International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307
Available Online at : jsrcseit.com
doi : https://doi.org/10.32628/CSEIT2612544

ARGUS Survey of AI Driven Personal Finance Analytics and Budget Forecasting

Gurusundaragunalan M[1], Dr. D. Mohanapriya[2], Dr. V. S. Anita Sofia[3]

[1]PG Student, Department of Computer Applications (PG), PSG College of Arts & Science, Coimbatore, Tamil Nadu, India

[2]Assistant Professor, Department of Computer Applications (PG), PSG College of Arts & Science, Coimbatore, Tamil Nadu, India

[3]Associate Professor, Department of Computer Applications (PG), PSG College of Arts & Science, Coimbatore, Tamil Nadu, India

ARTICLEINFO

Article History:

Accepted : 15 Sep 2026
Published: 13 Oct 2026

Publication Issue
Volume 12, Issue 5
September-October-2026

Page Number
262-277

ABSTRACT

Personal-finance software is useful only when it can move from records to decisions without hiding uncertainty. We reviewed 25 papers published in 2025 and 2026 and grouped their evidence into six connected activities: UNDERSTAND, ANTICIPATE, PERSONALIZE, ADVISE, EXPLAIN, and TRUST. For every paper, we recorded the study design, data, method, reported result, limitation, and relevance to ARGUS. The stated review window runs from 1 January 2025 to 30 September 2026, while the locked-corpus audit was completed on 22 September 2026. The study set contains 11 observational surveys, six human experiments, four audits of model output, two computational studies, and two reviews. This distribution matters because only P02 directly evaluates both personal transaction classification and budget prediction. The remaining evidence is strongest on advice, adoption, privacy, vulnerability, explanation, and trust. We therefore do not treat an adoption coefficient or a favorable user rating as proof that a budgeting model works. We also compare the literature with the inspected ARGUS implementation. The current project uses deterministic categorization, exact-money arithmetic, budget comparisons, and answers grounded in the user ledger; it does not contain a trained forecasting model. Archived project records report 205 passing automated tests, acceptable grounded responses for ten controlled questions, and correct amount and currency extraction from two clear synthetic receipts. These are engineering checks on specific cases, not estimates of population accuracy or financial benefit. The review supports an auditable design for personal-finance assistance, but it also shows that learned forecasting, personalization, and

Copyright © 2026 The Author(s) : This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

long-term household outcomes still require direct evaluation.

Keywords: Personal finance; literature survey; budget forecasting; financial advice; explainability; calibrated trust; ARGUS.

Introduction

Recording an expense is the easy part of personal-finance software. The harder question is what the record means for the next decision. A category affects a budget, a budget affects an alert, and an alert may lead to advice. If the first step is wrong, later stages can remain numerically consistent while still misleading the user. This dependency is why fluent explanations cannot compensate for weak financial evidence.

The 25 papers do not test one common system. Some report predictive scores, others study human reactions to advice, and several examine adoption or governance. Those outcomes are not interchangeable. Classification accuracy measures agreement with labels, whereas an adoption coefficient describes a relationship between survey constructs. Neither, on its own, demonstrates an improvement in household finances. Our synthesis keeps those boundaries visible before drawing design lessons.

ARGUS gives the review a practical reference point. Its codebase contains financial records, category summaries, budget monitoring, reviewed capture candidates, and a ledger-based question interface. We first examine what the papers actually establish and then compare those findings with the implemented project. The distinction is important: forecasting remains proposed work, and the present assistant refuses requests that its deterministic data paths cannot support.

The paper makes three contributions. It compares the 25 studies on a common evidence matrix, counts the main study designs and research stages, and traces selected research implications to inspected ARGUS modules and archived evaluation records. It does not introduce a forecasting model, and it does not present the supplied collection as an exhaustive search of recent financial-AI literature.

SURVEY METHODOLOGY

A. Review questions and scope

We used four review questions. RQ1 asks which data, computational methods, and evaluation designs have been used for intelligent personal finance. RQ2 examines how the evidence is distributed across understanding, anticipation, personalization, advice, explanation, and trust. RQ3 identifies the limitations that prevent a reported result from transferring directly to a household application. RQ4 compares the reviewed approaches with the mechanisms and evaluations that are already present in ARGUS.

Each supplied article is treated as one review unit and retains its identifier from P01 to P25. This is a purposive, locked corpus rather than a fresh systematic database search. Together, the PDFs contain 597 pages and 25 different file hashes. Different hashes confirm that the files are distinct, but they do not prove that participant samples or datasets are independent. IEEE numbers are assigned by first citation, while P-identifiers remain fixed for the evidence matrix.

B. Search sources and strategy

The project review note names Google Scholar, IEEE Xplore, SpringerLink, ScienceDirect, MDPI, arXiv, and the ACM Digital Library. It also lists topics such as expense categorization, notification capture, receipt OCR, explainability, and recommendations. The note does not contain database exports, dated query logs, or

hit counts. We can therefore report these services as sources considered by the project, but we cannot claim that every paper was retrieved from a particular database.

For a future update, a reproducible core query could combine personal finance, financial advice, robo-advice, and budget prediction with artificial intelligence, machine learning, language models, explainability, or trust. Transaction classification, receipt OCR, and notification detection deserve separate searches because they use different terminology and venues. This proposed query is a reconstruction for later work, not a record of the original search.

C. Eligibility and quality appraisal

The intended publication interval was 1 January 2025 to 30 September 2026, and the locked evidence audit was carried out on 22 September 2026. All supplied papers display a 2025 or 2026 publication year. We did not substitute acceptance dates or PDF creation dates for the first publication date. P15 carries a 2026 issue date with evidence of 2025 online publication; for papers whose exact online date is absent from the PDF, that date remains unresolved.

