although centralized cloud-based recommendation offers sufficient model capacity, it is constrained by latency and privacy bottlenecks. A pure terminal-side solution, while guaranteeing real-time performance, is limited by terminal computational power and cannot share knowledge globally. To address these issues, a three-tier collaborative recommendation architecture is proposed, comprising terminal-side lightweight inference, edge-side short-term modeling, and cloud-side federated aggregation, which separately addresses real-time responsiveness, personalization enhancement, and global evolution at the three levels. Let the learning resource repository be denoted as I , and the historical interaction sequence of user u at time t be denoted as $H_t^u = (i_1, i_2, \dots, i_k)$. The recommendation task is defined as selecting, from the candidate set I_{card} , the Top-K learning units that best match the user, under the constraints of the current context C_t^u and the strategy parameters P_t^u .

The recommendation model f_θ deployed on the terminal side directly undertakes the real-time inference task for the user. The model takes as input the contextual features C_t , strategy parameters P_t , historical interaction sequence H_t , and the embedding representation of the candidate item $item_j$, and outputs a preference score:

$$\hat{y}_{ij} = f_\theta(C_t, P_t, H_t, emb(item_j)) \quad (8)$$

To ensure that the model can perform a single inference within 50 milliseconds on mainstream smartphones, a two-stage compression strategy consisting of knowledge distillation and 8-bit integer (INT8) quantization is adopted. First, a teacher network $f_\theta^{student}$, with a parameter count approximately ten times that of the student network, is trained on a large-scale dataset in the cloud. Subsequently, knowledge is transferred to the student network $f_\theta^{student}$ through distillation. The distillation loss function is defined as:

$$L_{distill} = (1 - \alpha) \cdot L_{CE}(y_{true}, f_\theta^{student}(x)) + \alpha \cdot \tau^2 \cdot L_{KL} \left(\text{softmax} \left(\frac{f_\theta^{teacher}(x)}{\tau} \right) \right),$$

$$\text{softmax} \left(\frac{f_\theta^{student}(x)}{\tau} \right) \quad (9)$$

where, τ is the temperature parameter, which is used to soften the output probability distribution of the teacher network so that more dark knowledge can be transferred; in this work, τ is set to 3.0. $\alpha = 0.7$ is the weighting coefficient that balances the two loss terms. L_{CE} denotes the cross-entropy loss, and L_{KL} denotes the Kullback–Leibler divergence. After distillation, post-training quantization is performed on the student network, by which the 32-bit floating-point weights are quantized into 8-bit integers:

$$W_{int8} = \text{round} \left(\frac{W_{fp32} - \mu}{\sigma} \cdot 127 \right) \quad (10)$$

where, μ and σ are the mean and standard deviation of the weights to be quantized, respectively. This quantization strategy reduces the model size by approximately 76% and accelerates inference speed by a factor of three to four. After compression, the model size is controlled within 20 MB, enabling a single inference latency below 50 milliseconds on mid-to-low-end Android devices.

The lightweight model on the terminal side, constrained by its parameter count, has limited capacity to respond to short-term fluctuations in user interests. For example, when a learner develops a concentrated focus on project management topics due to a temporary work project, such an interest shift is difficult to capture rapidly using only the terminal-side model. The edge computing node here assumes the function of short-term interest modeling. Raw interaction logs are not uploaded from the terminal side; instead, model gradients, after being noised with differential privacy, are uploaded to the edge node. Taking the gradient update as an example, the noising mechanism is given as:

$$\tilde{g} = g + N(0, \sigma_{dp}^2 \cdot \mathbf{I}) \quad (11)$$

where, $N(0, \sigma_{dp}^2 \cdot \mathbf{I})$ denotes isotropic Gaussian noise, and σ_{dp} is jointly determined by the privacy budget ϵ and the gradient sensitivity Δ . In this work, ϵ is set to 2.0 to achieve a balance between privacy protection strength and model utility. A short-term interest vector $I_t^{short} \in \mathbb{R}^d$ is maintained at each edge node for each learner, with online incremental updates performed as follows:

