Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 20 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i20.63554

PAPER

Human–Artificial Intelligence Collaborative Decision-Making and Error Control in Mobile Accounting Systems

Lan Li(✉)

Hubei University of Education, Wuhan, China

lilan@hue.edu.cn

ABSTRACT

Human–artificial intelligence (AI) collaborative decision-making in mobile accounting systems is constrained by static division-of-labor boundaries and AI confidence bias, making automation efficiency and error control difficult to balance. A dynamic confidence threshold framework was proposed in this study. Confidence was calibrated through Platt scaling, reducing the expected calibration error to 0.031. An adaptive threshold function was constructed by integrating task complexity and historical error rates, and manual review task allocation was optimized by incorporating operator experience. Tasks were dynamically routed into three channels: automatic approval, human–AI collaborative correction, and in-depth manual review. Selective-risk statistical monitoring and drift detection mechanisms were also incorporated. Experimental results on a real-world accounting dataset and a mobile prototype system indicate that the auto-pass error rate was controlled at 4.2%, the automation rate reached 78.6%—an improvement of 13.3 percentage points over the optimal fixed threshold—accountant processing time was reduced by 23.0%, and cognitive load was reduced by 26.8%; favorable steady-state recovery was further exhibited under multiple perturbations. A systematic technical solution is thus provided for human–AI collaborative decision-making in mobile accounting scenarios.

KEYWORDS

human–artificial intelligence (AI) collaborative decision-making, mobile accounting, dynamic threshold, confidence calibration, error control

1 INTRODUCTION

The migration of accounting information systems from desktop platforms to mobile platforms has been accelerated by continuous improvements in the computational capabilities of mobile devices and the widespread deployment of 5G networks [1,2]. Core business processes, including voucher entry, invoice recognition, and online

Li, L. (2026). Human–Artificial Intelligence Collaborative Decision-Making and Error Control in Mobile Accounting Systems. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(20), pp. 155–168. https://doi.org/10.3991/ijim.v20i20.63554

Article submitted 2026-07-16. Revision uploaded 2026-09-16. Final acceptance 2026-09-17.

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

approval, are now supported by mobile accounting systems. However, operational scenarios have expanded from fixed office environments to uncontrolled spaces such as travel routes, and information acquisition has shifted from parallel browsing across multiple windows on large screens to serial touch-based operations on small screens. In addition, mobile scenarios are inherently characterized by limited screen size and variable operating environments [3]. Consequently, the traditional decision-making model—human-led with passive machine responses—is confronted with fundamental challenges. The cognitive load imposed on accountants in mobile settings is significantly increased, information retrieval and cross-verification efficiency are diminished, decision fatigue is accelerated, and error rates rise accordingly [4].

With the evolution of artificial intelligence (AI) technology, machines have gradually been transformed from automation tools into decision-making collaborators capable of autonomous judgment [5]. High accuracy has been achieved by optical character recognition, rule engines, and machine learning models in tasks such as invoice information extraction [6], expense compliance verification, and anomalous transaction detection [7]. However, the decision quality of AI systems is not constant; their confidence is found to fluctuate significantly with task type, data quality, input completeness, and other factors [8]. More fundamentally, raw probability scores output by deep learning models often exhibit systematic bias [9], such that a judgment claimed by a model to have 90% certainty may have an actual accuracy far below that value [10]. In mobile accounting scenarios, convenient desktop-based cross-validation tools are unavailable to accountants, and judgments concerning the correctness of AI outputs are highly dependent on confidence signals provided by the system [11]. If such signals are themselves unreliable, the decision quality of human–AI collaboration cannot be guaranteed.

Substantial progress has been made in existing research on human–AI collaborative decision-making [12], yet three notable deficiencies remain when mobile accounting scenarios are addressed. First, although probability-based intervention thresholds have been introduced into adaptive digital systems to support real-time error prediction and interface control, such thresholds are generally determined under predefined experimental conditions and are not dynamically adjusted according to task complexity and historical error rates [13]. In addition, manual review tasks lack an allocation mechanism matched to operator experience. These limitations are especially pronounced in mobile scenarios characterized by significant task heterogeneity and environmental dynamism [14]. Second, AI confidence calibration remains almost unexplored in accounting human–AI collaboration research. Although mature approaches such as Platt scaling and temperature scaling have been developed in machine learning [9], they have not yet been introduced into human–AI collaborative decision-making frameworks for accounting information systems. The direct use of uncalibrated confidence scores for task allocation gives rise to systematic misallocation risk. Third, although validation and fallback mechanisms have been incorporated into feedback-controlled AI architectures in other digital application domains, error control in accounting information systems still lacks a systematic design spanning the entire human–AI collaboration process [15]. Existing studies have largely reduced it to post hoc audit sampling [16]. From selective risk control for automatic approval to analysis of human–AI collaborative correction records to root-cause induction in in-depth manual review, a closed-loop feedback system covering the full chain has yet to be established.

To address these deficiencies, a mobile accounting human–AI collaborative decision-making and error-control framework based on a dynamic confidence threshold is proposed. In this framework, raw confidence scores of AI models are calibrated by Platt scaling; an adaptive threshold function integrating task complexity

and historical error rates is constructed; manual review task allocation is optimized by incorporating operator experience; and tasks are dynamically routed into three channels: automatic approval, human–AI collaborative correction, and in-depth manual review. Selective-risk statistical monitoring, threshold drift detection driven by the sequential probability ratio test, and a correction-record-driven feedback learning mechanism are introduced, forming a closed-loop error-control system covering the entire decision-making chain. Through experiments on a real-world accounting dataset and a mobile prototype system, the performance of the framework is systematically evaluated across three dimensions: decision performance, user experience, and long-term robustness.