A paper was eligible when its full text was readable, its scholarly source could be identified, its topic addressed a review question, and its method or synthesis could be extracted. Personal-finance studies were directly relevant. Corporate planning was retained only when it offered adjacent evidence about uncertainty or prediction. Duplicate papers, unavailable full text, work outside the period, unrelated topics, and unsupported promotional material were exclusion grounds. None of the 25 supplied items was removed at this stage, which does not tell us how earlier candidates may have been screened.

We assessed quality by reading the sample source, evaluation design, outcome measure, reproducibility details, and limits on generalization. Weak evidence was not converted into an arbitrary score; it was used to narrow the claim drawn from that paper. One reviewer coded the corpus and checked the entries against the PDFs. Independent double coding and an inter-rater reliability estimate were outside this revision.

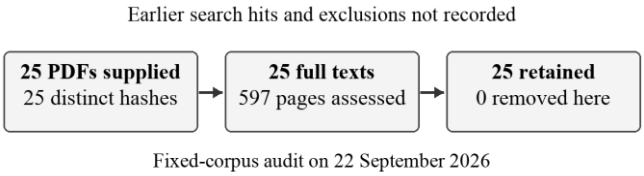


Fig. 1. Verifiable study-selection flow

RELATED WORK

A. Advice performance and financial understanding

Baeckstrom and Matkovskyy place professional advisers and several large language model configurations in the same vignette-based comparison [1]. The study finds projection in human advice and configuration-dependent conservatism in model advice. Its wealth results come from simulated returns and fee assumptions, so they describe consequences under the experiment rather than realized client performance.

Yang et al. provide the clearest technical match to ARGUS's forecasting objective [2]. Their DFTS model learns from transaction text and temporal patterns while training classification and budget prediction together. The result is relevant, but transfer is not automatic: the dataset language, financial setting, label design, and validation procedure differ from those of a new deployment.

Two survey studies broaden the question from prediction to behavior. Olajide et al. use national survey data to relate willingness to use AI advice, financial knowledge, and reported behavior [3]. Antonoudi and Ludwig examine the use of AI tools, financial planners, and combined help among public employees [4]. Both identify useful associations, but their observational designs cannot show that the technology caused the reported behavior.

B. Personalization, vulnerability and adoption

Bonelli changes the digital-twin profile and conversational interface while keeping the financial

advice constant [5]. Participants can perceive greater personalization and greater privacy concern at the same time. Zeka and Johns reach a related caution through a prompt audit: general-purpose models do not consistently adjust plans for consumer vulnerability [6]. A detailed profile and a sympathetic tone therefore cannot be assumed to produce suitable advice.

Takemeyang et al. connect digital financial literacy, customer value, and adoption intention [7]. Chawla et al. add a less comfortable finding: within respondents willing to use AI advice, miscalibrated knowledge is associated with fraud loss [8]. Sungkarungsri and Kiattisin study transparency, perceived intelligence, trust, and advocacy in mobile financial advice [9]. Read together, the studies favor informed use over indiscriminate adoption.

C. Explanation and hybrid advice

Brunetti et al. use eye tracking to see how an adviser label changes attention to loan costs [10]. Ioannou et al. vary explanation quality and decision risk [11]. These designs treat explanation as something that can redirect attention and alter judgment, not simply as extra text on a screen. Kumta et al. compare human, AI, and hybrid portfolio advice [12]; the encouraging results remain bounded by a small sample and a simulated long-term task.

D. Acceptance, uncertainty and model limitations

Szopinski et al. model acceptance of AI financial tools through an extended technology-acceptance framework [13]. Zhang, by contrast, evaluates fuzzy financial decision support on corporate data [14]. The first informs how users may approach a service; the second illustrates rule-based reasoning under uncertainty. Neither result validates household budget forecasting in ARGUS.

Wang et al. examine how feelings of technological vulnerability shape adoption of AI advisers [15]. Schlosky and Raskie review ChatGPT responses to 21 personal-finance scenarios and still find omissions despite improvement across model versions [16]. Tang and Son focus on the perceived value of digital-human advisers [17], while Sharma and Priya connect FinTech use with inclusion and reported well-being [18]. Perceived value is meaningful, but it is not the same outcome as measured financial improvement.

E. Trust dynamics, resistance and bias

Han and Ko follow trust across repeated advice and show how one error can damage it; a later explanation can help repair confidence [19]. Verma et al. study resistance to robo-advice [20], and Bennetts and Ludwig identify distinct preference groups for AI and professional retirement help [21]. These papers make a uniform onboarding strategy difficult to justify. Users differ, and a system also needs a visible route for correcting mistakes.

The model-audit papers show why output quality needs more than a surface check. Winder et al. report several forms of risk amplification in generated investment portfolios [22]. Bennetts and Ludwig find subtler differences in language across racial identifiers in planner-client messages [23]. Fairness review must therefore cover the recommendation and the way it is communicated.

F. Reviews and the integration problem

Namyslo et al. organize design requirements for robo-advisory systems [24], whereas Sai et al. review generative-AI applications and risks across finance [25]. Both widen the design discussion, but they remain secondary evidence. We did not count the studies inside those reviews as additional members of the 25-paper corpus.

We use the six-stage taxonomy to organize evidence, not to claim that the literature has tested one complete architecture. UNDERSTAND concerns the meaning of financial records. ANTICIPATE covers future conditions. PERSONALIZE adds goals and circumstances; ADVISE compares actions; EXPLAIN makes the reasoning inspectable; and TRUST concerns justified reliance, privacy, adoption, and failure. Many papers touch several stages, so the matrix assigns one dominant stage only for counting.