$$I_{t+1}^{short} = \beta \cdot I_t^{short} + (1 - \beta) \cdot g(\text{interaction}_t) \quad (12)$$

where, $\beta \in (0, 1)$ is the decay factor, which controls the forgetting rate of historical interests, and is set to 0.8 in this work; $g(\cdot)$ is the encoding function that maps a single interaction behavior into an incremental vector in the interest space. When the maximum prediction confidence of the terminal-side model across all candidate items falls below a threshold δ , i.e., $\max_j \hat{y}_j < \delta$, a collaborative inference request is initiated from the terminal side to the edge node. The edge node performs supplementary computation based on I_t^{short} and returns the enhanced recommendation result. This on-demand collaboration mechanism strictly controls the terminal-edge communication frequency to boundary cases where confidence is insufficient, thereby obtaining the benefit of edge modeling capability while avoiding the network and energy overhead caused by continuous communication.

At the cloud layer, no raw user data is stored. The core function of the cloud layer is to aggregate model updates from each edge node via federated learning, so that continuous evolution of the global model is driven. Each edge node e periodically

uploads its local model parameters θ_e to the cloud. The global model is aggregated in the cloud using the federated averaging algorithm:

$$\Theta_{new}^{global} = \sum_{e=1}^E \frac{n_e}{N} \cdot \theta_e \quad (13)$$

where, E is the total number of edge nodes, n_e is the number of users covered by the e -th edge node, and $N = \sum_{e=1}^E n_e$ is the total number of users in the entire system.

This aggregation is performed as a weighted average weighted by user count, ensuring that model updates from large-scale nodes are assigned an appropriate proportion in the global model. The aggregated global model Θ_{new}^{global} is distributed back to each edge node, and after distillation and compression at the edge nodes, it is further delivered to the terminal devices under their jurisdiction. The above upload and distribution processes are executed on a daily cycle, with communication overhead fully controllable. Throughout this evolutionary closed loop, the cloud only accesses aggregated model parameters from the edge, and the edge nodes only receive gradient information that has been noised with differential privacy; the raw interaction data remain locked on the terminal devices from the beginning to the end, thereby eliminating, at the architectural level, the privacy risks associated with centralized data collection.

A rational division of labor across temporal scales is formed by the three-tier collaboration mechanism. The local inference closed loop operates on a cycle of hundreds of milliseconds, ensuring immediate responsiveness for every interaction. The edge-enhanced closed loop is triggered on demand, intervening only when terminal-side confidence is insufficient, thereby balancing personalization gains with communication efficiency. The global evolution closed loop is executed on a daily basis, driving continuous model optimization without interfering with the normal operation of online services. The strategy parameters P_t output by the terminal-side context-awareness module serve as input features to the recommendation model f_θ , enabling content filtering to be guided simultaneously by the user's long-term preferences and current scenario constraints. The edge-side short-term interest vector I_t^{short} provides enhancement signals when terminal-side confidence is inadequate, compensating for the expressive limitations of the lightweight model. Cloud-side federated aggregation ensures that the entire system continuously evolves as the user base grows. These three components are organically connected, jointly constituting a comprehensive recommendation solution tailored for adult education scenarios.

5 EXPERIMENTAL EVALUATION

5.1 Experimental setup

The experiments were conducted using real usage logs from an adult online learning platform, covering the period from September 2024 to February 2025. After anonymization, the dataset comprised 186,432 interaction records from 1,247 learners. The learners were aged between 24 and 47 years, with professional backgrounds spanning information technology, finance, education, and manufacturing. The interaction records included fields such as learning content identifier, learning duration, timestamp, anonymized geographic location, and device model. The dataset was partitioned into training, validation, and test sets at a ratio of 8:1:1, with time windows split by day to prevent data leakage.