2 METHODOLOGY

2.1 Overview of the system framework

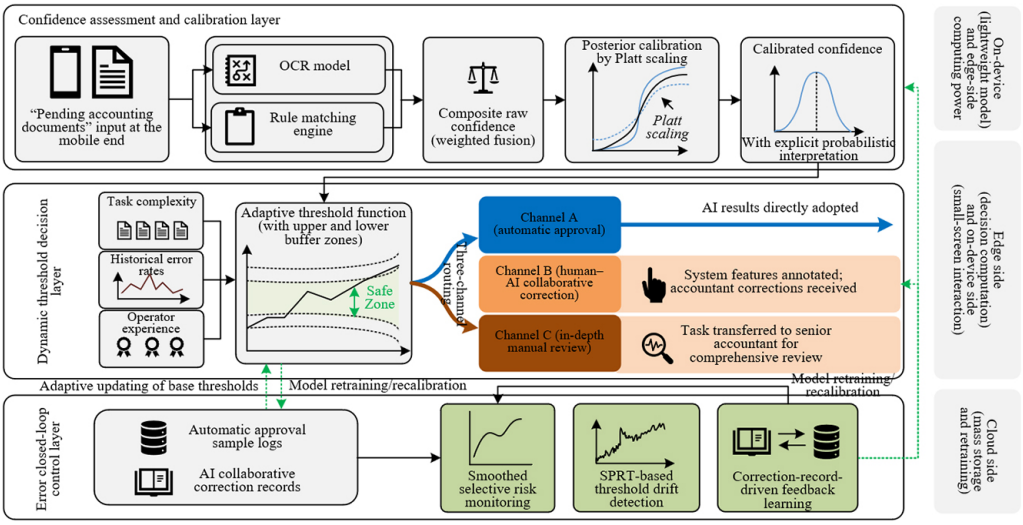


Fig. 1. Overall framework diagram of dynamic-threshold-based human–artificial intelligence collaborative decision-making and error control for mobile accounting

As illustrated in Figure 1, a human–AI collaborative decision-making framework composed of three layers—confidence assessment and calibration, dynamic threshold decision-making, and error closed-loop control—is constructed. In the first layer, for each accounting document, optical character recognition confidence and rule-matching confidence are fused into a composite confidence score, and the raw score is posterior-calibrated by means of Platt scaling, so that a reliable probabilistic interpretation is conferred on the confidence. In the second layer, based on the calibrated confidence, an adaptive threshold is computed in real time by incorporating task complexity and historical error rates; tasks are routed into three channels—automatic approval, human–AI collaborative correction, and in-depth manual review—and manual review tasks are matched according to operator experience. In the third layer, a closed-loop mechanism of decision-making, feedback, and tuning is formed through selective-risk statistical monitoring, sequential probability ratio test-driven threshold drift detection, and correction-record-driven feedback learning. The three layers are tightly coupled through data flows: input for the decision layer is provided by confidence outputs; feedback data for the closed-loop layer are constituted by routing results and manual

review conclusions; and the confidence assessment model and threshold function parameters are revised through feedback from the calibration results of the closed-loop layer so that the system is continuously optimized during sustained operation.

2.2 Confidence assessment and calibration model

For each pending accounting document x, whether the document-level final processing result is consistent with the manual review conclusion is taken as the correctness event, and the composite confidence is used to estimate the probability of this event. A composite confidence score $C_{raw}(x) \in [0, 1]$ is first output by the system, and this score is synthesized by a weighted combination of sub-confidences from two dimensions. $C_{ocr}(x)$ is the weighted average of recognition confidences of key fields such as invoice amount, date, and tax identification number by the optical character recognition module. $C_{rule}(x)$ is the quantified output of the degree of match between the current document and a preset accounting compliance rule base by the rule matching module; it is computed by filtering a validation rule set adapted to the current voucher type from the rule base, and a standardized score is generated after item-by-item comparison of each core element through pattern matching. The two are combined by weighted summation to obtain the raw composite confidence:

$$C_{raw}(x) = \omega_1 C_{ocr}(x) + \omega_2 C_{rule}(x) \tag{1}$$

The weight coefficients ω_1 and ω_2 satisfy $\omega_1 + \omega_2 = 1$ and are dynamically adjusted according to task type: in invoice recognition tasks, the weight of ω_1 is increased, whereas in tasks dominated by rule compliance verification, ω_2 is dominant.