Table I lets readers compare focus, method, and setting without collapsing unlike outcomes into a

ranking. The numerical results are discussed later with their denominators and limitations. A large survey does not repair a weak causal design, just as a high score on one dataset does not establish performance on another task.

COMPARATIVE ANALYSIS OF EXISTING STUDIES

Table I preserves the P01-P25 labels alongside the IEEE citations. Its purpose is descriptive: it shows what each paper studied, how it was studied, and where the result applies. Sections VI and VII carry the numerical findings and the limits needed to interpret them.

TABLE I. P01–P25 COMPARATIVE STUDY MATRIX

Ref.	Primary Focus	Main Approach	Application Context
P01 [1]	Financial advice comparison	Human adviser and nine LLM configurations; vignette experiment and portfolio simulation	Personal portfolio advice and long-term wealth
P02 [2]	Transaction classification and budget prediction	Chinese BERT, amount-weighted attention, BiLSTM and TCN; multitask learning	Personal transaction management and budgeting
P03 [3]	AI advice and financial behavior	NFCS cross-sectional regression with financial-knowledge interactions	Consumer financial behavior
P04 [4]	Technology-enabled retirement advice	Logistic regression comparing AI, planner and combined help	Retirement saving among public employees
P05 [5]	Personalization and privacy	Randomized 2 × 2 financial-digital-twin and conversational-AI experiment	Robo-advisory and investment allocation
P06 [6]	Vulnerability and bias in AI planning	Vulnerability-varied prompt audit with qualitative content and discourse analysis	Consumer financial planning
P07 [7]	Literacy, value and AI adoption	Customer survey and PLS-SEM	AI-driven financial-service adoption in Cameroon
P08 [8]	AI advice willingness and fraud loss	NFCS multivariate logistic regression with knowledge miscalibration	Consumer fraud risk and financial advice
P09 [9]	Trustworthy AI and adoption	Survey and structural equation modeling	AI financial advisory in Thai mobile banking
P10 [10]	Attention and adviser source	Loan-choice eye-tracking experiment and mediation analysis	Consumer loan decision support
P11 [11]	Explainability and financial risk	Controlled online experiment with adviser, explanation and risk conditions	AI-supported investment decisions
P12 [12]	Human, AI and hybrid advice	Quasi-experiment, portfolio task and Monte Carlo simulation	Long-term portfolio advice

Ref.	Primary Focus	Main Approach	Application Context
P13 [13]	Acceptance of AI financial tools	Extended technology-acceptance survey and PLS-SEM	Consumer financial decision-making in Poland

TABLE I. P01-P25 COMPARATIVE STUDY MATRIX (CONTINUED)

Ref.	Primary Focus	Main Approach	Application Context
P14 [14]	Financial planning under uncertainty	Fuzzy inference, adaptive weighting and defuzzification; ARIMA, SVM and RF comparison	Corporate financial planning and crisis warning
P15 [15]	AI adviser adoption and vulnerability	US survey with mediation and moderation analysis	Consumer AI financial-adviser adoption
P16 [16]	Quality of ChatGPT financial advice	Expert qualitative assessment of 21 scenarios; prompt and model comparison	General personal-finance advice
P17 [17]	Perceived value of digital-human advisers	Survey and SEM using a stimulus-organism-response model	Digital-finance adviser use in mainland China
P18 [18]	FinTech, inclusion and well-being	Survey and PLS-SEM	AI-enabled FinTech use in North India
P19 [19]	Trust formation, error and repair	Two repeated-advice experiments with accuracy and explanation manipulations	Human-AI financial advisory
P20 [20]	Resistance to robo-advisers	Innovation-resistance survey and SEM	Indian retail-investor robo-advisory
P21 [21]	Preference segments for AI retirement advice	Latent class analysis	Employer-sponsored retirement-plan advice
P22 [22]	LLM investment-bias amplification	Four systematic portfolio-generation studies with risk and performance analysis	Private-investor portfolio advice
P23 [23]	Racial bias in financial communication	Structured ChatGPT generation with content and discourse analysis	Planner-client diversification emails
P24 [24]	Robo-advisory design requirements	Systematic literature review and requirement synthesis	Investment robo-advisory design
P25 [25]	Generative AI applications in finance	Narrative review and finance case studies	Consultation, markets, risk and financial documents

AI AND COMPUTATIONAL TECHNIQUES

A. Semantic and temporal learning

P02 combines a pretrained Chinese BERT encoder, amount-weighted attention, a bidirectional LSTM, and temporal convolutions [2]. The text representation helps distinguish similar transaction descriptions, while the sequence components capture patterns across time. Because classification and budget prediction are trained together, the model tests whether the tasks can share useful information. Its multilingual appendix concerns backbone transfer and is not a full replication of the architecture.

A prospective forecast can use only information available before the prediction target. Bidirectional encoding is legitimate inside an observed history window, but it creates leakage if future records enter that window. Applying this approach to ARGUS would require chronological splits, user-disjoint testing, and an explicit inventory of when every feature becomes available. P02's published score does not establish those conditions for ARGUS data.

B. Fuzzy systems and conventional baselines

P14 converts financial indicators into fuzzy values, evaluates rules with adaptive weights, and then converts the result back into a decision [14]. The intermediate rules are inspectable, although the membership functions and weights still need empirical justification. The study compares its method with ARIMA, support vector machines, random forests, and conventional fuzzy logic. Since the data concern firms and crisis warnings, the result is only adjacent to household budgeting.