Four types of schemes were selected as baseline comparisons. Cloud-only is a traditional cloud-centric centralized scheme in which all inference is performed in the cloud and the terminal side is only responsible for data collection and interface rendering. Edge-only places all inference at the edge nodes, with no model deployed on the terminal side. Local-only is a pure terminal-side scheme without edge collaboration or federated updates. No-context removes the context-awareness module from local-only and performs recommendations based solely on historical interactions, serving to quantify the independent contribution of contextual information. The evaluation dimensions covered response latency, terminal-side resource consumption, recommendation accuracy, cold-start performance, and communication efficiency. Three tiers of terminal devices were used in the tests: the high-end device was the Google Pixel 8 Pro, the mid-range device was the Xiaomi 12, and the low-end device was the Redmi Note 11. The edge nodes were configured with Intel Xeon E-2288G processors and 32 GB of memory, while the cloud server was equipped with an NVIDIA A100 graphics processing unit.

5.2 End-to-end response latency comparison

Response latency directly determines the smoothness of mobile interaction. Experiments were conducted under four network conditions—Wi-Fi, 5G, 4G, and poor network—to measure the end-to-end latency of each scheme from user interaction trigger to interface rendering completion. Each scheme was tested 1,000 times repeatedly, and the P50, P95, and P99 latencies were recorded.

Table 1. End-to-end response latency comparison under different network conditions (ms)

Scheme	Wi-Fi			5G			4G			Poor Network		
	P50	P95	P99	P50	P95	P99	P50	P95	P99	P50	P95	P99
Cloud-only	124	187	245	142	218	289	218	374	512	683	1236	1854
Edge-only	78	115	156	89	134	182	135	208	287	345	612	893
Local-only	42	68	91	43	70	94	45	74	98	46	79	106
No-context	41	66	89	42	68	91	44	72	96	45	77	103
Proposed method	43	69	93	44	71	95	46	75	100	47	81	109

As shown in Table 1, the cloud-only scheme achieves a P95 latency of 187 ms under Wi-Fi, which is still within an acceptable range; however, under 4G networks, this latency rises to 374 ms, and under poor network conditions, it reaches 1,236 ms. In this latency composition, network round-trip transmission is the dominant factor. Edge-only shifts inference to the network edge, significantly shortening the physical path reducing the poor-network P95 to 612 ms, yet the impact of network fluctuation still persists. For local-only and no-context, because all inference is performed locally on the terminal, the network state introduces almost no perturbation to their response latency. The proposed scheme, due to the additional execution of the context-awareness engine, exhibits a slight increase in latency compared to local-only. Under poor network conditions, the P95 latency is 81 ms, representing a 93.4% reduction compared to cloud-only, and only a 4 ms increase compared to local-only. This increment has no perceptible impact on user experience, whereas the accuracy gain brought by contextual information is substantially significant.

5.3 Terminal-side resource consumption evaluation

Resource consumption determines the deployment feasibility of the system on mainstream devices. Each scheme was run on three tiers of Android devices, with memory usage, central processing unit utilization, and battery consumption continuously monitored by Android Profiler over a two-hour simulated usage scenario.

Table 2. Terminal-side resource consumption comparison (two-hour simulated usage)

Scheme	High-End Device			Mid-Range Device			Low-End Device		
	Memory (MB)	Central Processing Unit (%)	Battery (mAh)	Memory (MB)	Central Processing Unit (%)	Battery (mAh)	Memory (MB)	Central Processing Unit (%)	Battery (mAh)
Cloud-only	176	18.4	342	163	21.7	387	158	24.3	421
Edge-only	89	12.1	204	94	14.6	241	102	17.2	278
Local-only	124	15.3	267	131	18.1	302	147	22.6	356
No-context	96	10.7	198	103	13.2	234	118	17.4	287
Proposed method	118	14.8	254	126	17.5	291	142	21.8	344

