Vol. 8, No. 2, April 2026

Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

ACTE: A Pilot Feasibility Evaluation of a Mastery-Aware Task Recommender for Mobile Language Learning in Real-World Contexts

Yudhy Setyo Purwanto[1], Rahmat Gernowo[2], Dinar Mutiara Kusumo Nugraheni[3]

[1]Doctoral Program of Information Systems, School of Postgraduate Studies, Diponegoro University, Indonesia
[2]Department of Physics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia
[3]Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Semarang, Indonesia
[1]Department of Informatics, Faculty of Energy Telematics, Institut Teknologi PLN, Jakarta, Indonesia

Received:
December 2, 2025
Revised:
March 11, 2026
Accepted:
March 31 2026
Published:
April 12, 2026

Corresponding Author:
Author Name*:
Yudhy Setyo Purwanto
Email*:
y.purwanto@itpln.ac.id

DOI:
10.63158/journalisi.v8i2.1554

Abstract. Mobile-assisted language learning (MALL) apps often present generic activities that ignore the semantic meaning of real-world places and provide limited skill-specific, mastery-based progression. This pilot feasibility study introduces the Adaptive Contextual Task Engine (ACTE), a lightweight on-device recommender that personalizes tasks using location semantics, CEFR-aligned modules, mastery status, and performance timing. ACTE was evaluated with 10 university students aged 18–23 in a simulated café environment to balance ecological validity and experimental control. Participants completed three A2 speaking tasks, the System Usability Scale (SUS), and a five-item relevance questionnaire. Results showed a mean SUS score of 72.0, exceeding the benchmark of 68. Participants rated task appropriateness for the café at 4.3/5 and real-life usability at 4.7/5, while 90% agreed that the tasks reflected authentic language use. Qualitative feedback confirmed contextual authenticity but indicated the need for clearer scoring explanations. These findings suggest that ACTE offers a practical, privacy-conscious, and replicable framework for situated MALL by linking semantic place affordances with mastery-based progression in controlled real-world simulations.

Keywords: mobile-assisted language learning (MALL), context-aware task recommendation, CEFR-based task design, mastery learning, situated learning

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

Language learning does not flourish in isolation; it thrives in the flow of everyday life. Whether ordering coffee, asking for directions, or discussing a museum exhibit, meaningful language use emerges from real-world contexts where words carry purpose, social weight, and cultural nuance. Mobile-assisted language learning (MALL) has emerged as a promising bridge between virtual practice and real-world application, yet its effectiveness depends critically on how well it aligns with the situated nature of language use [1], [2], [3], [4]. Despite decades of research affirming the value of situated cognition, many MALL applications continue to treat learning as a context-free transaction: flashcards on a screen, scripted dialogues in a void, disconnected from the places where language actually lives [5], [6], [7].

Contemporary MALL systems often leverage location data as a technical trigger, using GPS coordinates to push content, but rarely engage with the semantic meaning of places. A café, a hospital, or a train station is more than a latitude and longitude; it is a social space with predictable interactions, roles, and linguistic demands. Early location-based MALL systems used GPS primarily as a trigger mechanism, delivering content when users entered predefined geofenced zones, but rarely considered the semantic affordances of those spaces [8], [9], [10]. More recent work has begun to incorporate place semantics through points of interest (POIs) or user-tagged locations [24], [30], [31], yet tasks often remain generic (e.g., 'practice vocabulary') rather than tied to the communicative demands of the setting (e.g., 'request a refill' in a café). Without this semantic grounding, location-awareness remains superficial, failing to align task design with the affordances of real environments [11], [12], [13], [14], and consequently, learners receive tasks that may be linguistically appropriate but contextually irrelevant, practicing medical vocabulary while standing in a bookstore, for instance, undermining the very immersion these systems claim to support.

Moreover, most MALL apps operate with a one-size-fits-all approach to skill development, cycling indiscriminately through speaking, listening, reading, and writing without allowing learners to focus on a single modality at a time. This lack of skill granularity overlooks the cognitive reality that learners often seek to improve specific competencies in

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

response to immediate needs, such as mastering transactional speaking before a trip abroad [15], [16], [17]. Research shows that L2 learners often experience skill-specific challenges (particularly in speaking) and benefit from focused, contextual rehearsal [18], [19], [20], yet few systems support intentional, user-driven skill selection within real-world settings. Adaptive task sequencing in MALL has largely relied on two paradigms: knowledge tracing models (e.g., Bayesian networks) and collaborative filtering. While powerful in data-rich environments, these approaches require extensive historical interaction logs and are ill-suited to lightweight, privacy-conscious mobile applications [21], [22], [23]. Moreover, they often optimize for engagement or prediction accuracy rather than pedagogical soundness. In contrast, mastery learning, originating in Bloom's work, emphasizes that learners should demonstrate proficiency before advancing. Recent implementations in digital learning [24], [25], [26] show that mastery-based progression improves long-term retention, yet these models are rarely integrated into mobile language contexts where tasks are fleeting and context-bound.

The Common European Framework of Reference for Languages (CEFR) provides a widely adopted standard for describing language proficiency, but its practical implementation in MALL remains inconsistent. Many apps label content with CEFR levels without ensuring alignment in task complexity, linguistic demands, or cognitive load [27], [28], [29]. A few studies have attempted structured mappings, such as linking A2 listening tasks to 'identifying main ideas in short announcements' [30], but these mappings are seldom built into algorithmic logic or tied to real-world scenarios. Consequently, CEFR labels often function as marketing tags rather than operational design constraints, limiting their utility for personalized, context-aware delivery. Efforts to bridge context and pedagogy have emerged in augmented reality (AR) and immersive MALL. For instance, AR applications overlay vocabulary labels on real-world objects [31], [32], and conversational agents simulate service interactions in virtual cafés [33], [34], [35]. While promising, these approaches often require high-bandwidth connections, specialized hardware, or controlled environments, reducing their accessibility for everyday, on-the-go learning. Furthermore, they tend to simulate contexts rather than leverage actual ones, missing opportunities for authentic, unscripted engagement with real places and people [36], [37].

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Another strand of research explores 'just-in-time' or 'opportunistic' learning, where systems detect teachable moments based on user behavior or environment [38], [39]. However, these systems frequently prioritize when to prompt over what to deliver, and when content is selected, it is often drawn from flat, undifferentiated pools. The critical layer of task appropriateness (such as matching module semantics, skill focus, learner level, and place affordances) is rarely addressed algorithmically. Godwin-Jones [5], [6] observes that 'The next frontier in MALL is not more data, but more meaning: tasks that feel inevitable because they fit the moment.' This observation resonates with sociocultural theories of learning [40], [41], [42], [43] and ecological perspectives on language acquisition [44], [45], which position real-world places (e.g., cafés, markets, transit hubs) not as passive backdrops, but as active sites of linguistic and cultural meaning. Learning, in this view, happens not in isolation, but in the ordinary moments of everyday life.

Despite advances in mobile-assisted language learning, a critical gap remains: most context-aware systems treat location as mere coordinates rather than semantic places with cultural and communicative affordances [6], [46], while adaptive recommenders often prioritize engagement metrics over mastery progression [21], [22]. Few systems integrate where the learner is with what they have mastered to deliver pedagogically grounded, non-intrusive practice. This disjunction between technical context-awareness and pedagogical relevance limits not only learning effectiveness but also learner agency and authenticity.

To address this gap, we propose the Adaptive Contextual Task Engine (ACTE), a lightweight, on-device recommender framework that selects language tasks based on semantic place understanding, CEFR-aligned modules, and mastery status. ACTE's novelty lies in its 7-module badge architecture per skill–location–level combination and its time-sensitive scoring function that rewards both accuracy and readiness without gamification or cloud dependency. This paper reports a pilot feasibility evaluation of ACTE with ten university students in a simulated café environment, assessing usability (via SUS) and perceived contextual relevance. Our contribution is threefold: (1) a reproducible rule-based algorithm for context-aware task selection, (2) empirical evidence that semantic alignment enhances perceived authenticity in MALL, and (3) a privacy-conscious design

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

model that demonstrates how intelligent systems can support situated learning without behavioral surveillance, offering a scalable foundation for future work in educational information systems.