ARIMA, support vector machines, and random forests appear as baselines in the reviewed studies. Their inclusion in a paper does not mean ARGUS implements them. Any move from corporate prediction or a published benchmark to personal-finance data would need a new task definition and a fresh validation.

C. LLM advice and model-output auditing

The language-model papers examine advice, explanations, plans, and portfolios rather than one shared endpoint. Across P01, P06, P16, P22, and P23, the problems include preference projection, limited treatment of vulnerability, missing information, portfolio concentration, and biased phrasing [1], [6], [16], [22], [23]. These findings support controlled prompt suites, recorded model versions, and checks on the financial output. Prompt refinement by itself does not establish suitability.

Ledger arithmetic and natural-language explanation belong to different trust boundaries. Verified services need to supply the amount, period, currency, and ownership constraints. A language model may help present those facts, but it should not recreate them from prose. The inspected ARGUS route remains deterministic, and its response-time measurements cannot be used as evidence about an external model.

TABLE II. TECHNIQUES AND THEIR EVIDENCE ROLES

Technique and support	Use and principal limitation
BERT, BiLSTM, TCN [2]	Joint transaction meaning and temporal prediction; training data and transfer must be validated.
Fuzzy inference [14]	Inspectable uncertainty rules; corporate-domain evidence and rule tuning limit transfer.
ARIMA, SVM, RF [2], [14]	Reported comparison baselines; not deployed ARGUS algorithms.
LLM generation [1], [6], [16], [22], [23]	Advice and communication; prompt sensitivity, bias and unsupported assertions.
Regression [3], [4], [8]	Associations between advice, behavior and risk; no causal treatment effect from cross-sectional data.

Technique and support	Use and principal limitation
SEM / PLS-SEM [7], [9], [13], [17], [18], [20]	Latent adoption/value relationships; measurement quality and selection affect interpretation.
Latent class analysis [21]	Segment preferences for help; classes are sample-dependent, not universal user types.
Deterministic ARGUS rules	Exact calculations and traceable output; limited vocabulary, fixed thresholds, no learned forecasting.

D. Statistical methods are not product models

Regression, structural equation models, and latent-class analysis describe associations or groups in a sample. They are research methods, not ready-made recommendation engines. For example, a preference class may suggest an onboarding hypothesis, but production assignment would still require validation, a clear purpose, and evidence of benefit. A statistically significant path coefficient is not an application accuracy score.

E. Explainability and hybrid computation

The explanation studies measure information quality, visual attention, reliance, and recovery after error [10], [11], [19]. That work is different from attributing features inside a neural predictor. Current ARGUS explanations arise from ledger entries and explicit rule factors, so labels such as SHAP or LIME would be inaccurate. A future hybrid could join a learned forecast with deterministic calculations, provided that the explanation is checked against both. That remains proposed work.

SURVEY RESULTS AND FINDINGS

A. Extraction and counting procedure

For each article, the extraction sheet stores bibliographic identity, sample or dataset, method, findings, limitations, and relevance to ARGUS. We then assigned one main study-design label and one main taxonomy stage. Percentages use all 25 papers as the denominator. The labels describe the central research question, not every method or secondary theme mentioned by the authors.

TABLE III. PRIMARY EVIDENCE DESIGNS

Design	Papers	n (%)
Observational surveys	03, 04, 07, 08, 09, 13, 15, 17, 18, 20, 21	11 (44)
Human experiments	01, 05, 10, 11, 12, 19	6 (24)
Model-output audits	06, 16, 22, 23	4 (16)
Computational models	02, 14	2 (8)
Reviews	24, 25	2 (8)

Survey designs account for 11 of the 25 papers (44%). Six papers (24%) use controlled or quasi-experimental human tasks, usually brief or simulated. Four (16%) audit generated model outputs. Only two (8%) evaluate a computational prediction or decision model, and one of those uses corporate data. The remaining two papers are reviews. The corpus is therefore much richer on attitudes and advice than on deployed household forecasting.

B. Taxonomy coverage

The dominant-stage coding yields one UNDERSTAND paper, one ANTICIPATE paper, three PERSONALIZE papers, five ADVISE papers, three EXPLAIN papers, and twelve TRUST papers. TRUST includes adoption, resistance, vulnerability, bias, and governance, which explains its large share. The count does not mean that twelve papers verify the technical trustworthiness of a finished system.

A single dominant label inevitably hides overlap. P02 is placed under UNDERSTAND because transaction

interpretation anchors its joint model, even though it also predicts budgets. P14 occupies ANTICIPATE but uses corporate data. For that reason, the stage chart is best read as a map of emphasis rather than six separate and equally developed research areas.

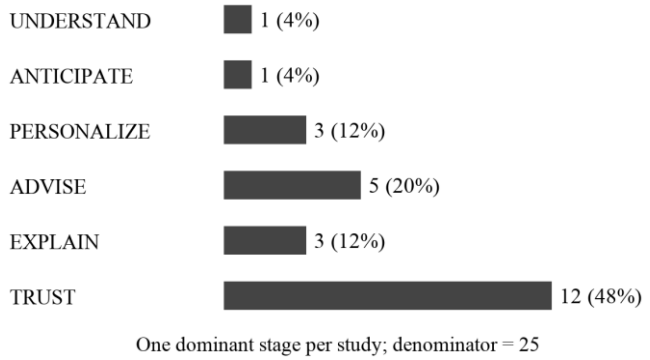


Fig. 2. Dominant taxonomy focus, n=25

C. Datasets and outcome comparability

The evidence comes from very different sources: transaction records in P02, corporate time series in P14, questionnaires, short decision tasks, and generated portfolios or messages. P03 and P08 use the same NFCS wave. P04 and P21 both describe 2,000 public employees and may overlap. We therefore do not add participant totals across papers. Artifact availability is also uneven: P02 provides a supplementary benchmark, P14 mixes public and confidential inputs, and P22 reports public code and data.