As shown in Table 2, the cloud-only scheme, with only the collection and rendering modules running on the terminal side, might be expected to exhibit the lowest memory and central processing unit usage; however, the continuous overhead of network communication and data serialization results in the highest battery consumption instead, reaching 421 mAh on the low-end device, which is 22.4% higher than that of the proposed scheme. This indicates that the energy consumption of the wireless communication module cannot be neglected in mobile scenarios. Edge-only achieves the lowest resource consumption after offloading the computational load, yet the previous latency experiments have already revealed its availability shortcomings under poor network conditions. No-context is the most resource-economical scheme, but its recommendation accuracy is limited. Due to the deployment of the context-awareness engine and the lightweight recommendation model, the proposed scheme exhibits a moderate increase in resource consumption compared to no-context, with memory usage increased by 24 MB and battery consumption increased by 57 mAh on the low-end device, which remains within the acceptable range for mainstream devices. On mid-range and high-end devices, central processing unit utilization is controlled at 17.5% and 14.8%, respectively, without interfering with the normal operation of other tasks on the device.

5.4 Recommendation accuracy comparison

Recommendation accuracy serves as the definitive metric for evaluating the core intelligence level of the system. Offline evaluation was conducted on the test set, where Precision@K, Recall@K, and NDCG@K were calculated for each scheme.

As shown in Figure 2, the offline evaluation results exhibit a clear performance gradient. Cloud-only achieves the highest accuracy among the baselines (excluding the proposed scheme), benefiting from the large model deployed in the cloud, with Precision@10 reaching 0.284. The accuracy of edge-only is slightly lower

than that of cloud-only, because edge node resources, while superior to those of terminals, are still insufficient to host a model of comparable scale. Local-only, constrained by model capacity, shows a Precision@10 that is 14.4% lower than that of cloud-only. No-context achieves the lowest accuracy, with Precision@10 only 0.208, indicating that a pure interest model lacking context awareness struggles to cope with the preference drift exhibited by adult learners across different scenarios. The proposed scheme achieves Precision@10 = 0.311, outperforming all baselines, with a relative improvement of 9.5% over cloud-only and 49.5% over no-context. A consistent pattern is observed for NDCG@10; the proposed scheme attains 0.352, which is 10.7% higher than cloud-only's 0.318, indicating that the top-ranked recommendation results are more relevant and that the ranking quality is superior.

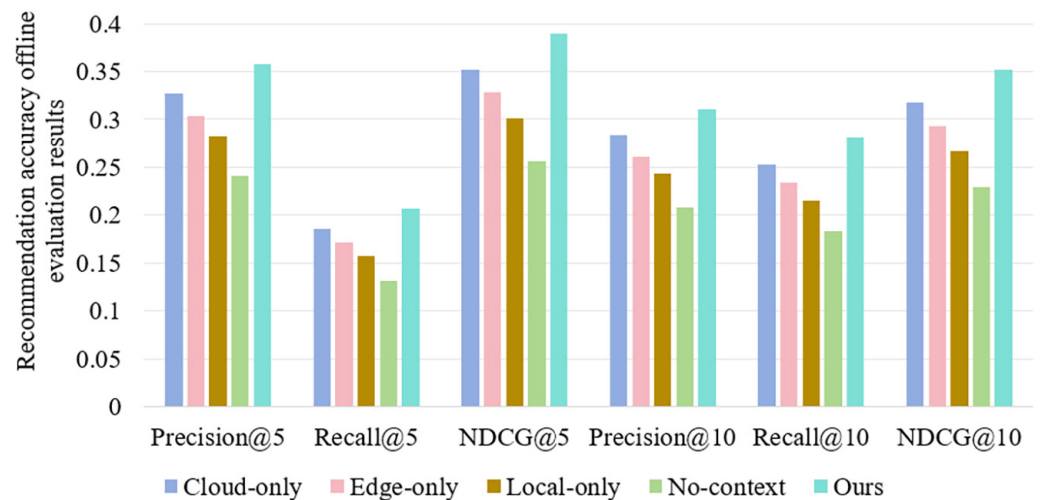


Fig. 2. Offline evaluation results of recommendation accuracy

5.5 Online performance and cold-start evaluation

Online A/B testing randomly assigned real users into five groups, each corresponding to one of the schemes. Over a two-week period, click-through rate, learning completion rate, and average daily usage duration were recorded. For the cold-start experiment, a total of 89 users with fewer than 10 interaction records were screened from the dataset, and the click-through rate of the first 10 recommendations and Precision@10 were evaluated for each scheme.