Systematic bias generally exists in the raw confidence scores C_{raw} output by deep learning models, manifesting as overconfidence or underconfidence; if such scores are directly used for human–AI task allocation, systematic misallocation risk will result. Posterior calibration of the raw scores is performed using Platt scaling. To avoid divergence of the logit transformation when C_{raw} is 0 or 1, let $\tilde{C}_{raw} = min\{max(C_{raw}, \varepsilon), 1 - \varepsilon\}$, and then, by solving scaling parameters α and β through maximum likelihood estimation on a calibration dataset, the raw score is mapped to the calibrated confidence:

$$C_{cal}(x) = \frac{1}{1 + exp\left[-\left(\alpha log \frac{\tilde{C}_{raw}(x)}{1 - \tilde{C}_{raw}(x)} + \beta\right)\right]} \tag{2}$$

A reliable probabilistic interpretation is conferred on confidence scores by calibration: among predictions with a calibrated confidence of c, the actual proportion that are correct should approximate c. Calibration quality is quantified by the expected calibration error. Prediction samples are partitioned into M equal-width confidence bins B_m, and the expected calibration error is defined as:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \tag{3}$$

where, $acc(B_m)$ is the actual accuracy of samples in the m-th bin, and $conf(B_m)$ is the average confidence of samples in that bin. For a well-calibrated model, the expected calibration error should approach zero. The calibrated confidence $C_{cal}(x)$

obtained after calibration is used as the foundational input for all subsequent threshold decisions. For notational simplicity, it is uniformly denoted as C hereinafter.

2.3 Dynamic threshold decision mechanism

Fixed confidence thresholds have usually been adopted in existing studies as the basis for human–AI task allocation, and significant heterogeneity of accounting tasks across dimensions such as complexity and risk level has thereby been neglected. A multi-factor adaptive threshold function is constructed so that the automatic approval boundary can be dynamically adjusted according to task characteristics. The base threshold for the k-th cycle, $T_{base,k}$, is jointly determined by an online-updated base term $T_{0,k}$ and two task adjustment terms $T_{base,k} = clip(T_{0,k} + \lambda\sigma_h + \mu\phi(\kappa), T_{min}, T_{max})$, where $T_{0,1} = 0.70$ is the initial base value of the system; σ_h is the historical error rate of the task type, computed based on a historical voucher database stored at edge nodes, and a higher error rate yields a stricter threshold, with coefficient $\lambda > 0$; $\phi(\kappa)$ is a task complexity function, in which κ is comprehensively quantified by the number of document fields, the number of rule-matching entries, and monetary amount magnitude, and $\phi(\kappa) \in [0,1]$ is monotonically increasing, with coefficient $\mu > 0$. The weight coefficients λ and μ are determined through multiple regression analysis of historical data. Operator experience does not directly alter the automatic approval threshold and is used only for manual review task matching; time pressure within an accounting cycle is used only for prioritization of the manual review queue and does not lower the automatic approval risk threshold.

Based on the calibrated confidence C and the dynamic threshold $T_{base,k}$, three-channel routing decisions are executed. An upper buffer zone $\delta_{upper} = 0.05$ and a lower buffer zone $\delta_{lower} = 0.10$ are set, and the routing rules are defined as follows: when $C \geq T_{base,k} + \delta_{upper}$, the task is routed into the automatic approval channel, and the AI processing result is directly adopted; when $T_{base,k} - \delta_{lower} \leq C < T_{base,k} + \delta_{upper}$, the task is routed into the human–AI collaborative correction channel, and the task is pushed to the mobile end for rapid review by an accountant; when $C < T_{base,k} - \delta_{lower}$, the task is routed into the in-depth manual review channel, and comprehensive review by a senior accountant is performed. For tasks routed into the human–AI collaborative correction channel, matching is further performed according to operator experience and task complexity; when current operator experience is insufficient to cover task complexity, the task is transferred to in-depth manual review. Frequent channel switching caused by minor fluctuations in confidence near the threshold is effectively avoided by the buffer zones, and damage to user experience and system stability from pattern oscillation is prevented. On the mobile-end interface of the human–AI collaborative channel, only the three fields with the lowest confidence and the three fields with the greatest rule discrepancies are displayed so that the cognitive load of the accountant is minimized.

The base threshold term is continuously and adaptively updated through an online learning mechanism during system operation. Let R_k be the actual error rate of the automatic approval channel in the k-th cycle, R_{target} be the target error rate, and $\Delta E_k = R_k - R_{k-1}$ be the error rate change trend. The base term update rule is $T_{0,k+1} = clip[T_{0,k} + \alpha(R_k - R_{target}) + \beta\Delta E_k, T_{0,min}, T_{0,max}]$, where $\alpha > 0$ and $\beta > 0$ are learning rate parameters. After $T_{0,k}$ is updated in each cycle, $T_{base,k}$ is calculated by incorporating the historical error rate and complexity of tasks in that cycle. When the actual error rate is higher than the target value, the base threshold is raised to tighten the automatic approval criterion; otherwise, the base threshold is moderately lowered to increase the automation rate. At the same time, the error rate change trend term ΔE_k

is introduced so that an early response can be made when the error rate deteriorates rapidly, and excessive threshold elevation that would sacrifice automation efficiency is avoided when the error rate continuously improves. By the above adaptive mechanism, the system can respond to data distribution changes, and consistency between threshold updates and risk control objectives can be maintained.

2.4 Error closed-loop control mechanism

To monitor the selective risk of the automatic approval channel, a smoothed error rate estimation based on the add-one rule is adopted. Let e be the number of samples in the calibration data that are automatically approved but actually erroneous, n be the total number of automatically approved samples, and p_{target} be the target error rate. The smoothed estimator is given by $\hat{p}_{smooth} = (e+1)/(n+2)$. The system continuously compares $\hat{p}_{smooth}$ with p_{target}, and once $\hat{p}_{smooth} > p_{target}$, threshold increase or model recalibration is triggered. This smoothed estimator is used for online empirical risk monitoring and is not interpreted as a finite-sample confidence upper bound.