2. ADAPTIVE CONTEXTUAL TASK ENGINE (ACTE) ALGORITHM DESIGN

The Adaptive Contextual Task Engine (ACTE) is a lightweight algorithm designed to answer a deceptively simple question: *Given that a learner is ready and present in a meaningful place, what language task should they be offered next?* ACTE does not determine readiness, that role is reserved for external mechanisms or explicit user initiation. Instead, ACTE operates only when a "stay-state" is confirmed (e.g., the learner is stationary in a café for at least 60 seconds), ensuring that recommendations are made not just in context, but in a context of *attentive presence*. This separation of concerns (readiness detection versus task selection) allows ACTE to remain focused on pedagogical appropriateness without overloading its logic with sensory fusion or behavioral inference.

At its core, ACTE is built upon a structured module taxonomy that aligns language tasks with real-world semantics. The taxonomy is organized as a four-dimensional space: Location Type × Language Skill × CEFR Level → Module Pool. For any given location (e.g., *café*), skill (e.g., *speaking*), and CEFR level (e.g., *A2*), ACTE draws from a fixed set of seven contextual modules, each representing a distinct communicative scenario: *Ordering Menu*, *Polite Complaint*, *Splitting the Bill*, *The Taste Critic*, *Person Description*, *Opinion Sharing*, and *Experience Journaling*. Each module is pre-annotated with metadata including category (e.g., *Transactional*, *Descriptive*), expected duration (d_{ref}), modality constraints (e.g., requires microphone), and semantic affinity to location types. This design ensures that task selection is not arbitrary but grounded in the affordances of everyday life. As we can see from Figure 1, the ACTE algorithm progresses through four stages (input, selection, performance, and scoring) to deliver contextually resonant tasks while embodying principles of contextual resonance, skill granularity, mastery momentum, and pedagogical transparency.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

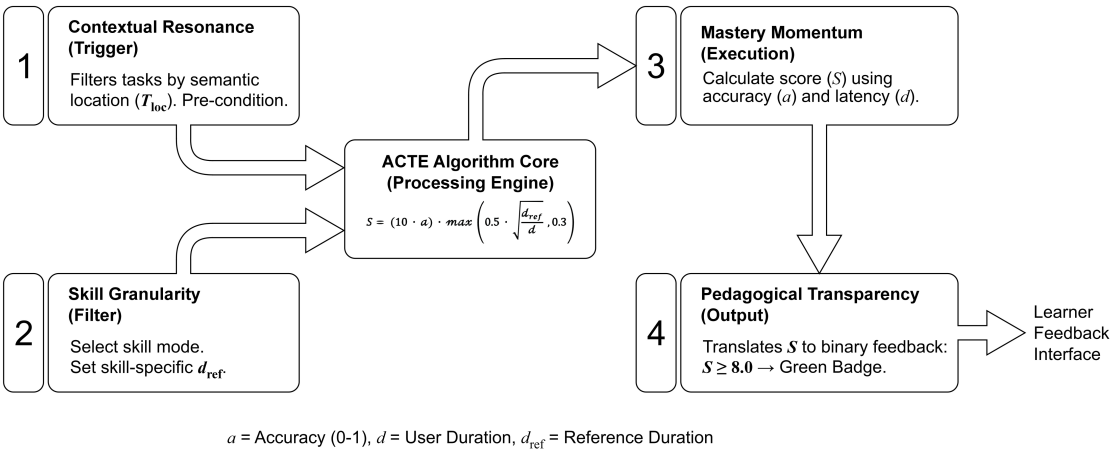

Figure 1. ACTE Algorithm Relation to 4 Principles of the App.

Table 1 illustrates how semantic place affordances are operationalized into seven distinct speaking tasks for an A2 learner in a café context.

Table 1. ACTE Module Taxonomy – A2 Speaking in a Café

Module Name	Category	Duration (sec)	Typical Task
Ordering Menu	Transactional	8	*"Order one coffee and a croissant, please."*
Polite Complaint	Transactional	8	*"Complain politely that your coffee is too cold."*
Splitting the Bill	Social	10	*"Ask your friend how to split the bill evenly."*
The Taste Critic	Expressive	12	*"Describe why you like or dislike this café."*
Person Description	Descriptive	12	*"Describe your friend sitting across from you."*
Opinion Sharing	Expressive	10	*"Say whether you prefer tea or coffee and why."*
Experience Journaling	Reflective	15	*"Briefly recall your last visit to a café."*

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Task selection in ACTE follows a mastery-aware, interleaving-sensitive logic we refer to as the *Picking Game*. For a target skill s, location ℓ, and CEFR level L, let $\mathcal{M} = m_1, \ldots, m_7$ denote the module pool, where each m_i has a binary mastery state $\Phi_i \in 0,1$ (0 = not mastered, 1 = Green Badge earned) and a semantic category c_i. The algorithm computes a *PickScore* for each candidate:

$$\text{PickScore}(m_i) = 10 \cdot (1 - \Phi_i) + 2 \cdot \mathbb{I}[c_i \neq c^{\text{prev}}] + \epsilon_i \quad (1)$$

Where:

c^{prev} is the category of the last completed task,

$\mathbb{I}[\cdot]$ is the indicator function, and

$\epsilon_i \sim$ Uniform $(0,0.1)$ introduces min. randomness to prevent deterministic loops.

Non-mastered tasks are heavily prioritized (weight = 10), while category switching is gently encouraged (weight = 2) to implement the pedagogical principle of interleaving [47]. The module with the highest PickScore is selected for delivery.

Once a task is attempted, ACTE evaluates performance through a dual-component Scoring Race that balances linguistic accuracy and behavioral readiness. Let:

$a \in [0,1]$ denote task accuracy (computed skill-specifically, as detailed below)

d the actual completion duration in seconds,

d_{ref} denote the reference duration for the task type (e.g., 8 sec. for *Ordering Menu*)

The final score is computed as:

$$S_{\text{final}} = (10 \cdot a) \cdot \max\left(0.5 \cdot \sqrt{\frac{d_{\text{ref}}}{d}}, 0.3\right) \quad (2)$$

The square-root penalty gently discounts slow responses, reflecting hesitation or cognitive load, without overly penalizing thoughtful pauses. A soft floor of 0.3 ensures that even very slow but accurate attempts retain partial credit. Mastery is awarded only if $S_{\text{final}} \geq 8.0$, at which point Φ_i is set to 1 and a Green Badge is displayed. This threshold

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

ensures that progression reflects both correctness and fluency, aligning with CEFR's emphasis on *effective communication*.

Accuracy computation is skill-adaptive and designed for on-device feasibility:

a) Speaking tasks: $a = 0.6 \cdot$ ASR confidence $+ 0.4 \cdot$ keyword coverage, where keyword sets are defined per module (e.g., {"coffee", "please", "one"} for *Ordering*).
b) Listening and reading tasks: accuracy is derived from task-specific correctness (e.g., exact match for multiple-choice, semantic similarity via lightweight Sentence-BERT for short answers).
c) Writing tasks: $a = 0.7 \cdot$ task completion $+ 0.3 \cdot$ linguistic quality, with completion verified via regex/keyword rules and quality assessed through on-device grammar heuristics and CEFR-aligned vocabulary lists.

All processing is designed to occur locally on the device, minimizing data transmission and protecting learner privacy. Table 2 provides the reference durations and skill-specific accuracy formulas that ensure fair and consistent performance evaluation across task types.

Table 2. Reference Durations and Accuracy Formulas by Task Type

Task Type	Duration (sec)	Accuracy (a) Computation
Speaking – Transactional	8	$0.6 \cdot$ ASR conf. $+ 0.4 \cdot$ keyword_coverage
Speaking – Expressive	12	$0.6 \cdot$ ASR conf. $+ 0.4 \cdot$ semantic_completeness
Listening – MCQ	6	1 if correct, 0 otherwise
Reading – Short Answer	10	Cosine similarity (response, model_answer)
Writing – Guided Email	14	$0.7 \cdot$ task completion $+ 0.3 \cdot$ grammar_score

Finally, ACTE embodies a philosophy of *pedagogical transparency*: it reveals progress without pressure. The Green Badge system makes mastery visible but optional, learners may ignore it or engage with it as a quiet guide. There are no countdown timers, leaderboards, or loss aversion mechanics. Instead, ACTE offers a calm, consistent rhythm: *practice, reflect, earn, move forward*. In doing so, it operationalizes the vision of Contextual Immersion Learning, not as a system that teaches, but as a companion that

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

notices, suggests, and respects the learner's journey through the ordinary moments of language use. The ACTE workflow is illustrated in Figure 2 below.