Only P02, or 4% of the corpus, directly studies personal transactions together with budget prediction. Adding P14 raises the number of computational-model papers to two, but its crisis-warning task concerns firms. This is enough to motivate technical experiments, not enough to describe learned forecasting as an established household solution.

TABLE IV. SELECTED PUBLISHED MODEL RESULTS

Study / task	Reported result	Boundary
P02 classification [2]	96.3% accuracy; 95.8% F1	52,847-record primary dataset
P02 budget [2]	6.2% MAPE	Study-specific forecasting setup
P14 warning [14]	88.2% accuracy; 5.6% false alarms	Corporate crisis warning

We do not pool the reported scores. Classification accuracy in P02, budget MAPE in the same study, and crisis-warning accuracy in P14 refer to different targets and errors. Most other papers report associations, perceptions, experimental effects, or qualitative judgments. Table IV records published literature values and does not present them as ARGUS results.

RESEARCH GAPS

The main gap is visible in the distance between component results and household outcomes. The corpus contains many short tasks and model outputs, but little evidence about months of use with irregular income, incomplete accounts, or changing obligations. We still do not know whether a lower prediction error reliably prevents missed payments, improves a cash buffer, or reduces harmful financial stress.

TABLE V. FINDINGS TO GAPS AND RESEARCH NEEDS

Corpus finding	Gap and required evidence
One direct personal prediction study (P02)	Transfer across languages, institutions and irregular incomes needs chronological, user-disjoint benchmarks.

Corpus finding	Gap and required evidence
11 observational surveys	Adoption intention is overrepresented. Prospective usage and objective outcomes are needed.
Prompt/version-sensitive advice (P01, P06, P16, P22, P23)	Repeated model audits, suitability checks and subgroup analysis need to accompany deployment.
Privacy concern in rich personalization (P05)	Perception studies should be complemented by measured retention, deletion and consent-withdrawal behavior.
Explanation effects vary (P10, P11, P19)	Measure comprehension and appropriate override, not trust or satisfaction alone.
Shared or overlapping samples (P03/P08, P04/P21)	Track dataset lineage before pooling respondents or treating findings as independent replication.

A second gap concerns ordinary failure cases. Receipts can be blurred, merchants can be ambiguous, categories can conflict, and an account history can be incomplete. A useful system needs to reveal these conditions and ask for review instead of converting uncertainty into a precise answer. Research should measure the work imposed on users when they correct those failures.

Fairness is the third gap. Output audits can identify different wording or recommendations, but they rarely follow the financial consequences for the affected person. Future studies need outcome measures, intersectional samples, and explicit attention to vulnerability. Equal error rates may still produce unequal harm when households have very different financial buffers.

RELEVANCE TO ARGUS

We use ARGUS as an inspected implementation case, not as proof that every reviewed idea has been realized. The main checkout was examined on 22 September 2026 at the recorded revision. The comparison distinguishes code that exists, behavior documented in archived evaluations, and capabilities that remain planned. This avoids turning a research roadmap into a statement about current product behavior.

TABLE VI. TAXONOMY TO IMPLEMENTATION MAPPING

Stage	Current ARGUS boundary
UNDERSTAND	Normalized keyword categorization, user correction precedence, reviewed OCR/notification candidates. No trained BERT classifier.
ANTICIPATE	Budget comparisons and period-change insights. Unsupported assistant forecasts are limited; no trained cash-flow forecaster verified.
PERSONALIZE	Owner-scoped facts, user corrections, budgets and temporal context. No learned digital twin or latent-class recommender.
ADVISE	Bounded suggestions with evidence, alternatives and trade-offs. No portfolio optimizer or autonomous money movement.
EXPLAIN	Rule factors, actual totals, period and uncertainty. No SHAP/LIME attribution implementation claimed.
TRUST	Validation, ownership, minimized context and review before capture persistence. No longitudinal trust or fairness outcome established.

Several papers explain why the project favors inspectable mechanisms. P02 motivates stronger categorization and future experiments with forecasting. P05 makes profile visibility and privacy controls relevant, while P06 and P23 justify checks for vulnerability and communication bias. P19 supports visible correction after errors. These studies guide design choices, but none evaluates the ARGUS implementation itself.

The implemented baseline is deliberately narrower than the product vision. Its value is that later learned components can be compared with deterministic behavior whose inputs and limits are known. A more complex model is worthwhile only if it improves a defined task without weakening ownership checks, exact-money handling, review, or traceability.

ARGUS METHODOLOGY AND ALGORITHMS

A. Architecture and verified workflow

The code separates the React Native client, API services, persistence, and provider adapters. Persisted financial records are controlled by the API, and capture inputs pass through validation before they become review candidates. Categorization, confidence, budget checks, and grounded question answering operate as explicit services. This structure keeps financial calculations outside presentation code and makes the main decision paths easier to inspect.