During adaptive updating, the dynamic threshold may drift imperceptibly due to gradual changes in data distribution. A drift detection service based on the Wald sequential probability ratio test is deployed, with the per-instance error indicator variable $z_i \in \{0,1\}$ of automatically approved samples taken as observations, where an error is recorded as 1 and a correct decision is recorded as 0. Let the acceptable error rate be $p_0 = R_{target}$, and let the abnormal error rate at which an alarm is to be triggered be $p_1 = p_0 + \Delta p$, where $\Delta p > 0$ is a preset minimum detectable risk increment. The test hypotheses are H_0: $p = p_0$ and H_1: $p = p_1$. After n automatically approved samples, the log-likelihood ratio is:

$$\Lambda_n = \sum_{i=1}^{n} [z_i \log(p_1/p_0) + (1-z_i)\log((1-p_1)/(1-p_0))] \tag{4}$$

When $\Lambda_n \geq A$, H_0 is rejected and an alarm is triggered; when $\Lambda_n \leq B$, H_0 is accepted and operation is continued; otherwise, sampling is continued. The alarm thresholds A and B are determined by a preset type I error probability α_{SPRT} and type II error probability β_{SPRT}, typically taken as $A = \log[(1-\beta_{SPRT})/\alpha_{SPRT}]$ and $B = \log[\beta_{SPRT}/(1-\alpha_{SPRT})]$. Under given constraints on the two error probabilities, a lower average detection delay can be achieved by the sequential probability ratio test so that an alarm is obtained promptly when a persistent exceedance of the error rate is indicated.

In the human–AI collaborative correction channel, every modification made by an accountant to an AI result is completely recorded by the system. These correction records serve both as a basis for immediate quality traceability and as feedback signals driving continuous model optimization. The correction rate $Corr_j$ of field j is defined as the ratio of the number of records in which the field is corrected to the total number of records containing the field. The correction rate of each field is continuously monitored by the system, and when $Corr_j$ exceeds a preset threshold, retraining of the submodel corresponding to the field or updating of associated rules is triggered. Correction records are also fed into an incremental learning data pool, and model fine-tuning is periodically performed. The three mechanisms—smoothed selective-risk monitoring, sequential probability ratio test-driven threshold drift detection, and correction-record-driven feedback learning—jointly constitute a full-chain error closed-loop control system covering pre-event monitoring, in-event alarm, and post-event optimization, so that the error control capability of the system is continuously enhanced as operating time accumulates.

2.5 Mobile prototype system architecture

To verify the feasibility and effectiveness of the proposed framework in actual mobile scenarios, a mobile accounting prototype system based on a device–edge–cloud three-tier architecture was developed. On the device side, that is, mobile devices, lightweight optical character recognition and rule-matching models optimized through TensorFlow Lite are run so that basic confidence assessment can still be completed under unstable network or offline scenarios. A human–AI collaborative interaction interface is also provided, through which calibrated confidence information is displayed to accountants and correction input is received. On the edge side, edge computing nodes are deployed, where the full confidence assessment model and dynamic threshold computation engine are deployed, recent historical voucher data and error records are stored, and online adaptive updating of thresholds is performed. On the cloud side, a central server is used, by which persistent storage of all data and audit logs is handled, model training and recalibration processes are run, and drift detection and alarm services based on the sequential probability ratio test are deployed. In the human–AI collaborative correction interface, only the three fields with the lowest confidence and the three fields with the greatest deviation from the rule base are presented to accountants, and each field is annotated with the calibrated confidence score and accompanied by green–yellow–red color coding so that cognitive load in mobile small-screen scenarios is minimized. The specific content and time consumption of each modification operation by accountants are fully recorded as input data for correction-record-driven feedback learning in error closed-loop control.

3 EXPERIMENTAL DESIGN AND VALIDATION

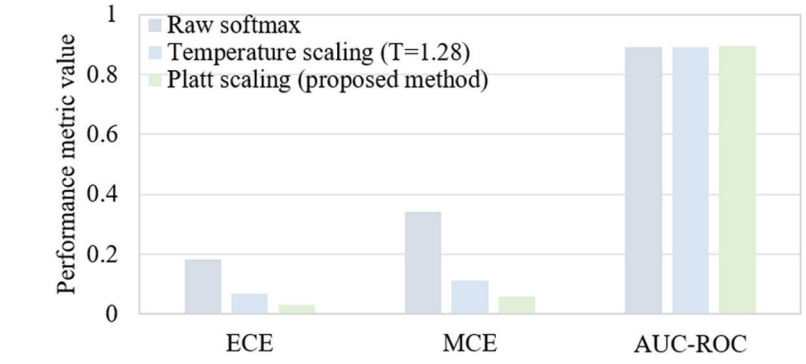
Five progressive experiments were designed to comprehensively validate the effectiveness of the proposed framework. In Experiment 1, the improvement in AI output probability quality produced by the confidence calibration module was validated. In Experiment 2, the overall performance of the dynamic threshold strategy was compared with that of multiple fixed-threshold baselines on a static dataset. In Experiment 3, the convergence characteristics and parameter sensitivity of the online adaptive threshold updating algorithm were examined. In Experiment 4, the effects of the framework on accountants' cognitive load and decision efficiency in mobile operation scenarios were evaluated through a controlled user experiment. In Experiment 5, multiple system interference scenarios were simulated to validate the robustness and self-recovery capability of the error closed-loop control mechanism during long-term operation. All experiments were implemented based on the same core dataset and prototype system.