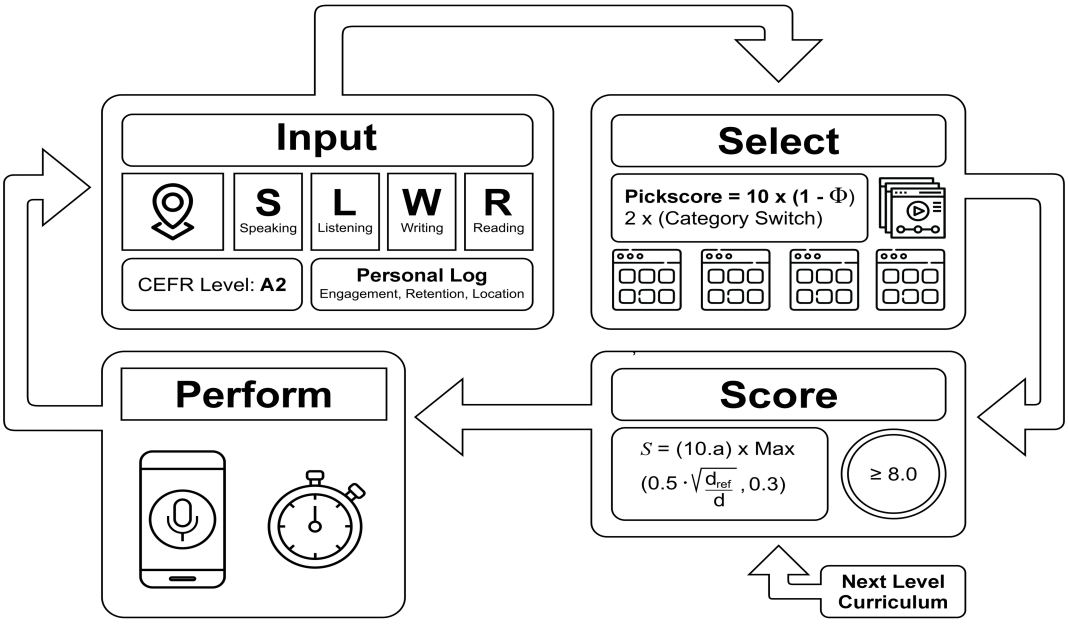

Figure 2. ACTE Workflow

3. METHODOLOGY

3.1. Research Design and Workflow

This study employs a pilot feasibility evaluation to assess the usability, perceived relevance, and functional integrity of the Adaptive Contextual Task Engine (ACTE). As a formative investigation, the goal was not to test hypotheses but to gather preliminary evidence on whether ACTE's design principles (contextual resonance, skill granularity, mastery momentum, and pedagogical transparency) translate into a coherent and meaningful user experience. This approach aligns with design science research paradigms in educational technology, where iterative build-and-evaluate cycles are essential before large-scale deployment [48], [49], [50].

The research followed a five-stage workflow as shown in Figure 3. The details as follow.

1) Algorithm Design: ACTE's core logic was specified, including module taxonomy, PickScore selection, and time-sensitive scoring functions.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

2) Module Construction: Seven speaking tasks were designed for A2-level learners in a café context, with reference durations and accuracy formulas defined.
3) Android Implementation: A functional prototype was developed using Android Studio, integrating on-device speech recognition and scoring logic.
4) Pilot Evaluation: Ten participants completed three ACTE-selected tasks in a controlled simulated environment and provided quantitative and qualitative feedback.
5) Data Analysis: Descriptive statistics and thematic analysis were applied to usability scores, relevance ratings, and open-ended responses.

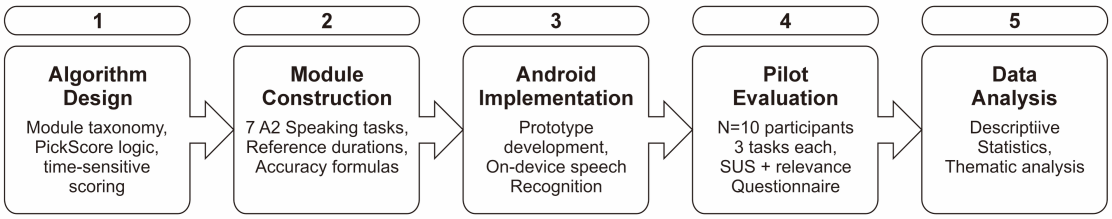

Figure 3. Pilot Test Work Flow

This systematic workflow ensures methodological transparency and reproducibility, allowing other researchers to replicate or extend the study.

3.2. Participants Recruitment and Profile

Ten university students were recruited from engineering programs at a private university in Indonesia. Participants were aged 18 to 23 (M = 20.0), with six female and four male participants, all reporting Bahasa Indonesia as their first language. Eight participants also speak English as a second or third language, while two reported regional languages (e.g., Javanese, Sundanese) as additional linguistic resources.

English proficiency was self-assessed using the CEFR Self-Assessment Grid [49], a validated instrument widely used in MALL research for its accessibility and reliability [28], [51]. Self-ratings ranged from A2 (n = 5) to B2 (n = 1), with the majority at A2 or B1. Participants were selected via convenience sampling with purposive emphasis on linguistic diversity and mobile device familiarity. Informed consent was obtained prior to participation, and all data were anonymized using participant codes (P01–P10). The result can be seen in Table 3.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

TABLE 3. Participant Demographics (N = 10)

Characteristic	Value
Age range	18–23 years
Mean age	20.0 years
Gender	6 Female, 4 Male
Native language	100% Bahasa Indonesia
Additional languages	80% English, 30% local languages (e.g., Javanese, Sundanese)
Self-rated CEFR level	A2: 50% (n=5), B1: 40% (n=4), B2: 10% (n=1)

3.3. ACTE Framework: Design Rationale and Parameter Selection

ACTE is proposed as a lightweight, context-aware recommender framework for mobile language learning. The framework integrates four core components:

3.3.1. Module Taxonomy Design

The seven modules (*Ordering Menu*, *Polite Complaint*, *Splitting the Bill*, *The Taste Critic*, *Person Description*, *Opinion Sharing*, *Experience Journaling*) were selected through an iterative design process involving:

1) Literature review of common café-based communicative scenarios in MALL research [52],
2) Expert consultation with two language instructors to ensure pedagogical appropriateness for A2 learners,
3) Pilot testing with three non-participant students to validate task clarity and cultural relevance.

Each module was assigned a reference duration (d_{ref}) based on:

1) Typical utterance length for the task type (e.g., 8 seconds for transactional exchanges, 12–15 seconds for expressive tasks),
2) CEFR A2 speaking benchmarks for fluency and complexity [51],
3) Preliminary timing tests with native speakers to establish realistic baselines.

3.3.2. Scoring Parameter Justification

The PickScore weights (10 for non-mastered tasks, 2 for category switching) were selected to:

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1) Strongly prioritize mastery progression (weight = 10 ensures unmastered tasks are always selected first),
2) Gently encourage interleaving (weight = 2 provides variety without disrupting mastery focus),
3) Align with spaced repetition and varied practice principles in second language acquisition [53], [54], [55].

The mastery threshold ($S_{\text{final}} \geq 8.0$) was chosen to:

1) Require high accuracy (≥80% on the 10-point scale),
2) Balance fluency (timing penalty) with correctness (accuracy score),
3) Reflect CEFR's emphasis on effective communication rather than perfection [54].

The soft floor (0.3) in the scoring function ensures that slow but accurate attempts retain partial credit, preventing excessive penalty for thoughtful pauses while still encouraging readiness.

3.3.3. Expert and Pre-Pilot Validation

To ensure pedagogical and cultural appropriateness, the seven modules underwent two rounds of validation prior to the main pilot:

1) Expert consultation: Two certified English instructors (5+ years teaching A2 learners in Indonesia) reviewed all modules for linguistic appropriateness, CEFR alignment, and task clarity. Key feedback included:
   a) Adding *Polite Complaint* to address a common real-world need in Indonesian café culture,
   b) Simplifying vocabulary in *Experience Journaling* to match A2 lexical resources,
   c) Ensuring all prompts avoided culturally specific references unfamiliar to urban Indonesian youth.
2) Pre-pilot testing: Three non-participant university students (A2–B1 English proficiency) completed all seven tasks in a simulated environment. Feedback focused on:
   a) Task clarity: All prompts were understood without ambiguity,
   b) Cultural relevance: Scenarios felt authentic to Indonesian café experiences,

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

c) Timing: Reference durations were adjusted upward by 1–2 seconds for expressive tasks based on observed hesitation patterns.