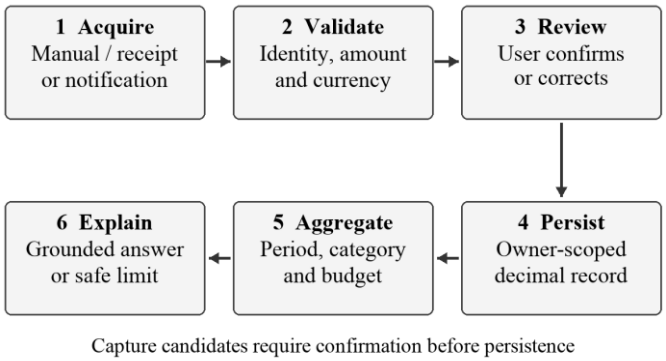


Fig. 3. Implemented ARGUS processing workflow

B. Transaction categorization

ExpenseCategorizationEngine normalizes the validated merchant and description text before matching it. An owner-specific correction is checked first and receives the fixed score 0.99. If no correction exists, the engine counts category-keyword matches and selects the largest count. With k matches, the score is min(0.92, 0.70 + 0.11k); an unrecognized transaction becomes OTHER with score 0.35. Scores below 0.75 remain subject to confirmation. These numbers are hand-set review signals, not calibrated probabilities or measured accuracy. The method is transparent, but substring ambiguity and limited dictionary coverage remain practical constraints.

C. Exact-money insight and budget rules

Currency values stay as decimal strings until they are aligned to a common scale and converted to integers. Period change is reported in basis points as q = trunc(10,000 x (C - P) / P), where C and P are the aligned current and previous totals. When P is zero, no percentage is produced. This prevents the system from presenting an undefined comparison as a meaningful increase.

InsightGenerator emits a period-change message when the absolute change reaches 1,000 basis points, or 10%. Its confidence is fixed at 0.88 for five or more transactions and 0.76 otherwise. BudgetSuggestionEngine acts only when a category budget exists and confirmed spending exceeds it; the response includes the amount, limit, alternatives, and trade-offs. Both mechanisms use design constants. Neither predicts a future budget from learned parameters.

D. Grounded financial questions

The assistant routes a normalized question through explicit intent patterns, resolves the time period, and loads authenticated financial facts. Supported requests cover spending totals, categories, largest expenses, comparisons, and budget status. Prediction requests, credential requests, ambiguous periods, and missing confirmed records produce bounded responses rather than invented values. The service identifies itself as local and rule based.

The provider interface is an extension boundary, not evidence of an external language model. In the

current path, financial answers come from backend ledger facts and exact calculations. A future provider may help phrase an explanation, but the source amounts, ownership checks, and unsupported-request rules still need to remain authoritative.

Algorithm 1. Concise implementation-derived procedure

Validate identity, ownership, and input schema.
If capture input exists, extract a candidate and persist it only after confirmation.
Normalize transaction text; apply an owner correction before generic keyword matching.
Resolve the requested period and load confirmed ledger facts and budgets.
Compare exact decimal amounts; emit trend or budget review messages only at defined thresholds.
Route the question to a supported fact path; limit unsupported requests and return evidence.

Source trace: ai-core categorization-engine, decimal, confidence, insight-generator, and budget-suggestion-engine; API intelligence and grounded-assistant services; and the capture services. The procedure summarizes their main branches. It does not mean that every request executes every step, and it is not evidence that the rules were trained or calibrated during this revision.

ARGUS SYSTEM EVALUATION AND RESULTS

A. Evaluation provenance and environment

The Phase 6 Evaluation Results document records checks run on 26 and 27 August 2026. We reproduce those archived observations rather than rerun them. The record does not describe a new user study, and later code inspection cannot be treated as a replication of the August measurements.

The recorded host was an Apple M2 Mac with 16 GiB of memory running macOS 26.5.2. The stack included Node.js 24.19.0, pnpm 11.9.0, local PostgreSQL, Expo SDK 57.0.16, and React Native 0.86.2. Mobile checks covered an Android emulator and a development APK; this environment is relevant when interpreting the timing and platform claims.

TABLE IX. ARCHIVED FUNCTIONAL AND SAMPLE RESULTS

Measure	Recorded outcome and interpretation
Automated checks	API 135; mobile JS 54; AI core 12; native notification 4. Total 205 passed. Software checks, not model accuracy.
Line coverage	API 80.11%; AI core 91.26%; mobile 52.43%. Measured before final focused tests were added.
Ask ARGUS	10/10 controlled questions acceptable; 2/2 unsupported requests limited; 0 grounding errors after fixes. Local-rule-based provider.
Receipt OCR	2/2 clear generated samples had correct amount/currency; 1 low-contrast sample failed with HTTP 422. No general OCR accuracy claim.
Notifications	Six non-transaction classes rejected; one debit and one credit became review candidates. Deterministic cases, not population sampling.
Fresh-process launch	n=5; mean 1,586.6 ms; median 1,662 ms; nearest-rank p95 1,798 ms. Metro/API available; emulator observation.

The two OCR examples were generated PNG receipts and do not form a representative receipt dataset. Capture remained review-first: extracted records and notification candidates required confirmation, and tests checked user ownership and permission boundaries. The results show behavior on the stated examples, not general OCR accuracy.

B. Local API timing

Seven authenticated API operations were sampled ten times each in sequence. At n=10, the nearest-rank p95 is simply the slowest observed request, so it is a descriptive maximum rather than a stable estimate of tail latency. The local network, database state, warm-

up behavior, and single host also limit comparison with production performance.

TABLE X. ARCHIVED LOCAL API ROUND TRIPS

Operation (n=10)	Mean ms	Median ms	p95 ms
Dashboard	3.74	3.14	9.27
Transaction list	6.33	4.24	24.24
Analytics	7.08	4.56	22.46
Budget utilization	7.15	4.26	23.24
Recommendations	3.13	2.72	5.25
Reminders	3.06	2.79	5.34
Ask ARGUS	12.79	10.36	32.91

Ask ARGUS timing and the ten-question grounding check came from the same controlled evaluation. They are related observations, not independent replications. The archived report also records setup details and known limitations that must travel with the headline numbers.