3.1 Validation of confidence calibration effectiveness

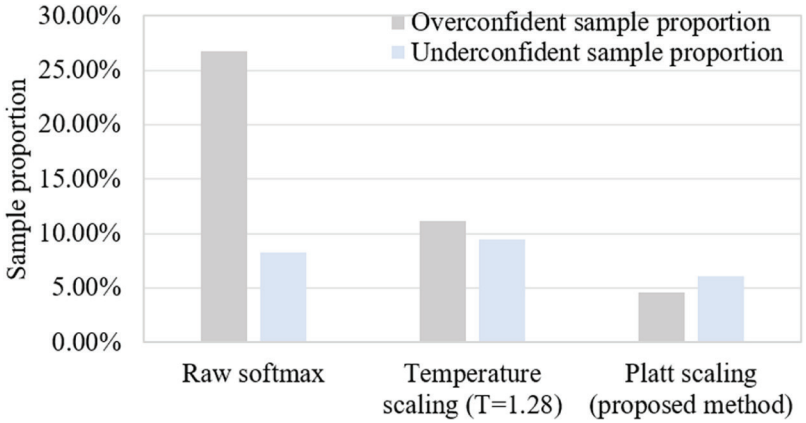
In this experiment, an optical character recognition model fusing ResNet-50 and the convolutional recurrent neural network, together with an extreme gradient boosting-based rule matching model, was adopted as the basis. Raw softmax confidence scores output by the two models were weighted and fused according to task type to obtain a raw composite confidence C_{raw}. Three comparison conditions were established: raw softmax confidence without calibration; temperature scaling, in which the temperature parameter T was optimized by grid search on the calibration set using negative log-likelihood loss; and Platt scaling, the proposed

method, in which the scaling parameters α and β were solved by maximum likelihood estimation on the calibration set.

As shown in Figure 2, the expected calibration error of raw softmax confidence was as high as 0.182, the maximum calibration error reached 0.341, and overconfident samples accounted for 26.7%, indicating severe overconfidence in the model. The expected calibration error was reduced to 0.067 by temperature scaling, a decrease of 63.2%, which verified the effectiveness of scaling calibration; however, systematic underconfidence bias remained in the medium-to-low confidence range. The expected calibration error was further compressed to 0.031 by the Platt scaling adopted in the present framework, representing reductions of 82.9% relative to the original model and 53.7% relative to temperature scaling, and the proportion of overconfident samples was reduced to 4.6%. Differences in the area under the receiver operating characteristic curve under the three conditions were small, and no obvious degradation in model discriminative ability caused by calibration was observed. While overall discriminative ability was maintained by Platt scaling, confidence scores were made more consistent with actual accuracy, thereby providing a reliable foundation for subsequent dynamic threshold decision-making.



a) Comparison of the ECE, MCE, AUC-ROC



b) Comparison of overconfident and underconfident sample proportions

Fig. 2. Performance comparison of different confidence calibration methods

3.2 Performance comparison between dynamic and fixed thresholds

In this experiment, the complete three-channel routing system was run on the test set, and four comparison conditions were established: a fixed threshold $T = 0.70$;

a fixed threshold $T = 0.80$; a grid-search-optimized fixed threshold, under the constraints that overall accuracy be maximized and the auto-pass error rate not exceed 5%; and the dynamic threshold strategy of the proposed framework. In the dynamic strategy, the task threshold was computed in real time as:

$$T_{base,k} = clip(T_{0,k} + \lambda\sigma_h + \mu\phi(\kappa), T_{min}, T_{max}) \quad (5)$$

where, the regression coefficients were fitted on the training set as $\lambda = 0.42$ and $\mu = 0.35$, and the base term $T_{0,k}$ was adjusted according to the auto-pass error rate of the previous cycle by the online update rule.

Experimental results are presented in Figure 3 and Table 1. It can be observed that although an automation rate of 72.5% was obtained under the fixed threshold $T = 0.70$, an auto-pass error rate of 6.8% was reached, exceeding the 5% risk-control upper limit. When the threshold was raised to 0.80, the error rate was reduced to 4.3%, but the automation rate dropped sharply by 14.1 percentage points. Under the grid-search-optimal threshold $T^* = 0.74$, an automation rate of 65.3% was achieved while the risk-control constraint was satisfied. With the dynamic threshold strategy of the proposed framework, the lowest auto-pass error rate of 4.2% was attained, while the automation rate reached 78.6%, an improvement of 13.3 percentage points over the grid-search optimum. This dual optimization is attributed to differentiated thresholds implemented by the dynamic threshold according to different task characteristics: for low-risk routine travel expense documents, the threshold was lowered to 0.66–0.68, whereas for high-risk business entertainment expense documents, the threshold was raised to 0.76–0.79. At the mobile end, only approximately one fifth of documents needed to be processed by accountants, whereas approximately one third needed to be processed under the optimal fixed threshold.