All suggested refinements were incorporated before the main pilot evaluation. The full validation protocol and anonymized feedback summaries are available from the corresponding author upon request.

3.4. Prototype Implementation and Standardization

A functional Android prototype of the ACTE-enabled mobile application was developed using Visual Studio Code 1.109.5 (user setup), Flutter SDK 3.35.5, Dart 3.9.2, with Java JDK 21.0.4 (LTS), Android SDK/API level: Platform Android-36 (Booth-Tools version 36.1.0). The app was designed to simulate a café environment while maintaining strict control over experimental conditions.

3.4.1. Simulated Café Environment and Device Standardization

To ensure consistency across participants, the pilot evaluation was conducted in a controlled laboratory setting configured to simulate a café environment. A single Android smartphone (Samsung Galaxy A54, Android 14) was provided by the researcher and used by all participants. The device's microphone gain was fixed at 75% to ensure consistent audio input levels across sessions.

The laboratory space was arranged to evoke a café ambiance: participants sat at a table with the smartphone placed 30 cm in front of them, while a secondary monitor displayed a continuous video loop of a typical urban Indonesian café interior (tables, customers, barista activity). Ambient audio (65 dB) featuring coffee machine sounds, light chatter, and background music was played through room speakers to enhance contextual immersion. This multi-sensory simulation (visual (monitor), auditory (ambient sound), and spatial (table setting)) aimed to approximate real-world café conditions while maintaining experimental control.

All speech recognition was processed on-device using Android's SpeechRecognizer API (language set to English US). Failed recognitions (e.g., no speech detected) were logged and excluded from analysis. This controlled setup eliminated device variability and

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

ensured that all participants experienced identical environmental and technical conditions.

Figure 4. Laboratory setup simulating a café environment: (a) smartphone with ACTE app, (b) monitor displaying café video loop and ambient audio playback

3.4.2. Task Delivery and Scoring

For each participant, the app:

1) Initialized ACTE with A2 speaking modules and all $\Phi_i = 0$ (no prior mastery),
2) Selected the first task using PickScore (Equation 1),
3) Displayed the contextual prompt and recorded the user's spoken response,
4) Computed S_{final}using the scoring function (Equation 2),
5) Displayed immediate feedback ("8.6 → Green Badge earned!" or "7.2 → Try again later"),
6) Updated $\Phi_i$if mastery was achieved,
7) Repeated for two additional tasks (total of 3 tasks per participant).

System logs recorded task scores, durations, badge outcomes, and any technical errors.

3.5. Data Collection Instruments

Data collection included both quantitative and qualitative measures.

3.5.1. System Usability Scale (SUS)

The SUS, a 10-item validated instrument [56], was administered post-session to assess overall usability. SUS remains the gold standard for lightweight usability evaluation in mobile contexts due to its reliability, brevity, and normative benchmarks [57]. Scores were

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

computed following standard protocol: odd items scored as (response − 1), even items as (5 − response), summed and multiplied by 2.5 to yield a 0–100 scale.

3.5.2. Task Relevance Scale

A custom 5-item relevance questionnaire captured participant perceptions using 5-point Likert scales:

1) The tasks felt appropriate for a café setting,
2) I can imagine using these language skills in real life,
3) The scoring (e.g., "8.6 / 10") helped me understand how well I did,
4) The "Green Badge" system motivated me to do my best,
5) The time limit felt fair and realistic.

3.5.3. Open-Ended Feedback

Three qualitative prompts invited reflection:

1) What did you like most about the tasks or the app?
2) What could be improved?
3) Any additional thoughts or suggestions?

All responses were collected via Google Forms, and system logs captured performance metrics. No voice recordings were stored beyond the moment of scoring.

3.6. Data Analysis Procedures

Quantitative data were analyzed using descriptive statistics. SUS scores and relevance ratings were summarized using means, standard deviations, and the percentage of participants rating 4 or 5 ("Agree" or "Strongly Agree"). Qualitative responses underwent thematic analysis using an inductive coding approach, consistent with Braun and Clarke's [58], [59] reflexive method. Two researchers independently identified recurring themes (e.g., "contextual realism," "timing concerns"), resolved discrepancies through discussion, and grouped exemplar quotes. All analyses were conducted using Microsoft Excel to ensure transparency and reproducibility.

3.7. Ethical Considerations and Pilot Scope

This pilot study involved minimal risk and was exempt from full ethical review under Indonesian research guidelines for educational technology studies with anonymized data.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Informed consent was obtained from all participants, who were informed they could withdraw at any time without penalty. Scope Limitations: This evaluation represents an initial feasibility pilot with a small sample (N = 10), a single simulated location (café), and one skill focus (A2 speaking). While these constraints limit generalizability, they are appropriate for a formative study aimed at validating ACTE's core logic and user experience before broader deployment. Future work will expand to diverse locations, multiple skills, and real-world field testing.

4. RESULTS AND DISCUSSION

4.1. Pilot Feasibility Evaluation Outcomes

This subsection presents the empirical outcomes of the pilot feasibility evaluation of the Adaptive Contextual Task Engine (ACTE). The evaluation involved 10 university students, each of whom completed three ACTE-selected A2 speaking tasks in a simulated café environment, resulting in a total of 30 task attempts. After completing the tasks, participants responded to the System Usability Scale (SUS) and a five-item relevance questionnaire, and they also provided brief open-ended comments about their experience. The findings reported in this subsection are presented descriptively and focus on usability, contextual relevance, task performance, and participant feedback.

The usability findings indicate that ACTE achieved a mean SUS score of 72.0 with a standard deviation of 12.3, placing the system above the commonly used benchmark score of 68 for acceptable usability [61]. Individual SUS scores ranged from 50.0 to 90.0, showing that most participants perceived the system positively, although some variation in user experience was observed. More specifically, six out of ten participants reported SUS scores above 70, suggesting that the system was generally seen as usable within the simulated learning scenario. This overall comparison is illustrated in Figure 6, which shows that the mean SUS score exceeded the benchmark threshold and therefore met the expected level for acceptable usability in a pilot study context.

Perceived relevance was measured using a 5-point Likert scale, where 1 represented Strongly Disagree and 5 represented Strongly Agree. The questionnaire examined whether the tasks matched the simulated café setting, whether the language felt useful

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

for real-life communication, and whether key system elements such as scoring, badges, and time limits were meaningful to users. As summarized in Table 4 and visually presented in Figure 5, the strongest ratings were associated with contextual fit and practical usefulness. The statement "I can imagine using these language skills in real life" received the highest mean score of 4.7 with a standard deviation of 0.48, and 90% of participants selected Agree or Strongly Agree. The statement "The tasks felt appropriate for a café setting" received a mean of 4.3 with a standard deviation of 0.67, with 80% of participants expressing agreement. In Figure 5, these two items appear as the highest-rated dimensions, highlighting the strong perceived connection between the selected tasks, the café scenario, and authentic language use.

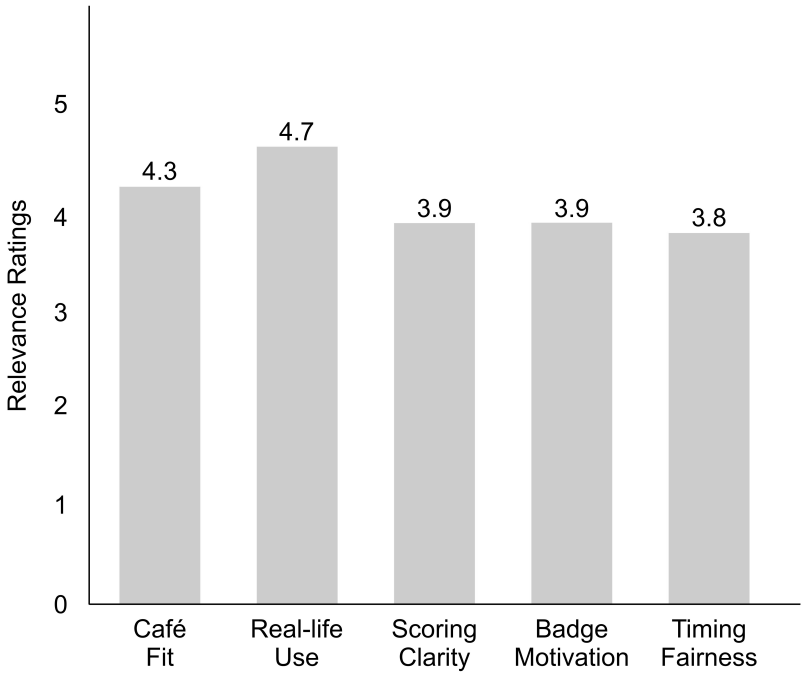

Figure 5. Relevance Scores Ratings

The remaining three relevance items received moderately positive responses, with somewhat greater variation across participants. The item "The scoring (e.g., '8.6/10') helped me understand how well I did" obtained a mean score of 3.9 with a standard deviation of 0.74, and 70% of respondents agreed. The item "The 'Green Badge' system motivated me to do my best" also produced a mean of 3.9, though with a higher standard deviation of 0.88, indicating greater differences in how participants experienced the motivational effect of the badge mechanism. Only 60% of respondents agreed with this statement. A