C. What the evaluation establishes

Within their stated cases, the archived checks support exact-money calculations, protected routes, owner-scoped services, review-first capture, grounded answers, and selected security controls. This is evidence of functional correctness for a bounded test set. It is useful for regression, but it is not a claim about every user, device, or financial situation.

The evaluation did not measure forecast MAE, RMSE, or MAPE because the project has no trained forecasting model. It also did not establish classifier generalization, financial-benefit effects, long-term retention, demographic fairness, or explanation comprehension. Those questions need different datasets and study designs.

D. Platform and temporal limits

The evaluated APK was a development-client debug build connected to local services. An iOS export was recorded, but no iOS device run or Android-notification equivalent was demonstrated. The measurements belong to the August environment and should be repeated after material dependency, device, database, or network changes.

DISCUSSION

The review reveals an imbalance. The personal-finance vision is broad, but direct technical evidence is concentrated in a small part of the corpus. Adoption and trust are studied often; only P02 directly combines personal transaction classification with budget prediction. That makes social acceptance important, but it leaves the central forecasting claim less established than the surrounding interface research.

The six stages help show where evidence enters the process. Financial meaning comes first, followed by any projection of future conditions. Goals and constraints then shape the options, and explanations connect those options back to evidence. Trust is not a final screen or a satisfaction score; it changes as users inspect results, correct errors, and experience the system over time.

Some apparent contradictions disappear once the outcome is named. Personalization can increase perceived relevance yet fail to predict adoption in another sample. Human advice can attract attention while AI or hybrid advice performs better on a simulated task. These findings address different mechanisms and populations, so combining them into a single verdict about 'AI advice' would be misleading.

ARGUS's rule-based design is a useful baseline precisely because its limitations are visible. It can calculate and explain supported ledger questions, but it cannot learn seasonal spending, infer future shocks, or replace professional judgment. A learned extension needs to beat this baseline on a defined workload while preserving ownership, exact-money arithmetic, correction, and auditability.

A. Threats to validity

The review is vulnerable to selection bias because it starts with a supplied corpus and lacks the historical

search log. The original audit also ended before the nominal 30 September date. Publication years are visible, but exact first-publication dates are not equally documented in every PDF. Coding was performed by one reviewer, and the dominant-stage labels simplify papers that cross several themes.

The implementation evidence has a separate boundary. It consists of archived measurements, controlled examples, one development host, an emulator, and a later source audit. The source audit confirms what code was present; it does not rerun the evaluation. Neither evidence stream supports claims about field effectiveness or population-level financial outcomes.

CONCLUSION

This survey reviewed 25 recent papers with an explicit extraction scheme, a comparative matrix, a technique analysis, and descriptive counts. The corpus contains 11 observational surveys, six human experiments, four model-output audits, two computational studies, and two reviews. Its strongest coverage concerns advice, adoption, explanation, and trust; direct evidence for household transaction forecasting is much thinner.

The practical conclusion is modest. Financial assistance benefits from traceable calculations, visible uncertainty, user correction, and bounded advice. ARGUS already implements deterministic categorization, exact-money comparisons, and ledger-grounded responses, but forecasting remains future work. The next research step is to test whether learned components improve a clearly defined task without weakening privacy, ownership, or auditability.

FUTURE RESEARCH DIRECTIONS

The first priority is a logged literature update covering the completed date window. Query strings, database exports, screening decisions, publication-date evidence, and duplicate lineage need to be retained. Additional papers should be added through that process rather than selected only because they support the proposed system.

The second priority is a consented personal-finance benchmark with chronological and user-disjoint splits. Simple historical baselines need to be reported beside any learned method. Evaluation should include classification by category, forecast calibration, shortfall detection, and performance for irregular-income and incomplete-account cases.

A third study should test explanations through what users can do after reading them. Useful outcomes include comprehension, detection of a seeded error, and appropriate override. Comparing a concise evidence-led explanation with a longer conversational version would help separate information value from tone and length.

Longer field studies are also needed. With informed consent, they could measure use, correction burden, budgeting behavior, and financial outcomes across several income cycles. Privacy controls deserve direct testing as well: deletion, retention, permission withdrawal, and recovery after an incorrect recommendation should be treated as study outcomes.

Evidence availability: the literature analysis is limited to P01-P25. Implementation statements come from the inspected ARGUS source, and system measurements come from the archived Phase 6 evaluation record. No new participant study, forecasting experiment, or production deployment was conducted for this revision.

References

[1]. Y. Baeckstrom and R. Matkovskyy, "Financial advice behaviour: humans versus AI," Journal of Corporate Finance, vol. 101, art. 103044, 2026, doi: 10.1016/j.jcorpfin.2026.103044.

[2]. Z. Yang, B. Biibosunov, S. Biibosunova, G. Kalmanbetova, and X. Wang, "A deep learning framework for automated transaction classification and budget prediction in personal financial management," Discover Artificial

Intelligence, vol. 6, art. 885, 2026, doi: 10.1007/s44163-026-02025-1.

[3]. O. Olajide, T. Chimbane, and Y. Lu, "Artificial Intelligence Use for Financial Advice and Financial Behaviors: The Role of Financial Knowledge," Financial Planning Review, vol. 9, e70045, 2026, doi: 10.1002/cfp2.70045.

[4]. E. Antonoudi and E. T. Ludwig, "Technology-Enabled Financial Help-Seeking: Evidence of AI and Financial Planners on Retirement Saving Behavior," Financial Planning Review, vol. 9, e70043, 2026, doi: 10.1002/cfp2.70043.