The online metrics in Table 3 validate the effectiveness of the offline conclusions in real-world environments. The proposed scheme achieves a click-through rate of 21.4%, which is 2.7 percentage points higher than that of cloud-only. The learning completion rate reflects the degree of match between the recommended content and the actual needs of users; the proposed scheme ranks first with 48.6%, compared to 42.3% for cloud-only, with the gap sustained throughout the two-week test period. The difference in average daily usage duration is even more pronounced, with the proposed scheme reaching 22.7 minutes, while no-context is only 12.1 minutes. This is attributed to the context-awareness engine, which proactively pushes appropriately matched content during fragmented time periods, thereby reducing the decision cost for users from launching the application to actually engaging in learning.

In the cold-start scenario, the no-context scheme effectively degrades to random recommendation, with a cold-start CTR@10 as low as 8.3%. Cloud-only, lacking personalization signals, achieves a cold-start CTR@10 of 14.6%. The proposed scheme attains a cold-start CTR@10 of 18.2% and a Precision@10 of 0.214, with the advantage derived directly from the context-awareness engine: when historical user preferences are absent, the current context becomes the only available signal, and the strategy parameters P_t provide scenario constraints to the model, thereby avoiding completely blind recommendations. In subsequent recommendations for cold-start users, as interactions accumulate, the update of the edge-side short-term interest vector I_t^{short} enables the accuracy to approach that of normal users after approximately 15 interactions.

Table 3. Online performance and cold-start scenario evaluation results

Scheme	Click-Through Rate (CTR, %)	Learning Completion Rate (%)	Average Daily Duration (min)	Cold-Start CTR@10 (%)	Cold-Start Precision@10
Cloud-only	18.7	42.3	18.4	14.6	0.168
Edge-only	16.9	38.7	16.2	13.1	0.152
Local-only	15.8	36.1	14.7	11.8	0.134
No-context	13.2	31.5	12.1	8.3	0.097
Proposed method	21.4	48.6	22.7	18.2	0.214

6 CONCLUSION AND FUTURE WORK

In response to the technical challenges imposed by the mutually constraining requirements of real-time responsiveness, accurate recommendation, and privacy preservation in mobile learning scenarios for adult education, a terminal-edge-cloud collaborative intelligent mobile interaction system was proposed and implemented. At the terminal level, a context-awareness engine based on fuzzy inference was designed, through which millisecond-level dynamic output of micro-interaction strategies was accomplished via multidimensional fused inference on temporal, spatial, and device-status information, enabling the system to perceive contextual changes and adapt interaction modalities adaptively even without network dependency. At the recommendation level, a three-tier collaborative architecture was constructed, comprising terminal-side lightweight inference, edge-side short-term modeling, and cloud-side federated aggregation. At the edge nodes, user short-term interests were maintained via differentially private gradients to compensate for the limited expressive capacity of the terminal-side model; at the cloud side, continuous evolution of the global model was achieved through federated averaging, while the raw user data never left the terminals throughout the entire process. Future work will explore the application of lightweight Transformers to terminal-side context modeling so that contextual representations can be learned directly from data. Asynchronous federated learning frameworks will be investigated to support deployment on larger-scale heterogeneous terminals. The architecture will also be extended to intelligent recommendation of multimodal content, including video and interactive simulations.