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

similar pattern was observed for the time-constraint item, "The time limit felt fair and realistic," which received the lowest mean score of 3.8 with a standard deviation of 0.79, again with 60% agreement. This pattern is also visible in Figure 5, where the bars for scoring clarity, badge motivation, and time fairness are lower than those for real-life usability and contextual appropriateness, showing that participants responded somewhat less strongly to these supporting features than to the contextual design of the tasks themselves.

Table 4. Task Relevance Ratings (N = 10)

Statement	Mean	SD	% Agree
The tasks felt appropriate for a café setting.	4.3	0.67	80%
I can imagine using these language skills in real life.	4.7	0.48	90%
The scoring (e.g., "8.6/10") helped me understand how well I did.	3.9	0.74	70%
The "Green Badge" system motivated me to do my best.	3.9	0.88	60%
The time limit felt fair and realistic.	3.8	0.79	60%

The task performance data provide additional descriptive evidence regarding how participants interacted with ACTE during the pilot session. Across the 30 total attempts, 19 attempts reached or exceeded the mastery threshold, corresponding to a 63% badge attainment rate. The mean final score across all attempts was 7.9 with a standard deviation of 1.2, placing the average just below the system's mastery cut-off of 8.0. This suggests that many responses approached the expected level even when mastery was not formally achieved. At the same time, the standard deviation indicates some spread in performance, which is reasonable given the small sample and possible differences in fluency, response preparation, and familiarity with scenario-based speaking tasks. The combined pattern of near-threshold average performance and a majority of successful attempts suggests that the task difficulty was manageable while still allowing observable variation across participants.

Task timing data further show how participants managed the response window imposed by the system. The average completion duration was 9.2 seconds with a standard deviation of 2.1, which was consistently above the built-in reference durations for several task types, including the 8-second benchmark used for transactional prompts. This indicates that participants generally required more time than the system expected in

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

order to formulate and produce their spoken responses. Although all participants were able to complete the interaction cycle, the timing data suggest that the response window may have been somewhat demanding for some users, particularly in tasks requiring a full sentence or more deliberate lexical retrieval. This pattern is consistent with the more moderate rating for time fairness shown in Figure 5, where the time-limit item received one of the lowest mean scores among the relevance measures.

The open-ended feedback complements the questionnaire and performance results by showing how participants described their experience in their own words. Several comments reflected strong positive perceptions of authenticity and simplicity. For example, one participant stated, "The tasks felt exactly like what I'd say in a real café" (P02), while another commented, "I liked that it was simple, just speak and see if you got it" (P07). A third participant noted that "The Green Badge felt satisfying but not pushy" (P05), suggesting that the reward mechanism was noticeable but not overly intrusive. These comments align with the higher ratings for contextual appropriateness and real-life usability displayed in Figure 5, reinforcing the view that participants considered the system meaningful and easy to engage with in the simulated café setting.

At the same time, the qualitative feedback identified several areas where participants wanted greater clarity. One participant remarked, "I wish I knew whether my score was low because of speed or words" (P05), indicating uncertainty about how the system combined multiple performance factors into a final score. Another participant observed that "Six seconds feels rushed for a full sentence" (P02), which corresponds with the relatively lower fairness rating for time limits shown in Figure 5. A third participant suggested, "Maybe show what I said vs. what I should have said" (P08), pointing to a desire for more explicit corrective or comparative feedback after task completion. These responses add descriptive detail to the questionnaire findings by showing which aspects of the user experience participants felt required refinement.

An important descriptive finding is that no participant reported confusion with the core interaction sequence, namely prompt → speak → score → badge. This indicates that the essential operational flow of ACTE was understandable across all participants in the pilot. Even where participants suggested improvements to scoring transparency or time

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

fairness, the underlying interaction logic itself did not appear to present a barrier. This is consistent with the overall usability profile summarized in Figure 6, where the mean SUS score remained above the benchmark for acceptable usability.

Overall, the pilot feasibility evaluation produced a coherent pattern of findings across usability scores, relevance ratings, performance data, and open-ended feedback. ACTE achieved an acceptable mean usability score, as illustrated in Figure 6, and strong ratings for contextual appropriateness and real-life applicability, as shown in Figure 5. The system also recorded a 63% mastery rate across all task attempts and elicited generally positive comments regarding authenticity and interaction simplicity. The more moderate responses related to scoring explanation, badge motivation, and time fairness provide additional descriptive detail regarding how participants experienced the system during the pilot evaluation.

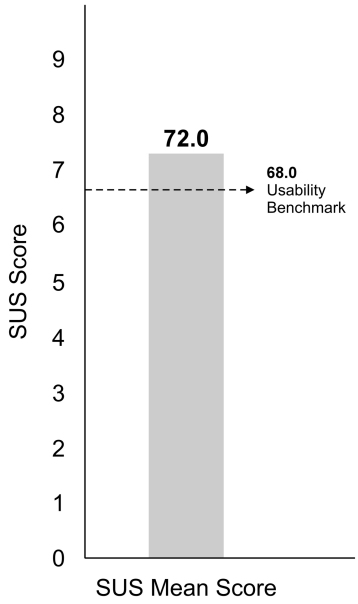

Figure 6. SUS Mean Score

4.2. Discussion

This section interprets the pilot feasibility findings of the Adaptive Contextual Task Engine (ACTE) in relation to usability, contextual relevance, task performance, and learner feedback. Given the small sample size and the use of a simulated café environment, the discussion is framed as a set of preliminary insights rather than definitive conclusions.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Even within these boundaries, the findings suggest that ACTE's design principles, particularly semantic place alignment, lightweight mastery tracking, and non-intrusive feedback, show promise for context-aware mobile-assisted language learning (MALL).

A central finding of the pilot was the strong positive response to contextual appropriateness and real-life usability. As shown in Figure 5, the highest-rated questionnaire items were "I can imagine using these language skills in real life" (M = 4.7) and "The tasks felt appropriate for a café setting" (M = 4.3). These results suggest that learners were highly responsive to the semantic match between task content and place context. This is important because ACTE was designed on the premise that location should not function merely as a technical coordinate, but as a source of communicative meaning that shapes what kinds of language tasks are pedagogically appropriate. This pattern is consistent with Godwin-Jones' view that emerging MALL environments should emphasize authentic, culturally meaningful interaction rather than technical novelty alone [6]. ACTE operationalizes this by aligning language practice with the affordances of a socially recognizable setting, in this case, the café as a site for ordering, requesting, clarifying, and responding politely. In this respect, the findings also support the argument that contextual intelligence in language learning should attend to semantic and social fit rather than relying only on raw environmental signals or device sensing [60]. The qualitative feedback reinforces this interpretation. Participants repeatedly described the tasks as "real," "natural," or close to what they would actually say in such a setting. This suggests that, even in a simulated environment, learners can experience a strong sense of authenticity when the discourse situation itself is plausible and culturally intelligible.