[5]. M. I. Bonelli, "Financial Digital Twins and Conversational AI in Robo-Advisory: Evidence from a Scenario-Based Randomized Experiment," FinTech, vol. 5, art. 57, 2026, doi: 10.3390/fintech5030057.

[6]. B. Zeka and R. Johns, "What Could Possibly Go Wrong? A Consumer-Centric Approach to Exploring Vulnerability and Algorithmic Bias in AI-Driven Financial Planning," Financial Planning Review, vol. 9, e70040, 2026, doi: 10.1002/cfp2.70040.

[7]. A. Takemeyang, H. J. Ketuma, and T. A. Akpor, "The mediating role of customer value proposition in the relationship between digital financial literacy and AI-driven finance adoption," Discover Artificial Intelligence, vol. 6, art. 1029, 2026, doi: 10.1007/s44163-026-01329-6.

[8]. I. Chawla, M. Joseph, K. White, and C. W. Scantling, "AI for Financial Advice, Fraud Loss, and the Moderating Effect of Financial Knowledge Miscalibration," International Journal of Financial Studies, vol. 14, art. 137, 2026, doi: 10.3390/ijfs14060137.

[9]. P. Sungkarungsri and S. Kiattisin, "How trustworthy AI fosters adoption and user advocacy in financial advisory services," Discover Artificial Intelligence, vol. 6, art. 455, 2026, doi: 10.1007/s44163-026-01451-5.

[10]. M. Brunetti, M. G. Ceravolo, V. Farina, L. Fattobene, U. Filotto, and L. Leonelli, "The human touch: Evidence from an eye-tracking experiment on financial advice," Finance Research Letters, vol. 94, art. 109664, 2026, doi: 10.1016/j.frl.2026.109664.

[11]. A. Ioannou, M. M. Tavalaei, D. Vatn, and P. Mikalef, "The Role of Explainability in AI-Driven Financial Decision Making," Information Systems Frontiers, 2026, doi: 10.1007/s10796-026-10727-1.

[12]. S. S. Kumta, N. Panakaje, S. M. R. Parvin, and A. M. Nikzad, "Can you smell what AI is advising? Human, machine, or both: A comparative experimental study on financial advice," Acta Psychologica, vol. 264, art. 106530, 2026, doi: 10.1016/j.actpsy.2026.106530.

[13]. T. Szopinski, M. Buszko, and M. Porada-Rochon, "Acceptance of AI-based tools in consumer financial decision-making: An application of the extended technology acceptance model," PLOS ONE, vol. 21, no. 3, e0344901, 2026, doi: 10.1371/journal.pone.0344901.

[14]. R. Zhang, "Fuzzy decision support system for financial planning and management," Discover Artificial Intelligence, vol. 6, art. 128, 2026, doi: 10.1007/s44163-025-00824-6.

[15]. Z. Wang, R. Yuan, B. Li, V. Kumar, and A. Kumar, "An Empirical Study of AI Financial Advisor Adoption Through Technology Vulnerabilities in the Financial Context," Journal of Product Innovation Management, vol. 43, pp. 14-30, 2026, doi: 10.1111/jpim.12795.

[16]. M. T. T. Schlosky and S. Raskie, "ChatGPT as a Financial Advisor: A Re-Examination," Journal of Risk and Financial Management, vol. 18, art. 664, 2025, doi: 10.3390/jrfm18120664.

[17]. Y. Tang and H. Son, "How Perceived Value Drives Usage Intention of AI Digital Human

Advisors in Digital Finance," Systems, vol. 13, art. 973, 2025, doi: 10.3390/systems13110973.

[18]. V. Sharma and B. Priya, "Bridging the gap: AI-powered FinTech and its impact on financial inclusion and financial well-being," Discover Artificial Intelligence, vol. 5, art. 290, 2025, doi: 10.1007/s44163-025-00465-9.

[19]. J. Han and D. Ko, "Trust Formation, Error Impact, and Repair in Human-AI Financial Advisory: A Dynamic Behavioral Analysis," Behavioral Sciences, vol. 15, art. 1370, 2025, doi: 10.3390/bs15101370.

[20]. B. Verma, M. Schulze, D. Goswami, and K. Upreti, "Artificial intelligence attitudes and resistance to use robo-advisors: exploring investor reluctance toward cognitive financial systems," Frontiers in Artificial Intelligence, vol. 8, art. 1623534, 2025, doi: 10.3389/frai.2025.1623534.

[21]. C. R. Bennetts and E. Ludwig, "The desire for AI advice in retirement plans: A latent class analysis," Financial Planning Research Journal, 2025, doi: 10.2478/fprj-2025-0007.

[22]. P. Winder, C. Hildebrand, and J. Hartmann, "Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors," PLOS ONE, vol. 20, no. 6, e0325459, 2025, doi: 10.1371/journal.pone.0325459.

[23]. C. R. Bennetts and E. T. Ludwig, "Echoes of Bias: An Analysis of ChatGPT in Financial Planner-Client Dialogs," Financial Planning Review, vol. 8, e70006, 2025, doi: 10.1002/cfp2.70006.

[24]. N. M. Namyslo, D. Jung, and T. Sturm, "The state of robo-advisory design: A systematic consolidation of design requirements and recommendations," Electronic Markets, vol. 35, art. 20, 2025, doi: 10.1007/s12525-025-00762-2.

[25]. S. Sai, K. Arunakar, V. Chamola, A. Hussain, P. Bisht, and S. Kumar, "Generative AI for Finance: Applications, Case Studies and Challenges," Expert Systems, vol. 42, e70018, 2025, doi: 10.1111/exsy.70018.