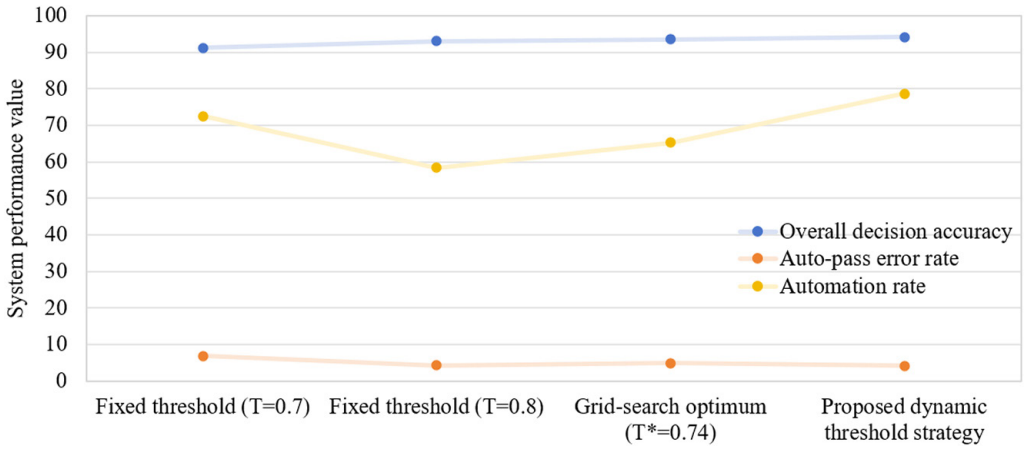


Fig. 3. System performance comparison under different threshold strategies

Table 1. Sample distribution under different threshold strategies

Threshold Strategy	Sample Distribution (Automatic Approval/ Collaborative Correction/in-Depth Manual Review)
Fixed threshold $T = 0.70$	72.5%/18.3%/9.2%
Fixed threshold $T = 0.80$	58.4%/25.6%/16.0%
Grid-search optimum $T^* = 0.74$	65.3%/22.1%/12.6%
Proposed dynamic threshold strategy	78.6%/14.2%/7.2%

3.3 Convergence and parameter sensitivity of online adaptive threshold updating

In this experiment, a Monte Carlo simulation method was adopted, and a simulated data stream spanning 50 consecutive accounting cycles was generated according to the real data distribution. The initial threshold was set as $T_0 = 0.70$, and the target error rate was set as $R_{target} = 5\%$. Three hyperparameter combinations were established: a low learning rate ($\alpha = 0.05$, $\beta = 0.02$), a medium learning rate ($\alpha = 0.10$, $\beta = 0.05$, the default setting of the proposed framework), and a high learning rate ($\alpha = 0.20$, $\beta = 0.10$). Each group was independently run 30 times, and the average was taken.

Table 2. Convergence performance of adaptive threshold updating under different learning rate configurations

Learning Rate Configuration	Convergence Cycles	Steady-State Threshold Mean	Steady-State Auto-Pass Error Rate	Steady-State Threshold Fluctuation (Standard Deviation)	Overshoot
Low learning rate ($\alpha = 0.05$, $\beta = 0.02$)	24	0.738	5.04%	0.008	0.6%
Medium learning rate ($\alpha = 0.10$, $\beta = 0.05$)	14	0.742	4.96%	0.021	1.2%
High learning rate ($\alpha = 0.20$, $\beta = 0.10$)	7	0.756	5.11%	0.047	3.8%

As shown in Table 2, 24 cycles were required for convergence under the low learning rate configuration, and the cumulative suboptimal decision loss during this period cannot be neglected. Although convergence was achieved within 7 cycles under the high learning rate configuration, the steady-state auto-pass error rate reached 5.11%, slightly exceeding the target, and the magnitude of threshold fluctuation was nearly 6 times that under the low learning rate configuration. Excessive fluctuation would cause frequent changes in the human–AI division-of-labor boundary, leading to process confusion. Under the medium learning rate configuration, the best balance was achieved, with convergence in 14 cycles, a steady-state error rate of 4.96%, and a fluctuation magnitude of 0.021; moreover, its conservative direction of deviation (slightly below the target value) is more preferable in risk management. This configuration was adopted in all subsequent experiments.

3.4 User experiment: Evaluation of mobile cognitive load and decision efficiency

Thirty-two practitioners with more than two years of practical experience were recruited through a provincial accounting association and randomly assigned to an experimental group and a control group, with 16 participants in each group. No significant differences in gender, age, years of practice, or mobile device usage habits were found between the two groups by analysis of covariance ($p > 0.05$). The complete prototype system integrating the dynamic threshold mechanism and calibrated confidence display was used by the experimental group. A fixed-threshold ($T = 0.070$) and traditional binary decision system was used by the control group, in which confidence was displayed as raw Softmax scores without calibration.

Review tasks for 200 test documents were completed by each participant on an iPad Pro, divided into four stages, and task order was balanced by a Latin square design.

As shown in Table 3, an error detection rate of 94.8% was achieved by the experimental group, significantly higher than the 92.1% achieved by the control group ($p = 0.040$), and a significantly lower false alarm rate was observed (6.2% vs. 8.7%, $p = 0.046$). These results indicate that limited cognitive resources of accountants can be concentrated on samples genuinely requiring human judgment under the complete human–AI collaborative framework. The average processing time was only 45.2 seconds in the experimental group, compared with 58.7 seconds in the control group, representing a saving of 23.0% ($p < 0.001$), which is consistent with the higher automation rate in the experimental group and the display of only key discrepancy fields in the human–AI collaborative interface. The National Aeronautics and Space Administration Task Load Index (NASA-TLX) total score was 38.5 in the experimental group, significantly lower than 52.6 in the control group ($p < 0.001$, $d = -1.40$); the mental demand and effort subscales were reduced by 14.6 and 15.2 points, respectively. The system usability scale score was 82.3 (excellent) in the experimental group versus 71.5 (lower end of good) in the control group, and the trust rating was 5.7 versus 4.8 ($p = 0.008$), indicating better user acceptance for the complete system.