The usability results further support the feasibility of the ACTE framework. The mean SUS score of 72.0, illustrated against the benchmark in Figure 6, exceeded the commonly accepted threshold of 68 for acceptable usability [61]. This indicates that, overall, participants found the system reasonably easy to understand and use during the pilot session. The fact that no participant reported confusion with the core interaction flow, namely prompt → speak → score → badge, is also significant. In feasibility work, clarity of interaction is fundamental, especially in speaking-based systems where cognitive load can increase quickly if navigation, instructions, or feedback become unclear. At the same time, the spread of SUS scores from 50.0 to 90.0 indicates that user experience was not

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

uniform across the sample. This variation may reflect differences in familiarity with mobile learning tools, confidence in spoken English, or comfort with timed response tasks. Even so, the overall usability outcome suggests that ACTE's lightweight and relatively minimalist interface is a workable basis for further development. In line with broader work on digital learning systems, usability at this level is not only a matter of interface convenience, but also a precondition for learner trust and sustained use [61].

A more mixed pattern emerged in relation to scoring clarity. Although the corresponding questionnaire item received a moderately positive rating (M = 3.9, 70% agreement), it was clearly weaker than the ratings for contextual fit and real-life applicability, as also visible in Figure 5. The open-ended responses help explain this gap. Several participants indicated that they were uncertain about how their scores were determined, particularly whether lower scores reflected timing issues, word choice, or both. This points to an explainability issue rather than a rejection of the scoring system itself. Hakimi et al. argue that interpretability and transparency are increasingly important in digital educational environments because learners need to understand how system-generated outcomes relate to their own actions and progress [61]. When feedback remains opaque, its instructional value is reduced. This is especially relevant in mastery-oriented language learning, where learners benefit not only from a score, but from a clear sense of why they received that score and how they might improve. A modest enhancement, such as displaying separate indicators for response speed and language adequacy, could make the system more transparent without sacrificing its minimalist design. Such an adjustment would also align with user-centered educational technology principles, which emphasize that useful feedback should support self-regulation, reflection, and learner agency [39].

The issue of timing fairness deserves particular attention because it emerged in both the questionnaire results and the task performance data. The statement "The time limit felt fair and realistic" received the lowest mean rating (M = 3.8), and only 60% of participants expressed agreement. This more cautious response corresponds with the performance data, which showed an average completion time of 9.2 seconds, exceeding the reference duration used for several transactional prompts. Together, these results suggest that the system's temporal expectations may have been somewhat demanding for at least part of

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

the sample. This tension reflects a broader challenge in speaking-oriented MALL design: the need to encourage fluency without imposing time constraints that feel cognitively or emotionally burdensome. The CEFR values spontaneous communication, but it also recognizes that lower-level learners often need additional processing time to formulate accurate and appropriate utterances [28]. In technology-mediated speaking tasks, even modest time pressure can heighten self-consciousness and reduce performance quality, particularly when learners are still developing fluency [16]. ACTE's timing mechanism was intended as a gentle calibration of performance rather than a hard cutoff, yet the findings suggest that fixed reference times may not equally suit all learners. This implies that future iterations may benefit from adaptive timing based on proficiency, task type, or prior performance history, thereby preserving the pedagogical value of fluency-sensitive scoring while reducing perceptions of unfairness.

The responses to the Green Badge system also provide an important insight into learner motivation. The item measuring whether the badge system motivated participants to do their best produced a moderate mean score (M = 3.9) and one of the highest standard deviations, indicating more variation in how learners experienced this feature. However, the qualitative comments suggest that the badge system functioned less as a gamified reward and more as a quiet acknowledgment of progress. This distinction matters. ACTE was intentionally designed to avoid heavy extrinsic reward structures such as points, levels, or leaderboards, and instead to represent mastery in a restrained and low-pressure way. This design stance aligns with self-determination theory, which proposes that learners are more likely to sustain engagement when motivation is linked to competence and autonomy rather than external incentives [62]. It also resonates with Pegrum's argument that "calm" mobile learning design, characterized by low-friction, non-intrusive interaction, may be more educationally sustainable than attention-grabbing gamification [36]. In this light, the moderate motivational score should not necessarily be interpreted as a weakness. Instead, it may reflect the fact that the badge system was experienced as informational rather than competitive. The challenge for future refinement is therefore not to make the system more game-like, but to make the meaning of badges more visible, so that learners can more clearly connect them to mastery achievement and progress across modules.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The task performance data further suggest that the current mastery threshold is reasonably calibrated, though still open to refinement. Across 30 task attempts, 19 met the mastery requirement, producing a 63% badge attainment rate, while the overall mean score (7.9) fell just below the threshold of 8.0. This pattern is actually encouraging in a pilot context because it suggests that the scoring model is neither too lenient nor unrealistically strict. In mastery-based systems, learners should face goals that are sufficiently challenging to remain meaningful but still attainable enough to encourage persistence [47]. ACTE appears to be approaching that balance. The fact that the mean score was close to the mastery cutoff suggests that the system was sensitive enough to differentiate performance without making success unattainable. At the same time, the moderate badge rate indicates that learners were able to experience achievement while still encountering room for improvement and retry. This is pedagogically desirable in mastery learning environments, where progression should reflect demonstrated competence rather than mere task completion [47]. The performance data therefore support the broader design rationale of ACTE as a mastery-oriented rather than purely activity-based system.

These findings also highlight a useful contrast between ACTE and more notification-driven microlearning approaches. In some mobile learning designs, system engagement is driven by reminders, prompts, or timed nudges intended to increase interaction frequency. While such strategies may improve short-term participation, they may also risk attentional fatigue or resistance if learners experience them as intrusive [46]. ACTE instead follows a pull-based model, in which the learner encounters contextually relevant practice when the environment semantically calls for it and when the learner is ready to engage. The positive ratings for contextual appropriateness and practical relevance suggest that this design choice may be particularly suitable for situated language learning, where timing and place meaning are integral to the pedagogical experience. Rather than maximizing exposure frequency through alerts, ACTE attempts to maximize the meaningfulness of each encounter. The pilot findings suggest that this may be a fruitful direction for context-aware MALL, especially when the goal is to support transferable communicative competence rather than mere app engagement [46].

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Several limitations should be considered when interpreting these results. First, the sample size (N = 10) was intentionally small, which limits generalizability and prevents strong inferential claims. Second, the café context was simulated, not naturally occurring, meaning that the interaction lacked some of the distractions, interruptions, and social pressures that characterize real public language use. Third, the evaluation focused only on A2 speaking tasks in a single location type. As a result, the present findings cannot yet be generalized to other skill domains, proficiency levels, or semantic environments. These constraints are appropriate for a pilot feasibility study, but they also define the scope of the conclusions that can be drawn. Future work should therefore test the system across a broader range of contexts, including different real-world locations, additional language skills, and longer-term use patterns.

Even with these limitations, the discussion points to several design implications. The results suggest that semantic alignment should remain central to context-aware MALL design, since learners responded most positively to features that linked tasks to authentic place-based communication. They also indicate that feedback transparency is a priority area for improvement, particularly if the system is to support metacognitive awareness and self-directed progress [39], [61]. The findings on timing suggest that adaptive temporal calibration may be more appropriate than fixed response windows, especially for lower-level learners working on spontaneous speaking [16], [28]. The moderate but meaningful response to the badge system suggests that quiet mastery signaling may be preferable to overt gamification when the goal is to foster competence without pressure [36], [62]. Finally, the acceptable usability level shown in Figure 6 supports the idea that a lightweight, on-device, privacy-conscious system can still deliver pedagogically relevant personalization without relying on intrusive tracking or overly complex infrastructure [60].

The pilot findings suggest that ACTE is a feasible and pedagogically promising framework for situated MALL. The strongest support for the system comes from the convergence of quantitative and qualitative evidence around contextual authenticity and practical language relevance, as reflected in Figure 5 and in participant feedback. The main areas requiring refinement concern the explainability of scoring and the fairness of timing constraints, both of which directly affect how learners interpret and trust the system.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

For a pilot feasibility study, this is a constructive outcome. It indicates that the core concept, semantically aligned, mastery-based, privacy-conscious task recommendation, is sound enough to justify broader testing, while also identifying clear design targets for the next stage of development.

5. CONCLUSION

This pilot feasibility study demonstrates that the Adaptive Contextual Task Engine (ACTE) is a usable and contextually relevant task recommender for mobile-assisted language learning. With a mean SUS score of 72.0 and 90% of participants agreeing that tasks reflected real-life language use, ACTE shows preliminary evidence of supporting situated speaking practice through semantic place alignment and mastery-aware progression. However, these findings are limited to a small sample (N = 10), a single simulated café environment, and A2-level speaking tasks. The moderate ratings for scoring clarity (M = 3.9) and timing fairness (M = 3.8) also indicate areas for refinement, particularly in feedback transparency and adaptive duration calibration. While ACTE's lightweight, on-device architecture offers a privacy-conscious alternative to data-intensive recommender systems, broader validation is required across diverse locations, multiple language skills, and real-world field conditions before generalizing its effectiveness. Future research should investigate ACTE's scalability, longitudinal impact on learning outcomes, and applicability to other domains requiring situated skill development.