7 REFERENCES

- [1] A. Rof, A. Bikfalvi, and P. Marques, “Exploring learner satisfaction and the effectiveness of microlearning in higher education,” *The Internet and Higher Education*, vol. 62, 2024. <https://doi.org/10.1016/j.iheduc.2024.100952>
- [2] N. Moussa, M. M. Abouelnour, and T. Saali, “Digital library and adult education: Assessing acceptance, utilization, and adaptability challenges in AI-driven and mobile-enabled library services,” *International Journal of Interactive Mobile Technologies*, vol. 19, no. 24, pp. 137–148, 2025. <https://doi.org/10.3991/ijim.v19i24.59111>
- [3] B. A. Kumar and B. Sharma, “Context aware mobile learning application development: A systematic literature review,” *Education and Information Technologies*, vol. 25, no. 3, pp. 2221–2239, 2019. <https://doi.org/10.1007/s10639-019-10045-x>
- [4] C. P. Gumbheer, K. K. Khedo, and A. Bungaleea, “Personalized and adaptive context-aware mobile learning: Review, challenges and future directions,” *Education and Information Technologies*, vol. 27, no. 6, pp. 7491–7517, 2022. <https://doi.org/10.1007/s10639-022-10942-8>
- [5] Y. Li and L. Peng, “Emotion-aware mental health intervention strategies on mobile computing platforms,” *International Journal of Interactive Mobile Technologies*, vol. 19, no. 17, pp. 101–114, 2025. <https://doi.org/10.3991/ijim.v19i17.57857>
- [6] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020. <https://doi.org/10.1109/MSP.2020.2975749>
- [7] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021. <https://doi.org/10.1007/s11263-021-01453-z>
- [8] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, “A comprehensive survey on model compression and acceleration,” *Artificial Intelligence Review*, vol. 53, no. 7, pp. 5113–5155, 2020. <https://doi.org/10.1007/s10462-020-09816-7>
- [9] G. Rong, Y. Xu, X. Tong, and H. Fan, “An edge-cloud collaborative computing platform for building AIoT applications efficiently,” *Journal of Cloud Computing*, vol. 10, no. 1, pp. 1–14, 2021. <https://doi.org/10.1186/s13677-021-00250-w>
- [10] B. A. Kumar, B. Sharma, and E. Y. Nakagawa, “Architectural support for context-aware mobile learning applications,” *Education and Information Technologies*, vol. 27, no. 3, pp. 3723–3741, 2021. <https://doi.org/10.1007/s10639-021-10771-1>
- [11] Q. Wang, H. Yin, T. Chen, J. Yu, A. Zhou, and X. Zhang, “Fast-adapting and privacy-preserving federated recommender system,” *The VLDB Journal*, vol. 31, no. 5, pp. 877–896, 2021. <https://doi.org/10.1007/s00778-021-00700-6>
- [12] M. Asad, S. Shaukat, E. Javanmardi, J. Nakazato, and M. Tsukada, “A comprehensive survey on privacy-preserving techniques in federated recommendation systems,” *Applied Sciences*, vol. 13, no. 10, p. 6201, 2023. <https://doi.org/10.3390/app13106201>

8 AUTHORS

Lin Chen holds a bachelor’s degree and works as a Lecturer at Baoding Open University, focusing on adult education law. She studied at Hebei Normal University from 1998 to 2002 and obtained her Bachelor’s degree in 2002. She has been working at the School of Continuing Education, Baoding Open University, since 2003. She has published five academic papers, including one in a core journal, and presided over three department-level research projects. She has been engaged in adult education

teaching throughout her career. Her research interests cover the reform and intelligent development of adult teaching models (E-mail: chen_lin770617@163.com).

Linping Han holds a bachelor's degree and works as an Associate Professor at the School of Continuing Education, Baoding Open University. Her research directions include food engineering for adult education, tourism, and hotel management. She received her undergraduate education at Hebei Normal University from 1998 to 2002 and earned her Bachelor's degree in 2002. She has been employed by Baoding Open University since 2002. From 2010 to 2013, she pursued a Master of Engineering degree at the College of Food Science and Engineering, Hebei Agricultural University, and graduated in 2013. She has published three papers, one of which was released in a core journal. She has long been committed to adult education and teaching. Her research interests lie in the research and reform of adult teaching models (E-mail: hanlinping66666688@163.com).