Table 3. Comparison of performance and subjective experience between the experimental group and the control group in mobile review tasks

Evaluation Dimension	Indicator	Experimental Group	Control Group	Effect Size (Cohen's d)	Statistical Test
Objective performance	Error detection rate	94.8%	92.1%	0.76	$t(30) = 2.14$, $p = 0.040$
	False alarm rate	6.2%	8.7%	–0.74	$t(30) = -2.08$, $p = 0.046$
	Average processing time (s/document)	45.2	58.7	–1.35	$t(30) = -3.82$, $p < 0.001$
	Human–AI collaborative correction accuracy	87.6%	82.3%	0.70	$t(30) = 1.98$, $p = 0.057$
Subjective workload	National Aeronautics and Space Administration Task Load Index (NASA-TLX) total score	38.5	52.6	–1.40	$t(30) = -3.95$, $p < 0.001$
	Mental demand subscale	35.2	49.8	–1.30	$t(30) = -3.67$, $p = 0.001$
	Effort subscale	40.1	55.3	–1.21	$t(30) = -3.41$, $p = 0.002$
System experience	System usability scale	82.3	71.5	1.18	$t(30) = 3.35$, $p = 0.002$
	Trust rating (1–7 Likert)	5.7	4.8	1.01	$t(30) = 2.85$, $p = 0.008$

3.5 Robustness validation of the error closed-loop control mechanism under interference scenarios

On the longitudinal simulation platform of Experiment 3, three types of interference were injected at cycles 20, 35, and 45, respectively: optical character recognition data distribution drift (recognition error rate increased from 3.2% to 8.5%); emergence of new expense categories (digital service fees accounting for 15%,

with a rule-matching degree of only 62%); and operator replacement (a senior accountant with $\tau = 0.92$ replaced by a novice with $\tau = 0.67$).

Table 4. Response performance of the error closed-loop control mechanism under three interference scenarios

Interference Type	Drift Detection Alarm Delay (Cycles)	New Steady-State Adjustment Period (Cycles)	Peak Error Rate	Steady-State Error Rate After Recovery	Recall Improvement of Frequently Corrected Fields
Optical character recognition data distribution drift	2	6	8.7%	4.9%	—
Emergence of new expense categories	3	8	7.4%	5.2%	28.6%
Operator replacement	1	4	6.8%	4.6%	—

As shown in Table 4, under the optical character recognition drift scenario, an alarm was triggered by the sequential probability ratio test detection at cycle 2, the threshold was adjusted to a new steady state over six cycles, and the error rate after recovery was 4.9%. The emergence of new expense categories was the most challenging interference: an alarm was triggered at cycle 3; because confidence for the new expense category was low (mean 0.58), most samples were directly routed into the in-depth manual review channel; the peak error rate was exposed with a delay; threshold adjustment required eight cycles; and the steady-state error rate after recovery was 5.2%, slightly exceeding the target upper limit by 0.2 percentage points. However, two fields, service content description and billing cycle, were identified from correction record analysis, with correction rates as high as 34.2% and 28.7%, respectively; after targeted submodel retraining was triggered, the recall of frequently corrected fields was improved by 28.6%. In the operator replacement scenario, after a change in the operator experience parameter $\Psi(\tau)$ was recognized by the system, manual review tasks were rematched; preventive task allocation adjustment was completed as early as cycle 1; a new steady state was reached in four cycles; and the error rate after recovery was 4.6%. Across the three scenarios, alarms or task adjustments occurred within 1–3 cycles, and adjustment to a new steady state required 4–8 cycles, indicating favorable robustness and adaptive recovery capability of the framework. In the new expense category scenario, the error rate after recovery remained slightly above the 5% target value.

4 CONCLUSION

To address the static division-of-labor boundaries and confidence bias faced by human–AI collaborative decision-making in mobile accounting systems, a collaborative decision-making and error-control framework based on dynamic confidence thresholds was proposed. In this framework, confidence calibration was achieved through Platt scaling; a multi-factor adaptive threshold function and a three-channel routing decision mechanism were constructed; and selective-risk statistical monitoring, sequential probability ratio test-driven threshold drift detection, and correction-record-driven feedback learning were integrated, forming a full-chain error closed-loop control system. Experimental results indicate that an automation rate of 78.6% was achieved by the dynamic threshold strategy under an auto-pass

error rate of 4.2%, representing an improvement of 13.3 percentage points over the grid-search-optimal fixed threshold. The online threshold updating algorithm converged to a steady state within 14 cycles. In the user experiment, the average processing time for accountants was shortened by 23.0%, and cognitive load was reduced by 26.8%. Under three interference scenarios—optical character recognition data drift, emergence of new expense categories, and operator replacement—a new steady state was entered by the system within four to eight cycles. In the optical character recognition drift and operator replacement scenarios, recovery to within the 5% target error rate was achieved, whereas in the new expense category scenario, the error rate after recovery was 5.2%, close to but still slightly above the target value. A theoretical foundation and technical solution are thereby provided for the systematic engineering deployment of human–AI collaborative decision-making in mobile accounting scenarios, and the core method can also be extended to the design of intelligent decision systems in other professional fields at the mobile end. Large language model-driven accounting semantic understanding and confidence assessment, as well as the application of multi-agent collaborative decision-making frameworks in financial scenarios, will be explored in future work.