REFERENCES

[1] B. Choi, "Playful Assessment: A Game-Based Approach to Assessing Teachers' Competency for the Wise Integration of Technology in the Classroom," *University of Connecticut*, Storrs, CT, USA, 2017.

[2] J. S. Brown, A. Collins, and P. Duguid, "Situated cognition and the culture of learning," *Educ. Res.*, vol. 18, no. 1, pp. 32–42, 1989, doi: 10.3102/0013189X018001032.

[3] B. Morgan, "Situated cognition and the study of culture: An introduction," *Poet. Today*, vol. 38, no. 2, pp. 213–233, 2017, doi: 10.1215/03335372-3868421.

[4] R. Pederson, "Situated learning: Rethinking a ubiquitous theory," *J. Asia TEFL*, vol. 9, no. 2, pp. 123–148, 2012.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[5] R. Godwin-Jones, "Emerging technologies: Mobile apps for language learning," *Lang. Learn. Technol.*, vol. 15, no. 2, pp. 2–11, 2011.

[6] R. Godwin-Jones, "Emerging spaces for language learning: AI bots, ambient intelligence, and the metaverse," *Lang. Learn. Technol.*, vol. 27, no. 2, pp. 6–27, 2023, doi: 10.64152/10125/73501.

[7] R. Godwin-Jones, "Distributed agency in second language learning and teaching through generative AI," *Lang. Learn. Technol.*, vol. 28, no. 2, pp. 5–31, 2024, doi: 10.64152/10125/73570.

[8] A. Bahari, X. Zhang, and Y. Ardasheva, "Establishing a computer-assisted interactive reading model," *Comput. Educ.*, vol. 172, no. September 2020, p. 104261, 2021, doi: 10.1016/j.compedu.2021.104261.

[9] H. Robles, K. Burden, and K. Villalba, "A socio-cultural approach to evaluating and designing reading comprehension apps for language learning," *Int. J. Mob. Blended Learn.*, vol. 13, no. 1, pp. 18–37, 2021, doi: 10.4018/IJMBL.2021010102.

[10] Y. J. Lee and P. Roger, "Cross-platform language learning: A spatial perspective on narratives of language learning across digital platforms," *System*, vol. 118, no. October, p. 103145, 2023, doi: 10.1016/j.system.2023.103145.

[11] A. Kukulska-Hulme and O. Viberg, "Mobile collaborative language learning: State of the art," *Br. J. Educ. Technol.*, vol. 49, no. 2, pp. 207–218, 2018, doi: 10.1111/bjet.12580.

[12] O. A. Al-Smadi, R. A. Rashid, H. Saad, Y. H. Zrekat, S. S. L. A. Kamal, and G. I. Uktamovich, "Artificial Intelligence for English Language Learning and Teaching: Advancing Sustainable Development Goals," *J. Lang. Teach. Res.*, vol. 15, no. 6, pp. 1835–1844, 2024, doi: 10.17507/jltr.1506.09.

[13] M. Mihaylova, S. Gorin, T. Reber, and N. Rothen, "A meta-analysis on mobile assisted language learning applications reveals moderate learning benefit and significant publication bias," pp. 1–52, 2020, [Online]. Available: https://osf.io/preprints/ux93y/

[14] Y. Zhou and M. Zhou, "A meta-analysis on mobile-assisted vocabulary learning: Do mobile applications help?," *ReCALL*, vol. 38, pp. 75–93, 2025, doi: 10.1017/S0958344025100335.

[15] Y. Bai, "A Mixed Methods Investigation of Mobile-Based Language Learning on EFL Students' Listening, Speaking, Foreign Language Enjoyment, and Anxiety," *SAGE Open*, vol. 14, no. 2, pp. 1–19, 2024, doi: 10.1177/21582440241255554.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[16] K. Tabuchi, S. Kobayashi, S. T. Fukuda, and Y. Nakagawa, "A Case Study on Reducing Language Anxiety and Enhancing Speaking Skills Through Online Conversation Lessons," *Technol. Lang. Teach. Learn.*, vol. 6, no. 3, pp. 1–21, 2024, doi: 10.29140/tltl.v6n3.1497.

[17] B. Olovson and M. Seigler, "Overcoming Skill-Specific Language Learning Anxiety: Research-Based Tools," in *Room For All at the Table*, 1st ed., B. M. Burke, Ed., Ashland, VA 23005: CSCTFL, 2020, pp. 47–74.

[18] Z. Zhou, "A Systematic Literature Review on the use of Mobile-assisted Language Learning (MALL) for Enhancing Speaking Skills in Chinese EFL context," *Int. J. Front. Sociol.*, vol. 3, no. 15, pp. 12–24, 2021, doi: 10.25236/ijfs.2021.031502.

[19] R. Li, "Effects of mobile-assisted language learning on foreign language learners' speaking skill development," *Lang. Learn. Technol.*, vol. 28, no. 1, pp. 1–26, 2024, doi: 10.64152/10125/73553.

[20] T. Jantakoon, T. Jantakun, K. Jantakun, W. Pongpanich, R. Pasmala, P. Wannapiroon, and P. Nilsook, "The effectiveness of artificial intelligence in English instruction for speaking and listening skills: A meta-analysis," *Contemp. Educ. Technol.*, vol. 17, no. 4, p. ep596, 2025, doi: 10.30935/cedtech/17310.

[21] S. Bourekkache and O. Kazar, "Mobile and adaptive learning application for english language learning," *Int. J. Inf. Commun. Technol. Educ.*, vol. 16, no. 2, pp. 36–46, 2020, doi: 10.4018/IJICTE.2020040103.

[22] B. Garg and D. S. Pant, "Effectiveness Of Adaptive Mobile-Assisted Language Learning (Mall) For Efl Vocabulary Acquisition Among Undergraduate Students," *Int. J. Environ. Sci.*, vol. 11, no. 18, pp. 2223–2231, 2025, doi: 10.64252/pqpf2j74.

[23] A. Bahari, "Affordances and challenges of technology-assisted language learning for motivation: A systematic review," *Interact. Learn. Environ.*, vol. 31, no. 9, pp. 5853–5873, 2023, doi: 10.1080/10494820.2021.2021246.

[24] A. Bobunova, M. Sergeeva, and E. Notina, *Integrating Computer-Assisted Language Learning into ESL Classroom: Formation of Moral and Aesthetic Values*. 2021. doi: 10.18178/ijiet.2021.11.1.1484.

[25] E. S. H. Haataja, A. Tolvanen, H. Vilppu, M. Kallio, J. Peltonen, and R. L. Metsäpelto, "Measuring higher-order cognitive skills with multiple choice questions –potentials and pitfalls of Finnish teacher education entrance," *Teach. Teach. Educ.*, vol. 122, 2023, doi: 10.1016/j.tate.2022.103943.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[26] A. T. Lidyasari, I. Rachmawati, A. Da Costa, and P. Wanyi, "How are the Cognitive, Affective, and Psychomotor Levels of Primary School Learners Living in Suburban Area of Yogyakarta based on Career Development?," *J. Prima Edukasia*, vol. 10, no. 2, pp. 130–137, 2022, doi: 10.21831/jpe.v10i2.48061.

[27] T. H. Phuc and T. T. Nghi, "Examining the Impact of Mobile Apps on Language Teaching and Learning in a Public University: An Experimental Study," *Int. J. Linguist. Lit. Transl.*, vol. 3, no. 11, pp. 55–67, 2023, doi: 10.32996/ijllt.2023.6.6.12.

[28] N. Figueras, D. Little, and B. O'Sullivan, *Aligning Language Education with the CEFR: A Handbook*, vol. 5, no. April. 2022. doi: 10.37546/jaltsig.cefr5-1.

[29] A. Z. Amiruddin, M. T. A. Ghani, W. A. A. W. Daud, N. H. Hussein, N. H. M. Yamin, and Y. S. Abdullah, "Empowering language learners: Innovations in CEFR-based language learning applications," *Edelweiss Appl. Sci. Technol.*, vol. 9, no. 2, pp. 352–359, 2025, doi: 10.55214/25768484.v9i2.4486.