5 REFERENCES

[1] A. C. Yoshikuni, R. Dwivedi, R. G. Dultra-de-Lima, C. Parisi, and J. C. T. Oyadomari, "Role of emerging technologies in accounting information systems for achieving strategic flexibility through decision-making performance: An exploratory study based on North American and South American firms," *Global Journal of Flexible Systems Management*, vol. 24, no. 2, pp. 199–218, 2023. https://doi.org/10.1007/s40171-022-00334-9

[2] M. Shakir, M. J. Al Farsi, I. R. Al-Shamsi, B. Shannaq, and T.-H. Ghilan Al-Madhagy, "The influence of mobile information systems implementation on enhancing human resource performance skills: An applied study in a small organization," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 13, pp. 37–68, 2024. https://doi.org/10.3991/ijim.v18i13.47027

[3] L. Cabañero, R. Hervás, I. González, J. Fontecha, T. Mondéjar, and J. Bravo, "Characterisation of mobile-device tasks by their associated cognitive load through EEG data processing," *Future Generation Computer Systems*, vol. 113, pp. 380–390, 2020. https://doi.org/10.1016/j.future.2020.07.013

[4] M. Westphal, M. Vössing, G. Satzger, G. B. Yom-Tov, and A. Rafaeli, "Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance," *Computers in Human Behavior*, vol. 144, Art. no. 107714, 2023. https://doi.org/10.1016/j.chb.2023.107714

[5] Y. Zhang, "Machine learning-based personalized learning path decision-making method on intelligent education platforms," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 16, pp. 68–82, 2024. https://doi.org/10.3991/ijim.v18i16.51009

[6] H. T. Ha and A. Horák, "Information extraction from scanned invoice images using text analysis and layout features," *Signal Processing: Image Communication*, vol. 102, Art. no. 116601, 2022. https://doi.org/10.1016/j.image.2021.116601

[7] M. Papík and L. Papíková, "Detecting accounting fraud in companies reporting under US GAAP through data mining," *International Journal of Accounting Information Systems*, vol. 45, Art. no. 100559, 2022. https://doi.org/10.1016/j.accinf.2022.100559

[8] J. Kokina, S. Blanchette, T. H. Davenport, and D. Pachamanova, "Challenges and opportunities for artificial intelligence in auditing: Evidence from the field," *International Journal of Accounting Information Systems*, vol. 56, Art. no. 100734, 2025. https://doi.org/10.1016/j.accinf.2025.100734

[9] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: A survey on how to assess and improve predicted class probabilities," *Machine Learning*, vol. 112, no. 9, pp. 3211–3260, 2023. https://doi.org/10.1007/s10994-023-06336-7

[10] S. A. Balanya, J. Maronas, and D. Ramos, "Adaptive temperature scaling for robust calibration of deep neural networks," *Neural Computing and Applications*, vol. 36, no. 14, pp. 8073–8095, 2024. https://doi.org/10.1007/s00521-024-09505-4

[11] K. Okamura and S. Yamada, "Adaptive trust calibration for human-AI collaboration," *PLoS ONE*, vol. 15, no. 2, Art. no. e0229132, 2020. https://doi.org/10.1371/journal.pone.0229132

[12] W. Xiong, H. Fan, L. Ma, and C. Wang, "Challenges of human–machine collaboration in risky decision-making," *Frontiers of Engineering Management*, vol. 9, no. 1, pp. 89–103, 2022. https://doi.org/10.1007/s42524-021-0182-0

[13] M. Savenko, "A systems engineering framework for real-time user error prediction and adaptive user experience control," *Journal of Engineering Management and Systems Engineering*, vol. 5, no. 3, pp. 319–339, 2026. https://doi.org/10.56578/jemse050303

[14] M. Adam, C. Diebel, M. Goutier, and A. Benlian, "Navigating autonomy and control in human-AI delegation: User responses to technology-versus user-invoked task allocation," *Decision Support Systems*, vol. 180, Art. no. 114193, 2024. https://doi.org/10.1016/j.dss.2024.114193

[15] I. Kyianytsia, D. Yatsiuk, H. Aldankova, O. Horobets, V. Slipchenko, and V. Dobrovolskyi, "A cyber-physical system approach to adaptive visual branding in cultural institutions," *Journal of Engineering Management and Systems Engineering*, vol. 5, no. 3, pp. 340–353, 2026. https://doi.org/10.56578/jemse050304

[16] J. W. Freiman, Y. Kim, and M. A. Vasarhelyi, "Full population testing: Applying multidimensional audit data sampling (MADS) to general ledger data auditing," *International Journal of Accounting Information Systems*, vol. 46, Art. no. 100573, 2022. https://doi.org/10.1016/j.accinf.2022.100573

6 AUTHOR

Lan Li graduated from Zhongnan University of Economics and Law with a Master's degree in Management. She currently works at the School of Economics and Management, Hubei University of Education. Her main research areas are the theory and practice of financial management and data asset risk management (E-mail: lilan@hue.edu.cn).