[30] W. C. Fang, H. C. Yeh, B. R. Luo, and N. S. Chen, "Effects of mobile-supported task-based language teaching on EFL students' linguistic achievement and conversational interaction," *ReCALL*, vol. 33, no. 1, pp. 71–87, 2021, doi: 10.1017/S0958344020000208.

[31] J. Wu, H. Jiang, and S. Chen, "Augmented reality technology in language learning: A meta-analysis," *Lang. Learn. Technol.*, vol. 28, no. 1, pp. 1–23, 2024, doi: 10.64152/10125/73596.

[32] B. Chen, Y. Wang, and L. Wang, "The Effects of Virtual Reality-Assisted Language Learning: A Meta-Analysis," *Sustain.*, vol. 14, no. 6, 2022, doi: 10.3390/su14063147.

[33] E. Ericsson and S. Johansson, "English speaking practice with conversational AI: Lower secondary students' educational experiences over time," *Comput. Educ. Artif. Intell.*, vol. 5, no. April, p. 100164, 2023, doi: 10.1016/j.caeai.2023.100164.

[34] B. Khosrawi-Rad, H. Rinn, D. Augenstein, D. Markgraf, and S. Robra-Bissantz, "Designing Pedagogical Conversational Agents in Virtual Worlds," *Lect. Notes Informatics (LNI), Proc. - Ser. Gesellschaft fur Inform.*, vol. P-338, pp. 181–186, 2023, doi: 10.18420/delfi2023-29.

[35] R. R. Divekar, H. Lepp, P. Chopade, A. Albin, D. Brenner, and V. Ramanarayanan, "Conversational Agents in Language Education: Where They Fit and Their Research Challenges," *Commun. Comput. Inf. Sci.*, vol. 1499 CCIS, pp. 272–279, 2021, doi: 10.1007/978-3-030-90179-0_35.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[36] M. Pegrum and A. Palalas, "Attentional literacy as a new literacy: Helping students deal with digital disarray," *Can. J. Learn. Technol.*, vol. 47, no. 2, pp. 1–18, 2021, doi: 10.21432/CJLT28037.

[37] M. Pegrum and Y. J. Lan, "Extended reality (XR) in language learning: Developments and directions," *Lang. Learn. Technol.*, vol. 27, no. 3, pp. 1–5, 2023, doi: 10.64152/10125/73528.

[38] Y. M. Fromm and D. Ifenthaler, "Designing adaptive learning environments for continuing education: Stakeholders' perspectives on indicators and interventions," *Comput. Hum. Behav. Reports*, vol. 16, no. August, 2024, doi: 10.1016/j.chbr.2024.100525.

[39] E. A. Oliveira, P. de Barba, and L. Corrin, "Enabling Adaptive, Personalised and Context-aware Interaction in a Smart Learning Environment: Piloting the Icollab System," *Australas. J. Educ. Technol.*, vol. 37, no. 2, pp. 1–23, 2021, doi: 10.14742/AJET.6792.

[40] J. P. Lantolf, M. E. Poehner, and S. L. Thorne, "Sociocultural Theory and L2 Development," *Theor. Second Lang. Acquis.*, pp. 223–247, 2020, doi: 10.4324/9780429503986-10.

[41] J. P. Lantolf, "Sociocultural Theory Learning Language Second Introduction to the Special Issue," *Natl. Fed. Mod. Lang. Teach.*, vol. 78, no. 4, pp. 418–420, 2008.

[42] J. P. Lantolf, "The sociocultural approach to second language acquisition," *2nd ed. New York, NY, USA: Routledge,* 2010, pp. 24–47.

[43] J. P. Lantolf and M. E. Poehner, "Sociocultural theory and classroom second language learning in the East Asian context: Introduction to the special issue," *Mod. Lang. J.*, vol. 107, pp. 3–23, 2023, doi: https://doi.org/10.1111/modl.12816.

[44] C. Kramsch and Contributors, *Language Acquisition and Language Socialization: Ecological Perspectives*. London, U.K.: Continuum, 2003.

[45] S. V. Steffensen and C. Kramsch, "The Ecology of Second Language Acquisition and Socialization," in *Language Socialization, Encyclopedia of Language and Education*, 2017, pp. 1–16. doi: 10.1007/978-3-319-02327-4_2-1.

[46] H. T. Nguyen Tran, W. Jou She, S. Kajimura, and Y. Shibuya, "Assessing Notification Timing Strategies for Improved Micro-Learning Engagement," *IEEE Access*, vol. 13, pp. 64534–64545, 2025, doi: 10.1109/ACCESS.2025.3559521.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[47] D. Rohrer, "Interleaving helps students distinguish among similar concepts . Educational Psychology Interleaving Helps Students Distinguish among Similar Concepts," *Educ. Psychol. Rev.*, vol. 24, pp. 355–367, 2012.

[48] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Q. Manag. Inf. Syst.*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.

[49] A. K. Carstensen and J. Bernhard, "Design science research–a powerful tool for improving methods in engineering education research," *Eur. J. Eng. Educ.*, vol. 44, no. 1–2, pp. 85–102, 2019, doi: 10.1080/03043797.2018.1498459.

[50] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007, doi: 10.2753/MIS0742-1222240302.

[51] B. Kremmel, K. Eberharter, E. Konrad, and ..., "The CEFR Companion Volume: Opportunities and challenges for language assessment," *Didakt. slawischer ...*, no. 3, 2023, doi: 10.48789/2023.

[52] K. Leitner and B. Kremmel, "Avoiding Scoring Malpractice: Supporting Reliable Scoring of Constructed-Response Items in High-Stakes Exams," *Challenges in Language Testing Around the World*. Springer Singapore, pp. 127–145, 2021. doi: 10.1007/978-981-33-4232-3_10.

[53] X. Chen *et al.*, "Modeling question difficulty for unbiased cognitive diagnosis: A causal perspective," *Knowledge-Based Syst.*, vol. 294, no. March 2023, p. 111750, 2024, doi: 10.1016/j.knosys.2024.111750.

[54] N. U. C. Mustaffa and S. N. Sailin, "A Systematic Review of Mobile-Assisted Language Learning Research Trends and Practices in Malaysia," *Int. J. Interact. Mob. Technol.*, vol. 16, no. 5, pp. 169–198, 2022, doi: 10.3991/ijim.v16i05.28129.

[55] Q. Xu and J. C. Richardson, "The Impact of Social Media and Gamification of a Mobile Vocabulary Learning App: Self-Regulation and Learning Persistence," *Online Learn. J.*, vol. 28, no. 4, pp. 4–32, 2024, doi: 10.24059/olj.v28i4.4592.

[56] J. Brooke, "SUS: A 'Quick and Dirty' Usability Scale," in *Usability Evaluation In Industry*, no. January 1996, 1996, pp. 189–194. doi: 10.1201/9781498710411-35.

[57] K. Ishaq, N. Azan, F. Rosdi, A. Abid, and Q. Ali, *Usability of Mobile Assisted Language Learning App*. 2020. doi: 10.14569/ijacsa.2020.0110145.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[58] V. Braun and V. Clarke, "Reflecting on reflexive thematic analysis," *Qual. Res. Sport. Exerc. Heal.*, vol. 11, no. 4, pp. 589–597, 2019, doi: 10.1080/2159676X.2019.1628806.

[59] G. Konstantinos, "Thematic analysis: A practical guide," *Eur. J. Psychother. Couns.*, vol. 26, no. 3–4, pp. 461–464, 2024, doi: 10.1080/13642537.2024.2391666.

[60] N. Tran, S. Kajimura, and Y. Shibuya, "Location- and Physical-Activity-Based Application for Japanese Vocabulary Acquisition for Non-Japanese Speakers," *Multimodal Technol. Interact.*, vol. 7, no. 3, 2023, doi: 10.3390/mti7030029.

[61] L. Hakimi, R. Eynon, and V. A. Murphy, *The Ethics of Using Digital Trace Data in Education: A Thematic Review of the Research Landscape*, vol. 91, no. 5. 2021. doi: 10.3102/00346543211020116.

[62] J. C. Dunn and C. Zimmer, "Self-determination theory," *Routledge Handb. Adapt. Phys. Educ.*, vol. 55, no. 1, pp. 296–312, 2020, doi: 10.4324/9780429052675-23.